SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(141st Edition)

SEPTEMBER 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 Forty-First** 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 **September, 2024**. 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 Compact Slot-Based Bi-Directional UHF RFID Reader Antenna for Far-Field Applications, Amit Birwal , Senior Member, IEEE, Akash Shakya, Saurav, Shalini Kashyap, and ***6.1Kamlesh Patel,*** Microwave Characterization ECE

2. A comprehensive analysis on cost, combustion, performance and emissions for diesel engine fueled with preheated esters of waste cooking oil, ***6.1 Prem Shanker Yadav, 6.Saket Kumar,*** and ***3.Raghvendra Gautam,*** Automobile Engineering, DTU

3. A framework for evaluating the barriers to adopting Industry 4.0 in Indian SMEs: an approach of best-worst method, ***8.1 Roshan Kumar, 3.Pravin Kumar***, Rajesh Kumar Singh, Anurika Vaish, & Garima Sharma, Machanical Engineering, DTU

4. A Hybrid Deep Learning Model for Sentiment Analysis of IMDB Movies Reviews, ***3.Rohit Beniwal, 8.1 Arpit Kumar Dinkar, 8.1 Arun Kumar,*** and ***8.1 Anurag Panchal,*** Computer Science, Computer Engineering, DTU

5. A hybrid metaheuristic optimised ensemble classifer with self organizing map clustering for credit scoring, ***3.Indu Singh***, D. P. Kothari, ***6.S. Aditya***, ***8.1 Mihir Rajora***, Charu Agarwal, and ***8.1 Vibhor Gautam***, Computer Science Engineering, Electronics and Communication, DTU

6. A Noise and Edge extraction-based dual-branch method for Shallowfake and Deepfake Localization, ***6.Deepak Dagar, 3.Dinesh Kumar Vishwakarma***, Information Technology, DTU

7. A Novel Acetylcholinesterase-Based Electrochemical Biosensor Using g-C3N4@MoS2 Nanohybrid for the Detection of Trichlorfon, ***7.1Srijita Chatterjee, 7.1 Harshita Singh, 6.Divya Hudda, 6.Sweety,*** and ***3.Devendra Kumar***, Applied Chemistry, DTU

8. A Schif base hydrogel of oxidized okra gum and carboxymethylated chitosan: a biocompatible and biodegradable injectable system for drug delivery in wound care, ***6.Jitendra Kumar, 3.Roli Purwar***, Applied Chemistry, DTU

9. Addressing Bias in Text-to-Image Generation: A Review of Mitigation Methods, ***7.1 Shah Prerak***, Computer Science & Engineering, DTU

10. Advanced Deep Learning Approaches for Fault Detection and Localization in Power Distribution Systems, **8.1 Nitin Sharma**, 6.Pankaj, ***3.J.N Rai, a***nd ***3.S.T. Nagarajan,*** Electrical Engineering, DTU

11. Advanced Tor-NonTor Traffic Classification, ***8.1 Khushi***, ***3.Dr. Jitendra Nath Rai***, Electrical Engineering, DTU

12. Advancing sentiment classifcation through a population game model approach, ***6.Neha Punetha, 3.Goonjan Jain***, Applied Mathematics, DTU

13. An improved Network Intrusion Detection System using a Novel Hybrid Genetic Coati-Pelican Optimization Algorithms with an Ensemble Fuzzy c-means and CNN-RF, ***3.Indu Singh, 8.1 Eric S. Moore, 8.1 Ebrima Jallow,*** and ***8.1Alhagie S. Jobe***, Computer Science & Engineering, DTU

14. ANN Controlled Single-Phase Microinverter for Off-Grid Solar Application, ***7.1 Ajay Kumar Sahu***, Ramnarayan Patel, and Lalit Kumar Sahu, Power System Electrical Engineering, DTU

15. Artificial Intelligence in Gastrointestinal Bleeding Analysis for Video Capsule Endoscopy: Insights, Innovations, and Prospects (2008-2023), Tanisha Singh, Shreshtha Jha, Nidhi Bhatt, ***6.1 Palak Handa***, ***3.Nidhi Goel*** , and ***3.S. Indu,*** Electronics and communication Engineering, DTU

16. BERT-based RNN for Effective Detection of Depression with Severity Levels from Text Data, ***6.Pavi Saraswat***, ***3.Rohit Beniwal***, Computer Science and Engineering, DTU

17. Biodegradable, biocompatible, and self-healing, injectable hydrogel based on oxidized Azadirachta indica gum and carboxymethyl chitosan through dynamic imine-linkage for biomedical application, ***6.Jitendra Kumar***, ***3.Roli Purwar***, Applied Chemistry, DTU

18. Decentralized Safe and Scalable Multi-Agent Control under Limited Actuation, Vrushabh Zinage, ***8.1 Abhishek Jha*** , Rohan Chandra,and Efstathios Bakolas, Machanical Engineering, DTU

19. Derivation Analysis and Control of Multiport Flyback Converter with Lyapunov Function-Based Controller in Renewable Energy Systems Considering Circuit Parasitics, Shabana Urooj, ***3.Dheeraj Joshi***, ***6.Ashutosh Gupta***, and Amal BaQais, Electrical Engineering, DTU

20. Design and Control of DVR Based on Adaptive Bateman Polynomial for Power Quality Improvement,***6.Kanchan Bala Rai, 3.Narendra Kumar, and 3.Alka Singh***, Electrical Engineering, DTU

21. Design and Implementation of BIST logic for High Speed and Energy Efficient Carry Select Adder(CSLA), ***6.Garima Solanki, 8.1 Sourav Agarwal, 8.1 Tushar Mishra, 8.1Vansh***, and Khandelwal, Electrical Engineering, DTU

22. Development and evaluation of gelatin/polyacrylamide/carboxymethyl tamarind kernel gum hydrogel for delivery of ampicillin sodium, ***7.1Tanuja Kumari, Nitin, 6.Priyanka Meena, a***nd ***3.Sudhir G. Warkar***, Applied Chemistry, DTU

23. Development of fexible piezoelectric nanogenerator based on PVDF/KNN/ZnO nanocomposite flm for energy harvesting application, ***6.Komal Verma***, ***8.Aman Kumar***, and ***3.Richa Sharma***, Applied Physics, DTU

24. Development of improved functional neural network based autoregression models for power quality improvement, ***3.Alka Singh, 8.1Srishti Singh***, Electrical Engineering, DTU

25. Does spirituality at work promote work-to-family enrichment among Indian female employees?, ***3.Naval Garg a***nd , Nidhi Sharma, DSM, DTU

26. Effect of Dy3+ ions on physical, structural and thermal properties of ZnNaPbB glasses Manjeet, ***6.Anu*** , Nisha Deopa , Anand Kumar, and Rishi Pal Chahal, Applied Physics, DTU

27. Efect of load frequency and amplitude of displacement on soft synthetic rock joints under cyclic shear loads S. M. Mahdi Niktabar, K. Seshagiri Rao, and ***3.Amit Kumar Shrivastava***, Jiří Ščučka,Civil Engineering, DTU

28. Effect of promoters in hydrates based carbon dioxide capture: A review, ***6.Aseem Dubey,*** and ***3.Akhilesh Arora***, Mechanical Engineering, DTU

29. Emergence of Topological Phase and Non-Trivial Surface States in Rare-Earth Semimetal GdSb with Pressure, ***7.Nidhi, 6.Ramesh Kumar***, Ramesh K. Bibiyan, and ***3.Mukhtiyar Singh***, Applied Physics, DTU

30. Enhancing Image Translation with Conditional Generative Adversarial Networks, ***8.1 Shobhit Verma, 8.1 Srishti Arora, 8.1 Tanvi Kayat,*** and ***3.Anurag Goel,*** Computer Science and Engineering, DTU

31. Evolution of transformer-based optical flow estimation techniques: a survey, ***6.Nihal Kumar, 3.Anil Singh Parihar,*** and ***3.Om Prakash Verma*** CSE, ECE, DTU

32. Experimental review on solid oxide fuel cell-based hybrid power generation system: Performance, economic, and environmental analysis, Anil Kumar Yadav, ***3.Anil Kumar***, Shailendra Sinha, Samarjit Singh, and S. K. Yadav, Mechanical Engineering, DTU

33. Fabrication of Lead-Free PVDF/KNNLTS/MWCNT Piezoelectric Nanogenerator: Role of MWCNT in the Piezoelectric Performance of Nanogenerator for Energy-Harvesting Application, ***6.Komal Verma, 6.Aman Kumar*** and ***3.Richa Sharma,*** Applied Physics, DTU

34. FLOW THROUGH STEAM TURBINE CASCADE FOR ROUGHNESS ANALYSIS, ***3.Manjunath K***., Ajeet Singh Sikarwar, and ***3.Naushad Ahmad Ansari***, Mechanical Engineering, DTU

35. Forecasting Electric Vehicle Sales in India using Singular Spectrum Analysis***, 8.1 YASH AGGARWAL***, DR. ROMA DEBNATH, School of Management (Economics), DTU

36. Ge–As–Se–Te chalcogenide based rib-waveguide for highly coherent on-chip mid-infrared supercontinuum generation: design and analysis, 6***.Deepak Garg, 6.Akash Khamaru,*** *a*nd ***3.Ajeet Kumar***, Applied Physics, DTU

37. Ge'ez Digit Recognition Model Based on Convolutional Neural Network, 3.Ruchika Malhotra, & ***6.Maru Tesfaye Addis***, Software Engineering, CSE, DTU

38. Gender Fairness Analysis in LLMs using Zero-Shot and Few-Shot Experiments, ***8.1 Keshav Gupta, 8.1 Kartikeya Jain,*** and ***8.1Lakshay Solanki***, CSE, DTU

39. GSVAELP: integrating graphSAGE and variational autoencoder for link prediction ***6.Fatima Ziya, 3.Sanjay Kumar***, Computer Science, DTU

40. Identifying Currency Crises in India: Empirically Weighted EMP Indices with Multiple Static Thresholds and Markov Regime-Switching Models Using Time-Varying Thresholds, ***6.Sanjay Kumar, 3. Dr. Nand Kumar***, Humanities, DTU

41. Improved ML based Load Forecasting using TimeSplit Convolutional Neural Network, ***8.1Muskan, 8.1Muskan, 8.1Pragati Sattavan,*** and ***3.JN Rai***, Electrical Engineering, DTU

42. Investigating the Effect of Sliding Window Length for EEG-based Anxiety Classification, 6.Shikha***, 3. Divyashikha Sethia,*** and ***3.S. Indu***, CS, Software Engg., ECE, DTU

43. Investigation of circularly polarized MIMO antenna with enhanced isolation for sub-6 GHz application***, 6.Parul Kansal, 3.Anup Kumar Mandpura,*** and ***3.Narendra Kumar***, Electrical Engg., DTU

44. Investigations and Multi-response Optimization of Process Parameters for SS316L Cold Metal Transfer-Wire Arc Additive Manufactured Samples, ***6.Sudeep Kumar Jain , 3.Qasim Murtaza,*** and ***3.Pushpendra Singh,*** Mechanical Engineering, DTU

45. K-BEST subspace clustering: kernel-friendly block-diagonal embedded and similarity-preserving transformed subspace clustering, Jyoti Maggu, ***3.Anurag Goel***, Computer Science, DTU

46. Machine learning approach for evaluating soil liquefaction probability based on reliability method, Sophia Khatoon, Kishan Kumar, Pijush Samui, Laith Sadik, and ***3.1 Sanjay Kumar Shukla***, Civil Engg., DTU

47. MASSFormer: Mobility-Aware Spectrum Sensing using Transformer-Driven Tiered Structure, Dimpal Janu, Sandeep Mandia, ***6.1 Kuldeep Singh***, and ***6.1 Sandeep Kumar,*** Computer Vision and ECE, DTU

48. Measurement of Heavy Metals Content in Suspended Sediment of Ganges River Using Atomic Absorption Spectrometry, ***8.1 M. Aggarwal***, ***3.S. Anbukumar*** and ***3.T. Vijaya Kumar,*** Civil Engg. DTU

49. Mental Health and Stress Prediction Using NLP and Transformer-based Techniques, ***7.1 Abhishek Pandey, 3.Sanjay Kumar***, CSE, DTU

50. Modelling Persistence and Asymmetry in Currency Volatility: A Multi-Country Application of FIGARCH Models, ***6.Sanjay Kumar, 3.Dr. Nand Kumar***, Humanities, DTU

51. Optical characteristics, Judd-Ofelt analysis of enhanced luminescence by flux in thermally stable, novel Eu3+- doped BaZr(PO4)2 phosphor for indoor lighting applications, ***6.Shweta Yadav , 3.A.S. Rao,*** & ***3.Deshraj Meena***, Applied Physics, DTU

52. Optimization of Lamps Configuration for a Large-scale Indoor Solar Simulator for PV Panels and Solar Collectors, Al Yaqdhan Al Umairi, Khalid Al Sultani, Abdullah Al Shuaili, Aisha AI Adawi, Afzal Husain, Mohd Zahid Ansari, and ***3.Mohd Zunaid***, ,Mechanical Engg., DTU

53. Optimization of Wearable Biosensor Data for Stress Classification Using Machine Learning and Explainable AI, ***6.SHIKHA , 3.DR. DIVYASHIKHA SETHIA*** , and ***3.Prof. S. Indu,*** Computer Science, Software science, ECE, DTU

54. Optimizing Carrier Concentration for Enhanced Thermoelectric Performance in AgSbS 2 Monolayer, ***6.1Stephen Shaiju Mathew, 6.Sangeeta, 6.Rajesh Kumar, 3.Mukhtiyar Singh,*** and ***Manish K. Kashyap***, Applied Physics, DTU

55. Predicting Breast Cancer Survival Rate Based on Genetic Data: A Machine Learning Approach, ***6.1 Saanya Yadav,***and ***3.Yasha Hasija***, Biotechnology, DTU

56. Prediction of aeration performance of different types of piano key weirs using different machine learning models, 8***. Niyati Narwala, 3.Nipun Bansal,*** Deepak Singh, and ***3.Munendra Kumar***, Computer Science, Civil Engg. DTU

57. Realizing In-Memory Computing using Reliable Diferential 8T SRAM for Improved Latency, ***6.AYUSH DAHIYA, 3. Poornima Mittal,*** and ***3. Rajesh Rohilla***, ECE, DTU

58. Resp-BoostNet: Mental Stress Detection from Biomarkers Measurable by Smartwatches using Boosting Neural Network Technique, ***6. SANJAY KUMAR, 8.*** **ANSHUMAN RAJ CHAUHAN, 8.1 AKHIL**, AKSHI KUMAR, and GUANG YANG, Computer Science, DTU

59. Response surface method-based hydraulic performance optimization of a single-stage centrifugal pump, ***6.Durvesh Yadav, 3.Raj Kumar and 3. Singh K. Manjunath***, Mechanical Engg., DTU

60. Review on design, development, and implementation of an unmanned aerial vehicle for various applications, 3.***Mubina Shekh, 3.Sushila Rani***, and Rituparna Datta, Machanical Engineering, DTU

61. rGO-Embedded Polymer Nanocomposite Layer for Improved Performance of Triboelectric Nanogenerator***, 6. Shilpa Rana, 3. Bharti Singh***, Applied Physics, DTU

62. Sentiment Classification on Suicide Notes Using GPT, Bi-LSTM and CNN, ***7.1 Akash Bansal, 3.Rohit Beniwal***, Computer Science, DTU

63. SOME CHALLENGES AND ISSUES FOR WATER TRANSFER FROM GANGA-BRAHMAPUTRA BASINS TO CAUVERY-PENNAR BASINS IN INDIA, ***3.1 S.K. MAZUMDER,*** Civil Engg. DTU

64. Structural and spectroscopic features of single-phase Dy3+ activated BiYWO6 phosphor for luminescent device applications, ***6.Indrajeet Maurya, 3.M. Jayasimhadri***, Applied Physics, DTU

65. Supercontinuum generation in Ga-Sb-S chalcogenide- based PCF using optofluidic approach***, 6.Deepak Garg, 6.Akash Khamaru, and 3.Ajeet Kumar***, Applied Physics, DTU

66. Stacked Denoising Variational Auto Encoder Model for Extractive Web Text Summarization, ***6.1 Madhuri Yadav,*** and ***3.Rahul Katarya***, Computer Science Engineering, DTU

67. Synthesis and Assessement of Acacia Gum-Based Hydrogel as a Promising Novel Biopolymeric Matrix for Delivery of Ciprofloxacin, ***8.Anshu***, ***7.1 Garima Tomar***, ***6.Priyanka Meena,*** and ***3.Sudhir G. Warker,*** Applied Chemistry, DTU

68. Tailoring pH-sensitive carboxymethyl tamarind kernel gum-based hydrogel for an efficient delivery of hydrophobic drug indomethacin***, 6. Priyanka Meena, 3. Poonam Singh,*** and ***3.Sudhir G. Warkar***, Applied Chemistry, DTU

69. Tex-ViT: A Generalizable, Robust, Texturebased dual-branch cross-attention deepfake detector, ***6. Deepak Dagar, 3.Dinesh Kumar Vishwakarma,*** Information Technology, DTU

70. The Effectiveness of Advance Deep Learning Architectures for Classification of Stress using Raw EEG Data, ***8.Ayush Tibrewal, 6.Shikha,*** and ***3.Divyashikha Sethia***, Applied Physics, CSE and Software Engg., DTU

71. The Evolution of Currency Regimes in Selected Countries from the Americas, Europe, Asia, and Africa: A Comparative Analysis from the Late 19th Century to 2023, ***6.Sanjay Kumar, 3.Nand Kumar***, Humanities, DTU

72. The potential of magnesium oxide nanoparticles (MgONPs) in the transesterifcation of lipids produced by Rhodotorula minuta***, 3.Richa Srivastava***, Niti Srivastava, · SonamPreet Kaur, Kamal Kant and Kumar Gaurav, Applied Chemistry, DTU

73. Transit Photometry for Estimating the Velocity of Exoplanets and specific Defence Applications, ***8.Harshita Bose***, Applied physics,DTU

74. Ultrasonic Assisted Synthesis of CuFe2O4-Ag infused Gum Hydrogels Nanocomposite for photocatalytic Degradation of Organic Dye from Wastewater Meenu · Pratibha Sharma · ***3.Roli Purwar,*** Navita Sharma, Pooja Rawat, · Seema R. Pathak · Atul Thakur · Hyunook Kim · Monu Verma · Sudip Majumder and Chandra Mohan Srivastava, Applied Chemistry, DTU

75. Unveiling Anomalies: A Review of Anomaly Detection Through Lens of Explainable AI, ***7.1 Shikhar Asthana, 3.Anurag Goel***, CSE, DTU

76. Unveiling Toxic Tendencies of Small Language Models in Unconstrained Generation Tasks, ***8.1 Lakshay Chandra, 3.Seba Susan, 8.1 Dhruv Kumar***, and ***8.1Krishan Kant***, IT, DTU

77. Volatility Clustering and Long- Memory -a FIGARCH Analysis of Selected Currency Pairs ***6.Sanjay Kumar***, ***3.Dr. Nand Kumar***, Humanities, DTU

78. "You're Just So Weird and Crazy": Recounting My Lived Experiences as a PhD Scholar in India, 6***.1.Elixir Sagar,*** Delhi School of Management, DTU

79. ZCS Switched-Capacitor Cell Balancing Circuit with Bidirectional Buck-Boost Charging, ***6.Archit Chauhan, 3.Sudarshan K. Valluru,*** and ***3.Vanjari Venkata Ramana,*** Electrical Engineering, 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*

A Compact Slot-Based Bi-Directional UHF RFID Reader Antenna for Far-Field Applications

Amit Birwal, *Senior Member, IEEE*, Akash Shakya, Saurav, Shalini Kashyap, *Member, IEEE*, and Kamlesh Patel, *Senior Member, IEEE*

***Abstract*—This research introduces a novel circularly polarized compact antenna designed for universal ultrahigh-frequency (UHF) radio-frequency identification (RFID) handheld readers for Bi-directional RFID Far-field Applications. The antenna features a microstrip feed positioned at the center opposite a slot-based square ground. The ground plane is perturbed to include a thin horizontal and vertical stub on the left side, along with a thick rectangular slot at the right side of the square ground plane to achieve circular polarization. The simulated antenna provides a 3-dB axial ratio bandwidth (ARBW) of 42 MHz (831–873 MHz), a 10 dB impedance bandwidth of 15% (814–945 MHz), and a peak gain of 5.0 dBi. The antenna is fabricated on both layers of an affordable FR4 substrate, measuring 81 × 81 × 1.6 mm^3 and its measurement results are in close agreement with simulated. The application of this antenna is made with a commercial UHF RFID reader module. The obtained read range and field of view confirm that this proposed antenna is a promising option for compact universal UHF RFID handheld reader applications and other Internet of Things (IoT) based applications.**

***Index Terms*—Axial ratio (AR), circular polarization (CP), RFID, reader antenna, and return loss.**

I. Introduction

RADIO-FREQUENCY identification (RFID) is an automated identification technology that utilizes radio waves to track and identify objects in various contexts such as access control, supply chains, inventory management, and traffic toll with other commercial use [1], [2]. RFID stands out from many other identification technologies for its numerous benefits, including high data rates, precision detection, and high reliability. RFID systems are classified based on their operational frequency bands, however, these are broadly categorized as low-frequency (LF) in the band 125–135 kHz (ISO 18000-2), high-frequency (HF) of 13.56 MHz (ISO 18000-3), ultrahigh-frequency (UHF) of 433.92 MHz (ISO 18000-7) and in the band 860–960 MHz (ISO 18000-6), and microwave frequency of 2.45/5.8 GHz (ISO 18000-4).

Manuscript received 5 July 2024; revised 25 August 2024; accepted 6 September 2024. Date of publication 10 September 2024; date of current version 17 September 2024. This work was supported in part by the Faculty Research Programme (FRP) of the Institute of Eminence (IoE) Scheme of the University of Delhi under Grant IoE/2023-24/12/FRP dated 31/08/2023. *(Corresponding author: Kamlesh Patel.)*

The authors are with the Department of Electronic Science, University of Delhi South Campus, New Delhi 110021, India (e-mail: amit.birwal@south.du.ac.in; akashshakya821@gmail.com; sauravprasher0@gmail.com; shalini90kashyap@gmail.com; kpatel@south.du.ac.in).

Digital Object Identifier 10.1109/JRFID.2024.3457691

Recently, RFID systems operating in the global UHF band (840 to 960 MHz) have become particularly popular due to their high data rates, rapid reading speeds, and large read ranges can typically range from 2-5 m for passive tags and up to 100 m for active tags [3], [4]. An RFID system consists of an RFID tag (which includes a tag antenna and an integrated circuit), an RFID reader (which comes with an external antenna), and a host computer to process the received data (Fig. 1). For development purposes, an Arduino microcontroller is used to interface the reader and host computer. In essence, the reader transmits a radio-frequency signal to the RFID tag via its antenna and then receives a modulated signal back from the tag. The RFID tag must be within the reader's designated read range to receive the signal from the reader's antenna [5]. In practical applications, the tag antennas in RFID systems are typically linearly polarized (LP) and are often positioned arbitrarily in orientations. To maintain reliable communication between RFID readers and tags, it is vital to use a reader antenna with circular polarization (CP) characteristics [6], which can reduce polarization loss. In the reported near-field and far-field multiservice antennas, the RFID reader antenna with CP radiation ensures the detection of the tag irrespective of orientation along with a strong and uniform magnetic field [2], [7]. In [8], a loop antenna array is designed for near-field and far-field applications using many microstrip collinear loop antenna with periodic radiating units. It produced a horizontal polarized omnidirectional radiation in 802 to 963 MHz with an enlarged near-field reading volume of 280 × 280 × 540 mm^3 and a far-field gain of 6.66 dBi.

In recent years, various CP reader antenna solutions, including microstrip patch slot antennas and stacked-architecture antennas, have been explored using different design techniques for UHF RFID applications. Most of these CP antennas are reported with impedance bandwidth (IBW), axial ratio bandwidth (ARBW) and gain. A measured 3-dB axial-ratio (AR) bandwidth in 903 – 909 MHz band (RO4003C) and 918-929 MHz band (FR4) is reported for an asymmetric circular shaped slotted microstrip patch antennas with slits are reported for CP radiation [9], while the wideband CP stacked slotted microstrip patch antenna operating in the 835-955 MHz band is reported in [10]. A low-cost antenna consisting of four sequentially rotated Inverted-F antennas (IFAs) was fed by a compact four-feed network and made on an FR-4 substrate [11]. This antenna reported a measured 3-dB AR bandwidth is 18.6% from 805 MHZ to 970 MHz over

2469-7281 © 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 September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

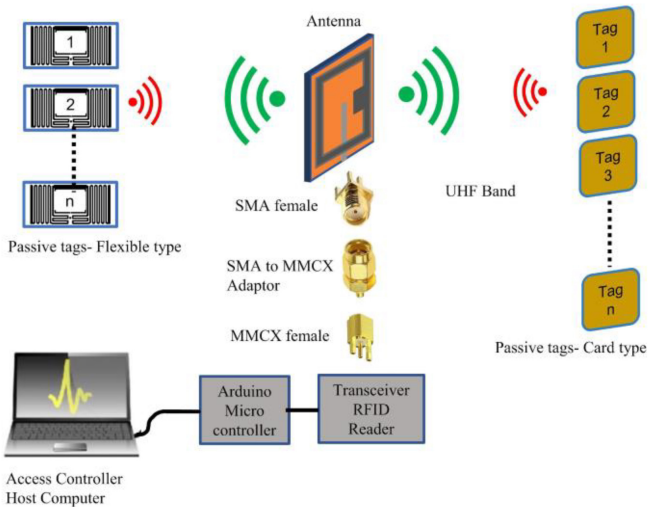


Fig. 1. External antenna with RFID system architecture.

the compact size of 95×100×13.6 mm^3 for handheld reader applications.

Also, a dual-band CP antenna is proposed with a compact size of 85 × 85 × 1.6 mm^3 in [12] for 915 MHz/2.45 GHz. A dual-band antenna with CP and two pairs of crossed dipoles is designed for the RFID reader [13], where the antenna directivity is improved by employing an artificial magnetic conductor (AMC). This results in a much wider ARBW of 25 MHz and IBW of 15% (814-945 MHz) with a radiation efficiency of more than 90%. In [14], the author has proposed the CP UHF RFID reader application, and the fabricated antenna on an FR-4 substrate is made of size 12×12×1.6 mm3, which is reported with a good ARBW, gain, and IBW. A dual-layer CP PIFA is proposed in [15] for application as an RFID reader in which two stacked PIFA are orthogonally oriented and excited synchronously with equal power by using a single probe. In [16], a dual-band CP circular patch antenna is proposed for UHF RFID and WLAN applications and comprised of two perpendicular slots on the circular patch and four slits on the ground layer with coaxial cable feed. AMC reflector is then placed to lower the back-lobe level and improve gain up to 5 dBi. A CPW-fed slot antenna consisting of an equilateral triangular patch shows S_{11} well below −23 dB and a wider impedance bandwidth of 130 MHz with the presence of two notches compared to S_{11} = −17.5 dB and 113.85 MHz bandwidth without notches [17]. A spherical patch disc monopole antenna was made with two-sided corner truncations and circular and (+) shaped slots for the RFID UHF 860-960 MHz band. It was integrated with an RFID reader board and an STM32L476RG microcontroller to interpret the tags [18]. For Car parking applications at 908 MHz, a low-cost RFID reader antenna was realized in a grounded rectangular form based on a slotted patch which consists of 5 etched slots on the upper rectangular-shaped metal side [19]. It shows a response of S_{11}= −25 dB and a gain of 1.6 dB.

Thus, the antenna attached to the reader is an acute component of RFID systems. Its poor performance in terms of polarisation loss, power limitations, and signal degradation can limit the entire system's performance. The research challenges for RFID reader antenna design are reliable detection of tag irrespective of its orientation, and compact compared to the linearly polarized antenna on the same low-cost substrate with moderate gain and high efficiency. So, for UHF RFID handheld readers, a compact, low-profile, and lightweight CP antenna improves the reliability of such systems. In addition, a few reports discussed the integration of the proposed UHF antenna with RFID reader or system, which is also important for coverage verification of reader antenna performance.

The objective of this research is, therefore, to design a compact reader antenna with a bi-directional pattern on a very low-cost material with a wide frequency band for the UHF RFID system. To achieve this objective a compact slot-based bi-directional UHF RFID Reader antenna is fabricated and analyzed for far-field applications. Section II describes the designed antenna and its simulated results, including S11, and axial ratio, with the parametric study of slot dimensions with the explanation of surface current distribution under various phases. Section III presents the measured results of the proposed UHF antenna, radiation patterns, and comparison with the simulated results. It also shows the novelty of this proposed antenna by comparing it with previously published works. Section IV examines the performance of the proposed antenna as a reader antenna by integrating it with a commercially available RFID reader in the UHF band (865-867 MHz) regulated for operation in India. The integration of the proposed antenna with the RFID reader using an Arduino microcontroller and host computer is also presented. The parameters, such as field of view (FOV) and read range, are analyzed under varying transmit power levels, while Section V provides concluding remarks.

II. Antenna Design

The RFID reader antenna proposed in this work features a two-layer design, where a simple microstrip line feeds the square loop in the ground plane perturbed with slots. The upper layer, side view, and lower layer are depicted in Fig. 2(a), along with the optimized dimensions for the UHF band (865-867 MHz). The antenna proposed is fed through an SMA connector connected to a microstrip feedline located on the underside of the FR4 substrate. The substrate has a loss tangent (δ) of 0.02, a relative permittivity (ε_r) of 4.4, and a dielectric thickness of 1.6 mm. A large rectangular slot of size (L_3 × W_3) i.e., 48 × 16.5 mm^2 is inserted at the right side of the primary square slot ground plane. The square ground slot is perturbed to achieve circular polarization (CP), and a pair of I-shaped microstrip line feedlines is placed with an angle of 90° between them on the left side of the ground plane, which results in generating orthogonal electric field components. The square dimension of the designed antenna is (L × W) i.e., 81×81 mm^2. The antenna design, optimization, and simulation are performed on the commercial software CST Microwave Studio. Table I provides the optimized dimensions for the final proposed antenna. Fig. 2(b) presents the fabricated antenna's top and bottom views.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

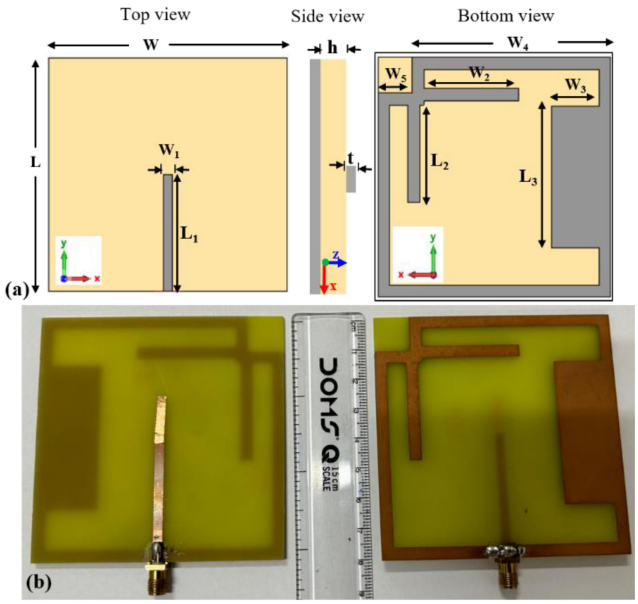


Fig. 2. Geometrical view of the proposed antenna (a) top, side, and bottom view, (b) top and bottom view of fabricated prototype.

TABLE I
OPTIMIZED DIMENSIONS OF THE PROPOSED UHF ANTENNA

Parameter	Value (mm)	Parameter	Value (mm)
L	81	W	81
L_1	50	L_2	33
L_3	48	W_1	3
W_2	33	W_3	16.5
W_4	69	W_5	12
H	1.6	t	0.035

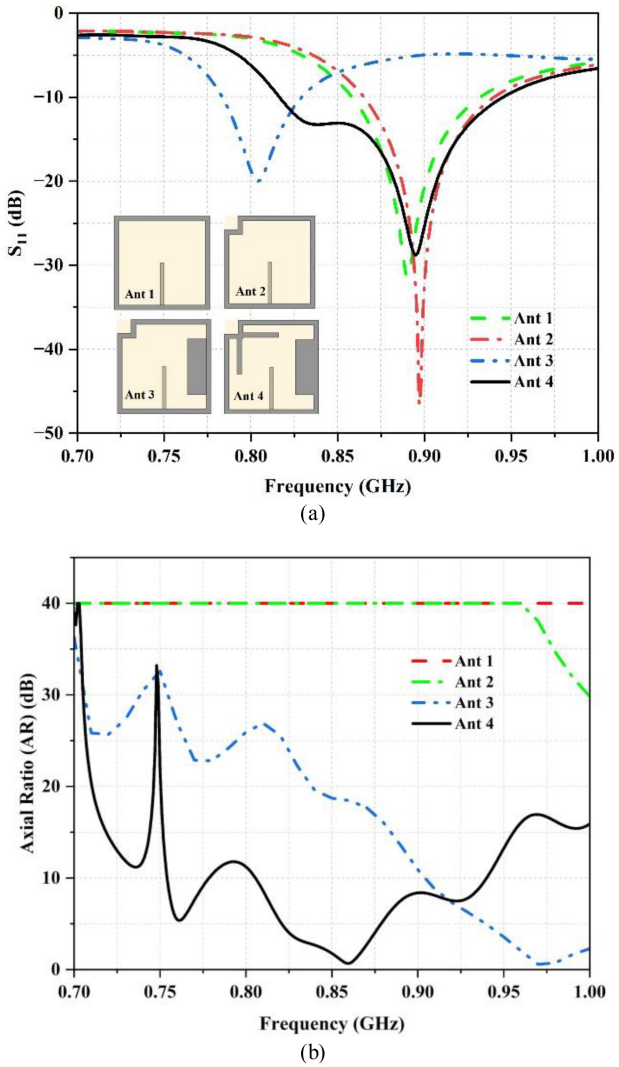


Fig. 3. Performance variation during the evolution stages of the proposed antenna (a) S_{11} parameters and (b) axial ratio.

The CP radiation in this antenna can be achieved by exciting two orthogonal electric field vectors, i.e., the horizontal component (E_H) and the vertical components (E_V), with equal amplitudes and a 90° phase shift. To illustrate the CP characteristics of the designed antenna, Fig. 3(a) outlines the simulated impedance during the antenna's evolutional stages, while Fig. 3(b) presents the ARBW for antennas in these stages.

Designing this antenna involves four stages, labeled Ant 1 through Ant 4, which is the final design. Initially, Antenna 1 features a simple grounded square loop of 81 mm fed by a microstrip line from the opposite side, which provides a narrow band resonance of around 870 MHz. Ant 1 offers a linear polarization of approximately flat 40 dB AR. In the second stage, to generate a CP, the symmetric square loop is perturbed from the left side with a square of side $W_5 = 12$ mm, forming Ant 2. With this modification, the impedance bandwidth improves up to −45 dB around 900 MHz, and AR of around 30 dB is observed at 1000 MHz. In the third stage, to fine-tune the CP bandwidth within the intended RFID band (840 MHz to 960 MHz) and significantly enhance the impedance bandwidth, a pair of I-shaped narrow slots L_2 and W_2, which are 90^o apart is connected at the left side edge of the ground plane, thus creating Ant 3. Adding these I-shaped stubs, shifts resonance frequency from 900 MHz to 800 MHz as the structure becomes electrically large. The 3-dB ARBW of 50 MHz is observed in the frequency band from 950 to 1000 MHz, which is far from the desired UHF RFID band. Then, in the fourth stage, for overall performance enhancement, a rectangular vertical stub ($L_3 \times W_3$) is added to the right of the square ground slot, leading to Ant 4, the proposed final design. Ant 4 demonstrates the IBW of 131 MHz (816–947 MHz) and ARBW of 41 MHz (833–874 MHz) including the desired UHF RFID band (865-867 MHz).

The parametric analyses are also performed on the proposed antenna to gain more detailed insights into the antenna optimization process. First, a parametric study was carried out on the two slots, which are responsible for generating circular polarization. The physical parameters analyzed include the dimensions ($L_2 = W_2$) of the I-shaped stubs, which are 90° apart, and the width (W_3) of the rectangular vertical stub ($L_3 \times W_3$). To assess the influence of these physical parameters on the performance of the proposed antenna, each parameter

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

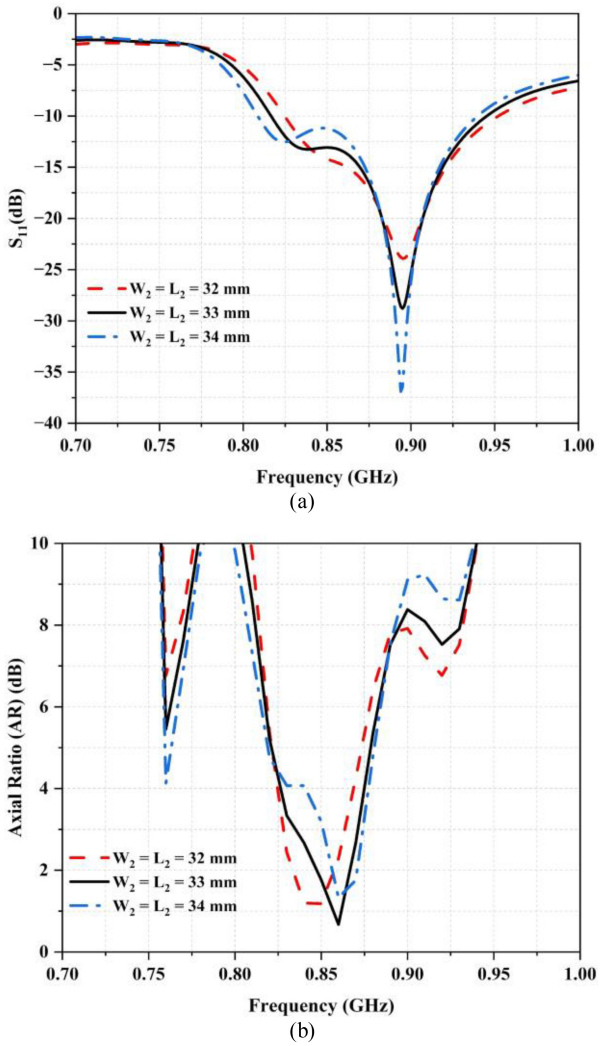


Fig. 4. Parametric variation of the proposed antenna (a) S_{11} (dB), (b) AR (dB) when the length of the slot $L_2 = W_2$ varied.

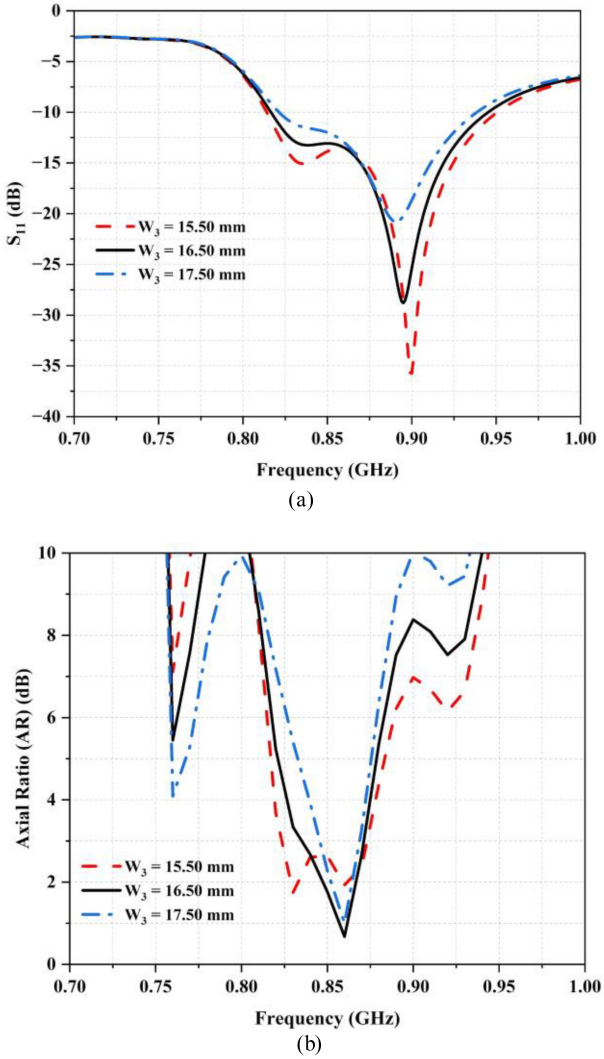


Fig. 5. Parametric variation of the proposed antenna (a) S_{11} (dB), (b) AR (dB) when the width of the slot W_3 varied.

is varied individually while keeping the other parameters constant mentioned in Fig. 2(a). The changes in IBW and AR are presented due to these parameters in Fig. 4 and Fig. 5, respectively.

In Fig. 4, the findings indicate that altering L_2 from 32 to 34 mm in steps of 1 mm affects the impedance matching of the antenna, but much not alter the start and end frequencies of the impedance bandwidth. Additionally, the analysis reveals that while L_2 influences the ARBW as much as for $L_2 = 33$mm, a significant AR of about 0.5 dB is observed at 865 MHz. Further increasing the L_2 deteriorates the AR, so the optimized value of $L_2 = W_2 = 33$ mm is chosen. As depicted in Fig. 5(a), increasing the value of W_3 leads to a fall in the resonant frequency.

When W_3 is set to 16.5mm, the ARBW shifts slightly to the desired lower frequencies at 860 MHz. The best ARBW of 42 MHZ (831-873 MHZ) is observed for $W_3 = 16.5$mm, as shown in Fig. 5(b). Table I shows the optimized dimensions of the proposed UHF antenna.

To confirm into the circular polarization (CP) mechanism of the proposed antenna, the surface current distribution on the antenna aperture was plotted using CST Microwave Studio and illustrated in Fig. 6. When a phase zero signal is applied, the dominant current vector, shown in black color, is seen along the positive X axis. As the phase angle shifts to 90°, the current vector reorients to the positive Y axis. For phase angles of 180° and 270°, the vector changes to the negative X and Y axes, respectively. Observing the rotation of this resulting current vector from the top of the antenna (+Z axis), it is clear that the rotation is counterclockwise. This counterclockwise rotation indicates Right-Hand Circular Polarization (RHCP).

III. Antenna Measurement Results

The proposed antenna is made on the FR4 substrate using the LPKF ProtoMat S100 PCB milling machine, and an SMA connector is connected at the optimized feed location on the microstrip line. The S-parameter (S_{11}) of the fabricated UHF antenna is measured using the Rohde & Schwarz ZVH8 vector network analyzer. Figs.7 (a-b) compares the simulation and measurement data of S_{11} and the axial ratio of the proposed UHF RFID antenna. A measured 10 dB IBW of 149 MHz

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

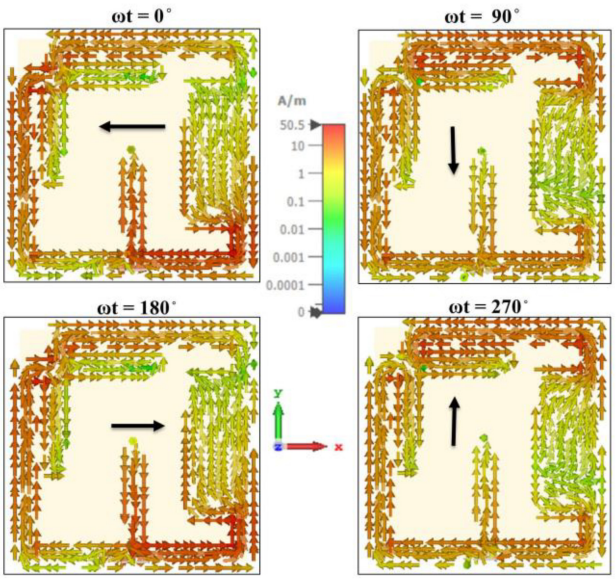


Fig. 6. Surface current distribution on the proposed antenna at 865 MHz at different phase instants.

(810.5-959.35 MHz) is obtained compared to the simulated one of 131 MHz (816-947 MHz). Also, the measured S_{11} shows a better value of about −60 dB in the 900 MHz band than the simulated value of about −30 dB. This is due to the addition of a small copper strip over the feedline visible in Fig. 2(b), which offered a slight shift in the resonant frequency with the perfect matching and wider BW for the fabricated antenna.

For performance investigation of the circularly polarized nature of the antenna, a rotating linearly polarized transmitting standard horn antenna is used, while the proposed UHF RFID antenna is in receive mode aligned in a similar direction. The frequency has been varied from 865 to 867 MHz to measure the received power on the R&S®FSL6 Spectrum Analyzer. Then, the placement of the proposed UHF antenna is rotated by 90°, and the received power is captured at the same frequency points. Two received powers are subtracted from each other to express the axial ratio (AR) in dB. The measured 3-dB axial-ratio bandwidth (ARBW) is 25 MHz (850-875 MHz), compared to the simulated ARBW of 42 MHz, as shown in Fig. 7(b).

The measured average gain of this proposed UHF antenna is approximately 5 dBi with a minimum gain fluctuation of about ±0.2dBi similar to the antenna in [2] and it is shown in Fig. 8.

Fig. 9 compares normalized simulated and measured radiation patterns at 865 MHz. The symmetry of the radiation patterns is observed at an angle (phi = 0° and 90°) in the simulated one, while asymmetric patterns are found in the both XZ- and YZ-planes due to a small extra length of copper feed line. Still, the proposed UHF RFID antenna provides a very stable radiation pattern in both broadside directions (±Z). These findings prove that the proposed antenna is well suited to cover all the required (865-867 MHz) frequency range needed for RFID operations in India.

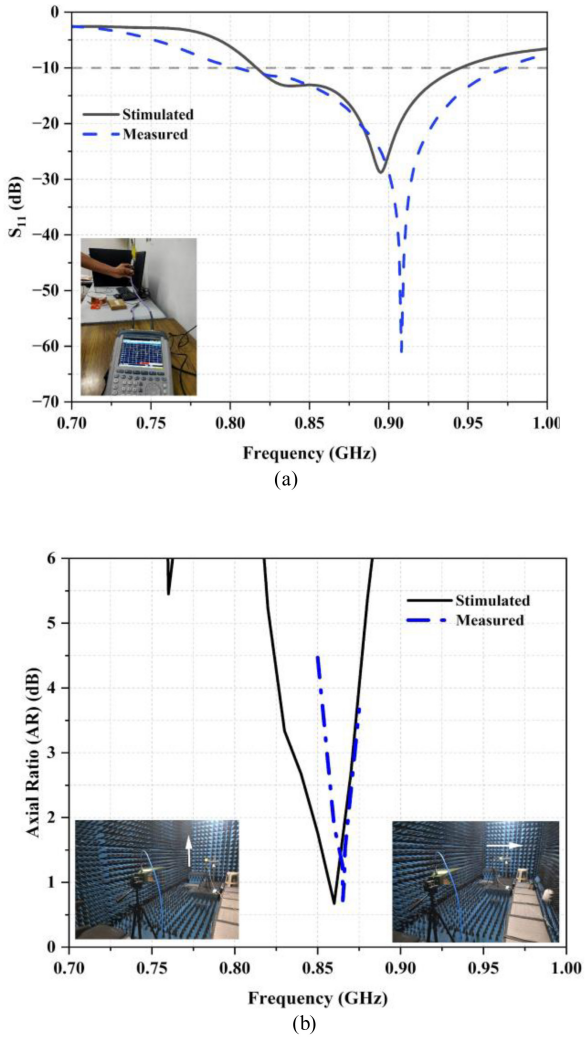


Fig. 7. Simulated and measured result (a) S_{11} (dB), (b) axial ratio (dB) of the proposed CP antenna.

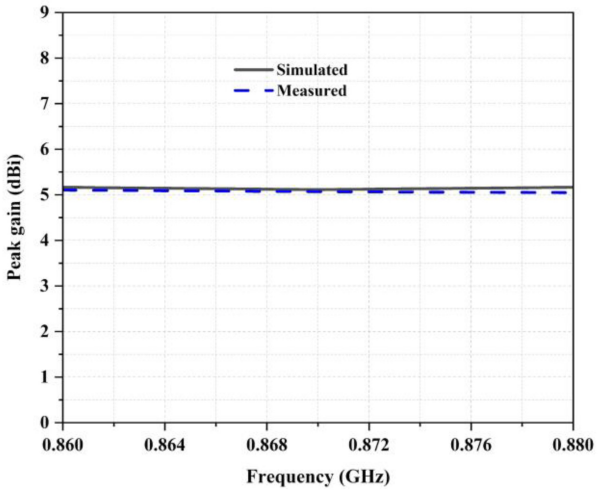


Fig. 8. Peak measured vs simulated gain of the proposed antenna around the UHF band.

In Table II, the reported features of previously published antennas at the RFID UHF band are compared. the following features of the proposed antenna are unique:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

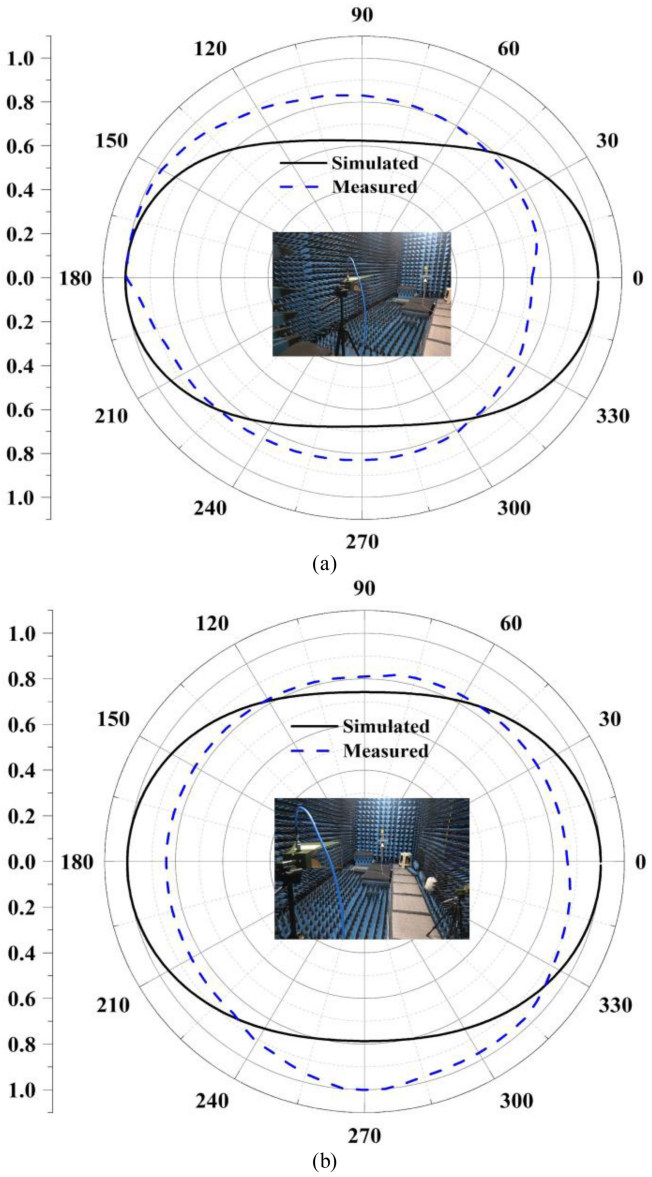


Fig. 9. Simulated and measured normalized 2D radiation pattern of the antenna at 865 MHz, (a) XZ-plane, (b) YZ-plane.

1) It is simple, compact, able to integrate, and offers wide impedance bandwidth in the 865-867 MHz frequency band.
2) It has a bi-directional radiation pattern with an almost constant gain of 5 dBi in the UHF RFID band.
3) It achieves CP radiation and high efficiency over a low-profile and low-cost substrate compared to most of the reported antennas made on the same substrate.
4) The efficacy, accuracy, and integration of the proposed RFID reader antenna are verified experimentally in the next section.

IV. Antenna Integration With RFID Reader

To evaluate the potential of the proposed UHF antenna for deployment with an RFID reader utilizing digital modulation and an Arduino Uno, a setup is illustrated in Fig. 10. It

TABLE II
Comparison of Antenna Features

Ref	Size (mm^3)	IBW (MHz)	AR Bandwidth (MHz)	Gain	Radiation Efficiency (%)	Substrate used
[6]	156×156×20	-	(818-964)	8.3 dBic	-	-
[11]	95×100×13.6	(748 to 1076)	(805-870)	3.1 dBic	67	FR-4
[12]	85×85×1.6	(871-1297) & (2160-2800)	(882-936) & (2350-2580)	(1.4-1.5) dBic	80	FR-4
[13]	104×104×0.813	(740-1100) & (2000-3050)	(770-1060) & (2230-2950)	9.0 & 15.0 dBic	95	Rogers 4003
[14]	120×120×1.6	(630-1103)	(818-1278)	4.0 dBic	-	FR-4
[15]	180×152×34.6	(417-449)	(430.3 -435.1)	3.8 dBic	>97 (Simu)	-
[16]	210×210×30	(903-927) & (2424-2472)	(902-950) & (2423-2470)	3.1 dBic & 6.25 dBic	57.14 & 74.3 1 (Simu)	FR-4
This work	**81×81×1.6**	**(816–947)**	**(831–873)**	**5.0 dBi**	**85-90 (Simu)**	**FR-4**

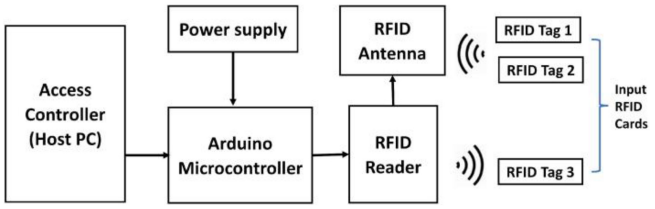
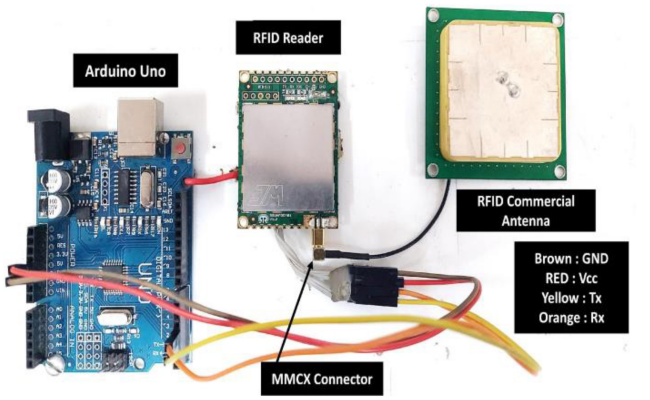


Fig. 10. Block Diagram of a setup utilizing UHF RFID Module and Arduino Uno.

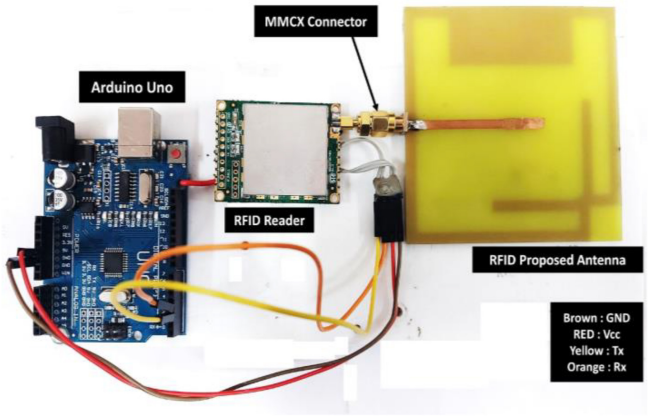
consists of RFID tags containing data chips, a UHF RFID module model no. RRUHFOEM01 from RapidRadio operating in 865 to 867 MHz and is compatible with ISO 18000-6C. This green module, measuring 34mm × 40.64mm with a thickness of 4mm, features an external antenna to transmit signals and receive data with a read range of up to 1 meter, and a power supply [20]. The external commercial RHCP RFID patch antenna is of dimensions 40mm × 40mm from ThingMagic EL6e Reader Module Antenna Kit, which has a gain of 3dBi and VSWR < 2 in 865 to 868 MHz. An optional host computer can process the tag data through Arduino Uno (ATmega328P microcontroller) using the universal asynchronous receiver/transmitter (UART) Protocol. The Tag UHF Smart card (Product Code: RRUHFSM01) can be read from distances up to 20+ feet, made of white PVC. This card has dimensions of $85\times54mm^2$ with a thickness of 0.8mm. It operates in an 865-867 MHz frequency band and follows the ISO 18000-6C: EPC Class1 Gen2 standard [21].

The experimental setup shown in Fig. 11, where, the commercial RFID antenna is connected to the UHF RFID reader module via a micro-miniature coaxial (MMCX) connector, while an MMCX to SMA adaptor is used to connect the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.



(a)



(b)

Fig. 11. (a) with Commercial Antenna (b) with Proposed UHF RFID Antenna.

fabricated RFID antenna with the reader, which is depicted in Fig. 11(b).

In the performance study, first, the power and frequency settings of the UHF RFID module are configured using the specific command codes. The details are mentioned in [22]. Subsequently, these command codes are utilized to identify and extract the Electronic Product Code (EPC) of RFID tags. An Electronic Product Code (EPC) is a universal identifier that gives a unique identity to a specific physical object [23]. When an EPC tag is scanned, the reader extracts all of the information stored in the RFID tag and displays it on the Arduino IDE serial monitor. When the tag is not detected, the response will be the first line, as shown in Fig. 12, and when the tag is detected, the response will be the last line, including the EPC code.

Now, the reading capabilities of two RFID antennas are investigated: the commercial RFID antenna and the proposed UHF RFID antenna. The read range was measured at different levels of transmitted power and plotted on a graph, as shown in Fig. 13, to visualize the effect of increasing transmitted power on the read range according to tag detection depicted in Fig. 12. According to Fig. 13, the proposed antenna has shown a significantly greater read range of 230 cm than the maximum 76 cm with the commercially available one due to different antenna characteristics for the same transmit power.

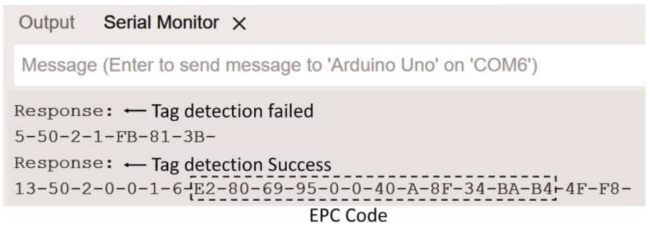


Fig. 12. Response of RFID Tag detection at the serial monitor of Arduino.

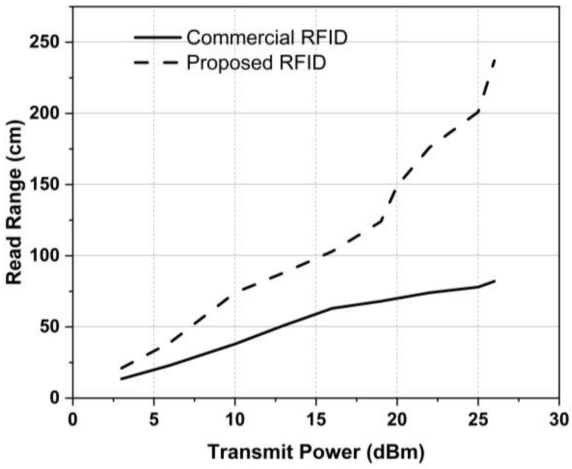


Fig. 13. Transmitted Power vs Read Range for both RFID antennas.

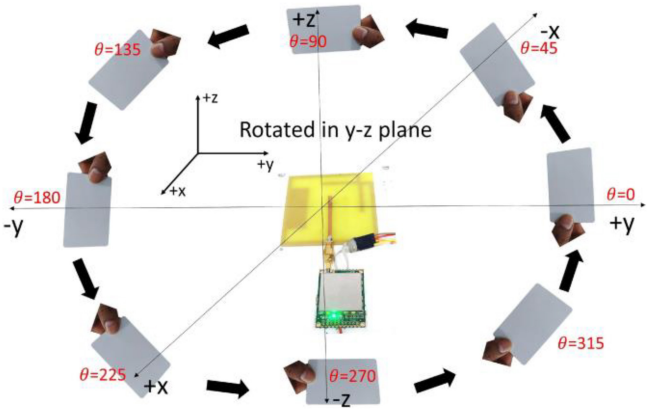


Fig. 14. Measurement setup for Field of View.

Now, to assess a 360° coverage of the antenna, the RFID tag is moved across eight different angular positions along the altitude axis, as depicted in Fig. 14. At each position, the maximum read range (MRR) is measured while keeping the RFID transmitting antenna stationary. The 360° coverage of the proposed UHF antenna is determined by varying its position in the transmit power levels from 3 dBm to 26 dBm, and the corresponding read ranges have been noted to investigate how transmit power affected coverage.

Fig. 15 illustrates that the commercial RFID antenna can read tags effectively only on the +Z axis. In contrast, the proposed RFID antenna exhibited enhanced performance by successfully reading tags in both the +Z and −Z axes, as depicted in Fig. 16. A similar symmetric coverage in the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

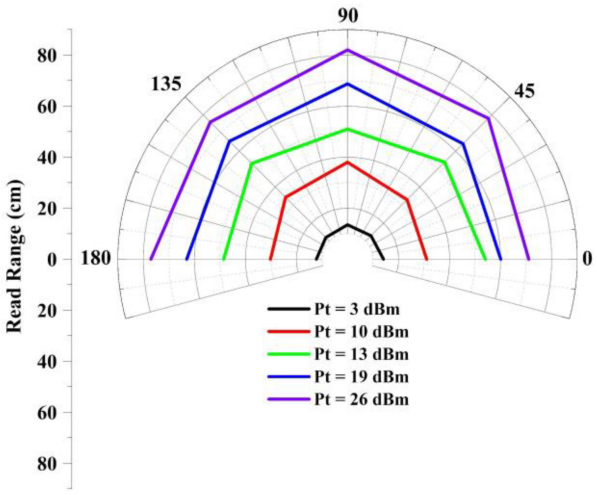


Fig. 15. Polar plot of Read Range of Commercial RFID antenna with different transmit powers.

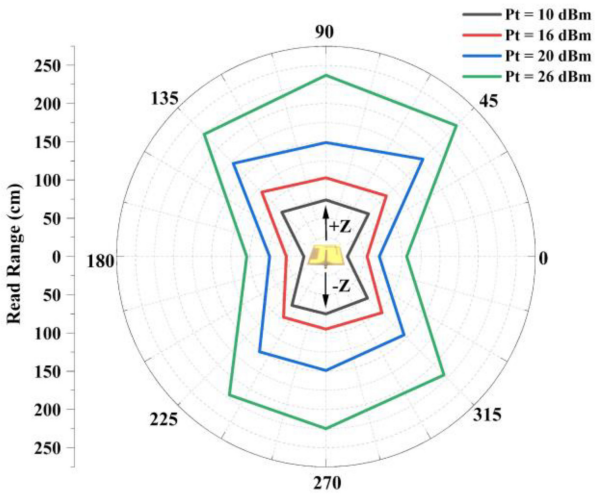
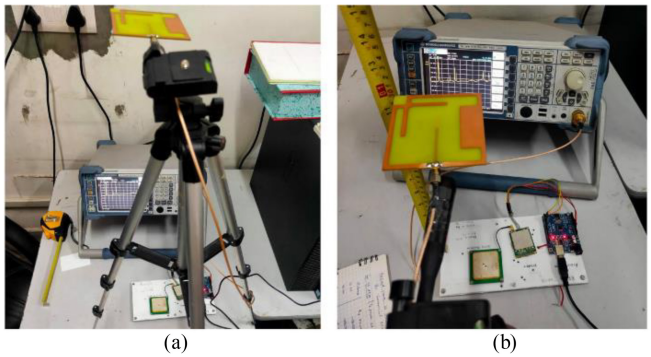


Fig. 16. Polar plot of Read Range of proposed RFID antenna with different transmit powers.



(a) (b)

Fig. 17. Measurement Set up for received power vs Read Range (a)Front view (b)Top view.

measured 2D-map was reported by integrating a linearly polarized Koch-shaped log-periodic dipole array UHF RFID antenna of >3 dBi gain with a Thinkfy-TR50 reader module and a personal digital assistant (PDA) [24]. Such antenna is

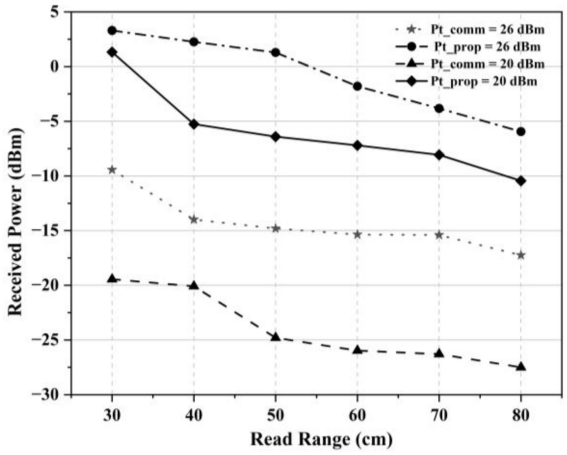


Fig. 18. Received Power vs Read Range for both RFID antennas.

useful in reading multiple tags without rotating the reader axis from $+Z$ to $-Z$ direction.

For measuring the received power in the different read ranges, the R&S®FSL6 Spectrum Analyzer was used alongside a custom-built setup, as shown in Fig. 17. Then, the received power is measured by increasing the distance of the receiving antenna (proposed RFID antenna) from the transmitting antenna (commercial/proposed RFID antenna) and shown in Fig. 18. The proposed UHF RFID antenna showed better power reception for the same read range due to higher gain and bi-directional radiation, so it is more sensitive and performs better at detecting and reading information from tags.

V. Conclusion

In this work, a novel reader antenna is introduced and analyzed experimentally, for universal UHF RFID handheld applications. The proposed UHF antenna features a simple microstrip feedline positioned at the center of a square ground slot. The ground plane is perturbed with a thin, thick rectangular slot at precise locations to achieve circular polarization. Also, due to the geometry of the proposed antenna, bi-directional radiation is achieved ($\pm Z$). The antenna exhibits the measured impedance bandwidth of 16.5% (810.5-959.5 MHz), a 3-dB axial ratio bandwidth of 25 MHz, and a consistent radiation gain of about 5 dBi across the universal UHF RFID band (840–960 MHz). The proposed UHF antenna achieves a maximum read range (MRR) of approximately 245 cm on both sides, which is sufficient for reliable detection. It also has a field of view of almost 360° with varying detection power. Thus, the measured performance metrics confirm that this antenna meets the requirements for universal UHF RFID handheld reader applications, offering wide axial ratio bandwidth and compact size of only $81 \times 81 \times 1.6$. mm^3.

References

[1] A. Ren, C. Wu, Y. Gao, and Y. Yuan, "A robust UHF near-field RFID reader antenna," *IEEE Trans. Antennas Propag.*, vol. 60, no. 4, pp. 1690–1697, Apr. 2012, doi: 10.1109/TAP.2012.2186254.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

[2] J. K. Pakkathillam, M. Kanagasabai, and M. G. N. Alsath, "Compact multiservice UHF RFID reader antenna for near-field and far-field operations," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 149–152, 2017, doi: 10.1109/LAWP.2016.2561960.
[3] A. Birwal, K. Patel, and S. Singh, "Circular slot-based microstrip circularly polarized antenna for 2.45-GHz RFID reader applications," *IEEE J. Radio Freq. Identif.*, vol. 8, pp. 10–18, 2024, doi: 10.1109/JRFID.2024.3354515.
[4] A. Birwal, V. Kaushal, and K. Patel, "Investigation of circularly polarized CPW fed antenna as a 2.45 GHz RFID reader," *IEEE J. Radio Freq. Identif.*, vol. 6, pp. 593–600, 2022, doi: 10.1109/JRFID.2022.3172691.
[5] W.-S. Chen and Y.-C. Huang, "A novel CP antenna for UHF RFID handheld reader," *IEEE Antennas Propag. Mag.*, vol. 55, no. 4, pp. 128–137, Aug. 2013, doi: 10.1109/MAP.2013.6645148.
[6] Z. N. Chen, X. Qing, and H. L. Chung, "A universal UHF RFID reader antenna," *IEEE Trans. Microw. Theory Techn.*, vol. 57, no. 5, pp. 1275–1282, May 2009, doi: 10.1109/TMTT.2009.2017290.
[7] D. Li, Q. Wang, Z. Duan, and W. Cao, "UHF RFID reader antenna for near-field and far-field operations," in *Proc. 13th Global Symp. Millimeter-Waves Terahertz (GSMM)*, Nanjing, China, 2021, pp. 1–3, doi: 10.1109/GSMM53250.2021.9511936.
[8] C. Song and Z. Wu, "A loop antenna array with an enlarged near-field interrogation zone for UHF RFID near-field and far-field applications," *IEEE Antennas Wireless Propag. Lett.*, vol. 21, no. 10, pp. 1985–1989, Oct. 2022, doi: 10.1109/LAWP.2022.3187340.
[9] Nasimuddin, Z. N. Chen, and X. Qing, "Asymmetric-circular shaped slotted microstrip antennas for circular polarization and RFID applications," *IEEE Trans. Antennas Propag.*, vol. 58, no. 12, pp. 3821–3828, Dec. 2010.
[10] Nasimuddin, X. Qing, and Z. N. Chen, "A wideband circularly polarized stacked slotted microstrip patch antennna," *IEEE Antennas Propag. Mag.*, vol. 55, no. 6, pp. 84–99, Dec. 2013.
[11] Q. Liu, J. Shen, H. Liu, Y. Wu, M. Su, and Y. Liu, "Low-cost compact circularly polarized directional antenna for universal UHF RFID handheld reader applications," *IEEE Antennas Wireless Propag. Lett.*, vol. 14, pp. 1326–1329, 2015, doi: 10.1109/LAWP.2015.2404033.
[12] F.-P. Lai, J.-F. Yang, and Y.-S. Chen, "Compact dual-band circularly polarized antenna using double cross dipoles for RFID handheld readers," *IEEE Antennas Wireless Propag. Lett.*, vol. 19, no. 8, pp. 1429–1433, Aug. 2020, doi: 10.1109/LAWP.2020.3004881.
[13] R. Xu et al., "Dual-band circularly polarized antenna with two pairs of crossed-dipoles for RFID reader," *IEEE Trans. Antennas Propag.*, vol. 69, no. 12, pp. 8194–8203, Dec. 2021, doi: 10.1109/TAP.2021.3083827.
[14] W. A. Ahmed and F. Quanyuan, "A novel compact CP antenna with wide axial ratio bandwidth for worldwide UHF RFID handheld reader," *Int. J. Antennas Propag.*, vol. 2019, pp. 1–9, May 2019. doi: 10.1155/2019/2497450.
[15] L. Sun, Y. Li, Z. Zhang, and M. F. Iskander, "Low-cost compact circularly polarized dual-layer PIFA for active RFID reader," *IEEE Trans. Antennas Propag.*, vol. 67, no. 1, pp. 681–686, Jan. 2019, doi: 10.1109/TAP.2018.2880093.
[16] S. Sarkar and B. Gupta, "A dual-band circularly polarized antenna with a dual-band AMC reflector for RFID readers," *IEEE Antennas Wireless Propag. Lett.*, vol. 19, no. 5, pp. 796–800, May 2020, doi: 10.1109/LAWP.2020.2980325.
[17] M. El Bekkali, P. Singh, M. El Bakkali, L. Kansal, and G. K. Sodhi, "A low profile CPW-fed slot antenna for UHF-RFID readers," in *Proc. Int. Conf. Comput. Commun. Inform. (ICCCI)*, Coimbatore, India, 2021, pp. 1–4, doi: 10.1109/ICCCI50826.2021.9402378.
[18] K. Mekki, O. Necibi, W. El May, K. Rabaani, and A. Gharsallah, "Design of a UHF/UWB monopole antenna integrated in an RFID reader board," in *Proc. IEEE 19th Int. Symp. Antenna Technol. Appl. Electromagn. (ANTEM)*, Winnipeg, MB, Canada, 2021, pp. 1–2, doi: 10.1109/ANTEM51107.2021.9518820.
[19] S. Sakouhi, L. Olympio, and A. Gharsallah, "Design of an RFID reader antenna in the UHF band for car park applications," in *Proc. 22nd Mediterr. Microw. Symp. (MMS)*, Sousse, Tunisia, 2023, pp. 1–5, doi: 10.1109/MMS59938.2023.10421170.
[20] "UHF RFID OEM module—RRUHFOEM01," 2023. [Online]. Available: https://rapidradio.co.in/product/uhf-rfid-oem-module-rruhfoem01/
[21] "RFID UHF smart card," 2023. [Online]. Available: https://rapidradio.co.in/product/rfid-uhf-smart-card/
[22] "RFID_implementation_program." 2024. [Online]. Available: https://github.com/Akash-shakya/RFID_Implementation_Program
[23] "EPC information," 2024. [Online]. Available: https://www.epc-rfid.info/
[24] H.-T. Hsu and T.-J. Huang, "A koch-shaped log-periodic dipole array (LPDA) antenna for universal ultra-high-frequency (UHF) radio frequency identification (RFID) handheld reader," *IEEE Trans. Antennas Propag.*, vol. 61, no. 9, pp. 4852–4856, Sep. 2013, doi: 10.1109/TAP.2013.2264451.



Amit Birwal (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Guru Gobind Singh Indraprastha University, New Delhi, India, in 2003, and the M.Tech. degree in microwave electronics from the University of Delhi, India, in 2006, where he is currently pursuing the Ph.D. degree. He worked at various levels in the development and test phases of telecom devices with Ericsson India. He is currently an Assistant Professor with the Department of Electronic Science, University of Delhi South Campus, India. His main research interests are designing and optimizing planar antennas for IoT- based systems, and communications systems.



Akash Shakya received the B.Sc. degree in electronics from the Bhaskaracharya College of Applied Sciences, University of Delhi, India, in 2020, and the M.Sc. degree in electronics from the University of Delhi, South Campus, India, in 2024. He worked on various embedded projects including home automation and health monitoring. His research interests include Embedded systems using arduino, raspberry Pi, Internet of Things, and signal processing.



Saurav received the B.Sc. degree in electronics from the Rajdhani College, University of Delhi, India, in 2022, and the M.Sc. degree in electronics from the University of Delhi, South Campus, India, in 2024. His research interests include Embedded systems using arduino, antenna designing, and optimization.



Shalini Kashyap (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Maharishi Dayanand University, Rohtak, India, in 2012, and the M.Tech. degree in electronics and instrumentation engineering from the YMCA University of Science and Technology, Faridabad, India, in 2015. She is currently pursuing the Ph.D. degree with the Department of Electronic Science, University of Delhi, South Campus, India. She worked as an Assistant Professor with the Satya College of Engineering and Technology, Palwal, India. Her main research interests are signal processing in RADAR systems, Internet of Things-based systems, and communications systems.



Kamlesh Patel (Senior Member, IEEE) received the M.Sc. degree in electronics from Rani Durgavati Vishwavidyalaya, Jabalpur, in 1999, and the M.Tech. degree in microwave electronics from the University of Delhi, India, in 2003. He is currently pursuing the Ph.D. degree with Delhi Technological University, Delhi. He was a Scientist with CSIR-National Physical Laboratory from 2004 to 2013. He is currently an Associate Professor with the Department of Electronic Science, University of Delhi South Campus, New Delhi. His research interests include microwave components, material characterization, and planar antennas for mobile communications.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:30:22 UTC from IEEE Xplore. Restrictions apply.

Clean Technologies and Environmental Policy
https://doi.org/10.1007/s10098-024-03017-4

ORIGINAL PAPER

A comprehensive analysis on cost, combustion, performance and emissions for diesel engine fueled with preheated esters of waste cooking oil

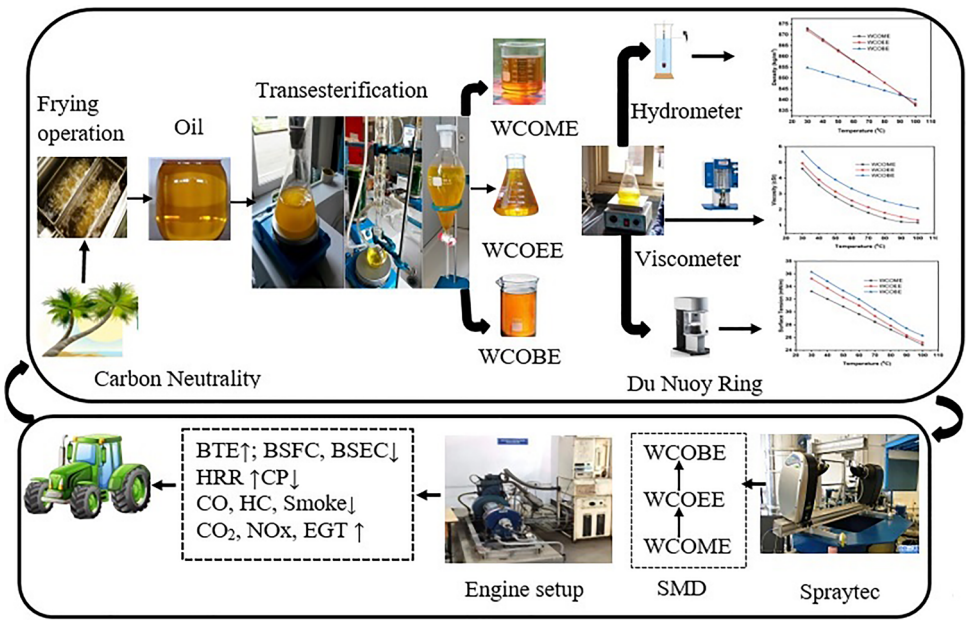
Prem Shanker Yadav[1,2] · Saket Kumar[2,3] · Raghvendra Gautam[2]

Received: 18 July 2024 / Accepted: 7 September 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

The current study examines the suitability of biodiesel made from used cooking oil (WCO) utilizing heated methyl ester, ethyl ester and butyl ester as alternative fuels for combustion ignition engines. However, there has been limited investigation into the butyl ester of waste cooking oil (WCOBE) as a fuel in both unheated and preheated circumstances. The economic feasibility of biodiesel production from WCO using methanol, ethanol and butanol was analyzed. The findings indicated that heated butyl ester and ethyl ester of WCO were produced at a cost that was 44.71% and 42.5% less than diesel fuel, respectively, for 1 kWh of energy. Further, the Sauter mean diameter was measured at temperatures between 30 and 100 °C using the Malvern Spraytec laser diffraction-based system. In the temperature range found, methyl ester had a smaller Sauter mean diameter than ethyl ester and butyl ester. Furthermore, the engine characteristics showed that preheated butyl ester exhibited a peak pressure that was 0.49% higher than unheated, whereas preheated ethyl ester showed a peak pressure that was 4.7% lower. Preheated esters emitted less CO, HC and smoke during peak load than unheated esters but more NO_x and CO_2 during peak load. The detail comparative results are included at the end of results and discussion.

Graphical Abstract



Keywords Biodiesel · Sauter mean diameter · Cost analysis · Diesel engine

Extended author information available on the last page of the article

Published online: 18 September 2024

Abbreviations

A_BDC	After bottom dead center
A_TDC	After bottom dead center
B_TDC	Before top dead center
B_BDC	Before bottom dead center
BMEP	Break mean effective pressure
BSEC	Brake-specific energy consumption
BSFC	Brake-specific fuel consumption
BTE	Brake thermal efficiency
CN	Cetane number
CO	Carbon monoxide
CO_2	Carbon dioxide
CP	Cylinder pressure
cSt	Centistoke
CV	Calorific value
DF	Diesel fuel
FFA	Free fatty acid
HC	Unburnt hydrocarbon
Hp	Horse power
HSU	Hartridge smoke unit
kWh	Kilowatt-hour
mN/m	Milli newton per meter
NO_x	Oxides of nitrogen
ppm	Part per minute
PTSA	P-toluenesulfonic acid
PWCOBE	Preheated waste cooking oil butyl ester
PWCOEE	Preheated waste cooking oil ethyl ester
PWCOME	Preheated waste cooking oil methyl ester
RPM	Revolution per minute
Rs	Indian currency (Rupee)
SMD	Sauter mean diameter
WCO	Waste cooking oil
WCOBE	Waste cooking oil butyl ester
WCOEE	Waste cooking oil ethyl ester
WCOME	Waste cooking oil methyl ester
w/w	Weight in weight

Greek letter

ρ	Density (kg/m^3)
μ	Viscosity (cSt)

Introduction

Energy requirements in the domains of agriculture, industry and transportation are incessantly augmenting; nonetheless, population growth and industrialization are slackening fossil fuel reserves and other concerned resources, paving a way for profound surge in petroleum prices (Yadav et al. 2018). Consequently, it is undeniably an immense challenge for researchers and scientists to find an alternative fuel to replace conventional options. Biodiesel, derived from locally accessible feedstock, has gained significant traction as a viable option for existing diesel engines. Its utilization brings multiple benefits to agriculture, as it is biodegradable and contributes to the growth of rural economies (Qiu et al. 2016). Moreover, being a renewable source, it offers a way to mitigate emissions of carbon and other harmful pollutants, such as hydrocarbons (HCs), carbon monoxide (CO), particulate matter (PM), benzene, ethyl, toluene, xylene (BTEX) and more (Agarwal 2007). The global acceptance of biodiesel can be attributed to its advantages over diesel fuel, such as improved lubricity, high cetane number, reduced aromatics and minimal sulfur content, and with an oxygen content of around 10–12% by weight, biodiesel effectively decreases the emission of harmful pollutants (Rakopoulos et al. 2006). Challenges to expected scientific progress in biodiesel production include the injector deposits, filter plugging, increase in NO_x emissions, (Yadav et al. 2021). Biodiesel can be produced from various feedstocks, including waste cooking oils, non-edible oils, animal fats and edible oils. However, currently, edible oil accounts for 95% of biodiesel production, which exacerbates raises production costs and food scarcity concerns. Consequently, it becomes imperative to pinpoint a feasible raw material for profitable production in the primary regions (Baskar and Aiswarya 2016). The biofuels derived from the transesterification reaction using methanol, ethanol and butanol are referred to as methyl ester, ethyl ester and butyl ester, respectively, and these esters possess distinct physicochemical properties. The rheological properties (density and viscosity) of esters influence the flow, pumping, energy density, air–fuel ratio, lubrications and most importantly spray quality. Butyl esters exhibit higher viscosity than methyl esters and ethyl esters and little less value of density. The atomization characteristics of fuel affects combustion, performance and emissions of diesel engine. The diesel engine exhibits performance issues when using alcohols due to their high latent heat of evaporation, low cetane number and low heat value. Therefore, alcohols and their components should be mixed with diesel fuel in appropriate proportions. The hydroxyl group (–OH) in alcohols is hydrophilic and polar, leading to phase separation at low temperatures. This phase separation can cause damage to the fuel system and storage components over the long term (Atmanli and Yilmaz 2018).

Han et al. (2014) examined the single injection parameters of ethyl oleate and methyl oleate in CRDI engine. The result showed that ethyl oleate and methyl oleate exhibited differences in mass cycle injection quantities with respect to injection pressure and injector energizing time. Diesel fuel was lower injection delay values than those of alkyl esters. They observed that methyl oleate showed smaller injection delay than ethyl oleate. Moreover, with the same energizing time and injection pressure, ethyl oleate and methyl oleate were injected more amount than diesel fuel. This is due to high densities of esters. Yadav et al. (2023) did investigation

Table 1 Comprehensive literature survey comparing the ethyl esters to methyl esters at maximum load

References	Biodiesel percent (vol.%)	Details of the Engine Under Test	Testing parameters	Performance		Emission			
				BSFC (kg/kW.h)	BTE (%)	NO_x (ppm)	HC (ppm)	PM (HSU)	CO (vol.% or ppm)
Alptekin et al. 2019)	WCO 100%	DI, 4S, RS: 4000 rpm, CR: 18.45: 1, RP: 77 kW	RPM: 1500 2000 2500	↓0.1% ↓0.7% ↓1.9%	↑ ↑ ↑	–	–	–	–
Baiju et al. 2009)	Karanja 20%	1C, 4S, WC, CR: 17.5:1, RS: 1500 rpm, RP: 5.9 kW	Load%:0 20 40 60 80 100	↑ ↑ ↑ ↑ ↑ ↑	↑ ↑ ↓ ↓ ↑ ↑	↑ ↑ ↑ ↑ ↑ ↑	–	–	↑ ↑ ↑ ↑ ↑ ↑
Lapuerta et al. 2008)	WCO 100%	4C, RS: 4000 rpm, 4S, CR: 18:1 CRDI, RP: 82 kW	BMEP: 2 4 6	↓ → →	↑ ↑ ↑	↓ ↓ ↓	↓ ↓ ↓	↓ ↓ ↑	NA
Makareviciene and Janulis 2003)	Rapseed oil 100%	4C, DI Diesel Audi 80 engine	Ester %: 25 50 75 100	–	–	↓ ↓ ↓ ↓	–	↑ ↑ ↑ ↑	↓ ↓ ↓ ↓
Sanli et al. 2015)	WCO 100%	Turbocharged, 6C, 4S, CR: 16.4:1, RS: 2400 rpm, RP: 136 kW	RPM: 1100 1400 1600	↓ ↓ ↓	↑ ↑ ↑	↑ ↓ ↓	↓ ↓ ↓	–	↓ ↓ ↓
Chavan et al. 2015)	WCO 100%	1C, 4S, RS: 1500 rpm, WC, CR: 17.5:1 RP: 3.5 kW,	Load (0-14 kg), CR (14,15,16,17,18)	–	–	↑ ↑ ↓ ↑	↓ ↓ ↓ ↓	–	↓ ↓ ↓ ↓ →
Jindal et al. 2010)	Jatropha 100%	1C, 4S, RP: 3.5 kW, RS: 1500 rpm, WC, CR: 17.5:1,	Load (kg): 3 6 9 12 15	↓ ↓ ↓ ↓ ↓	↑ ↑ ↑ ↑ ↑	↑ ↑ ↑ ↑ ↑	→ ↓	–	↓ ↓ ↓ ↓ ↓
Eliçin et al. 2019)	Rapeseed oil 100%	1C, WC, CR: 17.5:1 4S, RP: 5.5 kW	RPM:1500	↑	–	→	→	→	→

↑: Increase, ↓: Decrease, →: Same, RS: Rated speed, CR: Compression ratio, 1-cylinder: 1C, 4S: Four-stroke, RP: rated power, WC: water-cooled

on atomization characteristics of JOME and JOEE at different ambient pressure. They reported that Sauter mean diameter of JOEE was higher than JOME with ambient pressures. Ethyl esters, for instance, exhibit significantly lower iodine content, improved lubricity, higher oxidation stability and cold filter plugging, pour point and lower cloud point compared to methyl esters (Mendow et al. 2011; Reyero et al. 2015). The studies (Coniglio et al. 2013; May et al. 2005) suggested that both methyl ester and ethyl esters of any oil serve better in performance and emissions with low blend ratio (below B20). However, these low blend fuels are compatible with diesel engines without the need for any fuel modifications. These limitations are due to high rheological properties and low calorific value. The detailed review of ethyl esters versus methyl aster on rheological properties, combustion, performance and emissions is presented in Table 1. Similarly, butanol overcomes the limitations of lower alcohols by offering advantages such as lower volatility, higher heating value, enhanced safety, fewer ignition problems and enhanced inter-solubility (Ezeji et al. 2007; Dürre 2008). It exhibits a higher oxygen content, resulting in reduced soot formation. Additionally, its high heat of evaporation contributes to lower NO_x emissions and a subsequent decrease in combustion temperature (Rakopoulos et al. 2010). These mentioned advantages stimulate the substantial utilization of bio-based ethanol and butanol in biodiesel production and their subsequent utilization in combustion ignition engines. Despite the significant benefits of bio-ethanol, butanol with higher viscosity when blended with vegetable oil results in poor atomization, larger droplet

formation, heterogeneous air-fuel mixing, and reduced volatilization. These factors hinder its effectiveness in achieving better combustion, resulting in increased levels of HC, CO and carbon residue on the cylinder head, combustion chamber wall, piston, valves and injector nose (Hoang and Le 2019; Sidibé et al. 2010). These advantages demonstrate that butanol offers greater benefits compared to widely used biodiesel and ethanol. There are four methods to mitigate viscosity, namely heating, blending, emulsification and transesterification. Among these methods, heating is a cost-effective approach to mitigate the aforementioned issues. Preheating biodiesel has shown a significant decrease in viscosity and improved spray characteristics, performance, efficacy and combustion properties. Researchers have expressed optimism regarding the use of ethyl esters in diesel engines compared to methyl esters (Alptekin et al. 2019; Baiju et al. 2009; Lapuerta et al. 2008; Makareviciene and Janulis 2003; Sanli et al. 2015; Chavan et al. 2015; Jindal et al. 2010; Eliçin et al. 2019).

Novelty and outline of the current research

From this comprehensive review, it becomes evident that there are two significant research gaps noticed.

a. No previous studies have explored the effects of butyl ester of waste cooking oil on atomization, combustion, performance and emissions at diesel engine conditions. These analyses serve as contemporary tools for evaluating the comparative analysis among methyl ester, ethyl ester and butyl ester.
b. There is a notable absence of research on cost assessments per kilowatt-hour and the effect of unheated and preheated esters on combustion, performance and emission characteristics.

The study investigated the impact of preheating on viscosity, surface tension and density as well as the atomization characteristics of unheated ethyl ester, butyl ester and methyl ester of waste cooking oil. Furthermore, the engine characteristics of all esters (both unheated and preheated) were evaluated and compared with those of mineral diesel under varying brake mean effective pressure. The present study concludes by summarizing the previous findings under comparable conditions.

Materials and methods

Preparation of WCOME, WCOEE and WCOBE

The first crucial step in biodiesel production involves accurately determining the free fatty acid (FFA) content of the oil. Based on the FFA content (2 wt% ≥ FFA) determined during this initial step, the subsequent process is determined. When the FFA level is greater than 2 wt%, the esterification reaction is necessary to prevent saponification before proceeding to the transesterification process. Two key steps in the production process have been successfully completed:

1. Esterification and
2. Transesterification, along with an alkaline catalyst.

The biofuel generation from waste cooking oil with substantial FFA has also been explored in a study conducted by Sinha et al. (2008)

Acid-catalyzed esterification

The procedure entails the reaction of alcohol with glyceride (WCO oil), where the alcohol functions as a reactive component in the presence of the acid catalyst p-toluenesulfonic acid (PTSA) to mitigate the excess free fatty acid (FFA) content. To expedite the reaction time and enhance the reaction rate, the filtered oil is dehydrated. Initially, the oil was subjected to a heating step, reaching a temperature of around 120 °C, maintained for a duration of 30–40 min and then allowed to cool down to 60 °C. The purpose of this step is to expedite the processing rate and facilitate the evaporation of the moisture present in the filtered oil. Afterward, a premixed solution of methanol and catalyst is introduced to the oil. The resulting mixture is carefully stirred for around 2 h using a magnetic stirrer rotating at approximately 400 rpm. The alcohol-to-oil ratio is maintained at 6M:1, and the catalyst concentration is set at 0.5% (w/w). At the end of the reaction, the final product comprises impurities, excess methanol, catalyst and a mixture of esterified oil. These components are subsequently separated by allowing

Table 2 Transesterification parameters

Parameter	Butanol	Ethanol	Methanol
Catalyst concentration ratio %w/w of NaOH	1.5	1.5	0.5
Time of reaction	180 min	150 min	60 min
Temperature	70 °C	65 °C	60 °C
Methanol–oil ratio	15:1	12:1	6:1
Yield	88.5%	96.9%	98.4%

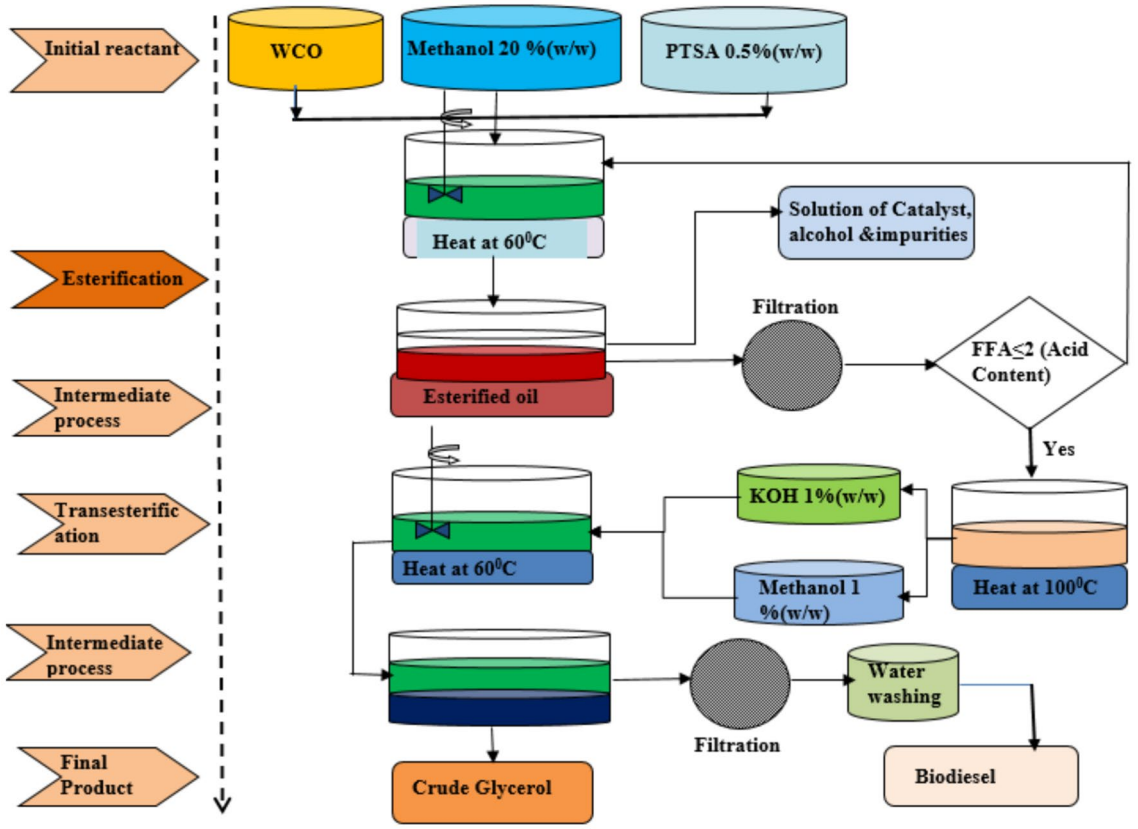


Fig. 1 Flowchart for generation of biodiesel (Yadav et al. 2023)

Table 3 Some fuel properties of butyl ester, ethyl ester and methyl ester of WCO

Property	Unit	WCOBE	WCOEE	WCOME	DF	EN 14214
Fatty acid methyl ester content	%(mass)	–	–	98.4	–	60 min
Viscosity (20 °C)	mm^2/s	5.5	5.1	4.6	2.98	3.5–5.0
Density (20 °C)	Kg/m^3	878.9	883.2	886.5	832.7	860–900
Cetane number	–	52	52	54	50	51 min
Flash point	°C	162	158	153	63	101 min
Sulfur content	mg/kg	6.3	4.5	2.4	5.7	10.0 max
Copper strip corrosion (3 h, 50 °C)	Degree of corrosivity	No 1	No 1	No 1	No 1	No 1 max
Iodine value	g I_2/100 g	110.01	107.11	103.01	9.82	120 max
Water content	mg/kg	406.5	417.8	430.2	48.83	500 max
Acid value	mg KOH/g	0.36	0.37	0.37	0.10	0.50 max
Linolenic acid methyl ester content	% (mass)	–	–	0.42	–	12.0 max
Methanol content	% (mass)	–	–	–	–	0.20 max
Monoglyceride	% (mass)	0.34	0.27	0.16	–	0.70 max
Diglyceride	% (mass)	0.15	0.11	0.08	–	0.20 max
Triglyceride	% (mass)	0.05	0.06	0.07	–	0.20 max
Free glycerin	% (mass)	0.08	0.08	0.08	–	0.20 max
Total glycerin	% (mass)	0.102	0.098	0.066	–	0.25 max
Cold filter plugging point	°C	6	5	−3	−15	–
Higher heating value	kJ/kg	40,213	40,051	39,869	45,962	–
Lower heating value	kJ/kg	37,680	37,550	37,334	43,213	–

the mixture to settle in a separate funnel for approximately 5 h. The oil, which contained water and excess alcohol, underwent a heating process using a rotary evaporator at a temperature of 100 °C for 1 h under vacuum conditions. Following this, the esterified oil was found to have an FFA content of 1.85%. The result aligned with the expected range derived from theoretical predictions. This confirmed that the subsequent transesterification process could proceed.

Base-catalyzed esterification

The final step involved the base-catalyzed transesterification to produce biodiesel. This reaction was relatively simple to perform since methanol had already been used to decrease the FFA level. A mixture was formed by combining the oil obtained from diverse esterification processes with a 20% w/w solution of methanol, 1% w/w of KOH pellets, with a ratio of catalyst to waste cooking oil considered to be 1:6M. The mixture was then heated at 1000 rpm for 120 min and transferred to a funnel for 12 h. In the final step, the end product obtained was waste cooking oil methyl ester (WCOME), together with unrefined glycerol. Due to its higher weight compared to the oil, the crude glycerol settled at the base and was subsequently filtered out. After the base-catalyzed transesterification process, the remaining esters underwent a cleaning process using hot water to obtain purified biodiesel. The reaction parameters for methanol, ethanol and butanol are given in Table 2. Additionally, a

Table 4 Production expenditure of methyl ester, ethyl ester and butyl ester of waste cooking oil as per Indian planning commission

Items	Value in Rs/ltr of WCO (10 ltr)	Value in Rs/ltr of WCO (10 ltr)	
a. Waste cooking oil in restaurant	13	21	https://auto.economictimes.indiatimes.com/news/oil-and-lubes/zomato-to-collect-used-cooking-oil-for-biodiesel/71691599 and Gupta and Agarwal (2016)
b. The logistics, coupled with the transportation			
i. Packaging and gathering	6.74	6.74	Plastic water bottle 10 L, INR 65 of market price
ii. Quality assessment	1.5	1.5	
iii. The movement of goods from the restaurant to the warehouse	12.24	12.24	The present price for transporting is INR 4 per liter
iv. The expenses associated with holding or storing	6.37	6.37	Carrying cost per liter of oil INR 5
Sum (a + b)	39.85	47.85	
c. Employee wage	3.39	3.39	The combined cost of managerial unskilled and administrative expenses is considered as INR 1 per liter for the entire alteration process
d. Transformation from a chemical	8, 11.31,12.63	8, 11.31,12.63	0.22 l methanol, ethanol and butanol with 0.0061 kg KOH utilized to produce 1 l each
e. Power consumption	9.00	9.00	Taking into account an electricity consumption of 0.84 units per 1 L of biodiesel
f. Downgrading in machinery and plant	2.6	2.6	Considering a 15% rate applied to the overall value
g. The interest cost on overall working capital	2.59	2.59	Annual rate of 15%
The ultimate processing cost (c + …. + g)	25.63, 28.94,30.26	25.63, 28.94,30.26	Here, the aggregate processing expense using methanol, ethanol and butanol is 25.63, 28.94 and 30.26, respectively
Total cost of biodiesel (a + …. + g)	65.47,68.78,70.10	73.48, 76.79,78.11	
Earning from by-products			
i) Gain from cake oil	15.00	15.00	Yadav and Sharma 2019; Singhal and Sengupta 2012)
ii) Revenue from glycerol	4.80	4.80	
Entire return from enduring product (i + ii)	19.80	19.80	
The aggregate price of WCOME, WCOEE, WCOBE	45.67, 48.98,50.30	53.68, 56.99, 58.31	

Table 5 Cost assessment of butyl ester, ethyl ester and methyl ester of WCO with diesel fuel

Parameters	Diesel	WCOME	WCOEE	WCOBE	PWCOME	PWCOEE	PWCOBE
Price per liter of fuel (INR)	89.89	53.68	56.99	58.31	–	–	–
ρ (kg/m^3)	830	881	872	867	–	–	–
i) Price (per kg)	108.30	60.93	65.36	67.25	–	–	–
The amount of heat needed to raise the fuel temperature to 80 °C (kJ)	–	–	–	–	63.8	64.35	65.45
ii) Per kg heating expense (INR)	–	–	–	–	0.124	0.125	0.127
Total expense post-heating (Rs/kg) (i + ii)	108.30	60.93	65.36	67.25	61.054	65.485	67.397
BSFC (kg/kWh)	0.185	0.2299	0.20	0.215	0.194	0.171	0.16

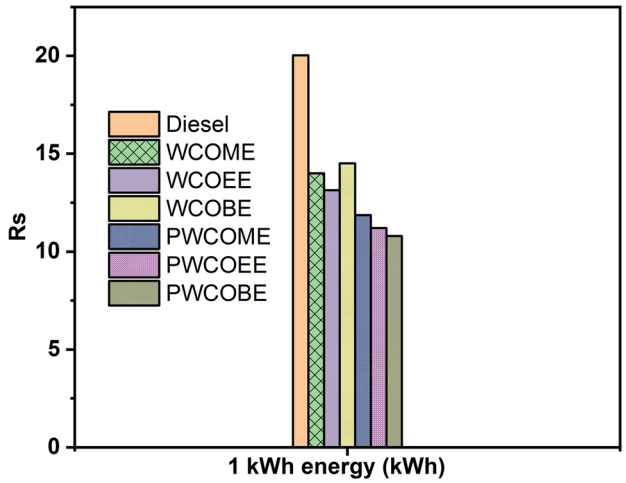


Fig. 2 INR versus 1 kWh

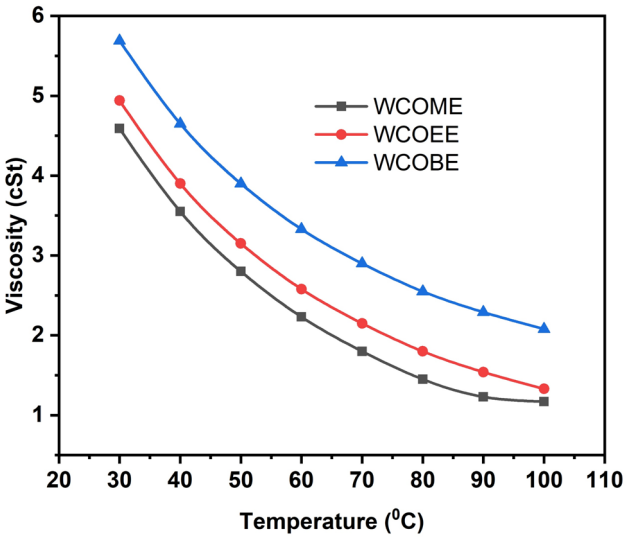


Fig. 3 Viscosity (cSt) versus temperature (°C)

parallel process was implemented to produce waste cooking oil ethyl esters (WCOEE) and waste cooking oil butyl esters (WCOBE). The biodiesel production schema is illustrated in Fig. 1.

The fuel qualities of the produced methyl, ethyl and butyl esters of waste cooking oil were tested using ASTM limits. Table 3 presents the specific qualities used for comparison. Upon analyzing the properties of the three ester types, it becomes evident that methyl esters exhibit superior qualities in comparison with the others. However, both ethyl and butyl esters still meet the specified parameters. It is worth mentioning that butyl esters have the highest viscosity among the three, which may potentially result in inconsistent spray generation. Despite minor deviations in a few attributes, the overall characteristics indicate that the biodiesel samples produced conform to the ASTM limitations. Therefore, these kinds of fuels have the potential to be regarded as viable alternatives for diesel engines.

Cost analysis

Table 4 displays the respective prices per liter for WCOME, WCOEE and WCOBE, which are INR 53.68, INR 56.99 and INR 58.31. It is important to note that there is a heating process involved in raising the temperature of WCOME, WCOEE and WCOBE from 25 to 80 °C. The extra energy loss contributes to an increase in prices when compared to non-heated fuel options. The calculation of the heat necessary to elevate the temperature of 1kg of WCOME from 25 to 80 °C can be determined using Eq. 1 (Ben Jabeur et al. 2018). Additionally, for the purpose of calculation, the assumed rate for the power tariff is INR 7 per kilowatt-hour (kWh) of electricity.

$$Q = mc_p\Delta t \tag{1}$$

where T, Cp and m are temperature in kelvin, specific heat in kJ/kg-k and mass of fuel in kg.

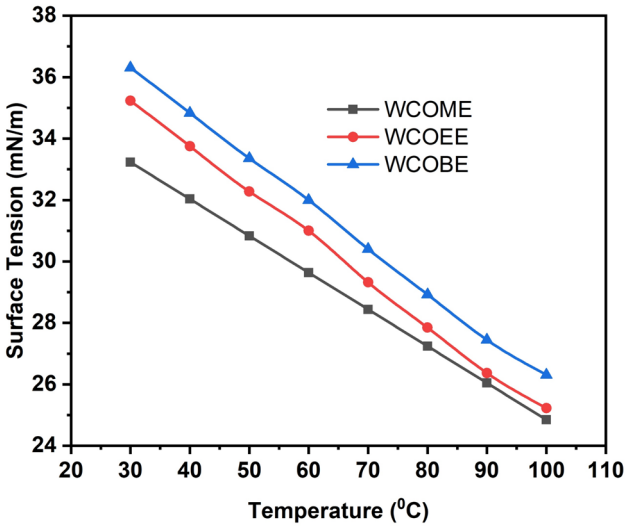


Fig. 4 Surface tension (mN/m) versus temperature (°C)

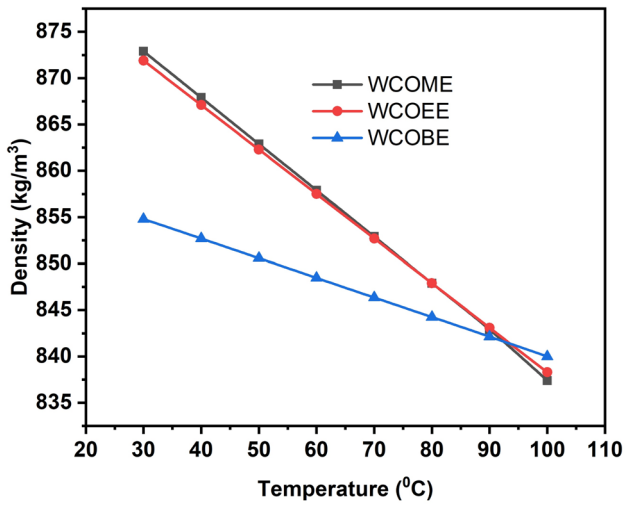


Fig. 5 Density (kg/m^3) versus temperature (°C)

The calculation of costs associated with generating 1 kilowatt-hour (kWh) of energy output, in accordance with Indian regulations, is as follows:

The authors used a C-Therm thermal effusivity sensor to test the specific heat of WCOME, WCOEE and WCOBE in order to calculate cp. These measurements revealed that the respective values were 1.16 kJ/kg/K, 1.17 kJ/kg/K and 1.19 KJ/kg/K. Furthermore, the theoretical analysis (Venu et al. 2020) used to validate the collected data revealed that they were 1.206 kJ/kg/K, 1.214 kJ/kg/K and 1.239 kJ/kg/K, respectively, with acceptable errors of 4.2%, 3.8% and 4.4% for further research. Here, the cost of used cooking oil has been calculated per liter using the standard rate established by restaurants. The price of diesel is highest in Table 5, while the price of WCOME is lowest and includes the price of heating oil. But diesel fuel has the highest cost per kWh of energy produced, and PWCOBE has the lowest as shown in Fig. 2. The use of either PWCOBE or PWCOEE fuel in an unmodified diesel engine can be a superior alternative to diesel fuel for agricultural applications, according to the authors.

Methods

The conventional processes employed earlier to improve the spray characteristics of fuel are like hydrotreating, heating of biofuel, blending with low viscous fuels etc. Among them, preheating is very easy and economical, and this process exhibits less change in chemical characteristics. Therefore, it is more suitable for unmodified diesel engines in agricultural-driven countries like India. The heating of butyl ester, ethyl ester and methyl ester of waste cooking oil results in decrease in the viscosity, surface tension and density to standards value of fuel. The measurement of viscosity, surface tension and density were taken using hydrometer having rage of 600–1000 kg/m^3, a cannon viscometer with tube with constant $k = 0.0328$ and Du Nouy ring method, respectively, from temperature range of 20–100 °C. The average of five trails of the experiment has considered with accuracy up to five decimal and error of 2.14%. The variations of viscosity, surface tension and density with temperature are presented in Figs. 3, 4 and 5, respectively. The results indicated that all rheological property of esters reduces with an increase in the temperature, and at 95 °C, the density of WCOME, WCOEE and WCOBE was 840.4 kg/m^3, 841.1 kg/m^3 and 839.63 kg/m^3, respectively. The viscosity reduced to 1.35cSt, 1.65cSt and 2.29cSt for WCOME, WCOEE and WCOBE, respectively. At higher temperatures, the increased molecular movement overcomes the attractive forces that cause resistance to flow. These forces include cohesive forces within the fluid, which are weaker when molecules are moving faster, and as temperature elevates, the kinetic energy of the molecules elevates too. This higher energy allows the molecules to move more freely and rapidly, reducing the strength of the intermolecular forces (like van der Waals forces) that contribute to viscosity (Kodate et al. 2020). Similarly, surface tension exhibited reduction to 25.045 nN/m, 26.03 mN/m and 26.43 mN/m for WCOME, WCOEE and WCOBE, respectively. Density, viscosity and surface tension of unheated diesel are 832.7 kg/m^3, 2.98 cSt and 28.89 mN/m which are utilized in engine testing. Similar observation is also suggested by study (Vedharaj et al. 2015). Hence, the experimental examination of atomization, combustion, performance and emission characteristics was performed at 95 °C of fuel temperature. Further, a comparative evaluation of unheated and preheated methyl ester, ethyl

ester, butyl ester of waste cooking oil and diesel fuel was evaluated using diesel engine.

Atomization characteristics

The setup used for the experiment involved a constant volume cylinder with measurements of 60 cm in diameter and 100 cm in length. Additionally, a common rail system was employed to transfer fuel from a high-pressure pump to the injector. To facilitate visual observation of the spray, a 150 mm optical circular window was incorporated into the stainless steel visual pressure chamber. The ambient pressure could be adjusted using the intake valve, while the ambient temperature was measured using a thermocouple. In the current study, a single-hole solenoid was utilized to actuate the fuel injector. The same-sized injector was also employed in the engine trial to deliver fuel, with a common rail system (CRI 3-18) used to control the injection

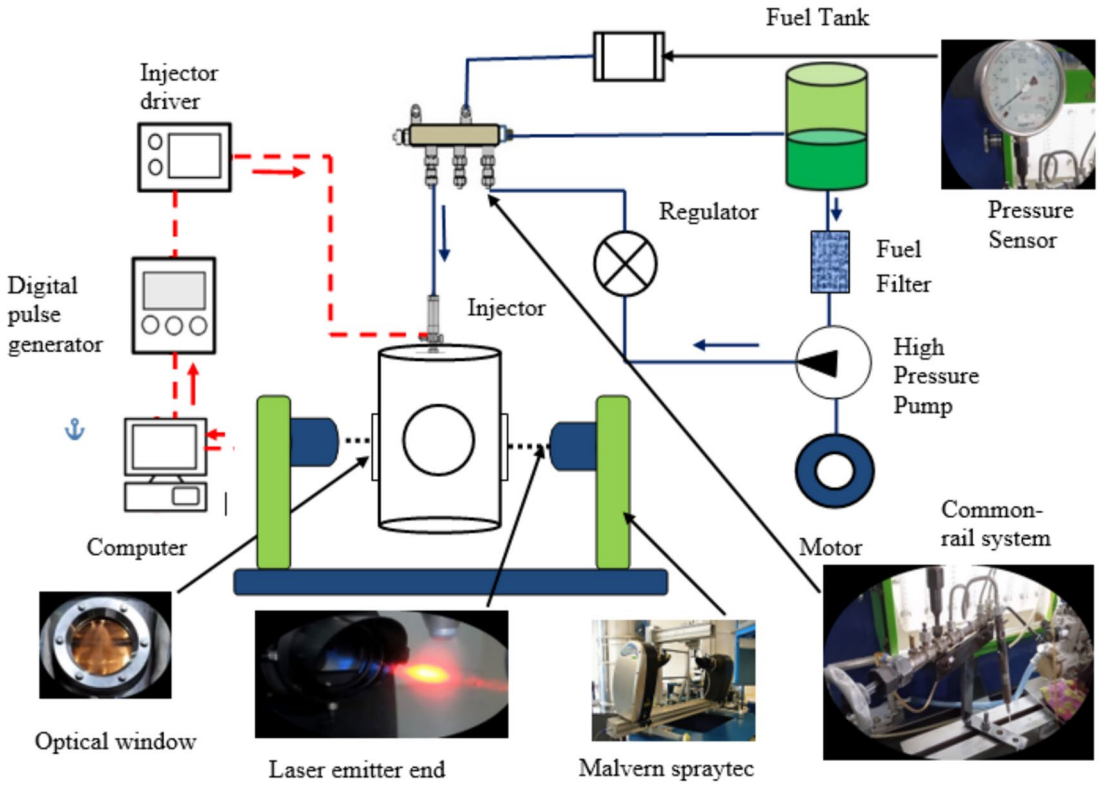


Fig. 6 Experimental configuration for spray analysis (Yadav et al. 2023)

Table 6 Malvern Spraytec instrument is characterized by the following specifications

Measurement specifications	Laser diffraction technique
Size constraint	0.1–200 µm
Current models	Fraunhofer approximation and Mie theory integrated with patented multiple scattering rectification
Lens capacity	750 mm: 2.0–2000 µm (Dv: 4 to 1500 µm)
Sensor	Detector array consisting of 35 log-spaced silicon diodes
Source of light	634.2 nm, 6 nW neon–helium laser
Operating scope	0.5 µm at 100 mm ˃ 1 m at 10 µm
Peak rate of acquisition	Regular mode: 1 Hz as fast mode:, 2.5 kHz as standard:, 10 kHz additional software feature key
Concentration range	Minimum transmission: 5%
Accuracy	˃ ± 1% on the Dv50
Repeatability/precision	˃ ± 1% at the Dv50

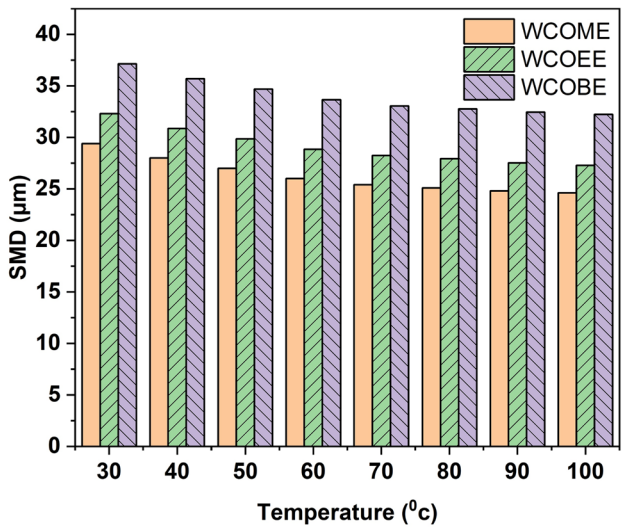


Fig. 7 SMD versus temperature (°C)

pressure. The activation of fuel injection was managed by a pulse generator/delay, which served the dual purpose of signaling the injector driver and controlling the opening and closing of the injector valve. The atomization characteristics of the test fuel were assessed using the Malvern Spraytec instrument, which functions using laser diffraction principles. This instrument measures the size and concentration of secondary atomized particles as they pass through the laser beam. The assumption is made that the droplets are spherical based on the scattering of the beam. It is important to consider that droplet size varies across the fuel jet's cross section in relation to the beam, owing to the spray pattern created by the injection system positioned between the beam and injector. To capture the maximum potential secondary atomization, the injector was adjusted accordingly within the constant volume chamber. The injection pressure was set to 220 bar to align with the fuel supply, as the engine used in the experiment was a field engine with a 220 bar injection pressure. The high-pressure pump, driven by a V-belted motor, is responsible for forcing the fuel into the common rail. Further, the injector driver connected the high-pressure pipeline to the injector, enabling fuel delivery at the desired pressure. Each test fuel's spray characteristics were predicted to persist for 200 µs, with a 12-µs injection duration. To allow for the full development of secondary atomization and facilitate clear comparisons, the data were recorded at half of the injection period (6 ms). Additionally, five trials of each test were conducted under low light conditions to minimize the impact of sunlight, and the final values were averaged. The Malvern Spraytec instrument utilized in this study has a size range of 0.1–2000 µm and provides continuous and rapid measurements. Figure 6 provides a depiction of the experimental setup, while relevant information from the Malvern Spraytec instrument is given in Table 6. An experiment was carried out to investigate the effects of fuel temperature (ranging from 30 to 100 °C) on diesel engine characteristics. The size of fuel droplets at a specific cross-sectional time can be used to calculate the SMD, which, in turn, can impact on vaporization rate and fuel combustion. Smaller SMD values generally result in higher vaporization rate and combustion. The data recorded at half of the injection period of the fuel are significant parameters that impact the size of droplets.

Figure 7 illustrates the changes in SMD for esters of waste cooking oil as a function of fuel temperature. At all fuel temperatures, WCOBE exhibited larger droplet sizes compared to the other test fuels. This results from its higher surface tension and viscosity, which hinder the generation of smaller droplets. As the fuel temperature increased, the SMD of all test fuels decreased. This can be attributed to the increased air drag, which overpowered the surface energy of the droplets. Moreover, this was noted that preheated methyl ester,

Table 7 Uncertainty, range and accuracy of Instruments

Device parameters/specification	Percentage uncertainties	Range	Accuracy
Pressure sensor	±0.45	0–200,000 psi	±0.45%
Motor	±0.20	0–5 hp	
Injector	±0.036	0.18 mm	
Pressure gauge	±0.45	0–100 MPa	1%
Injector driver	±0.17		±0.16%
Injection system	±0.30	0–220 MPa	±15 Pa
Fuel Filter	±.25	0–6.5 microns	±0.1
Fuel pump	±0.15	0–200 MPa	±10 Pa
Thermocouple	±0.25	95–9000c	0.70%
Pulse generator	±0.01	1 mHz to 1.0 MHz	±1%
Kinematic viscosity	±0.35	0.2–20,000 mm^2/s	
Sauter mean diameter	±0.30		±0.01%

ethyl ester and butyl ester resulted in smaller droplet sizes compared to their non-preheated counterparts. This can be explained by the fact that preheating reduced the surface tension, viscosity and density of the biodiesel. These findings align with previous studies (Hoang and Le 2019; Anis and Budiandono 2019), which have reported that biodiesel tends to yield larger SMD values.

<>The calculated parameters may be inaccurate due to changes in the configuration and instrument selection during testing, as well as variations in climatic conditions. Measurement errors can be categorized into random errors and bias errors, which can occur for various reasons. Bias errors, which are deviations not caused by chance, can be analyzed using calibration processes or past results. Random errors, on the other hand, result from variability and random variation, and can be decreased by performing more sample calculations. Methodological equations are used to assess experimental data errors. To evaluate experimental mistakes, instruments, sensors and machines must be utilized along with estimation and computations. Uncertainty analysis estimates the true value by taking the mean of repeated measurements. In this study, six



Fig. 8 Engine setup. Schematic diagram of engine setup

experiments were performed for a specific parameter, and the average was considered to assess errors. Various unknown uncertainties were identified using mathematical equations and established parameters. The uncertainty results were calculated using Holman's square root method, as demonstrated in Eq. 2 (Holman 2001). Table 7 shows the uncertainty values for various instruments, along with their range and accuracy.

$$U_E = \left\{ \left[\left(\frac{\partial E}{\partial X_1} \right) u_1 \right]^2 + \left(\frac{\partial E}{\partial X_2} \right) u_2 \right]^2 + \left(\frac{\partial E}{\partial X_3} \right) u_3 \Big]^2 + \cdots + \left(\frac{\partial E}{\partial X_n} \right) u_n \Big]^2 \right\}^{1/2} \quad (2)$$

Table 8 Specifications of test engine

Technical details of Kirloskar engine	
Engine	Kirloskar
Cubic capacity	0.78 lit
Number of cylinder	1
Model	TV1
CR	17.5:1
Brake power (kW/bhp)	5.9/8
Bore x stroke (mm)	95 × 110
Speed (rpm)	1500
Cooling method	Water-cooled
Opening of the inlet valve	4.5^0 B_TDC
Opening of the exhaust valve	4.5^0 A_TDC
Lubrication method	Forced Feed
Opening of exhaust valve	35.5^0 B_BDC
Opening of inlet valve	35.5^0 A_BDC
Fuel injection timing	26^0 B_TDC

In this context, E represents the primary variable influencing other variables, denoted as X_1……Xn. Here, X refers to the quantity used in the experiment, while u_1… u_n signify the uncertainties associated with the results. The overall uncertainty of the experiment was ultimately described as follows.

Total uncertainty $=\sqrt{}$ (Uncertainty (%) of (ambient temperature^2+SMD^2+ ambient pressure^2+ pressure sensor^2+ injection system^2 + kinematic viscosity^2+fuel pump^2)) =±1.632%

The total uncertainty, calculated at ±1.632%, falls within the range observed by earlier studies (Yadav and Gautam 2022; Raju et al. 2020).

Engine trial setup

The experimental procedure employed a one-cylinder, four-stroke stationary engine commonly used in agricultural sector and power generation in India. The engine operates on a water-cooled dynamometer with a rating of 7.5 KW. The dynamometer employs eddy current (or Foucault's current) to generate loading by engaging the shaft. The engine is rated at 5.9 KW of power and operates at a speed of 1500 rpm. The setup includes a data acquisition system and control panel that enables the monitoring and evaluation of various parameters in the combustion chamber, such as ambient pressure, fuel flow rate, load and temperature. These parameters are monitored at different locations within the combustion chamber. The actual engine setup and the schematic diagram are shown in Fig. 8. The dynamometer is coupled with a load cell utilizing a strain gauge and the load cell provides reliable and accurate engine load data to the control unit. For measuring the flow rate, a differential pressure transmitter with the model number EJA110-EMV-5A-93NN is employed. Additionally, to measure the combustion

Table 9 Principle of devices, accuracy, working range and uncertainty

Measurements	Principle of working	Accuracy	Range	Uncertainty (%)
Engine load	Strain gauge	±0.11 kg	0–25 kg	±0.25
Chamber pressure	Piezoelectric transducer	±0.25 bar	0–150 bar	±0.20
Engine speed	Magnetic type	±0.2 rpm	0–2000 rpm	±0.20
Temperature	thermocouple	±35 °C	0–1000 °C	±0.25
Crank angle encoder (CAD)	Optical	±1.25°	0–720°	±0.40
Time	Stopwatch	±0.12 s		±0.30
Fuel injection	Differential pressure transmitter	±0.10 mm	0–500 mm	±0.15
Carbon monoxide	Non-dispersive infrared (NDIR)	±0.01%	0–10% vol	±0.20
Hydrocarbon	Non-dispersive infrared (NDIR)	±50 ppm	0–20,000 ppm	±0.15
Smoke density	photochemical	±0.45%	0–100%	±0.15
Nitrogen oxide	Electrochemical	±50 ppm	0–5000 ppm	±0.45
Fuel consumption	Turbine flow type	–	–	±0.40
Air consumption	Level sensor	–	–	±0.30

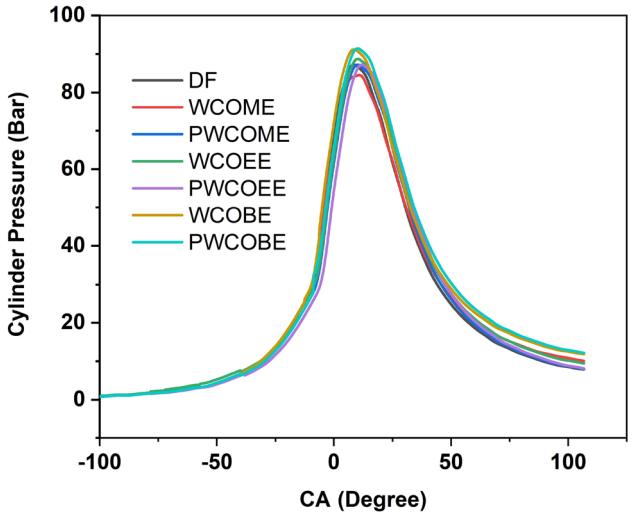


Fig. 9 Cylinder pressure (bar) versus crank angle

chamber pressure, a Kistler piezoelectric pressure transducer (model number 6613CA) is employed. Once the data are accumulated from the measurements, it is processed using the software "Engine soft" on the computer within the data acquisition system. The software-generated signal is subsequently converted into a readable format for better understanding. The smoke meter AVL-437 and di-gas analyzer AVL-444 were employed to evaluate the levels of harmful emissions, including HC, CO, smoke and NO originating from the test fuels.

Following that, the engine testing was carried out under diverse loading conditions, adhering to the specified standards of BIS-10000. The corresponding values of BMEP varied from 0 to 6.3 bar, depending on the loads. To ensure the strength of the data obtained in the experiments, the diesel engine was required to get ready until it reached a steady state, as indicated by the stabilization of the cooling water temperature. Comparative testing was conducted on unheated and preheated biodiesel, alongside neat diesel, to evaluate and compare their performance. Table 8 outlines the details of the engine configuration, while Table 9 provides a detailed examination of engine parameter uncertainties.

The overall level of experimentation uncertainty.

$$= \sqrt{\left[\text{uncertainty \% of } \left[(\text{BTE})^2 + (\text{B.P})^2 + (\text{BSEC})^2 + (\text{CO})^2 + (\text{HC})^2 + (\text{Smoke})^2 + (\text{NOx})^2\right]\right.}$$

$$= 1.81\%$$

Based on the analysis, the overall uncertainty estimate is 1.81%, which is less than 5%. This value appears to be consistent with the published research discussed in Sect. 2.4.

Result and discussion

In-cylinder pressure and heat release rate

Cylinder pressures play an important role in understanding the combustion behavior of fuels. Figure 9 illustrates the cylinder pressure versus crank angle for a fuel injection temperature of 950 °C, both preheated and unheated, at 75% load. The pressure curves for all esters of waste cooking oil (WCO) exhibit similar trends, with PWCOBE showing the highest peak pressure and WCOME the lowest. Preheating the esters leads to a reduction in rheological properties, resulting in improved spray characteristics, and the higher calorific value of PWCOBE enhances combustion characteristics. This study's rationale is also supported by reference (Hazar et al. 2019). The peak pressures of PWCOBE, WCOBE, WCOEE and WCOME are 10.03%, 9.52%, 4.90%

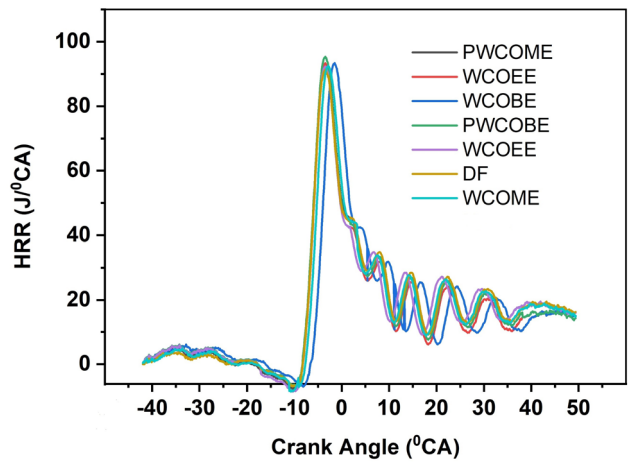


Fig. 10 HRR versus Crank angle

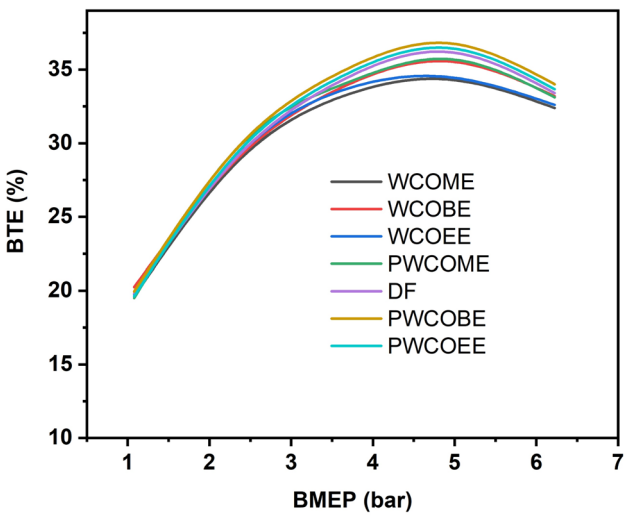


Fig. 11 BTE versus BMEP

and 2.58% higher than mineral diesel, respectively, while PWCOEE and PWCOME are 0.06% and 2.80% lower than mineral diesel, respectively. This variation might be due to the advanced injection timing, lower compressibility, higher density and higher calorific value of unheated esters of waste cooking oil (Viswanathan and Wang 2021).

It was observed that despite the lower volatility and higher viscosity of unheated esters, their ignition delay is noticeably shorter than that of traditional diesel fuel. This can be attributed to rapid and complex pre-flame chemical reactions that occur at high temperatures. Figure 10 illustrates the net HRR versus crank angle for preheated (at 90 °C) and unheated injection temperatures at 75% load. All esters of waste cooking oil exhibit similar combustion stages compared to mineral diesel, including ignition delay, premixed combustion, diffusion combustion and late combustion. The start of injection timing for the diesel engine was set at 260 BTDC. It was observed that the HRR was initially negative around a crank angle of − 140, indicating fuel vaporization during the ignition delay period (Perumal and Ilangkumaran 2017). Following this, the premixed fuel–air mixture ignites quickly and is then succeeded by diffusion combustion, where the burn rate is governed by the air–fuel mixing process. Preheated esters of waste cooking oil exhibited an earlier start of combustion compared to unheated esters and diesel fuel under all engine operating conditions. The heat release during premixed combustion is consistently higher for preheated fuels due to improved air–fuel mixing. The longer ignition delay causes fuel accumulation during premixed combustion, resulting in higher heat release. The peak HRR decreases as the carbon–hydrogen–oxygen (C–H–O) chain reduces in unheated esters. However, preheated esters of waste cooking oil demonstrate higher heat release rates compared to unheated esters.

This can be attributed to the reduction in rheological properties, which leads to improved spray characteristics, atomization and fuel–air mixing of the preheated fuel (Sanli et al. 2015). PWCOBE, PWCOEE and PWCOME exhibit 6.84%, 5.10% and 5.43% higher heat release rates than WCOBE, WCOEE and WCOME, respectively.

Performance

Brake thermal efficiency

Brake thermal efficiency can be described as the ratio of the engine's actual brake power to the chemical energy supplied by the fuel. It serves as a crucial parameter in assessing the impact of various fuels on engine performance. Figure 11 demonstrates that as the brake mean effective pressure (BMEP) increases, the brake thermal efficiencies of all the tested fuels improve. This increase is primarily attributed to a reduction in brake-specific fuel consumption values. Among the tested fuels, PWCOBE and PWCOEE exhibit higher thermal efficiencies compared to mineral diesel, while PWCOME, WCOBE, WCOEE and WCOME show lower efficiency than diesel fuel as BMEP rises. The higher thermal efficiencies of certain fuels compared to diesel can be attributed to their lower rheological properties, which enhance spray characteristics. Additionally, the presence of oxygen in esters of waste cooking oil leads to improved combustion (Yadav et al. 2023). It was observed that despite WCOME having a lower calorific value than WCOEE and WCOBE, preheated WCOME demonstrated higher thermal efficiency than WCOEE and WCOBE. This is due to the preheating of the fuel, which promotes better atomization characteristics in the combustion chamber (Agarwal and Dhar

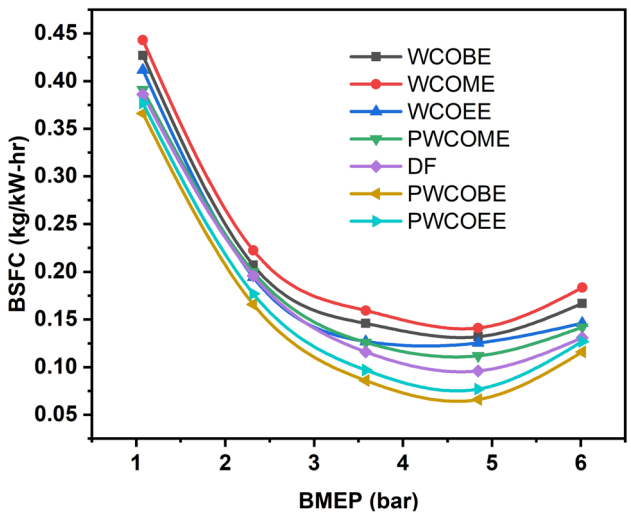


Fig. 12 BSFC versus BMEP

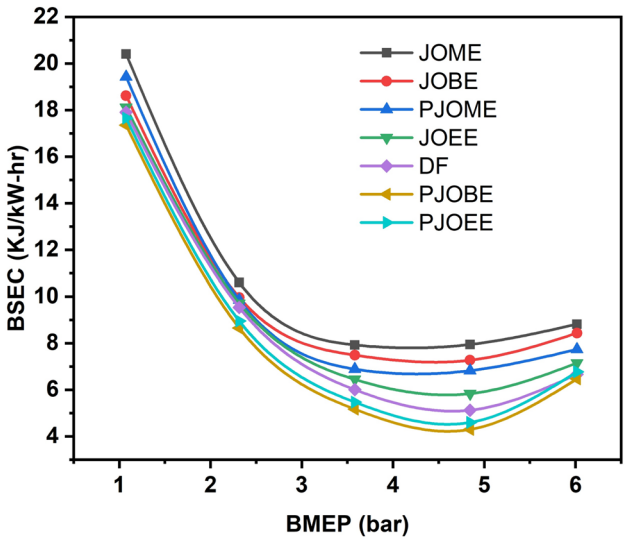


Fig. 13 BSEC versus BMEP

2010). The brake thermal efficiencies of preheated butyl and ethyl esters are 1.66% and 0.58% higher than mineral diesel, respectively. When comparing the ester fuels to each other, it is observed that the thermal efficiency of ethyl ester is higher than butyl ester, followed by methyl ester. Despite the higher heating content of butyl ester, the lower viscosity of ethyl ester results in better atomization and combustion characteristics.

BSFC serves as a fundamental parameter utilized to compare and evaluate the impact of different test fuels on engine performance, particularly in relation to fuel injection. Figure 12 presents the BSFC for the seven test fuels. The BSFC decreases for all test fuels as the BMEP increases. However, PWCOEE and PWCOBE exhibit lower BSFC than mineral diesel at all BMEP levels. This can be ascribed to higher turbulence levels within the cylinder and temperature with load, leading to enhanced combustion (Gautam et al. 2022). The test results indicate that PWCOEE and PWCOBE have BSFC values that are 7.56% and 13.51% lower than mineral diesel, respectively. On the other hand, PWCOME, WCOEE, WCOBE and WCOME have BSFC values that are 4.97%, 8.61%, 16.64% and 24.28% higher than mineral diesel. It is noteworthy that despite the higher calorific value of butyl ester compared to ethyl ester and methyl ester, ethyl ester exhibits lower BSFC due to better droplet disintegration in unheated conditions. Similarly, despite the lower calorific value of preheated methyl ester, it demonstrates lower BSFC than all unheated esters. Similar findings can be found in the literature (Hazar and Aydin 2010). This may be attributed to the reduction in rheological properties and improved spray and atomization characteristics due to heating. Another reason for the higher BSFC of unheated fuel is the increased fuel discharge for the same plunger displacement compared to preheated fuel, owing to higher density and lower bulk modulus (Chauhan et al. 2010; Chitsaz et al. 2015).

BSEC

This indicates the efficiency of extracting energy from fuel to generate power per unit. It is obtained by multiplying the fuel's calorific value by the brake-specific fuel consumption and is used to compare the calorific values of different fuels. Figure 13 demonstrates a clear inverse relationship between brake-specific energy consumption (BSEC) and the increase in BMEP. However, the BSECs of PWCOME, WCOEE, WCOBE and WCOME were higher than that of diesel fuel. Despite the higher calorific value of butyl ester compared to ethyl ester, the BSEC of ethyl ester is lower than that of butyl ester and higher than that of methyl ester in the unheated condition. The higher BSECs are caused by high viscosity, low calorific value and high boiling point (Ashok et al. 2018). All preheated esters exhibited higher brake-specific energy consumption than unheated esters, even with a higher calorific value, due to the dominance of spray characteristics and atomization characteristics.

Emissions

CO

Carbon monoxide (CO) emissions result from incomplete combustion, where the oxidation process fails to occur perfectly. The concentration of CO largely depends on the fuel/air mixture, especially during engine starting and acceleration when a rich mixture is needed (Sharma et al. 2005). Other causes of CO generation include larger droplets,

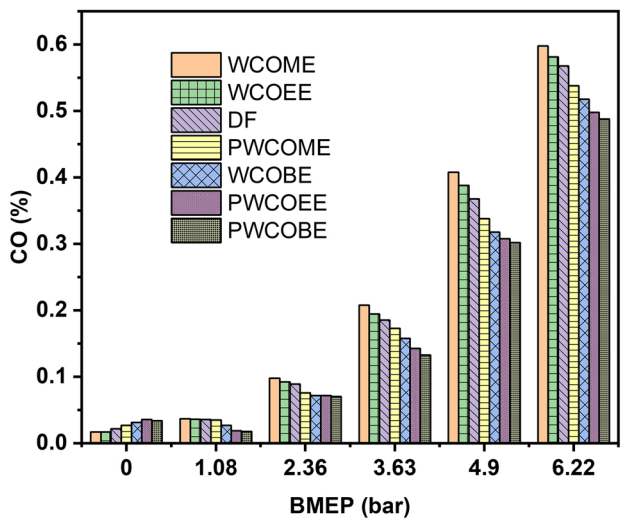


Fig. 14 CO versus BMEP

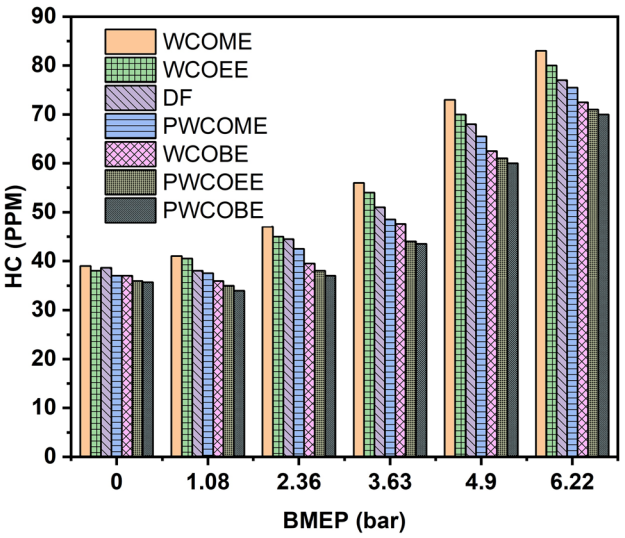


Fig. 15 HC versus BMEP

oxygen deficiency, insufficient turbulence and less swirl, which are responsible for improper air/fuel mixing and oxidation. Figure 14 illustrates the CO emissions of various test fuels at increasing load on the diesel engine. The CO emissions of the test fuels increased with brake mean effective pressure (BMEP). It was observed that, except for WCOME and WCOEE, all test fuels exhibited lower CO emissions than mineral diesel. Preheated butyl ester of waste cooking oil showed the lowest emissions among all test fuels. Preheating causes a reduction in the rheological properties of the fuel, and the ester of higher alcohol with higher oxygen content promotes the oxidation to convert CO to CO_2 (Gupta and Agarwal 2016; Viswanathan and Wang 2021; Hazar and Aydin 2010). Despite the higher viscosity of butyl ester, it exhibited lower emissions than mineral diesel and preheated methyl ester.

This is due to the availability of oxygen and higher calorific value, which promote chemical kinetics. The CO emissions of preheated butyl ester, ethyl ester and methyl ester were reduced by 7.09%, 17.90% and 13.12%, respectively.

HC

Hydrocarbon emissions (HC) consist of unburned fuel resulting from insufficient temperature near the cylinder wall, where the fuel–air mixture temperature is lower compared to the center of the cylinder. Hydrocarbons encompass various species, including alkenes, alkanes and aromatics, and are generally considered equivalent to CH_4 content. At low loads, the primary cause of hydrocarbon emissions is a lean fuel–air mixture. The slow flame speed hinders complete combustion throughout the power stroke, leading to a higher quantity of hydrocarbon emissions (Yamada et al. 2011; Zheng et al. 2008). Figure 15 illustrates the HC emissions of all test fuels and demonstrates that HC emissions increase with an increase in brake mean effective pressure for preheated, unheated and diesel fuel. The lowest HC emissions are observed for preheated butyl ester among all fuels at higher brake mean effective pressure (BMEP). This is attributed to the presence of oxygen molecules in butyl ester, which facilitates enhanced combustion in the engine cylinder (Karavalakis et al. 2009). It is worth noting that despite waste cooking oil-based esters (WCOBE) having higher viscosity than mineral diesel and preheated waste cooking oil-based esters (PWCOME), their HC emissions are still lower.

This can be attributed to the high reaction rate resulting from the availability of oxygen in butyl ester. However, preheating reduces viscosity, leading to improved atomization and combustion characteristics. The HC emissions of PWCOBE, PWCOEE, WCOBE and PWCOME are reduced by 9.09%, 7.79%, 5.84% and 1.94%, respectively, compared to diesel fuel, and increased by 3.89% and 7.79% for WCOEE and WCOME, respectively.

CO_2

CO_2 emissions are widely regarded as a crucial indicator of complete fuel combustion and represent one of the fundamental greenhouse gas emissions (Quadrelli and Peterson 2007). Figure 16 presents the evaluated CO_2 emissions for all test fuels. The results reveal that unheated esters exhibited elevated CO_2 emissions when operating at lower BMEP. However, with an increase in brake mean effective pressure, all preheated butyl, ethyl and methyl esters exhibited an increase in CO_2 emissions. These findings align with the study conducted by Sanli et al. (2015) and Bari et al. 2002). This phenomenon can be attributed to

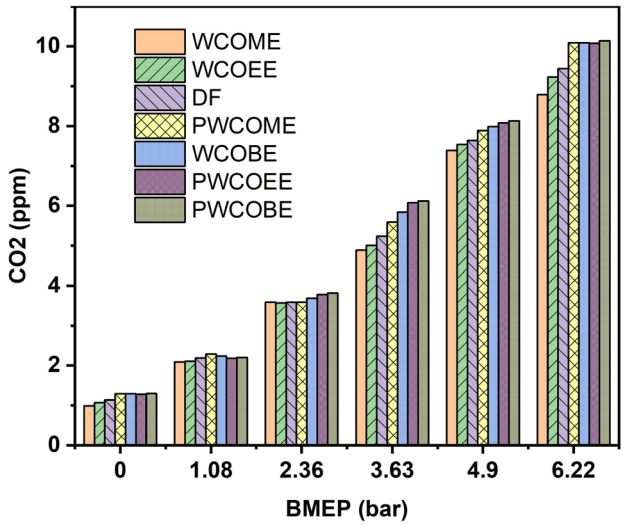


Fig. 16 CO_2 versus BMEP

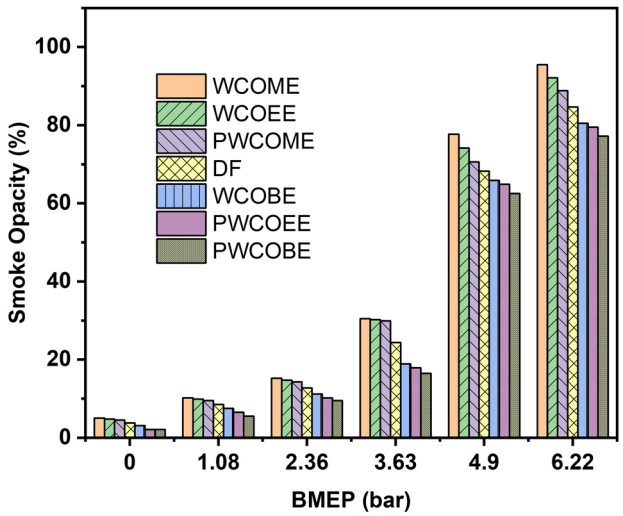


Fig. 17 Smoke opacity versus BMEP

the higher surface area-to-volume ratio of droplets and the presence of more oxygen content that facilitates the oxidation of CO to CO_2 during in-cylinder combustion (Chauhan et al. 2010). The preheated butyl ester of waste cooking oil exhibited the highest CO_2 emission, which was 8.45% higher than that of diesel fuel. In contrast, the unheated methyl ester of waste cooking oil exhibited 5.12% lower CO_2 emissions compared to diesel fuel. The increased carbon content of butyl ester relative to mineral diesel leads to a higher production of CO_2 during combustion, resulting in higher emissions of this greenhouse gas. For example, the chemical formula of diesel fuel is $C_{14}H_{25}$, while that of butyl ester biodiesel is $C_{22}H_{42}O_2$. The higher carbon content leads to higher fuel consumption, which in turn can result in elevated CO_2 emissions. Additionally, the presence of oxygen contributes to improved combustion quality by facilitating the oxidation of CO to CO_2.

Smoke opacity

Opacity refers to the extent to which smoke reduces or prevents the passage of light. The soot threshold influences the equivalence ratio and gas temperature. Conversely, high soot generation can occur due to limited oxygen availability and high ambient temperature (Tan et al. 2012). Figure 17 illustrates that the smoke opacity of test fuels is lower at low BMEP and increases with higher BMEP. However, PWCOBE, PWCOEE and WCOBE fuels generated less smoke compared to mineral diesel, while the remaining fuels exhibited higher levels of smoke than diesel fuel.

This phenomenon is attributed to better atomization characteristics and higher oxygen availability in the engine cylinder. Despite the improved spray characteristics of preheated methyl ester compared to unheated butyl ester, the smoke opacity of unheated butyl ester is higher than that of preheated methyl ester. This may be due to the higher kinematic reactivity toward the available oxygen in the ester. When comparing preheated and unheated fuels, the preheated butyl ester, ethyl ester and methyl ester exhibited 7.38%, 19.81% and 7.05% lower smoke opacity than unheated butyl, ethyl and methyl esters, respectively. These findings also suggest that preheating biodiesel could lead to lower C–C bond content with higher oxygen content, resulting in reduced smoke emissions (Utlu and Koçak 2008).

NO_x and EGT

NO_x formation is highly temperature-dependent. At higher combustion temperatures, the nitrogen in the air reacts with oxygen to form NO_x, primarily in the form of nitric oxide (NO) and nitrogen dioxide (NO_2). This is known as thermal NO_x. The formation rate of NO_x increases exponentially with temperature. Thermal NO_x occurs mainly through the Zeldovich mechanism, which involves a series of chemical reactions that occur in the flame front. Increased pressure also influences NO_x formation, though its effect is generally less significant than that of temperature. Higher pressure can lead to more efficient combustion, which might increase flame temperatures, indirectly promoting NO_x formation (Canakci 2009). Figure 18 illustrates that NO_x emissions from all ester fuels, along with diesel fuel, increase with brake mean effective pressure. Preheated butyl ester, followed by ethyl ester, methyl ester and unheated butyl ester,

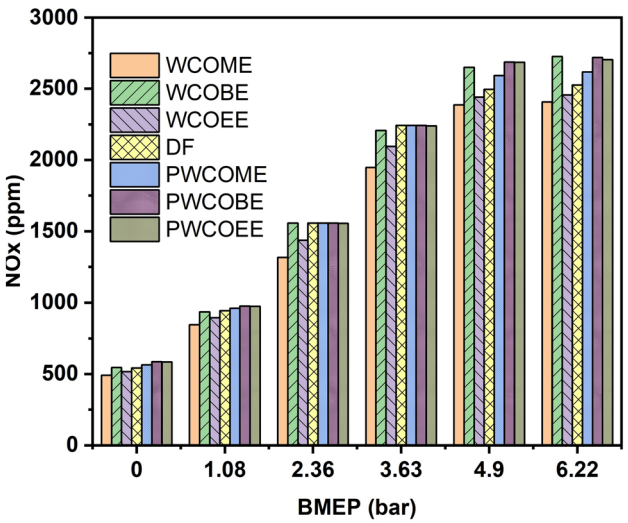


Fig. 18 NO_x versus BMEP

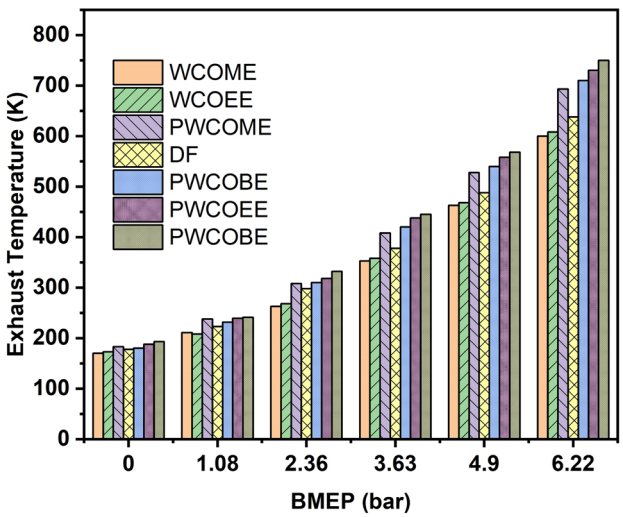


Fig. 19 Exhaust gas temperature versus BMEP

Table 10 Comparison of previous studies with present study (preheated compared to unheated or diesel)

Engine specification	Operating conditions	Oils	Analysis	Performance		Combustion		Emission			
				BTE (%)	BSFC (%)	HRR (%)	CP (%)	CO (%)	HC (%)	NO_x (%)	CO_2(%)
1C, CI engine, RP: 5.2 kW, 4S, RS: 1500 rpm, WC, CR: 17.5:1 (Kodate et al. 2020)	Different % of rated load (0,25, 50, 75, 100)	B0	At 100%load	7.9↑	4.7↓	13.3↑	0.1↑	0.9↓	2.36↓	0.92↑	4.9↑
		B30		2.9↑	5.1↓	11.2↑	1.1↑	1.11↓	2.5↓	1.2↑	2.3↑
		B50		4.1↑	5.8↓	9.33↑	1.4↑	3.5↓	4.6↓	3.1↑	3↑
		B100		9.1↑	6.5↓	8.56↑	3.9↑	8.1↓	10.11↓	6.1↑	5.8↑
		KOME (95^0C),									
1C, CI engine, RP: 5.2 kW, 4S, RS: 1500 rpm, WC, CR: 17.5:1 (Vedharaj et al. 2015)	Brake Power (kW) (1,2,3,4,5,6)	Diesel	At 100%load	–	–	–	–	–	–	↑	–
		CNSLME		–	–	–	–	–	–	↑	–
		CNSLME60		–	–	–	–	–	–	↑	–
		CNSLME80		20↑	20↓	30↑	↑	52↓	66↓	↑	–
1C, CI engine, RP: 5.9 kW, 4S, RS: 1500 rpm, WC, CR: 17.5:1 (Chauhan et al. 2010)	Different % rated load (0,20,40, 60,80,100)	PJO40	Average load. Compared to diesel fuel	8.21↓	–	–	–	–	–	1.61↑	–
		PJO60		7.26↓	–	–	–	–	–	21.9↑	–
		PJO80		3.50↓	–	–	–	–	–	27.7↑	–
		PJO100		2.87↓	–	–	–	–	–	31.8↑	23↑
4C, CI engine, 4S, WC, RP: 80hp, CR: 16.7:1 RS: 2000 rpm, (Hoang and Le 2019)	Engine speed (10.12,14,16,18,20) *100 rpm	PSCO80	At peak. Compared to diesel fuel	12.2↓	–	7.6↑	–	88.6↑	47.9↑	23.2↓	25.2↓
		PSCO105		9.3↓	–	2.1↑	–	23.1↑	23.4↑	8.9↓	14.6↓
		PSCO120		11.4↓	–	6.5↑	–	46.7↑	31.5↑	16.4↓	19.2↓
1C, CI engine RP: 5.2 kW, 4S, RS: 1500 rpm, WC, CR: 17.5:1 [present work]	BMEP (bar) (1,2,3,4,5,6)	WCOME	At average BMEP. Compared to diesel fuel	2.9↓	24.3↑	4.4↓	–	9.7↑	7.8↑	10.1↓	4.3↓
		WCOEE		2.1↓	8.6↑	2.9↓	–	4.0↑	3.9↑	5.1↓	2.5↓
		WCOBE		0.7↓	16.6↑	2.1↑	–	13.4↓	5.8↓	1.5↑	6.3↑
		PWOCME		0.5↓	5.0↑	2.1↑	–	7.4↓	1.9↓	1.7↑	4.3↑
		PWCOEE		0.6↑	7.6↓	3.8↑	–	17.5↓	7.8↓	3.3↑	8.1↑
		PWCOBE		1.7↑	13.5↓	5.4↑	–	20.5↓	9.1↓	3.4↑	8.9↑
Lombardini 6LD 400, 4S, WC, 1C, RP: 6.25 kW, CR: 17.5:1 RS: 3600 rpm, (Hazar and Sevinc 2019)	Engine speed (18.20,22,24,26,28,30) *100 rpm@100^0C	PDF	Average load	0.73↑	–	–	–	13.5↓	8.1↓	11.4↑	–
		H30		1.65↑	–	–	–	11.2↓	9.0↓	8.1↑	–
		H50		1.82↑	–	–	–	12.4↓	7.0↓	11.3↑	–

1C: Single cylinder; CR: Compression ratio, RP: rated power; 4S: four-stroke, WC: water-cooled; RS: rated speed ↑: increase; ↓: decrease

exhibited higher NO_x emissions than mineral diesel at all BMEP levels. This results from the increased cylinder temperature during combustion, which lead to increased NO_x emissions. The oxygen present in biodiesel aids in better combustion, resulting in higher temperatures. Additionally, the smaller droplets generated through preheating exhibit better combustion characteristics, leading to a significant temperature rise. The high cetane number of biodiesel promotes combustion by reducing the ignition delay, which in turn contributes to NO_x formation (Yesilyurt et al. 2020). The Zeldovich scheme mechanism suggests that the combustion chemistry of fuel can be coupled with the three main reaction chains involving O, O_2 and OH species (Turns 1996), as shown in Eqs. (3), (4), (5).

$$O + N_2 \leftrightarrow NO + N \tag{3}$$

$$N + O_2 \leftrightarrow NO + N \tag{4}$$

$$N + OH \leftrightarrow NO + H \tag{5}$$

It was observed that PWCOBE, PWCOEE, WCOBE and PWCOME emitted 4.46%, 4.22%, 3.04% and 2.17% higher NO_x emissions than diesel fuel at peak load. However, WCOME and WCOEE showed lower NO_x than diesel due to less heating value and improper atomization characteristics that degraded the combustion phenomena.

Figure 19 illustrates the disparity of exhaust gas temperature (EGT) with BMEP for pure diesel, preheated and unheated WCO. Figure 16 reveals a direct correlation between the increase in BMEP and the corresponding elevation in exhaust gas temperature (EGT) across all the tested fuels. The results emphasize that PWCOBE generated the highest exhaust gas temperature (732 K), whereas WCOME exhibited the lowest temperature (590 K) during peak load operations. Diesel fuel exhibited higher EGT than WCOME and WCOEE among all test fuel. Subpar atomization of biodiesel leads to reduced combustion effectiveness in unheated waste cooking oil.

The reduction in viscosity and density from preheating the fuel enhances air entrainment in the spray, resulting in better combustion. Additionally, the exhaust gas temperature exhibits lower values when operating at a stoichiometric air–fuel ratio compared to rich or lean air–fuel ratios. This is the primary reason for the better combustion of preheated esters of WCO than unheated esters of WCO, resulting in higher exhaust gas temperatures. These findings are consistent with (Albo et al. 2017). Table 10 offers a comprehensive analysis of the similarities and disparities between the current study's results and the outcomes reported in the literature.

Conclusion

The present study explores the influence of preheated methyl ester, ethyl ester and butyl ester of waste cooking oil biodiesel on atomization, engine characteristics of diesel engine. This study also emphasized on cost analysis of esters at unheated and preheated condition. The heating of fuel was performed through heat exchanger. The following findings were concluded.

1. The viscosity of methyl ester exhibited maximum reduction of 74.51% followed by ethyl ester and butyl ester of 73.1% and 63.44%, respectively. Consequently, SMD reduced accordingly 16.33%, 15.54% and 13.20% for methyl, ethyl and butyl ester, respectively.
2. The cost to generate 1 kWh energy was least for waste cooking oil butyl ester among all test fuels (diesel, unheated, preheated esters).
3. PWCOBE reported highest thermal efficiency and lowest brake-specific fuel consumption among all test fuel. This might be caused by the elevated calorific value, decreased rheological properties, oxygen availability promotes homogeneous combustion as compared to diesel fuel.
4. The CO, HC emissions and smoke opacity of WCOBE are lower than PWCOME. The CO_2 emissions of PWCOBE are higher than methyl ester of waste cooking oil.
5. The NO_x emission and EGT of preheated esters of waste cooking oil along with unheated butyl ester of WCO are higher than diesel fuel at peak load.

The results shows that heating is the most straightforward solution for addressing the drawbacks of using bio-oils in direct contact with diesel engines. Additional research is necessary to understand the corrosion phenomenon in engines and the accumulation of deposits in the combustion chamber and injector when using preheated bio-oils as an alternative fuel.

Acknowledgements The authors would like to thank Prof. Anh Tuan Hoang, Dong A University, Danang, Vietnam, for his valuable insights.

Author contribution PS and R wrote the main manuscript text. PS prepared Figs. 1–19. All authors reviewed the manuscript.

Funding This research received no external funding.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Agarwal AK (2007) Biofuels (alcohols and biodiesel) applications as fuels for internal combustion engines. Prog Energy Combust Sci 33(3):233–271

Agarwal AK, Dhar A (2010) Karanja oil utilization in a direct-injection engine by preheating. Part 1: experimental investigations of engine performance, emissions, and combustion characteristics. Proc Inst Mech Eng Part D J Automob Eng 224(1):73–84

Albo PG, Lago S, Wolf H, Pagel R, Glen N, Clerck M, Ballereau P (2017) Density, viscosity and specific heat capacity of diesel blends with rapeseed and soybean oil methyl ester. Biomass Bioenerg 96:87–95

Alptekin E, Sanli H, Canakci M (2019) Combustion and performance evaluation of a common rail DI diesel engine fueled with ethyl and methyl esters. Appl Therm Eng 149:180–191

Anis S, Budiandono GN (2019) Investigation of the effects of preheating temperature of biodiesel-diesel fuel blends on spray characteristics and injection pump performances. Renew Energy 140:274–280

Ashok B, Nanthagopal K, Vignesh DS (2018) Calophyllum inophyllum methyl ester biodiesel blend as an alternate fuel for diesel engine applications. Alex Eng J 57(3):1239–1247

Atmanli A, Yilmaz N (2018) A comparative analysis of n-butanol/diesel and 1-pentanol/diesel blends in a compression ignition engine. Fuel 234:161–169

Baiju B, Naik MK, Das LM (2009) A comparative evaluation of compression ignition engine characteristics using methyl and ethyl esters of Karanja oil. Renew Energy 34(6):1616–1621

Bari S, Lim TH, Yu CW (2002) Effects of preheating of crude palm oil (CPO) on injection system, performance and emission of a diesel engine. Renew Energy 27(3):339–351

Baskar G, Aiswarya R (2016) Trends in catalytic production of biodiesel from various feedstocks. Renew Sustain Energy Rev 57:496–504

Ben Jabeur R, Fguiri A, Jeday MR, Chekir H, Fatnassi H, Marvillet C (2018) Evaluation of the waste heat discharged from the Di-ammonium phosphate production plant: a case study. J Mater Cycles Waste Manag 20:1857–1866

Canakci M (2009) NO_x emissions of biodiesel as an alternative diesel fuel. Int J Veh Des 50(1–4):213–228

Chauhan BS, Kumar N, Du Jun Y, Lee KB (2010) Performance and emission study of preheated Jatropha oil on medium capacity diesel engine. Energy 35(6):2484–2492

Chavan SB, Kumbhar RR, Kumar A, Sharma YC (2015) Study of biodiesel blends on emission and performance characterization of a variable compression ratio engine. Energy Fuels 29(7):4393–4398

Chitsaz A, Mahmoudi SMS, Rosen MA (2015) Greenhouse gas emission and exergy analyses of an integrated trigeneration system driven by a solid oxide fuel cell. Appl Therm Eng 86:81–90

Coniglio L, Bennadji H, Glaude PA, Herbinet O, Billaud F (2013) Combustion chemical kinetics of biodiesel and related compounds (methyl and ethyl esters): Experiments and modelling—advances and future refinements. Prog Energy Combust Sci 39(4):340–382

Dürre P (2008) Fermentative butanol production: bulk chemical and biofuel. Ann N Y Acad Sci 1125(1):353–362

Eliçin AK, Ozturk F, Baran MF, Esgici R (2019) The use of rapeseed oil methyl and ethyl esters and of rapeseed oil-diesel mixtures as fuels for diesel engine. Fresenius Environ Bull 28(11):7915–7923

Ezeji T, Qureshi N, Blaschek HP (2007) Production of acetone–butanol–ethanol (ABE) in a continuous flow bioreactor using degermed corn and *Clostridium beijerinckii*. Process Biochem 42(1):34–39

Gautam R, Chauhan BS, Lim HC (2022) Influence of variation of injection angle on the combustion, performance and emissions characteristics of jatropha ethyl ester. Energy 254:124436

Gupta JG, Agarwal AK (2016) Macroscopic and microscopic spray characteristics of diesel and Karanja biodiesel blends (No. 2016-01-0869). SAE Technical Paper

Han D, Duan Y, Wang C, Lin H, Huang Z (2014) Experimental study on injection characteristics of fatty acid esters on a diesel engine common rail system. Fuel 123:19–25

Hazar H, Aydin H (2010) Performance and emission evaluation of a CI engine fueled with preheated raw rapeseed oil (RRO)—diesel blends. Appl Energy 87(3):786–790

Hazar H, Sevinc H (2019) Investigation of the effects of pre-heated linseed oil on performance and exhaust emission at a coated diesel engine. Renew Energy 130:961–967

Hazar H, Sevinc H, Sap S (2019) Performance and emission properties of preheated and blended fennel vegetable oil in a coated diesel engine. Fuel 254:115677

Hoang AT, Le AT (2019) A core correlation of spray characteristics, deposit formation, and combustion of a high-speed diesel engine fueled with Jatropha oil and diesel fuel. Fuel 244:159–175

Holman JP (2001) Experimental methods for engineers

https://auto.economictimes.indiatimes.com/news/oil-and-lubes/zomato-to-collect-used-cooking-oil-for-biodiesel/71691599

Jindal S, Nandwana BP, Rathore NS, Vashistha V (2010) Experimental investigation of the effect of compression ratio and injection pressure in a direct injection diesel engine running on Jatropha methyl ester. Appl Therm Eng 30(5):442–448

Karavalakis G, Stournas S, Bakeas E (2009) Light vehicle regulated and unregulated emissions from different biodiesels. Sci Total Environ 407(10):3338–3346

Kodate SV, Yadav AK, Kumar GN (2020) Combustion, performance and emission analysis of preheated KOME biodiesel as an alternate fuel for a diesel engine. J Therm Anal Calorim 141:2335–2345

Lapuerta M, Herreros JM, Lyons LL, García-Contreras R, Briceño Y (2008) Effect of the alcohol type used in the production of waste cooking oil biodiesel on diesel performance and emissions. Fuel 87(15–16):3161–3169

Makareviciene V, Janulis P (2003) Environmental effect of rapeseed oil ethyl ester. Renew Energy 28(15):2395–2403

May CY, Liang YC, Foon CS, Ngan MA, Hook CC, Basiron Y (2005) Key fuel properties of palm oil alkyl esters. Fuel 84(12–13):1717–1720

Mendow G, Veizaga NS, Sánchez BS, Querini CA (2011) Biodiesel production by two-stage transesterification with ethanol. Biores Technol 102(22):10407–10413

Perumal V, Ilangkumaran M (2017) Experimental analysis of engine performance, combustion and emission using pongamia biodiesel as fuel in CI engine. Energy 129:228e36. https://doi.org/10.1016/j.energy.2017.04.120

Qiu T, Guo X, Yang J, Zhou L, Li L, Wang H, Niu Y (2016) The synthesis of biodiesel from coconut oil using novel Brønsted acidic ionic liquid as green catalyst. Chem Eng J 296:71–78

Quadrelli R, Peterson S (2007) The energy–climate challenge: recent trends in CO_2 emissions from fuel combustion. Energy Policy 35(11):5938–5952

Raju VD, Venu H, Subramani L, Kishore PS, Prasanna PL, Kumar DV (2020) An experimental assessment of prospective oxygenated additives on the diverse characteristics of diesel engine powered with waste tamarind biodiesel. Energy 203:117821

Rakopoulos CD, Antonopoulos KA, Rakopoulos DC, Hountalas DT, Giakoumis EG (2006) Comparative performance and emissions study of a direct injection diesel engine using blends of diesel fuel with vegetable oils or bio-diesels of various origins. Energy Convers Manag 47(18–19):3272–3287

Rakopoulos DC, Rakopoulos CD, Giakoumis EG, Dimaratos AM, Kyritsis DC (2010) Effects of butanol–diesel fuel blends on the performance and emissions of a high-speed DI diesel engine. Energy Convers Manag 51(10):1989–1997

Reyero I, Arzamendi G, Zabala S, Gandía LM (2015) Kinetics of the NaOH-catalyzed transesterification of sunflower oil with ethanol to produce biodiesel. Fuel Process Technol 129:147–155

Sanli H, Canakci M, Alptekin E, Turkcan A, Ozsezen AN (2015) Effects of waste frying oil based methyl and ethyl ester biodiesel fuels on the performance, combustion and emission characteristics of a DI diesel engine. Fuel 159:179–187

Sharma M, Agarwal AK, Bharathi KVL (2005) Characterization of exhaust particulates from diesel engine. Atmos Environ 39(17):3023–3028

Sidibé SS, Blin J, Vaitilingom G, Azoumah Y (2010) Use of crude filtered vegetable oil as a fuel in diesel engines state of the art: literature review. Renew Sustain Energy Rev 14(9):2748–2759

Singhal R, Sengupta R (2012). Energy Security and Biodiesel: implications for land use and food security. Econ Polit Weekly 66–73

Sinha S, Agarwal AK, Garg S (2008) Biodiesel development from rice bran oil: transesterification process optimization and fuel characterization. Energy Convers Manag 49(5):1248–1257

Tan PQ, Hu ZY, Lou DM, Li ZJ (2012) Exhaust emissions from a light-duty diesel engine with Jatropha biodiesel fuel. Energy 39(1):356–362

Turns SR (1996) Introduction to combustion, vol 287. McGraw-Hill Companies, New York, p 569

Utlu Z, Koçak MS (2008) The effect of biodiesel fuel obtained from waste frying oil on direct injection diesel engine performance and exhaust emissions. Renew Energy 33(8):1936–1941

Vedharaj S, Vallinayagam R, Yang WM, Chou SK, Chua KJE, Lee PS (2015) Performance emission and economic analysis of preheated CNSL biodiesel as an alternate fuel for a diesel engine. Int J Green Energy 12(4):359–367

Venu H, Raju VD, Subramani L, Appavu P (2020) Experimental assessment on the regulated and unregulated emissions of DI diesel engine fuelled with Chlorella emersonii methyl ester (CEME). Renew Energy 151:88–102

Viswanathan K, Wang S (2021) Experimental investigation on the application of preheated fish oil ethyl ester as a fuel in diesel engine. Fuel 285:119244

Yadav PS, Gautam R (2022) Numerical and experimental analysis on spray characteristics of biodiesel (waste cooking oil) using pressure swirl atomizer. Environ Prog Sustain Energy 41(3):e13761

Yadav M, Sharma YC (2019) Transesterification of used vegetable oil using $BaAl_2O_4$ spinel as heterogeneous base catalyst. Energy Convers Manag 198:111795

Yadav M, Upadhyay S, Sharma YC (2018) Process optimization, the kinetics of production Jatropha curcus methyl ester, and its utilization in single cylinder diesel engine. Energy Convers Manag 160:364–374

Yadav PS, Said Z, Gautam R, Raman R, Caliskan H (2023) Novel investigation on atomization, performance, and emission characteristics of preheated jatropha oil methyl ester and ethyl ester. Energy 270:126870

Yadav P, Kumar N, Gautam R (2021) Improvement in performance of CI engine using various techniques with alternative fuel. Energy Sour Part A Recove Utiliz Environ Effects 1–27

Yamada H, Misawa K, Suzuki D, Tanaka K, Matsumoto J, Fujii M, Tanaka K (2011) Detailed analysis of diesel vehicle exhaust emissions: nitrogen oxides, hydrocarbons and particulate size distributions. Proc Combust Inst 33(2):2895–2902

Yesilyurt MK, Cesur C, Aslan V, Yilbasi Z (2020) The production of biodiesel from safflower (*Carthamus tinctorius* L.) oil as a potential feedstock and its usage in compression ignition engine: a comprehensive review. Renew Sustain Energy Rev 119:109574

Zheng M, Mulenga MC, Reader GT, Wang M, Ting DS, Tjong J (2008) Biodiesel engine performance and emissions in low temperature combustion. Fuel 87(6):714–722

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

Prem Shanker Yadav[1,2] · Saket Kumar[2,3] · Raghvendra Gautam[2]

✉ Prem Shanker Yadav
prem0jss@gmail.com

✉ Raghvendra Gautam

[1] Department of Mechanical Engineering, JSS Academy of Technical Education, Noida, India

[2] Centre for Advance Studies in Vehicle Diagnostic in Research in Automobile Engineering, Delhi Technological University, Delhi, India

[3] Department of Mechanical Engineering, Dronacharya Group of Institution, Greater Noida, India

ExLibris RapidILL

Rapid #: -23205803

CROSS REF ID: **6420164**

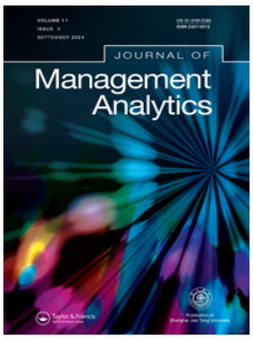
LENDER: **UKROH (Royal Holloway, University of London) :: Main Library**

BORROWER: **HUL (Harvard University) :: Widener Library**

TYPE:	Article CC:CCG
JOURNAL TITLE:	Journal of management analytics
USER JOURNAL TITLE:	Journal of Management Analytics
ARTICLE TITLE:	A framework for evaluating the barriers to adopting Industry 4.0 in Indian SMEs: an approach of best-worst method
ARTICLE AUTHOR:	Kumar, Roshan
VOLUME:	
ISSUE:	
MONTH:	
YEAR:	2024
PAGES:	1-33
ISSN:	2327-0012
OCLC #:	

Processed by RapidX: 9/27/2024 1:46:42 AM

This material has been supplied in accordance with UK copyright law.
It has been supplied for the purpose of non-commercial research or private study and in accordance with the terms laid out in the copyright declaration form signed when making your request:
•You may print out and retain a single paper copy for your own use.
•You must delete the electronic file once a satisfactory paper copy has been created.
•You may not make or distribute any further copies, either in paper or electronic format without the permission of the copyright owner.



Taylor & Francis
Taylor & Francis Group

Journal of Management Analytics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjma20

A framework for evaluating the barriers to adopting Industry 4.0 in Indian SMEs: an approach of best-worst method

Roshan Kumar, Pravin Kumar, Rajesh Kumar Singh, Anurika Vaish & Garima Sharma

To cite this article: Roshan Kumar, Pravin Kumar, Rajesh Kumar Singh, Anurika Vaish & Garima Sharma (12 Sep 2024): A framework for evaluating the barriers to adopting Industry 4.0 in Indian SMEs: an approach of best-worst method, Journal of Management Analytics, DOI: 10.1080/23270012.2024.2362645

To link to this article: https://doi.org/10.1080/23270012.2024.2362645

Published online: 12 Sep 2024.

Submit your article to this journal

Article views: 34

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tjma20

Journal of Management Analytics, 2024
https://doi.org/10.1080/23270012.2024.2362645

Taylor & Francis
Taylor & Francis Group
Check for updates

A framework for evaluating the barriers to adopting Industry 4.0 in Indian SMEs: an approach of best-worst method

Roshan Kumar[a], Pravin Kumar [a]*, Rajesh Kumar Singh[b], Anurika Vaish[c] and Garima Sharma[d]

[a]*Department of Mechanical Engineering, Delhi Technological University, Delhi, India;* [b]*Management Development Institute, Gurgaon, India;* [c]*Indian Institute of Information Technology, Allahabad, India;* [d]*Maharaja Agrasen Institute of Technology, Delhi, India*

(*Received 25 December 2022; revised 2 October 2023; accepted 29 May 2024*)

This study aims to highlight the need for Industry 4.0 in a manufacturing system and explore the importance of the barriers to adopting Industry 4.0 technologies in Indian SMEs. Many barriers to implementing Industry 4.0 were explored through a literature review. These barriers are prioritized using the Best-Worst Method (BWM). The framework has illustrated the approach to exploring the barriers and ranking these barriers based on feedback from industry experts. Some of the barriers such as lack of infrastructure, lack of financial resources, lack of government initiatives, high complexity, and cyber security and data ownership issues are observed to be very influential in SMEs to adopting Industry 4.0. The proposed framework can also be used in other industries for implementing Industry 4.0 technologies. Prioritizing and overcoming the barriers step-by-step may help the manager to digitalize the systems.

Keywords: Industry 4.0; barriers; best-worst method; manufacturing system; small and medium enterprises

1. Introduction

From the perspective of Indian Small and Medium Enterprises (SMEs), Industry 4.0 has huge implications. As per the estimation of McKinsey (2018), in the next 20 years, approximately 300 million well-educated workforces will be produced. In this situation, SMEs can create the required job opportunities by improving their productivity. Industry 4.0 requires less capital in comparison to Industry 3.0 because the replacement of old equipment is required with automated specialized machines in Industry 3.0. However, many SMEs have been using automated machines for the production of goods and services. Now, manufacturing establishments require retrofitting of existing infrastructure with the cyber-physical system (CPS) incorporating some new technologies like IoT, big data analytics, and robotics. The industry would need to integrate the manufacturing processes with cloud-based applications (Dutta et al., 2020).

The enactment of Industry 4.0 has emerged as an effective and influential method that brings technological improvement in the development of production, research, and management. There is a digital inclusion of modern practices of

*Corresponding author. Email: pravin.dce@gmail.com

© 2024 Antai College of Economics and Management, Shanghai Jiao Tong University

technical and digital facilities in the manufacturing processes and related services. Today is the era of agile manufacturing where digitalization and communication-based technologies and techniques play a vital role (Oztemel & Gursev, 2020). Through automation and data interchange, Industry 4.0 technology has been treated as a global solution for the manufacturing industry. The fourth industrial revolution brings digitalization in manufacturing, in which all aspects of production are integrated and computerized. The Internet of Things (IoT), robotic systems, additive 3D printing, and artificial intelligence are the parts of Industry 4.0, which can create a "smart factory." These technologies may be used to accomplish cyber-physical systems by combining the physical world with the virtual world. This can be done by observing physical activities, creating digital replicas of the physical world, and establishing decentralized manufacturing systems. In these systems, individuals, machines, and materials interact with each other (Rezqianita & Ardi, 2020).

Industry 4.0 has posed various problems for emerging economies such as China, Brazil, and India, where low labor costs would not be advantageous. In today's era of globalization, where the availability of the internet and 5G technology has emerged and is contributing to digitalization technologies. This has contributed to the development of the world's industry in a uniform manner (Kumar et al., 2020).

The government of India launched the "Make in India" program to promote the manufacturing ecosystem and brought a positive stance to the industrial sector. According to data provided by the Indian government, Gross Value Added in the manufacturing industry in India between 2012 and 2018 at a CAGR of 4.34 percent. Industry 4.0 is gradually being implemented through programs such as smart cities, digital India, etc. Using Industry 4.0 technologies, the aspects of cost, efficiency, value, flexibility, employee satisfaction, etc. can be enhanced. (Kamble et al., 2018).

In 2011, Hanover Fair put forward the idea of Industry 4.0 technologies, and the German government took it as an initiative to improve and develop its manufacturing environment in 2013 (Li, 2018). Industry 4.0 has been developed as an essential technology that overcomes the limitations of Industry 3.0 and boosts the development of manufacturing and technical processes. With the increase in population, there is a need to multiply the goods and services, ensuring longer life with adequate quantity and without compromising the quality. Due to the rapid influence of globalization, there is a need to develop a relationship between man and the machine, thus, the inclusion of Industry 4.0 is a need for rapid development. Artificial Intelligence along with the Internet of Things is a subset of Industry 4.0. It is a major tool in building the unbreakable relationship between workers and machines. It has to boost the development of industrial products catering to the needs of man and the environment (Laudante, 2017). Considering the importance of Industry 4.0 in SMEs, the management is facing a lot of problems in adopting Industry 4.0. These problems may be related to infrastructure, government support, and incentives, availability of technologies, financial support, etc. The following are the research questions, which are focused on this study:

RQ1: What are the barriers to adopting Industry 4.0 in Indian SMEs?

RQ2: How to prioritize the importance of these barriers, so that Industry 4.0 can be easily adopted in SMEs?

To find the answer to these research questions, this study formulates the following objectives:

(i) To investigate and explore the barriers to adopting Industry 4.0 technologies through a literature review and take opinions of the experts of SMEs about these barriers.
(ii) To analyze and rank the importance of these barriers, getting the inputs from the different industries.

The rest parts of the paper are organized as: Section 2 leads to the literature review; Section 3 discusses the research methodology; Section 4 represents the case analysis and results discussion, and Section 5 concludes the research with the scope of future research.

2. Literature review

SMEs play a major role in the network of large manufacturing organizations. They are mainly focused on producing and supplying the components to the OEM (Original Equipment Manufacturer). The demand for Industry 4.0 within SMEs is on the rise, as companies are focused on meeting the global competition in the markets. Most SMEs have concentrated their efforts to digitalize only a selected number of processes. Moving towards Industry 4.0 requires transparency in data acquisition and access, integration of supply chain entities, fostering a supportive environment and culture, and openness to integrating technology throughout the manufacturing process, yet SMEs struggle to achieve these goals (Menon et al., 2018). SMEs have more constraints compared to large manufacturing organizations in terms of level of flexibility and availability of resources, and stand out from their larger counterparts in manufacturing by their higher adaptability, limited resources, and specialization towards recess markets (Ghobakhloo, 2018). The adoption of Industry 4.0 can significantly enhance the performance of SMEs when government bodies, organizations, and their groups and associations work together to facilitate a successful digital shift. The primary goal for SMEs is profit maximization. Very few SMEs have recognized the need for digitalization to improve product quality or processes to fulfill the demands of Original Equipment Manufacturers (OEMs) (Arnold et al., 2017; Thakkar et al., 2012).

2.1. *Applications of Industry 4.0 in manufacturing systems*

The manufacturing industry has observed a great transformation in the era of Industry 4.0. Li (2018) observed the industrial advancement by implementing Industry 4.0 and highlighted the significant applications in the field of technological forecasting and their effect on the social environment. Laudante (2017) analyzed the manufacturing systems, ergonomically and suggested the needful changes with the application of Industry 4.0. With the emergence of digital transformation and cyber requirements, the manual-based work environment needs to be shifted to the internet and robot-based environment.

Bressanelli et al. (2018) highlighted the applicability of Industry 4.0 to overcome the challenges in the Product-Service Systems (PSS) Business model. Tao and Qi

(2019) proposed a framework for smart manufacturing using information technology. Dalenogare et al. (2018) highlighted the scope of Industry 4.0 in enhancing the performance of the industry. Rao and Prasad (2018) observed the impact of 5G technologies on Industry 4.0 and its functioning. Ang et al. (2017) studied the role of Industry 4.0 in the shipbuilding industry. They provided some significant inputs required for the smart manufacturing of the shipbuilding conforming to Industry 4.0. Qi and Tao (2018) compared the concept of industry 4.0 with the concept of big data in the field of smart manufacturing and came out with recommendations. Chong et al. (2018) conducted a study on the hybrid manufacturing process using digital technology and Industry 4.0.

Ivanov et al. (2019) outlined the role of digital technology in Industry 4.0 and recommended its integration for enhancing supply chain management. Zhang et al. (2019) suggested an IoT-based mechanism for a lean manufacturing system. Novais et al. (2019) conducted a systematic literature review on supply chain flexibility and mass customization. Lee et al. (2018) worked on the designing of an IoT-based smart warehouse system for convenient logistics. Durao et al. (2017) studied additive manufacturing in the context of localized spare parts production and distribution. Zanoni et al. (2019) also highlighted the applicability and contribution of additive manufacturing in supply chains. Srai and Lorentz (2019) worked on the concept of digitalization in the field of supply management and gave valuable suggestions. Kumar and Singh (2022) reviewed the clean and effective approach in Industry 4.0 applications using the Big data approach. Diez-Olivan et al. (2019) suggested the fusion of machine learning with data fusion to overcome the barriers in Industry 4.0. Ghadimi et al. (2019) suggested the use of multi-agent technology in enhancing Industry 4.0 in the supply chains. They highlighted the important point of an intelligent, sustainable supplier selection system. Some of the major applications of Industry 4.0 in manufacturing systems are summarized in Table 1.

2.2. *Barriers to adopting Industry 4.0 in Indian SMEs*

Deloitte (2014) highlighted the significant challenges/barriers in implementing Industry 4.0 and came out with a solution to overcome them. Zheng et al. (2021) conducted a literature review on the use of Industry 4.0 in the field of the manufacturing sector, and methods for its successful operations were recommended. In continuing the digital transformation, there is a need to study the products and services that evolved from Industry 4.0. These transformations are not simple to adopt as they possess challenges that are not easy to cater to. To explore these challenges "The resource-based view" theory is used in this study.

The resource-based view (RBV) is a widely accepted theory to manage the resources for competitive advantages explaining the firm's internal sources for sustained competitive advantage (SCA). For SCA, a firm is required to control the "valuable, rare, inimitable, and non-substitutable (VRIN)" resources and capabilities (Barney, 1991, 2002). This proposition is found in several connected studies, including the knowledge-based views (Grant, 1996), dynamic capabilities (Helfat & Peteraf, 2003; Teece et al., 1997), and. core competencies (Hamel & Prahalad, 1994). The RBV acknowledges the many elastic factors of production. Resource and capabilities development is part of long-term planning (Barney, 1991; Dierickx & Cool, 1989). In the implementation of Industry 4.0, resource plays an important role. It has been

Table 1. Application of Industry 4.0 in the Manufacturing Industry.

Technology	Application in Manufacturing Industry	References
Internet of Things (IoT)	supply chain optimization, asset tracking, inventory management, quality assurance, production process monitoring, and predictive maintenance	Soori et al. (2023)
Cyber-Physical System(CPS)	Predictive maintenance	Bokrantz et al. (2017)
	Mass customization and smart product design	Zawadzki and Żywicki (2016)
	Smart Monitoring (in operation, maintenance, and optimal scheduling)	Janakand Hadas (2015)
	Smart Controls (to manage various smart machines or tools physically)	Makarov et al. (2014)
Virtual reality and augmented reality (VR & AR)	Hybrid prototyping uses VR technology for additive manufacturing	Kolarevic (2004)
	CPS and AR are used to integrate the CAD & CAM software with a smart physical prototype system in real-time via 3D printing	Kolarevic (2004)
Big data and cloud computing	Data collection, processing, and utilization in manufacturing applications	Bi et al. (2023)
	Smart decision-making process consists of data-driven modeling and data-enabled predictive maintenance	Zhong et al. (2016)
	A comprehensive system based on real-time data for data-driven risk assessment	Niesen et al. (2016)
Cobot	Robotic operations involving human beings	Benotsmane et al. (2019), Chen (2017), Strandhagen et al. (2017)
	Manufacturing and automation	Ghobakhloo (2018), Strandhagen et al. (2017)
3D Printing	3D printing includes 3D scanners to capture the specific features and CAD software for subsequent printing	Chen et al. (2010)
	Real-time physical prototype development with the help of 3D printing utilizing CPS, AR and VR	Kolarevic (2004)

(*Continued*)

Table 1. Continued.

Technology	Application in Manufacturing Industry	References
Artificial Intelligence	Key enabling technologies for smart manufacturing systems	Chen et al. (2023), Xiang and Lee (2008)
	A range of artificial intelligence methodologies, including knowledge-based/expert systems, fuzzy logic, neural networks, genetic algorithms, and fuzzy logic, recommend suitable dispatching	Sabuncuoglu (1998) Álvarez and Díaz (2004), Chryssolouris and Subramaniam (2001), Grabot and Geneste (1994), Fox and Smith (1984)
Blockchain Technology	Data security for automation	Lu (2019)
	Smart procurement and sourcing	Srai and Lorentz (2019)
	Smart contract in energy systems	Mohamed et al. (2019)

observed that many barriers to adopting Industry 4.0 in SMEs are concerned with a lack of resources.

Hofmann and Rusch (2017) investigated the importance of Industry 4.0 and its applicability in future logistics. On similar grounds (Tang & Veelenturf, 2019), relate Industry 4.0 with the logistic-based problems. Many researchers have extensively investigated the barriers to the successful implementation of Industry 4.0 in various sectors, including the manufacturing sector, which are summarized in Table 2.

Based on the Literature review following research gaps have been identified:

(i) Very few SMEs have recognized the significance of digitalization, as they are obliged to enhance the quality of product/processes to fulfill the OEMs' demands.
(ii) SMEs differ from large manufacturing organizations in terms of level of flexibility, resources availability, a degree of inclination, and specialization toward niche markets.
(iii) SMEs are skilled in digitalizing only a few specific processes because they are focused on producing and supplying only limited components to OEM.

3. Research methodology

Some of the major applications of Industry 4.0 in manufacturing systems and the barriers to adopting Industry 4.0 are explored through an in-depth literature review. Many barriers were not very important for the SMEs, thus, only those barriers are incorporated in the analysis which was suggested by the experts. The expert members were also asked to suggest some barriers, which are not included in the discussion. They were asked to compare the Best (Most important) barriers with other barriers (BO) and other barriers to the Worst (Least Important) barrier (OW) on Saaty's 9-point rating scale. The BWM approach is used to rank the importance of the barriers so that effective allocation of resources and strategies can be made to overcome these barriers. The framework of the research methodology is shown in Figure 1.

Table 2. Barriers to implementing industry 4.0 in manufacturing systems.

References	Methodology	Barriers
Jankowska et al. (2023)	Literature review and empirical research	Insufficient capital expenditure, insufficient financial resources for staff training, Lack of intramural and R&D expenditures, lack of skills needed for automation process, lack of personnel, suboptimal administrative processes, inadequate form's strategy, need for HR recognition, lack of government support, regulatory barriers, external pressure for cost reduction
Senna et al. (2022)	Interpretive Structural Modeling	Increased level of investment, necessary adaptations at both organizational and operational stages, shortage of skilled workers, absence of a digital strategy, insufficient knowledge and data management systems, poor understanding of IoT benefits, lack of standardization initiatives, poor communication and IT infrastructure, concerns regarding safety, security, and privacy, insufficient integration and interoperability features, lack of regulatory framework, obstacle in adaptive retrofitting, lack of legal and contractual assurance
Cugno et al. (2021)	A mixed-method of interviews and case study approach	Inadequate information about Industry 4.0 technologies, insufficient know-how within companies, lack of skilled workforce, insufficient financial resources, scarcity of external funding, insufficient infrastructure, legal uncertainties, poor partnerships through engineering colleges and research institutes, lack of well-defined standards., resistance to organizational changes, sectors do not require investments in Industry 4.0
Kumar et al. (2021a)	Interpretive Structural Modeling and Analytic Network Process	Poor awareness, lack of a comprehensive framewok, shortage of skilled workers and lack of digital environment, poor physical facilities and infrastructure, lack of competencies and motivation, limited government support and initiatives, lack of sustaionable practices, no effective policy and protocol, poor acceptance, lack of circularity in design aspects, and fear of cultural changes

(*Continued*)

Table 2. Continued.

References	Methodology	Barriers
Kumar et al. (2021b)	Analytic Hierarchy Process, Elimination and Choice Expressing Reality	Lack of waste management, use of materials as energy, poor resource/ infrastructure quality, lack of awareness of Industry 4.0, and limited government support, resistance to change, poor market demand, poor management support, and short-term goals
Raj et al. (2020)	Grey Decision-Making Trial and Evaluation Laboratory Approach	The difference in economic benefit, limitations in value-chain integration, challenges breaches in security, irregularities in existing jobs, infrastructure shortage, digital skills shortage, implementation of quality data, internal digital culture and training, ineffective change management, and scarcity of resources
Kumar et al. (2020)	K-means clustering, Principal Component Analysis, and fuzzy analytical hierarchical process	Lack of IT Infrastructure, immatured data analytic techniques, missing standards, data insecurity, very high complexity, incompatibility of solutions/ devices/ retrofitting, huge investment, unavailability of partners and funding support, management support, poor government support, the unwarranted strain on the workforce, Lack of regulatory mechanism
Rezqianita and Ardi (2020)	Questionnaire survey	Lack of knowledge, inbuilt habits, difficultiy in strategy-formulation, poor knowledge management, poor infrastructure, lack of competence in leadership, incompetent workforce, inability to identify opportunities, poor commitment, unavailability of external expertise, limited time available for training, ecosystem policy, data security policy, poor government's infrastructure, no incentives, poor standards, threats cyber security, poor storage capacity
Chauhan et al. (2021)	Exploratory factor analysis using AMOS	Challenges at the elementary level, top management does not readily involve, inadaptability of new technology, improper digitalization, inadaptability of the employee for advancement in technology, data incompetency, improper tools for data interaction, and improper cost competency tool
Ozkan-Ozen et al. (2020)	Fuzzy Analytical Network Process	Inadequate information for industry 4.0, improper standards, required continued education and training of employees, required high investments, data sensitivity, data inadequately, data management, lack of management support for industry 4.0 transformation, inadequate environment for digitalization, inadequate environment for communication, unsupportive legal support system

Stentoft et al. (2021)	A statistical approach and questionnaire survey	Lack of standards, few human resources (manpower), absence of cyber security, inadequately qualified workers, lack of employee readiness, improper technology, and improper digitalization
Müller (2019)	Case study	Inadequate information for industry 4.0, improper standards, required continued education and training of employees, required high investments, data sensitivity, data inadequately, data management, lack of management support for industry 4.0 transformation, inadequate environment for digitalization, inadequate environment for communication, unsupportive legal support system
Türkeş et al. (2019)	Question-Answer-Based approach	Technical hindrances to the implementation of the fourth industrial revolution, unavailability of standards, data insensitivity, unavailability of proper techniques, unavailability of human resources,education and training of workers
Orzes et al. (2018)	Focus on group methodology	Increase in the cost related to manufacturing and supply of the products, untrustworthy of international trade due to large quantities, the deterioration in food nature and thus the minimized proper response, the challenging availability of resources due to the difficulty to access, a problem with the transport of the resource, a problem in imports of the resources and its legality, a problem of the allocation of resources, a problem of a factor of economy, inadequate supply of money and resources, variation in the price values, the proper value must be taken to keep consumable products safe, unavailability of the workers, the demand and supply behavior of the products, unavailability of food security, supply of high content food to workers
Kiel et al. (2017)	Questionnaire-based approach	Lack of knowledge of existing technology, organizational challenges, inconvenience in data handling, data insensitivity, accounting and costing in implementation, human resources

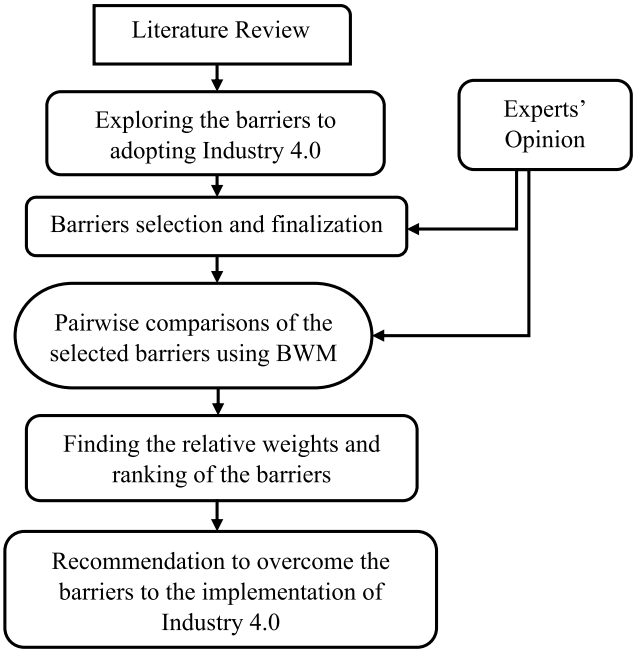


Figure 1. A framework for the research methodology.

3.1. *Best-worst method*

BWM determines the weights of the barriers to adopting Industry 4.0. It helps in minimizing the complexity and effort required to compare each criterion pairwise. Only the best (most significant) criterion is compared to other criteria, while the other criteria are compared to the worst (least significant) criterion. BWM provides more accurate findings in comparison to MCDM techniques (Rezaei, 2015; Rezaei, 2016). The following Mathematical steps are used in the BMW method to get the importance rating of the given criteria:

1. Identify the set of barriers to adopting Industry 4.0 in SMEs.
2. Identify the best and the worst barriers, based on the experts' opinions,
3. Rate the importance of the best barrier B compared to other barriers j and construct a vector X_B (best-to-other) Using Eq. (1)

$$X_B = [X_{B1}, X_{B2}, ---, X_{Bn}] \tag{1}$$

Where X_{BJ} shows the preference of the best barrier over other barriers J, and $X_{BB} = 1$.

4. Rate the importance of the barriers j compared to the worst barrier W on a 9-point rating scale and construct a vector X_W (other-to-worst) vector as mentioned in Eq. (2):

$$X_W = [X_{1W}, X_{2W}, ---, X_{nW}]^T \tag{2}$$

Where X_{JW} shows the preference of the other barriers j over the worst barrier W. It is clear that

$$X_{WW} = 1$$

5. Find the optimal weights of each barrier.

For finding the optimal weights of the different barriers, the maximum absolute differences $\left|\frac{W_B}{W_j} - x_{Bj}\right|\left|\frac{W_j}{W_W} - x_{jW}\right|$ for all j is to be minimized, which may be represented as Eq. (3):

$$\min_j \max\left\{\left|\frac{W_B}{W_j} - x_{Bj}\right|, \left|\frac{W_j}{W_W} - x_{jW}\right|\right\} \tag{3}$$

Subject to Wj ≥ 0, for all j

The model written above may be reproduced in the form of Eqs. (4-6)

min ξ

Subject to

$$\left|\frac{W_B}{W_j} - x_{Bj}\right| \leq \xi \quad \forall j \tag{4}$$

$$\left|\frac{W_j}{W_W} - x_{jW}\right| \leq \xi, \quad \forall j \tag{5}$$

$$\sum_j Wj = 1 \tag{6}$$

$$W_j \geq 0, \quad \forall j$$

W_j^* (optimal weights) and ξ*can be obtained by solving the LP model and using the consistency index (CI) as given in Table 3 (Rezaei, 2015; Rezaei, 2016).

The consistency ratio can be obtained using ξ* and the corresponding CI as shown in Eq. (7)

$$CR = \xi^*/CI \tag{7}$$

CR ≅ 0 is considered as high consistency; however, no threshold value for the CR.

Table 3. The consistency indices for the BWM.

X_{BW}	1	2	3	4	5	6	7	8	9
Consistency Index	0.00	0.44	1.00	1.63	2.30	3.00	3.73	4.47	5.23

4. Case analysis and results discussion

In this study, many experts were invited to participate in the screening and evaluation of the barriers to adopting Industry 4.0 in SMEs. A total of 12 experts gave their consent to the participation. Out of 12, 5 experts were from auto-component manufacturing, 3 experts were from electrical and electronic products manufacturing, and 4 experts were from the plastic moulding and manufacturing industries. They were asked about the barriers to adopting Industry 4.0 technology. Some barriers have already been explored through the literature review and were screened with the opinion of the experts. They shortlisted 25 barriers including some exclusive barriers from this study. These barriers are classified into five groups Government policy, Management initiatives, Human Resources/workforce, Infrastructure and financial resources, and Technology. The barriers incorporated in this study are shown in Table 4.

All 12 experts were asked to find the Best (most important) barrier and Worst Least important barrier among these five groups of barriers and the barriers in each group. Using BWM, the Best to Other barriers (BO) and Other to Worst (OW) barriers are compared on Saaty's 9-point rating scale. The intermediate tables of comparison and calculation based on BWM are shown in Appendix (Table A1-A18). With the help of the application of the BWM method, we obtained a global and local ranking of shortlisted barriers. The global and local ranking of the barriers of Industry 4.0 is given in Table 5.

The lack of infrastructure is the most important barrier to adopting Industry 4.0 in SMEs. It has been observed that SMEs are facing the problems of limited infrastructure. The second most important barrier is the lack of financial resources. Due to the unavailability of financial resources and high-quality infrastructures such as smart mobility, smart logistics, and digital technology, it becomes tough to adopt Industry 4.0 in the manufacturing industry.

The third important barrier is the lack of government initiative. The government is working proactively to create an ecosystem for the adoption of Industry 4.0. The fourth important barrier is the very high complexity of the technologies. Due to unavailability of the skilled workers, the complexity of technology becomes an important barrier. The fifth and sixth important barrier is concern about cybersecurity and data ownership issues and policy about data security. Cybersecurity and centralized data storage and control systems are assessment points to measure a company's readiness to adopt Industry 4.0. Because of the increased connection and usage of standard communications protocols that come with Industry 4.0, there is a greater need to secure key industrial systems from cybersecurity attacks. Similarly, all the other barriers are ranked as per the importance shown by the experts and using the BWM approach.

Managerial Implications: This study may help managers explore the barriers or challenges before implementing Industry 4.0. Many barriers in adopting Industry 4.0 have been explored through the literature review which may help the managers in finding the relevant barriers in the different manufacturing organizations. They can check the readiness of the organization for digitalization. They can extend the research after ranking the importance of the barriers in indexing organizations for readiness and they can prepare a readiness index using various quantitative models. Also, they can allocate their resources according to the importance of

Table 4. Barriers to implementation of Industry 4.0.

		Jankowska et al. (2023)	Senna et al. (2022)	Kumar et al. (2020)	Raj et al. (2020)	Kumar et al. (2021b)	Rezqianita and Ardi (2020)	Chauhan et al. (2021)	Ozkan-Ozen et al. (2020)
Government Policy (GP)	Lack of government initiatives (GP11)	√		√		√			
	Policy about data security (GP12)					√	√		
	Lack of Standards and Reference Architecture (GP13)				√	√	√		
	Regulatory Compliance issues (GP14)	√	√						
	Missing partners and funding support (GP15)								
Management Initiatives (MI)	Risk of misinvestment (MI21)		√	√					
	Structure/resistance to change (MI22)				√				
	Lack of waste management strategy (MI23)			√					
	Lack of management support (MI24)			√		√			√
	Lack of knowledge management systems (MI25)		√	√			√		
	Lack of planning skills and activities (MI26)	√							

(*Continued*)

Table 4. Continued.

		Jankowska et al. (2023)	Senna et al. (2022)	Kumar et al. (2020)	Raj et al. (2020)	Kumar et al. (2021b)	Rezqianita and Ardi (2020)	Chauhan et al. (2021)	Ozkan-Ozen et al. (2020)
Human Resources/ Workforce (HR)	Lack of trust between Partners (HR31)								
	Lack of Support from Customer/supplier (HR32)								
	Lack of competence of the workforce/Limited skilled workers (HR33)		√	√		√	√		√
	Lack of external experts (HR34)								
	Disruption to existing jobs (HR35)				√				
Infrastructure and Financial Resources (IFR)	Lack of Infrastructure (IFR41)			√	√		√		
	Lack of financial resources (IFR42)	√		√					
	Lack of a digital strategy alongside resource scarcity (IFR43)				√			√	

Technological Issues (TI)	Security and Privacy Issues (TI51)			√	√			√
	Seamless integration and compatibility issues (TI52)		√		√			
	Very high complexity (TI53)				√			
	Awareness about the potentialities of robot applications (TI54)							
	Challenges in ensuring data Quality (TI55)							
	Concerns about cybersecurity and data ownership issues (TI56)					√		

		Stentoft et al. (2021)	Horvath and Szabo (2019)	Müller (2019)	Türkeş et al. (2019)	Kamble et al. (2018)	Orzes et al. (2018)	Experts'Suggestions
Government Policy (GP)	Lack of government initiatives (GP11)						√	
	Policy about data security (GP12)					√		
	Lack of Standards and Reference Architecture (GP13)	√			√	√	√	
	Regulatory Compliance issues (GP14)					√		
	Missing partners and funding support (GP15)							√

(*Continued*)

Table 4. Continued.

		Stentoft et al. (2021)	Horvath and Szabo (2019)	Müller (2019)	Türkeș et al. (2019)	Kamble et al. (2018)	Orzes et al. (2018)	Experts'Suggestions
Management Initiatives (MI)	Risk of misinvestment (MI21)							
	Structure/resistance to change (MI22)							
	Lack of waste management strategy (MI23)							
	Lack of management support (MI24)			√			√	
	Lack of knowledge management systems (MI25)					√		
	Lack of planning skills and activities (MI26)							√
Human Resources/ Workforce (HR)	Lack of trust between Partners (HR31)						√	
	Lack of Support from Customer/supplier (HR32)						√	
	Lack of competence of the workforce/Limited skilled workers (HR33)	√						
	Lack of external experts (HR34)							√
	Disruption to existing jobs (HR35)							

Infrastructure and Financial Resources (IFR)	Lack of Infrastructure (IFR41)					√	
	Lack of financial resources (IFR42)	√	√			√	
	Lack of a digital strategy alongside resource scarcity (IFR43)						
Technological Issues (TI)	Security and Privacy Issues (TI51)				√		
	Seamless integration and compatibility issues (TI52)				√		
	Very high complexity (TI53)					√	
	Awareness about the potentialities of robot applications (TI54)					√	
	Challenges in ensuring data Quality (TI55)						√
	Concerns about cybersecurity and data ownership issues (TI56)		√				

Table 5. Local and Global Ranking of the barriers to implementation of Industry 4.0.

Barriers to Industry 4.0	Weight	Effects of Barriers on Industry 4.0	Local Weight	Local Ranking	Global Weight	Global Ranking
Government Policy (GP)	0.252	Lack of government initiatives (GP11)	0.339	1	0.085	3
		Policy about data security (GP12)	0.202	2	0.051	6
		Lack of standards and reference architecture (GP13)	0.100	5	0.025	17
		Regulatory compliance issues (GP14)	0.195	3	0.049	7
		Missing partners and funding support (GP15)	0.164	4	0.041	10
Management Initiatives (MI)	0.165	Risk of misinvestment (MI21)	0.185	3	0.031	14
		Structure/resistance to change (MI22)	0.091	5	0.015	22
		Lack of waste management strategy (MI23)	0.066	6	0.011	24
		Lack of management support (MI24)	0.126	4	0.021	21
		Lack of knowledge management systems (MI25)	0.296	1	0.049	8
		Lack of planning skills and activities (MI26)	0.235	2	0.039	11
Human Resources/ Workforce (HR)	0.109	Lack of trust between partners (HR31)	0.081	5	0.009	25
		Lack of support from customer/supplier (HR32)	0.119	4	0.013	23
		Lack of competence of workforce/limited skilled workers (HR33)	0.350	1	0.038	12
		Lack of external experts (HR34)	0.249	2	0.027	16
		Disruption to existing jobs (HR35)	0.202	3	0.022	19
Infrastructure & Financial Resources (IFR)	0.223	Lack of infrastructure (IFR41)	0.465	1	0.104	1
		Lack of financial resources (IFR42)	0.383	2	0.085	2
		Lack of a digital strategy alongside resource scarcity (IFR43)	0.152	3	0.034	13
Technology (T)	0.252	Security and privacy issues (T51)	0.188	3	0.047	9
		Seamless integration and compatibility issues (T52)	0.113	4	0.028	15
		Very high complexity (T53)	0.261	1	0.066	4
		Awareness about the potentialities of robot applications (T54)	0.093	5	0.023	18
		Challenges in ensuring data quality (T55)	0.085	6	0.021	20
		Concerns about cybersecurity and data ownership issues (T56)	0.260	2	0.065	5

barriers, so that they can digitalize the systems using Industry 4.0. The manager can select the most important and least important barriers very easily and these barriers can be compared with other barriers with less effort in comparison to other multi-criteria decision-making methods.

Theoretical Implications: This study may help the researchers to explore some more barriers to adopting Industry 4.0 in different industries. Some barriers may be dependent on or related to other barriers. These relationships may be explored through the questionnaire survey and statistical analysis. To overcome these barriers, this study may motivate the researchers to explore the enablers for the successful implementation of Industry 4.0.

5. Conclusions

In this study, an in-depth literature review was conducted to collect information regarding the different barriers to adopting Industry 4.0 in the manufacturing industry (SMEs). Industry experts were consulted to know their opinions regarding the importance of the barriers. These barriers were rated on a 9-point rating scale, and the Best-Worst method was used to find the local and global weightage of different barriers. It is observed that the most important barrier is the lack of infrastructure for proper implementation of the technologies concerned with Industry 4.0. With infrastructure, financial support becomes the second most important barrier. The third important barrier is the lack of government initiatives regarding the development of infrastructure, and financial support was observed. In addition, the implementation of the different kinds of technologies becomes very complex for SMEs. Therefore, the high level of complexity in the data management and implementation of Industry 4.0 was observed as the fourth important barrier. The fifth important barrier in the handling of big data, the cyber security and data ownership issue, was observed. Similarly, other barriers were also ranked.

The future scope of this research is as follows. This study is based on exploring and rating the barriers using BWM and is limited to SMEs. The same framework can be used in other industries for observing the challenges in the implementation of various technologies related to Industry 4.0. Also, readiness indexing can be made for the organizations for implementing Industry 4.0.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Pravin Kumar http://orcid.org/0000-0002-3378-7087

References

Alvarez, E., & Diaz, F. (2004). An application of a real-time scheduling system for turbulent manufacturing environments. *Robotics and Computer-Integrated Manufacturing*, *20*(6), 485–494.

Ang, J. H., Goh, C., Saldivar, A. A. F., & Li, Y. (2017). Energy-efficient through-life smart design, manufacturing, and operation of ships in an industry 4.0 environment. *Energies*, *10*(5), 610.

Arnold, C., Kiel, D., & Voigt, K. I. (2017). Innovative business models for the industrial internet of things. *BHM Berg-und Huttenmannische Monatshefte*, *162*(9), 371–381.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99–120.

Barney, J. B. (2002). *Gaining and sustaining competitive advantage*. Prentice-Hall.

Benotsmane, R., Kovacs, G., & Dudas, L. (2019). Economic, social impacts and operation of smart factories in industry 4.0 focusing on simulation and artificial intelligence of collaborating robots. *Social Sciences*, *8*(5), 143.

Bi, Z., Jin, Y., Maropoulos, P., Zhang, W. J., & Wang, L. (2023). Internet of things (IoT) and big data analytics (BDA) for digital manufacturing (DM). *International Journal of Production Research*, *61*(12), 4004–4021.

Bokrantz, J., Skoogh, A., Berlin, C., & Stahre, J. (2017). Maintenance in digitalised manufacturing: Delphi-based scenarios for 2030. *International Journal of Production Economics*, *191*, 154–169.

Bressanelli, G., Adrodegari, F., Perona, M., & Saccani, N. (2018). The role of digital technologies to overcome circular economy challenges in PSS business models: An exploratory case study. *Procedia Cirp*, *73*, 216–221.

Chauhan, C., Singh, A., & Luthra, S. (2021). Barriers to industry 4.0 adoption and its performance implications: An empirical investigation of emerging economy. *Journal of Cleaner Production*, *285*, 124809.

Chen, W., He, W., Shen, J., Tian, X., & Wang, X. (2023). Systematic analysis of artificial intelligence in the era of industry 4.0. *Journal of Management Analytics*, *10*(1), 89–108.

Chen, X., Wang, Y., & Yin, Z. (2010, August). *RFID based production and distribution management systems for home appliance industry*. 2010 IEEE international conference on automation and logistics, IEEE, pp. 177-182.

Chong, L., Ramakrishna, S., & Singh, S. (2018). A review of digital manufacturing-based hybrid additive manufacturing processes. *The International Journal of Advanced Manufacturing Technology*, *95*(5), 2281–2300.

Chryssolouris, G., & Subramaniam, V. (2001). Dynamic scheduling of manufacturing job shops using genetic algorithms. *Journal of Intelligent Manufacturing*, *12*(3), 281–293.

Cugno, M., Castagnoli, R., & Buchi, G. (2021). Openness to industry 4.0 and performance: The impact of barriers and incentives. *Technological Forecasting and Social Change*, *168*, 120756.

Dalenogare, L. S., Benitez, G. B., Ayala, N. F., & Frank, A. G. (2018). The expected contribution of industry 4.0 technologies for industrial performance. *International Journal of Production Economics*, *204*, 383–394.

Deloitte. (2014). Industry 4.0 challenges and solutions for the digital transformation and use of exponential technologies, https://www2.deloitte.com/ch/en/pages/manufacturing/articles/manufacturing-study- industry-4.html Retrieved on June 7, 2021.

Dierickx, I., & Cool, K. (1989). Asset stock accumulation and sustainability of competitive advantage. *Management Science*, *35*(12), 1504–1511.

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. Information. *Fusion*. , *50*, 92–111.

Durao, L. F. C., Christ, A., Zancul, E., Anderl, R., & Schutzer, K. (2017). Additive manufacturing scenarios for distributed production of spare parts. *The International Journal of Advanced Manufacturing Technology*, *93*(1), 869–880.

Dutta, G., Kumar, R., Sindhwani, R., & Singh, R. K. (2020). Digital transformation priorities of India's discrete manufacturing SMEs-a conceptual study in perspective of industry 4.0 . *Competitiveness Review: An International Business Journal*, *30*(3), 289–314.

Fox, M. S., & Smith, S. F. (1984). ISIS – a knowledge-based system for factory scheduling. *Expert Systems*, *1*(1), 25–49.

Ghadimi, P., Wang, C., Lim, M. K., & Heavey, C. (2019). Intelligent sustainable supplier selection using multi-agent technology: Theory and application for industry 4.0 supply chains. *Computers & Industrial Engineering*, *127*, 588–600.

Ghobakhloo, M. (2018). The future of manufacturing industry: A strategic roadmap toward industry 4.0. *Journal of Manufacturing Technology Management*, *29*(6), 910–936.
Grabot, B., & Geneste, L. (1994). Dispatching rules in scheduling dispatching rules in scheduling: A fuzzy approach. *The International Journal of Production Research*, *32*(4), 903–915.
Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal*, *17*(S2), 109–122.
Hamel, G., & Prahalad, C. K. (1994). *Competing for the future*. Harvard Business School Press.
Helfat, C. E., & Peteraf, M. A. (2003). The dynamic resource-based view: Capability lifecycles. *Strategic Management Journal*, *24*(10), 997–1010.
Hofmann, E., & Rusch, M. (2017). Industry 4.0 and the current status as well as future prospects on logistics. *Computers in Industry*, *89*, 23–34.
Horvath, D., & Szabo, R. Z. (2019). Driving forces and barriers of industry 4.0: Do multinational and small and medium-sized companies have equal opportunities? *Technological Forecasting and Social Change*, *146*, 119–132.
Ivanov, D., Dolgui, A., & Sokolov, B. (2019). The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research*, *57*(3), 829–846.
Janak, L., & Hadas, Z. (2015). Machine tool health and usage monitoring system: An initial analyses. *MM Science Journal*, *4*, 794–798.
Jankowska, B., Minska-Struzik, E., Bartosik-Purgat, M., Gotz, M., & Olejnik, I. (2023). Industry 4.0 technologies adoption: Barriers and their impact on Polish companies' innovation performance. *European Planning Studies*, *31*(5), 1029–1049.
Kamble, S. S., Gunasekaran, A., & Sharma, R. (2018). Analysis of the driving and dependence power of barriers to adopt industry 4.0 in Indian manufacturing industry. *Computers in Industry*, *101*, 107–119.
Kiel, D., Muller, J. M., Arnold, C., & Voigt, K. I. (2017). Sustainable industrial value creation: Benefits and challenges of industry 4.0. *International Journal of Innovation Management*, *21* (8), 1740015.
Kolarevic, B. (2004). Architecture in the digital age: design and manufacturing. Taylor & Francis. http://papers.cumincad.org/data/works/att/a3ac.content.07119.pdf Retrieved on June 21, 2022.
Kumar, P., & Singh, R. K. (2022). Application of industry 4.0 technologies for effective coordination in humanitarian supply chains: A strategic approach. *Annals of Operations Research*, *319*, 379–411.
Kumar, P., Singh, R. K., & Kumar, V. (2021b). Managing supply chains for sustainable operations in the era of industry 4.0 and circular economy: Analysis of barriers. Resources. *Conservation and Recycling*, *164*, 105215.
Kumar, S., Raut, R. D., Nayal, K., Kraus, S., Yadav, V. S., & Narkhede, B. E. (2021a). To identify industry 4.0 and circular economy adoption barriers in the agriculture supply chain by using ISM-ANP. *Journal of Cleaner Production*, *293*, 126023.
Kumar, S., Suhaib, M., & Asjad, M. (2020). Narrowing the barriers to industry 4.0 practices through PCA-fuzzy AHP-K means. *Journal of Advances in Management Research*, *18*(2), 200–226.
Laudante, E. (2017). Industry 4.0, innovation and design. A new approach for ergonomic analysis in the manufacturing system. *The Design Journal*, *20*(sup1), S2724–S2734.
Lee, C. K., Lv, Y., Ng, K. K. H., Ho, W., & Choy, K. L. (2018). Design and application of internet of things-based warehouse management system for smart logistics. *International Journal of Production Research*, *56*(8), 2753–2768.
Li, L. (2018). China's manufacturing locus in 2025: With a comparison of "made-in-China 2025" and "industry 4.0". *Technological Forecasting and Social Change*, *135*, 66–74.
Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. *Journal of Management Analytics*, *6*(1), 1–29.
Makarov, O., Langmann, R., Nesteresko, S., & Frank, B. (2014). Problems of the time deterministic in applications for process control from the cloud. *International Journal of Online Engineering*, *10*(4), 70–73.

Mckinsey . (2018). available at: https://www.mckinsey.com/business-functions/operations/our-insights/operations-blog/how-industry-40-can-help-india-escape-the-western-labor – productivity-paradox. (Retrieved on May 11, 2022).

Menon, S., Shah, S., & Coutroubis, A. (2018, November). *An overview of smart manufacturing for competitive and digital global supply chains*. 2018 IEEE international conference on technology management, operations and decisions (ICTMOD). IEEE, pp. 178-183.

Mohamed, N., Al-Jaroodi, J., & Lazarova-Molnar, S. (2019). Leveraging the capabilities of industry 4.0 for improving energy efficiency in smart factories. *IEEE Access*, *7*, 18008–18020.

Muller, J. M. (2019). Assessing the barriers to industry 4.0 implementation from a workers' perspective. *IFAC-PapersOnLine*, *52*(13), 2189–2194.

Niesen, T., Houy, C., Fettke, P., & Loos, P. (2016, January). *Towards an integrative big data analysis framework for data-driven risk management in industry 4.0*. 2016 49th Hawaii international conference on system sciences (HICSS). IEEE, pp. 5065-5074.

Novais, L. R., Maqueira, J. M., & Bruque, S. (2019). Supply chain flexibility and mass personalization: A systematic literature review. *Journal of Business & Industrial Marketing*, *34*(8), 1791–1812.

Orzes, G., Rauch, E., Bednar, S., & Poklemba, R. (2018, December). *Industry 4.0 implementation barriers in small and medium sized enterprises: A focus group study*. 2018 IEEE international conference on industrial engineering and engineering management (IEEM). IEEE, pp. 1348-1352.

Ozkan-Ozen, Y. D., Kazancoglu, Y., & Mangla, S. K. (2020). Synchronized barriers for circular supply chains in industry 3.5/industry 4.0 transition for sustainable resource management. Resources. *Conservation and Recycling*, *161*, 104986.

Oztemel, E., & Gursev, S. (2020). Literature review of industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, *31*(1), 127–182.

Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access*, *6*, 3585–3593.

Raj, A., Dwivedi, G., Sharma, A., de Sousa Jabbour, A. B. L., & Rajak, S. (2020). Barriers to the adoption of industry 4.0 technologies in the manufacturing sector: An inter-country comparative perspective. *International Journal of Production Economics*, *224*, 107546.

Rao, S. K., & Prasad, R. (2018). Impact of 5G technologies on industry 4.0. *Wireless Personal Communications*, *100*, 145–159.

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57.

Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, *64*, 126–130.

Rezqianita, B. L., & Ardi, R. (2020, June). *Drivers and barriers of industry 4.0 adoption in Indonesian manufacturing industry*. Proceedings of the 3rd Asia pacific conference on research in industrial and systems engineering 2020, pp. 123-128.

Sabuncuoglu, I. (1998). Scheduling with neural networks: A review of the literature and new research directtions. *Production Planning & Control*, *9*(1), 2–12.

Senna, P. P., Ferreira, L. M. D., Barros, A. C., Roca, J. B., & Magalhaes, V. (2022). Prioritizing barriers for the adoption of industry 4.0 technologies. *Computers & Industrial Engineering*, *171*, 108428.

Soori, M., Arezoo, B., & Dastres, R. (2023). Internet of things for smart factories in industry 4.0, a review. *Internet of Things and Cyber-Physical Systems*, *3*, 192–204.

Srai, J. S., & Lorentz, H. (2019). Developing design principles for the digitalisation of purchasing and supply management. *Journal of Purchasing and Supply Management*, *25*(1), 78–98.

Stentoft, J., Adsboll Wickstrom, K., Philipsen, K., & Haug, A. (2021). Drivers and barriers for industry 4.0 readiness and practice: Empirical evidence from small and medium-sized manufacturers. *Production Planning & Control*, *32*(10), 811–828.

Strandhagen, J. W., Alfnes, E., Strandhagen, J. O., & Vallandingham, L. R. (2017). The fit of industry 4.0 applications in manufacturing logistics: A multiple case study. *Advances in Manufacturing*, *5*(4), 344–358.

Tang, C. S., & Veelenturf, L. P. (2019). The strategic role of logistics in the industry 4.0 era. *Transportation Research Part E: Logistics and Transportation Review*, *129*, 1–11.

Tao, F., & Qi, Q. (2019). New IT driven service-oriented smart manufacturing: Framework and characteristics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(1), 81–91.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, *18*(7), 509–533.

Thakkar, J., Kanda, A., & Deshmukh, S. G. (2012). Supply chain issues in Indian manufacturing SMEs: Insights from six case studies. *Journal of Manufacturing Technology Management*, *23*(5), 634–664.

Turkes, M. C., Oncioiu, I., Aslam, H. D., Marin-Pantelescu, A., Topor, D. I., & Capusneanu, S. (2019). Drivers and barriers in using industry 4.0: A perspective of SMEs in Romania. *Processes*, *7*(3), 153.

Xiang, W., & Lee, H. P. (2008). Ant colony intelligence in multi-agent dynamic manufacturing scheduling. *Engineering Applications of Artificial Intelligence*, *21*(1), 73–85.

Zanoni, S., Ashourpour, M., Bacchetti, A., Zanardini, M., & Perona, M. (2019). Supply chain implications of additive manufacturing: A holistic synopsis through a collection of case studies. *The International Journal of Advanced Manufacturing Technology*, *102*(9), 3325–3340.

Zawadzki, P., & Zywicki, K. (2016). Smart product design and production control for effective mass customization in the industry 4.0 concept. *Management and Production Engineering Review*, *7*(3), 105–112.

Zhang, K., Qu, T., Zhou, D., Thürer, M., Liu, Y., Nie, D., Li, C., & Huang, G. Q. (2019). Iot – enabled dynamic lean control mechanism for typical production systems. *Journal of Ambient Intelligence and Humanized Computing*, *10*(3), 1009–1023.

Zheng, T., Ardolino, M., Bacchetti, A., & Perona, M. (2021). The applications of industry 4.0 technologies in manufacturing context: A systematic literature review. *International Journal of Production Research*, *59*(6), 1922–1954.

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering*, *101*, 572–591.

Appendix

Table A1. Best to Other vectors for five groups of criteria in the implementation of Industry 4.0.

Respondent No.	The Best (Most Important) barrier↓	Government Policy (GP)	Management Initiatives (MI)	Human Resources / Workforce (HR)	Infrastructure & Financial Resources (IFR)	Technology (T)
1	Technology (TI)	2	4	8	2	1
2	Technology (TI)	2	3	7	3	1
3	Technology (TI)	2	5	8	2	1
4	Infrastructure & Financial Resources (IFR)	2	5	9	1	2
5	Infrastructure & Financial Resources (IFR)	3	8	5	1	2
6	Government Policy (GP)	1	7	6	6	4
7	Human Resources / Workforce (HR)	2	8	1	4	5
8	Management Initiatives (MI)	6	1	3	4	4
9	Management Initiatives (MI)	4	1	6	4	3
10	Government Policy (GP)	1	8	5	6	4
11	Infrastructure & Financial Resources (IFR)	2	4	7	1	2
12	Technology (TI)	4	2	6	2	1

Table A2. Other to Worst vectors for five groups criteria to compare the barriers to implementation of Industry 4.0.

	Respondents											
	1	2	3	4	5	6	7	8	9	10	11	12
Worst (Least Important) barrier →	HR	HR	HR	HR	MI	MI	MI	GP	HR	MI	HR	HR
Government Policy: GP	6	5	5	6	2	7	4	1	3	8	5	3
Management Initiatives: MI	3	4	3	4	1	1	1	5	6	1	3	6
Human Resources / Workforce: HR	1	1	1	1	3	2	6	3	1	3	1	1
Infrastructure & Financial Resources: IFR	5	5	4	8	7	2	2	3	3	3	7	6
Technology: T	7	8	7	6	5	3	3	2	4	4	6	8

Table A3. Optimal weights of the five groups of criteria for implementation of Industry 4.0.

Respondent No.	Government policy (GP)	Management Initiatives (MI)	Human Resources / Workforce (HR)	Infrastructure & Financial Resources (IFR)	Technology (T)	Ksi*	CR
1	0.225	0.112	0.047	0.225	0.391	0.059	0.026
2	0.237	0.158	0.047	0.158	0.400	0.074	0.032
3	0.229	0.092	0.049	0.229	0.401	0.056	0.024
4	0.234	0.094	0.041	0.398	0.234	0.070	0.031
5	0.165	0.053	0.099	0.435	0.248	0.061	0.027
6	0.564	0.072	0.104	0.104	0.156	0.060	0.026
7	0.261	0.061	0.443	0.130	0.104	0.078	0.034
8	0.075	0.464	0.184	0.138	0.138	0.088	0.038
9	0.138	0.474	0.066	0.138	0.184	0.079	0.034
10	0.552	0.059	0.126	0.105	0.158	0.079	0.034
11	0.225	0.112	0.047	0.391	0.225	0.059	0.026
12	0.114	0.228	0.051	0.228	0.380	0.076	0.033
Average	**0.252**	**0.165**	**0.109**	**0.223**	**0.252**	**0.070**	**0.030**

Table A4. Best to other barriers related to Government Policy for implementation of Industry 4.0.

Respondent No.	Best (Most Important) Barrier↓	GP11	GP12	GP13	GP14	GP15
1	GP11	1	4	8	7	5
2	GP14	3	3	7	1	3
3	GP11	1	5	8	7	2
4	GP12	3	1	4	8	4
5	GP13	3	7	1	3	2
6	GP11	1	4	8	3	5
7	GP12	2	1	3	2	4
8	GP14	2	4	6	1	2
9	GP11	1	3	8	6	6
10	GP11	1	4	8	5	3
11	GP14	5	2	7	1	2
12	GP11	1	2	8	5	5

Table A5. Other to Worst vectors for Government Policy (GP) to compare the barriers to implementation of Industry 4.0.

	Respondents											
	1	2	3	4	5	6	7	8	9	10	11	12
Worst (Least Important) barrier→	GP 13	GP 13	GP 13	GP 14	GP 12	GP 13	GP 15	GP 13	GP 13	GP 13	GP 13	GP 13
Lack of government initiatives (GP11)	7	3	8	4	2	8	4	5	8	8	2	7
Policy about data security (GP12)	3	4	3	7	1	2	6	3	5	2	5	5
Lack of standards and reference architecture (GP13)	1	1	1	2	6	1	2	1	1	1	1	1
Regulatory compliance issues (GP14)	2	8	2	1	2	5	4	7	3	2	8	3
Missing partners and funding support (GP15)	3	4	6	4	4	2	1	5	3	5	6	2

Table A6. Optimal weights of the barriers related to Government Policy (GP) for implementation of industry 4.0.

Respondent No.	Lack of government initiatives GP11	Policy about data security GP12	Lack of standards and reference architecture GP13	Regulatory compliance issues GP14	Missing partners and funding support GP15	Ksi*	CR
1	0.556	0.158	0.068	0.091	0.127	0.078	0.034
2	0.167	0.167	0.056	0.444	0.167	0.056	0.024
3	0.488	0.109	0.054	0.078	0.271	0.054	0.024
4	0.187	0.477	0.140	0.056	0.140	0.084	0.037
5	0.149	0.064	0.415	0.149	0.223	0.032	0.014
6	0.498	0.143	0.053	0.191	0.115	0.074	0.032
7	0.214	0.357	0.143	0.214	0.071	0.071	0.031
8	0.222	0.111	0.056	0.389	0.222	0.056	0.024
9	0.534	0.205	0.057	0.102	0.102	0.080	0.035
10	0.498	0.143	0.053	0.115	0.191	0.074	0.032
11	0.092	0.230	0.048	0.400	0.230	0.061	0.026
12	0.465	0.265	0.057	0.106	0.106	0.065	0.028
Average	**0.339**	**0.202**	**0.100**	**0.195**	**0.164**	**0.065**	**0.028**

Table A7. Best to Other barriers related to Management Initiatives (MI) for implementation of Industry 4.0.

Respondent No.	Best (Most Important) barrier↓	MI21	MI22	MI23	MI24	MI25	MI26
1	MI26	8	7	7	2	3	1
2	MI25	3	6	8	7	1	4
3	MI25	5	5	8	6	1	2
4	MI26	6	8	6	2	3	1
5	MI25	2	7	8	6	1	3
6	MI25	3	4	7	5	1	3
7	MI21	1	3	7	3	3	5
8	MI21	1	2	5	3	3	4
9	MI25	4	7	8	5	1	3
10	MI26	3	8	4	5	2	1
11	MI25	2	6	7	8	1	2
12	MI26	2	8	3	2	2	1

Table A8. Over Worst vectors for Management Initiatives (MI) to compare the barriers to implementation of Industry 4.0.

	Respondents											
	1	2	3	4	5	6	7	8	9	10	11	12
Worst (Least Important) barrier→	MI 21	MI 23	MI 23	MI 22	MI 23	MI 23	MI23	MI 23	MI 23	MI 22	MI 24	MI 22
Risk of mis – investment (MI21)	1	4	3	3	4	5	8	7	4	5	5	3
Structure/resistance to change (MI22)	2	2	3	1	2	3	5	5	2	1	4	1
Lack of waste management strategy (MI23)	3	1	1	2	1	1	1	1	1	4	3	2
Lack of management support (MI24)	6	2	2	4	3	2	5	4	3	3	1	4
Lack of knowledge management systems (MI25)	5	7	8	3	6	7	4	3	8	6	8	4
Lack of planning skills and activities (MI26)	8	4	6	7	4	5	3	2	5	8	6	6

Table A9. Optimal weights of the barriers related to Management Initiatives (MI)for implementation of Industry 4.0.

Respondent No.	Risk of misinvestment: MI21	Structure/ resistance to change: MI22	Lack of waste management strategy: MI23	Lack of management support: MI24	Lack of knowledge management systems: MI25	Lack of planning skills and activities: MI26	Ksi*	CR
1	0.044	0.069	0.069	0.241	0.160	0.417	0.064	0.021
2	0.181	0.090	0.054	0.078	0.461	0.136	0.081	0.027
3	0.097	0.097	0.048	0.081	0.435	0.242	0.048	0.016
4	0.078	0.048	0.078	0.235	0.157	0.404	0.066	0.022
5	0.240	0.068	0.053	0.080	0.399	0.160	0.080	0.027
6	0.162	0.121	0.048	0.097	0.410	0.162	0.076	0.025
7	0.394	0.156	0.046	0.156	0.156	0.093	0.073	0.024
8	0.347	0.211	0.054	0.141	0.141	0.106	0.076	0.025
9	0.132	0.076	0.049	0.106	0.461	0.176	0.069	0.023
10	0.148	0.041	0.111	0.089	0.223	0.387	0.058	0.019
11	0.225	0.075	0.064	0.037	0.374	0.225	0.075	0.025
12	0.176	0.044	0.118	0.176	0.176	0.309	0.044	0.015
Average	**0.185**	**0.091**	**0.066**	**0.126**	**0.296**	**0.235**	**0.068**	**0.023**

Table A10. Best to Other barriers related to Human resources/Workforce (HR) for implementation of Industry 4.0.

Respondent No.	Best (Most Important) barrier↓	HR31	HR32	HR33	HR34	HR35
1	HR33	8	5	1	6	6
2	HR32	4	1	4	5	2
3	HR34	6	8	3	1	7
4	HR33	7	6	1	3	5
5	HR34	5	7	3	1	3
6	HR33	8	7	1	2	4
7	HR33	8	6	1	3	4
8	HR35	8	6	2	2	1
9	HR34	7	7	5	1	8
10	HR33	8	5	1	6	2
11	HR35	5	8	2	3	1
12	HR33	6	4	1	8	3

Table A11. Other to Worst vectors for Human Resources / Workforce (HR) to compare the barriers to implementation of Industry 4.0.

	Respondents											
	1	2	3	4	5	6	7	8	9	10	11	12
Worst (Least Important) barrier→	HR 31	HR 34	HR 32	HR 31	HR 32	HR 31	HR 31	HR 31	HR 35	HR 31	HR 32	HR 34
Lack of trust between partners: HR31	1	2	3	1	2	1	1	1	2	1	2	2
Lack of support from customer/ supplier: HR32	3	6	1	2	1	2	2	3	2	4	1	4
Lack of competence from the workforce/ limited skilled workers: HR33	8	2	5	8	4	8	7	6	3	8	6	8
Lack of external experts: HR34	2	1	8	4	7	6	4	6	7	2	4	1
Disruption to existing jobs: HR35	2	4	2	2	5	4	3	8	1	6	8	5

Table A12. Optimal weights of the barriers related to Human Resources / Workforce (HR) for implementation of Industry 4.0.

Respondent No.	Lack of trust between partners HR31	Lack of Support from customer/supplier HR32	Lack of competence from the workforce / limited skilled workers HR33	Lack of external experts HR34	Disruption to existing jobs HR35	Ksi*	CR
1	0.065	0.130	0.587	0.109	0.109	0.065	0.028
2	0.123	0.434	0.123	0.075	0.245	0.057	0.025
3	0.104	0.058	0.208	0.542	0.089	0.081	0.035
4	0.065	0.098	0.523	0.196	0.118	0.065	0.028
5	0.111	0.054	0.184	0.466	0.184	0.087	0.038
6	0.050	0.077	0.470	0.269	0.134	0.067	0.029
7	0.063	0.095	0.508	0.190	0.143	0.063	0.028
8	0.044	0.078	0.233	0.233	0.411	0.056	0.024
9	0.097	0.097	0.136	0.596	0.073	0.084	0.036
10	0.048	0.111	0.471	0.092	0.277	0.083	0.036
11	0.098	0.049	0.246	0.164	0.443	0.049	0.021
12	0.097	0.146	0.508	0.054	0.195	0.076	0.033
Average	**0.081**	**0.119**	**0.350**	**0.249**	**0.202**	**0.069**	**0.030**

Table A13. Best to Other barriers related to Infrastructure & Financial Resources (IFR) for implementation of Industry 4.0.

Respondent No.	Best (Most Important) barrier↓	IFR41	IFR42	IFR43
1	IFR41	1	5	8
2	IFR42	2	1	6
3	IFR41	1	4	7
4	IFR42	2	1	8
5	IFR41	1	4	8
6	IFR41	1	3	8
7	IFR42	3	1	7
8	IFR42	4	1	6
9	IFR43	2	5	1
10	IFR42	5	1	4
11	IFR41	1	2	7
12	IFR41	1	4	3

Table A14. Other to Worst vectors for Infrastructure & Financial Resources (IFR) to compare the barriers to implementation of Industry 4.0.

	Respondents											
	1	2	3	4	5	6	7	8	9	10	11	12
Worst (Least Important) barrier→	IFR 43	IFR 43	IFR 43	IFR 43	IFR 43	IFR 43	IFR 43	IFR 43	IFR 42	IFR 41	IFR 43	IFR 42
Lack of infrastructure: IFR 41	8	2	7	6	8	8	3	2	3	1	7	6
Lack of financial resources: IFR 42	3	6	2	8	3	2	7	6	1	6	4	1
Lack of a digital strategy alongside resource scarcity: IFR 43	1	1	1	1	1	1	1	1	6	2	1	2

Table A15. Optimal weights of the barriers related to Infrastructure & Financial Resources (IFR) for implementation of Industry 4.0.

Respondent No.	Lack of Infrastructure IFR41	lack of financial resources IFR42	Lack of a digital strategy alongside resource scarcity IFR43	Ksi*	CR
1	0.750	0.167	0.083	0.083	0.083
2	0.278	0.611	0.111	0.056	0.056
3	0.717	0.183	0.100	0.017	0.017
4	0.333	0.600	0.067	0.067	0.067
5	0.722	0.194	0.083	0.056	0.056
6	0.691	0.218	0.091	0.036	0.036
7	0.236	0.673	0.091	0.036	0.036
8	0.185	0.704	0.111	0.037	0.037
9	0.316	0.105	0.579	0.053	0.053
10	0.125	0.688	0.188	0.063	0.063
11	0.604	0.313	0.083	0.021	0.021
12	0.628	0.140	0.233	0.070	0.070
Average	**0.465**	**0.383**	**0.152**	**0.049**	**0.049**

Table A16. Best to Other barriers related to Technology (T) for the implementation of Industry 4.0.

Respondent No.	The Best (most important) barrier↓	T51	T52	T53	T54	T55	T56
1	T53	6	7	1	6	6	3
2	T56	4	4	2	8	4	1
3	T56	2	5	3	7	3	1
4	T53	5	6	1	8	5	2
5	T56	3	5	3	7	5	1
6	T51	1	2	4	5	8	4
7	T53	5	7	1	8	7	3
8	T51	1	3	3	7	6	2
9	T51	1	2	2	6	7	3
10	T53	6	7	1	6	6	2
11	T56	4	6	4	8	5	1
12	T54	5	4	3	1	8	5

Table A17. Other to Worst vectors for Technology (T) to compare the barriers to implementation of Industry 4.0.

	Respondents											
	1	2	3	4	5	6	7	8	9	10	11	12
Worst (Least Important) barrier→	T52	T54	T54	T54	T54	T55	T54	T54	T55	T52	T54	T55
Security and Privacy Issues: T51	3	4	5	3	4	8	3	7	7	2	4	4
Seamless integration and compatibility issues: T52	1	3	2	2	2	6	2	4	5	1	2	5
Very high complexity: T53	7	6	3	8	4	3	8	4	5	7	4	6
Awareness about the potentialities of robot applications: T54	3	1	1	1	1	2	1	1	2	2	1	8
Challenges in ensuring data quality: T55	3	3	2	3	2	1	2	2	1	3	3	1
Concerns about cybersecurity and data ownership issues: T56	5	8	7	6	7	3	5	6	3	5	8	3

Table A18. Optimal weights of the barriers related to Technology (T) for implementation of Industry 4.0.

Respondent No.	Security and Privacy Issues T51	Seamless integration and compatibility issues T52	Very high complexity T53	Awareness about the potentialities of robot applications T54	Challenges in ensuring data quality T55	Concerns about cybersecurity and data ownership issues T56	Ksi*	CR
1	0.094	0.055	0.475	0.094	0.094	0.188	0.088	0.029
2	0.113	0.113	0.225	0.042	0.113	0.394	0.056	0.019
3	0.210	0.084	0.140	0.049	0.136	0.382	0.037	0.012
4	0.097	0.081	0.435	0.048	0.097	0.242	0.048	0.016
5	0.161	0.097	0.161	0.054	0.097	0.430	0.054	0.018
6	0.409	0.227	0.114	0.091	0.045	0.114	0.045	0.015
7	0.112	0.080	0.488	0.052	0.080	0.187	0.073	0.024
8	0.376	0.144	0.144	0.046	0.072	0.217	0.057	0.019
9	0.360	0.198	0.198	0.066	0.047	0.132	0.035	0.012
10	0.085	0.052	0.438	0.085	0.085	0.255	0.072	0.024
11	0.136	0.091	0.136	0.051	0.109	0.477	0.068	0.023
12	0.105	0.132	0.175	0.439	0.044	0.105	0.088	0.029
Average Value	**0.188**	**0.113**	**0.261**	**0.093**	**0.085**	**0.260**	**0.060**	**0.020**

2024 Asia Pacific Conference on Innovation in Technology (APCIT)
Mysuru, India. Jul 26-27, 2024

A Hybrid Deep Learning Model for Sentiment Analysis of IMDB Movies Reviews

Rohit Beniwal
Department of Computer Science & Engineering
Delhi Technological University
New Delhi, India
rohitbeniwal@dtu.ac.in

Arun Kumar
Department of Computer Engineering
Delhi Technological University
New Delhi, India
techarun121@gmail.com

Arpit Kumar Dinkar
Department of Computer Engineering
Delhi Technological University
New Delhi, India
dinkerarp@gmail.com

Anurag Panchal
Department of Computer Engineering
Delhi Technological University
New Delhi, India
anuragpanchal68@gmail.com

2024 Asia Pacific Conference on Innovation in Technology (APCIT) | 979-8-3503-6153-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/APCIT62007.2024.10673659

Abstract **- Sentiment analysis shows the emotional tone of the digital text and it has gained a lot of attraction in diverse fields like Marketing and Advertising, Customer Service and Support, Product Development, Politics and Public Opinion, Financial Trading, Healthcare, Social Media Monitoring, a Recommender System, Reputation Management and many more. Moreover, Sentiment Analysis has been exploited in various movies, TV shows, web series, songs, and social media datasets. In this research paper, we use a novel hybrid approach for sentiment analysis on IMDB movie reviews using deep learning algorithms. The proposed hybrid deep learning model, i.e. Convolutional Neural Network (CNN) + Long short-term memory (LSTM) achieved an accuracy of 96.01%, which outperforms the results obtained via multiple deep learning models like logistic regression, pre-trained text blob model, and Convolutional Neural Network(CNN). The proposed model also outperforms the existing state-of-the-art literature. The state of art used in this paper explains and approaches sophisticated problems efficiently and effectively.**

Keywords: Binary Class Classification, CNN, Deep Learning, Hybrid Model, IMDB Reviews, LSTM, Sentiment Analysis

I. INTRODUCTION

The digital media era has seen a massive increase in the number of movie reviews that have been published on numerous platforms within the film business. A well-known website where people discuss and critique films is IMDb (Internet Movie Database). IMDb (Internet Movie Database) this online service-M renounces the official power of pros and has, they pan done with aplomb. IMDB incluties everything from scene indices to the ability to save and print lists to vole for future reference with millions of reviews for many types of movies, IMDb is a treasure trove waiting to be analyzed! It is important to define that sentiment analysis, being a subset of natural language processing (NLP), provides an organized way supporting the classification and recognition of sentiments expressed in textual text. Through sentiment analysing the IMDb movie reviews, academics and analysts can gain important intuition regarding how an audience feels about liking or disliking different films [1].

At the same time, sentiment analysis and other similar applications have shown that deep learning algorithms can accomplish impressive results, which sparked a new revolution in the history of natural language processing (NLP). "Recurrent neural networks (RNNs), Convolutional Neural Networks (CNNs) and transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers), demonstrated strong capabilities in drawing out implicit sentiment features from textual given inputs" [2].

Deep learning algorithms will be applied to investigate sentiment analysis in IMDb movie reviews. Cutting-edge neural network topologies and sophisticated pre-training techniques are used for classifying the analysis as positive, negative or neutral sentiment. This enables us to identify, e.g., factors causing sentiment polarity, phrases/events that can be associated with strong sentiments, and movements in the sentiment over time.

Our work is a step towards expanding the support base for sentiment analysis; investing in expanded studios, wider audience and accepting filmmakers. Stakeholders in the film industry can improve their decision-making processes regarding audience responses, creative decisions, and marketing methods by investigating the complex emotional dynamics found in IMDb movie reviews. The paper also explores the evolution of sentiment analysis algorithms, focusing on the adoption of Deep Neural Network (DNN) models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to remove the restrictions of older approaches [3].

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were established as a result of RNN constraints, such as disappearing and growing gradients in the long-term reliance Hussain [4]. Bidirectional GRU (Bi-GRU) and bidirectional LSTM (Bi-LSTM) models were also developed to account for the text analysis's past and subsequent contexts [5].

The research criticizes current deep learning models for Sentiment Analysis (SA) for focusing on specific issues while ignoring others. Chatterjee et al. [6], for example, used LSTM with pre-trained word embeddings to identify emotions, but they neglected to account for the relative weights of phrase components. Despite using bidirectional LSTM with CNN and attention algorithms, Xu et al. [7] did not address the accompaniment of little and more dependencies. Rezaeinia et al. [8] raised pre-trained word embeddings and employed CNNs; nevertheless, they neglected to take long-term dependencies or word importance into account.

To solve these shortcomings, a novel deep-learning model for divergence recognition in user comments is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

proposed in the [9]. The proposed method aims to extract both long and short intro sentence relations, prioritize significant features, consider bidirectional contextual relationships, and give different phrases in comments at different levels of attention [10]. By addressing a greater range of sentiment analysis issues, this approach seeks to improve existing models.

The suggested model was tested against six cutting-edge DNN-based and wordvec text categorization and sentiment analysis algorithms [11,12]. The comparison was done for both brief tweets and lengthy assessments. Based on standard performance metrics in the NLP and sentiment analysis areas, the suggested model outperformed the other six algorithms.

The study's primary contributions are:

- Introducing a sentiment analysis system using a novel deep learning architecture.
- Evaluating the suggested model's performance against current deep architectures and the existing state-of-the-art literature for sentiment analysis and review classification.

The following summarizes the remainder of the research paper's structure:

A brief review and survey of the past literature on artificial neural network(ANN) models for text categorization and sentiment analysis is given in Section 2. The theoretical and practical foundation of the suggested ANN model is presented in Section 3. Experiments and findings are presented in Section 4. Section 5 wraps up the work by concluding and making some recommendations for future lines of inquiry.

II. Related Work

Many hybrid ANN approaches can deal with complex data and perform accurate sentiment analysis more closely. However, different researchers adopt several different algorithms to perform sentiment analysis on movie review datasets, a summary of which is provided in the following subsections. These sub-sections present an insight into the different state-of-the-art done.

A. Sentiment Analysis for fuzzy semantic words based on feature assemble and CNN model

To enhance the SBELM model, Chaturvedi et al. [13] applied Bayesian networks to create BNELM. This enhancement replaces the computationally intensive Laplace estimates with a heuristic method, resulting in improved predictive accuracy. The framework integrates Convolutional Neural Networks (CNNs) and ELMs with Bayesian networks, with each component playing a certain role in the classification process. The CNN learns significant attributes from the training sentences, which are then fed into the BNELM for further processing. They got 89% accuracy using the BNELM model on the TASS dataset.

Phan et al. [14] highlighted the significance of incorporating sentiments, POS tags, and fuzzy sentiment phrases in enhancing word embedding techniques for TSA. To tackle these aspects, they suggested methodology combines feature ensemble modeling, deep learning algorithms, and a divide-and-conquer strategy. They improved the accuracy by 5%. Chaturvedi et al. [15] discussed the application of deep learning in extracting features from various modalities like text and video, and mapping them onto AffectiveSpace, a vector space model containing common sense concepts. AffectiveSpace is a product of SenticNet, a network where nodes symbolize concepts and edges show hierarchy. By reducing the dimensions of SenticNet, AffectiveSpace is created, enabling the identification of semantics for each concept. The model utilizes 24 primary emotions from the Hourglass model as centroids for clustering AffectiveSpace. They achieved 85% accuracy using CFSC, which is 5% more accurate than TDN.

B. Based on the deep learning approach

Ouyang et al. [16] employed a method that merges Word2vec and Convolutional Neural Networks (CNN) for sentiment analysis. Initially, Rezaeinia et al. [8] said Word2vec is utilized to transform words into vectors, which are then utilized to invent sentence vectors. These sentence vectors are subsequently fed into a CNN to categorize sentences into various sentiment labels. The CNN is executed using the Caffe deep learning framework, recognized for its effectiveness and flexibility.

They achieved 45.4% accuracy which is the highest from MV-RNN and RNN method's accuracy.

Jianqiang et al. [17] and Kumar et al. [18] discussed the significance of sentiment analysis (SA) in comprehending short texts (ST) posted on public network platforms like as Twitter and Facebook. SA plays a crucial role in evaluating opinions on diverse subjects like politics, movie reviews, etc. A hybrid algorithm for SA on IMDb movie reviews is suggested, which merges characteristics derived from statistical and lexicon methods. Several feature selection algorithms are utilized to eliminate irrelevant features, and the ultimate feature set is categorized using SVM, Naïve Bayes, KNN, and Maximum Entropy classifiers. They achieved 74.66% accuracy when using the RLPI feature selection method and 83.93% accuracy using the correlation method.

III. Proposed Work

A brief summary of the essential components that comprise our model is elaborated, where Section 3.1 delineates Long Short-Term Memory and GRU, and on the other hand, Section 3.2 presents a theoretical Convolutional Neural Network. After that, we will go over the theoretical description of the development of the model.

A. Long short-term memory

The detonating or vanishing issue that ordinary RNNs experience is tended to by the Long Short-Term Memory (LSTM) RNN, which infers its yield depends on the i/p from the ongoing time-step and the o/p of the past time-step. It sends the most recent o/p to the succeeding time step. Each LSTM unit is built from three non-linear entryways as shown in Figure 1: an input entryway (it), a disregard entryway (ft), and an output entryway (ot). The memory cell (ct) in each unit keeps up its state over subjective periods. Let σ represent the element-wise sigmoid function, tanh( ) the hyperbolic tangent function, and ⊙ the product. The forget gate uses the following equation to determine what information should be forgotten: it outputs a number [0,1].

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \mathbf{h}_{t-1} + \mathbf{U}_f \mathbf{x}_t + \mathbf{b}_f\right) \quad (1)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

where the input vector and hidden state are represented by xt and ht. Vector W displays the weight matrices of the gates or cell for the i/p xt and hidden state ht, while vector b indicates the bias vectors. The input gate computes and combines c^t to determine what fresh information needs to be stored.

$$\mathbf{i}_t = \sigma\,(\mathbf{W}_i\,\mathbf{h}_{t-1} + \mathbf{U}_i\,\mathbf{x}_t + \mathbf{b}_i) \quad (2)$$

$$\hat{\mathbf{c}}_t = \tanh\,(\mathbf{W}_c\,\mathbf{h}_{t-1} + \mathbf{U}_c\,\mathbf{x}_t + \mathbf{b}_c) \quad (3)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t \quad (4)$$

The decision of what parts of the cell state to o/p is made by the o/p gate.

$$\mathbf{o}_t = \sigma\,(\mathbf{W}_o\,\mathbf{h}_{t-1} + \mathbf{U}_o\,\mathbf{x}_t + \mathbf{b}_o) \quad (5)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\,(\mathbf{c}_t) \quad (6)$$

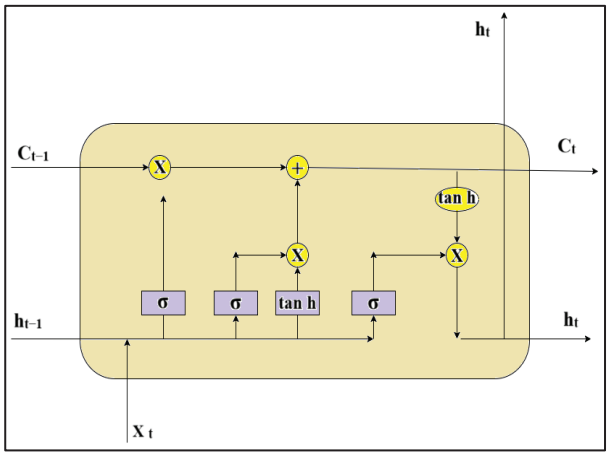


Fig. 1. LSTM Cell

A GRU, which is Gated Reccurent Network is like Long Short Term Memory and is operated on gated mechanism of input and output.

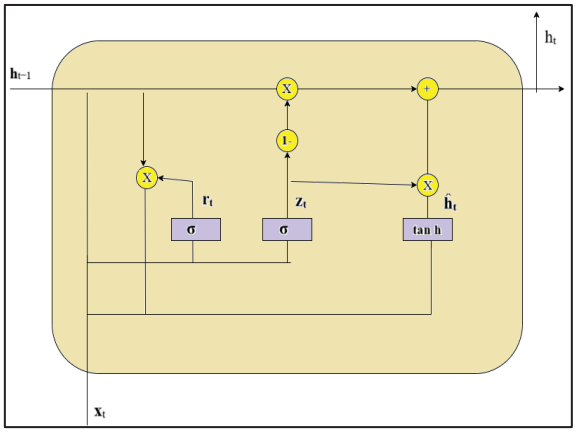


Fig. 2. GRU Cell

A GRU is a more straightforward variant of an LSTM, consisting of two gates, as shown in Figure 2, a reset gate (z) and an update gate (r), which combines i/p and forget gates. The reset and update are calculated using an LSTM-like method.

B. Convolutional Neural Network

CNNs are used in NLP applications to take out internal characteristics and are composed of many convolutional layers. These networks use linear filters to conduct convolution operations on the input information. It is standard procedure to input feature maps into a pooling or tested layer to reduce their dimension. A typical pooling technique is max-pooling.

The workflow of the proposed work is shown in Figure 3 with the help of the simple flowchart.

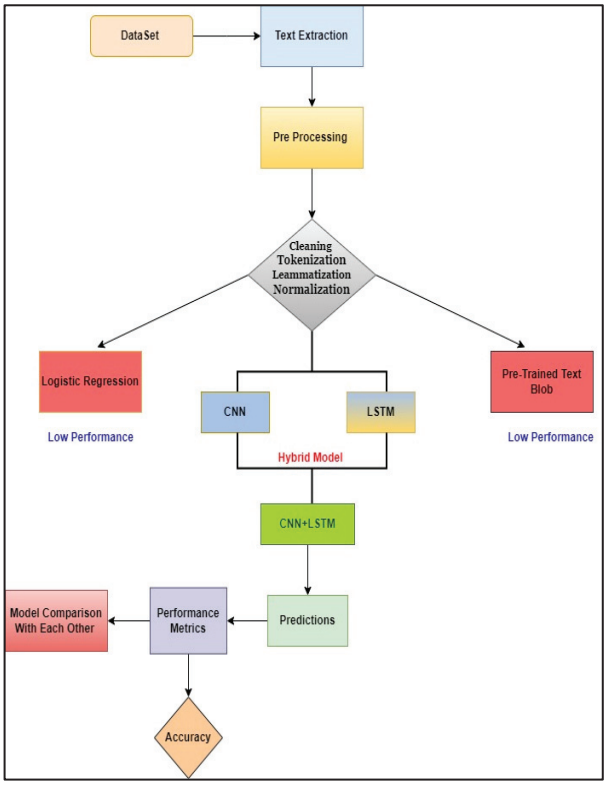


Fig. 3. Work Flow of the proposed work

C. Pre Processing

To extract important and meaningful features from the analysis text files that may be utilized for training, multiple methods are combined in this paper.

- Data Cleaning: The act of cleaning unstructured data, such as text. Any unnecessary characters, HTML tags, CSS tags, and tags were eliminated from the data.
- Tokenization: The process of replacing private information with one-of-a-kind identifying symbols while preserving all pertinent data attributes.
- Normalization: Remove punctuation, change all text to lowercase, and replace numbers with their corresponding words.
- Lemmatization and stemming: It eliminates or stems a word's final few characters, which frequently results in misspellings and wrong interpretations. Lemmatization reduces a word to its semantic base form, or lemma while taking the context into account.

D. Max Pooling Layer

The max-pooling mask has been used to both enhance and drop the convolutional layer's output. This is accomplished by applying the mask one at a time to the whole set of data, selecting just the utility with the highest weight and ignoring

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

the other values, hence drastically lowering the amount of data that is fed into the subsequent layer.

E. Output Dense Layer

At long last, Since the proposed show is related to the parallel classification issue, the sigmoid work is connected to the information gotten from the LSTM layer to produce a fragmentary esteem between 0 and 1, and after that, assess the classification based on this esteem. The final layer is entirely associated with a single yield hub. (For example, a review gets 0 points if it's negative and 1 point if it's positive).

IV. Results

We employed four approaches in this paper: two deep learning algorithms, one based on transfer learning, and one fundamental method.

A. Data Set Used

The IMDB dataset under consideration comprises 50,000 English-language movie reviews, with 25,000 +ve and 25,000 -ve ratings shown in Table 1 and Figure 4. File sizes have varied between 1 KB and 15 KB. There was no rating information in the text files. 80% of the dataset was used for training, and 20% was used for testing.

TABLE I. DATA SET SUMMARY

Reviews	No. of Positive Sentiments	No. of Negative Sentiments
50000	25000	25000

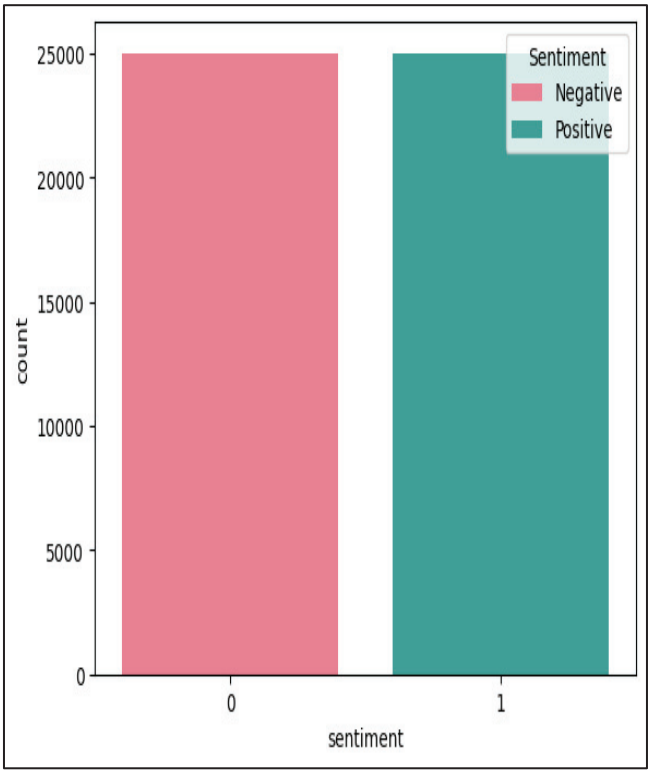


Fig. 4. Classification of Positive and Negative Movie Reviews

B. Fundamental Model: Logistic Regression

We are using 0.1 stratified data from our entire dataset. The result of the logistic regression is visualized by the confusion matrix in Figure 5.

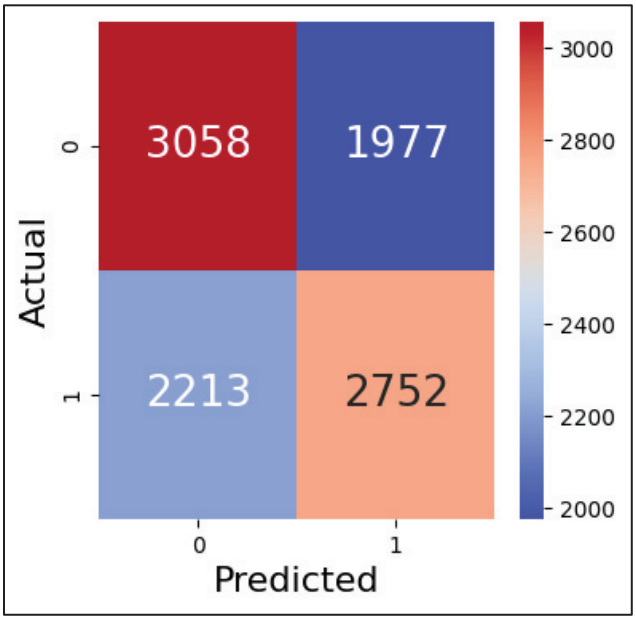


Fig. 5. Confusion Matrix for Logistic Regression

The Precision, Recall, and F1 score of logistic regression are presented Review-wise in Table 2.

TABLE II. METRICS FOR EVALUATING LOGISTIC REGRESSION

Review	Precision	Recall	F1-Score
0	0.58	0.61	0.59
1	0.58	0.55	0.57
		Accuracy	0.58

C. Transfer Learning: TextBlob Pre-Trained Model

Transfer Learning is a algorithm that makes use of a pre-trained model that was trained on very large data, but it is now being used for new problems by just doing hyperparameter tuning. Hence, we are using a text blob pre-trained model, which is trained on a large amount of text data. On applying and tuning the model, we got results, which are shown through the confusion matrix in Figure 6.

Text Blob performed better than the logistic regression, which is clearly shown in the confusion matrix.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

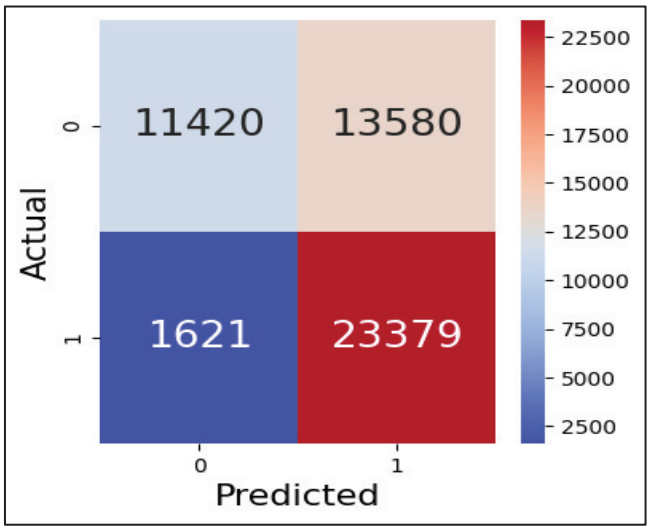


Fig 6. Confusion Matrix for Text Blob Pre-trained Model

The Precision, Recall, and F1 score of logistic regression are presented Review-wise in Table 3.

TABLE III. METRICS FOR EVALUATING TEXT BLOB PRE-TRAINED MODEL

	Positive	Negative
Actual	25000	25000
Predicted	36959	13041
	Accuracy	0.69598

D. N-gram Analysis

N-gram is a combination of n adjacent symbols in a particular order. Symbols can be syllables, letters, and alphabets.

1) 1-gram Analysis

When N=1, we call it a 1-gram analysis. We group only one symbol at a time and analyze the frequency and the same thing is shown in Figure 7.

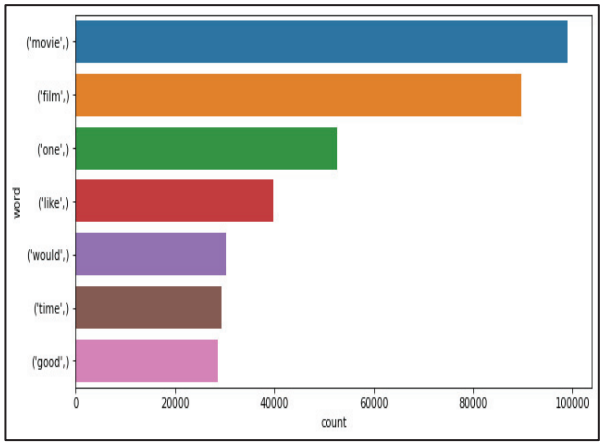


Fig. 7. Bar plot of 1 gram

2) 2-gram Analysis

When N=2, we call it a 2-gram analysis. Now we group two symbols at a time and analyze the frequency and the same thing is shown in Figure 8.

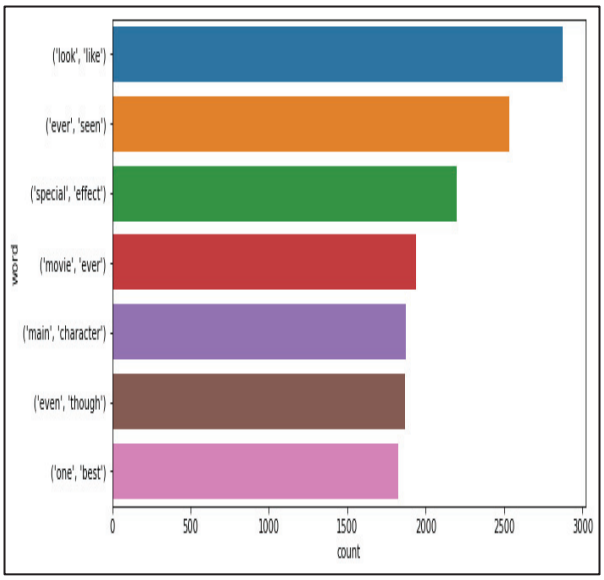


Fig. 8. Bar plot of 2 gram

3) 3-gram Analysis

When N=3, we call it as 3-gram analysis. Now we will group three symbols at a time and analyze the frequency and the same thing is shown in figure 9.

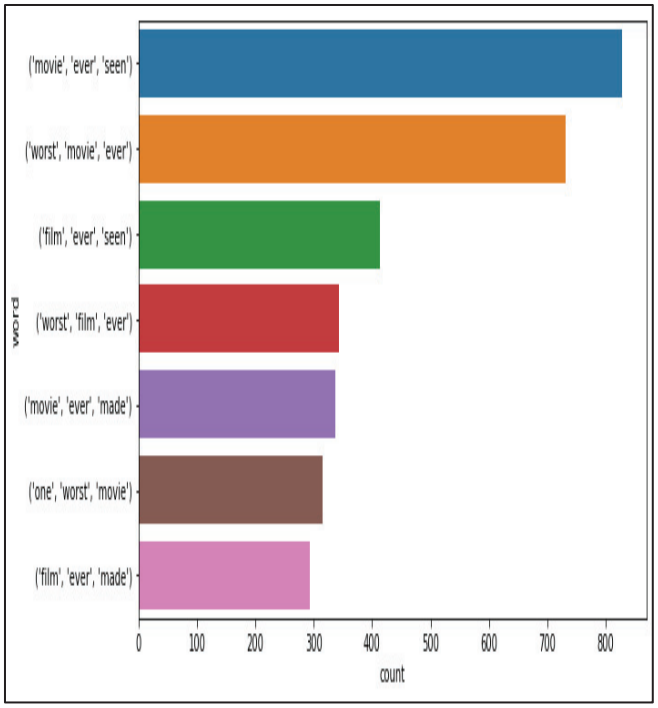


Fig. 9. Bar plot of 3 gram

E. Convolutional Neural Network

We employed CNN for sentiment analysis and the structure is composed of an Embedding layer with 150 input neurons following a Conv1D layer with 150 input neurons following the max pooling layer with 75 input neurons and these two layers are repeated again twice with a varying input number of neurons and at last, we proposed a dense layer with 18 input neuron and 1 output neuron. So, this was the whole structure of the CNN model proposed. Here, we presented tabularly the epoch output showing loss and accuracy after each epoch in Table 5.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

TABLE 4. CNN MODEL EPOCH-WISE LOSS AND ACCURACY

Epoch	Loss	Accuracy
1	0.6171	0.6226
2	0.5552	0.6744
3	0.5435	0.6814
4	0.5345	0.6871
5	0.5213	0.6969

Following this, loss and accuracy don't significantly alter; therefore, the model automatically stops. The CNN model's final accuracy is **69.69%**, which is better than that of the text blob.

F. Hybrid CNN+LSTM

We employ the hybrid approach of CNN + LSTM models and the structure of the model is given in Table 6, which comprises the max pooling, dense, and LSTM layer.

TABLE IV. CNN + LSTM MODEL DESCRIPTION

Layer (type)	Output Shape
Embedding_2 (Embedding)	(None, 150, 50)
Conv1d_6 (Conv1D)	(None, 150, 32)
Max_pooling1d_6 (MaxPooling1D)	(None, 75, 32)
Conv1d_7 (Conv1D)	(None, 75, 32)
Max_pooling1d_7 (MaxPooling1D)	(None, 37, 32)
Conv1d_8 (Conv1D)	(None, 37, 32)
Max_pooling1d_8 (MaxPooling1D)	(None, 18, 32)
Lstm_1 (LSTM)	(None, 100)
Dense_2 (Dense)	(None, 1)

Here, we presented tabularly the epoch output showing loss and accuracy after each epoch in Table 7.

TABLE V. CNN + LSTM MODEL EPOCH-WISE LOSS AND ACCURACY

Epoch	Loss	Accuracy
1	0.4441	0.7644
2	0.2486	0.9014
3	0.1875	0.9308
4	0.1166	0.9601

Following this, loss and accuracy don't significantly alter, therefore the model automatically stops.

G. Comparison of different state-of-the-art algorithms present in the proposed literature

According to the results, CNN_LSTM performed better than any other deep learning algorithm that was put into this research paper [19]. Therefore, the deep learning algorithms (CNN and CNN_LSTM) performed better than Text Blob, and logistic regression and visualization of performance are shown in Figure 10.

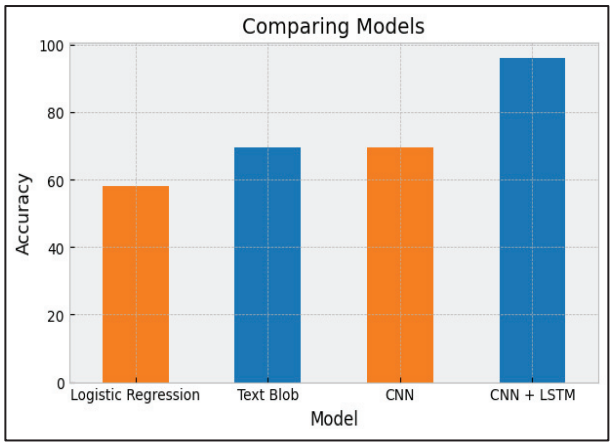


Fig. 10. Comparing algorithms presented in the proposed work

The CNN model's final accuracy is **96.01%,** which is the best among all models explained in this research paper.

H. Comparison of the proposed algorithm with the existing state-of-the-art literature

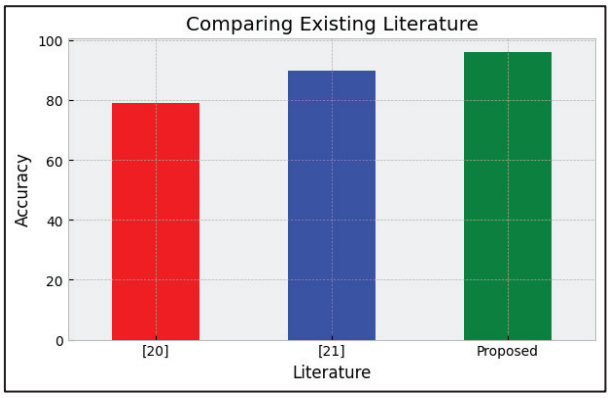


Fig 11. Comparing proposed work with existing state-of-the-art literature

The obtained accuracies were contrasted with those that are informed by [20,21] in previously published work using various machine learning algorithms and a dataset of IMDB film reviews.

V. CONCLUSION AND FUTURE WORK

The field of sentiment analysis is new, and deep learning advances have made it easier to analyze the polarity of sentiment expressed in reviews. The outcomes showed that, in comparison to more conventional machine learning algorithms, the neural network models (CNN, CNN_LSTM) in combination with other machine learning algorithms, produced superior and more promising results.

For future work, we can explore more logical and semantic data to be added to the sentiment analysis algorithms that are now in use. To break free from the limitations imposed by interdependence, unsupervised algorithms could be investigated. Emojis also require greater comprehension because they are essential components for expressing emotions. To improve the outcomes, other confused neural network topologies for word embedding and sentiment embedding characteristics should be investigated

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

and we can also use the pre-trained models to generate embeddings on textual data [22]. Then, we can use that to improve the results.

We can expand our dataset to Twitter and more review websites, which will create a large pool of words and relationships between words. We can also explore attention mechanisms with optimal attributes that will eventually lead to better results.

REFERENCES

[1] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294, Feb. 2021, doi: 10.1016/j.future.2020.08.005.

[2] M. Song, H. Park, and K. shik Shin, "Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean," *Inf Process Manag*, vol. 56, no. 3, pp. 637–653, May 2019, doi: 10.1016/j.ipm.2018.12.005.

[3] M. Usama, B. Ahmad, E. Song, M. S. Hossain, M. Alrashoud, and G. Muhammad, "Attention-based sentiment analysis using convolutional and recurrent neural network," *Future Generation Computer Systems*, vol. 113, pp. 571–578, Dec. 2020, doi: 10.1016/j.future.2020.07.022.

[4] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, Jan. 2018, doi: 10.1016/j.neucom.2017.10.010.

[5] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/j.neucom.2019.01.078.

[6] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding Emotions in Text Using Deep Learning and Big Data," *Comput Human Behav*, vol. 93, pp. 309–317, Apr. 2019, doi: 10.1016/j.chb.2018.12.029.

[7] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[8] S.M. Rezaeinia, R. Rahmani, A. Ghodsi, H. Veisi, Sentiment analysis based on improved pre-trained word embeddings, Expert Syst. Appl. 117 (2019) 139–147.

[9] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artif Intell Rev*, vol. 53, no. 4, pp. 2313–2339, Apr. 2020, doi: 10.1007/s10462-019-09770-z.

[10] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018. Hindawi Limited, 2018. doi: 10.1155/2018/7068349.

[11] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation." [Online]. Available: http://nlp.

[12] R. Liu, Y. Shi, C. Ji, and M. Jia, "A Survey of Sentiment Analysis Based on Transfer Learning," *IEEE Access*, vol. 7. Institute of Electrical and Electronics Engineers Inc., pp. 85401–85412, 2019. doi: 10.1109/ACCESS.2019.2925059.

[13] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, and E. Cambria, "Bayesian network based extreme learning machine for subjectivity detection," *J Franklin Inst*, vol. 355, no. 4, pp. 1780–1797, Mar. 2018, doi: 10.1016/j.jfranklin.2017.06.007.

[14] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," *IEEE Access*, vol. 8, pp. 1463014641,2020,doi:10.1109/ACCESS.2019.2963702.

[15] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit Lett*, vol. 125, pp. 264–270, Jul. 2019, doi: 10.1016/j.patrec.2019.04.024.

[16] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment Analysis Using Convolutional Neural Network," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, IEEE, Oct. 2015, pp. 2359–2364. doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.

[17] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," *IEEE Access*, vol. 6, pp. 23253–23260, Jan. 2018, doi: 10.1109/ACCESS.2017.277693

[18] H. M. Keerthi Kumar, B. S. Harish, and H. K. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 5, pp. 109–114, 2019, doi: 10.9781/ijimai.2018.12.005.

[19] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdiscipl. Rev.: Data Min. Knowl. Disco. 8 (4) (2018) e1253.

[20] N. G. Ramadhan and T. I. Ramadhan, "Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM," Sinkron, vol. 7, no. 1, pp. 39–45, Jan. 2022, doi: 10.33395/sinkron.v7i1.11204.

[21] M. R. Haque, S. Akter Lima, and S. Z. Mishu, "Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews," in 3rd International Conference on Electrical, Computer and Telecommunication Engineering, ICECTE 2019, Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 161–164. doi: 10.1109/ICECTE48615.2019.9303573.

[22] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning based text classification: a comprehensive review, 2020, arXiv preprint arXiv:2004.03705.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:30 UTC from IEEE Xplore. Restrictions apply.

Operational Research (2024) 24:57
https://doi.org/10.1007/s12351-024-00864-3

ORIGINAL PAPER

A hybrid metaheuristic optimised ensemble classifier with self organizing map clustering for credit scoring

Indu Singh[1] · D. P. Kothari[2] · S. Aditya[1] · Mihir Rajora[1] · Charu Agarwal[3] · Vibhor Gautam[4]

Received: 23 June 2023 / Revised: 18 August 2024 / Accepted: 23 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
Credit scoring is a mathematical and statistical tool that aids financial institutions in deciding suitable candidates for the issuance of loans, based on the analysis of the borrower's financial history. Distinct groups of borrowers have unique characteristics that must be identified and trained on to increase the accuracy of classification models for all credit borrowers that financial institutions serve. Numerous studies have shown that models based on diverse base-classifier models outperform other statistical and AI-based techniques for related classification problems. This paper proposes a novel multi-layer clustering and soft-voting-based ensemble classification model, aptly named Self Organizing Map Clustering with Metaheuristic Voting Ensembles (SCMVE) which uses a self-organizing map for clustering the data into distinct clusters with their unique characteristics and then trains a sailfish optimizer powered ensemble of SVM-KNN base classifiers for classification of each distinct identified cluster. We train and evaluate our model on the standard public credit scoring datasets—namely the German, Australian and Taiwan datasets and use multiple evaluation scores such as precision, F1 score, recall to compare the results of our model with other prominent works in the field. On evaluation, SCMVE shows outstanding results (95% accuracy on standard datasets) when compared with popular works in the field of credit scoring.

Keywords Machine learning · Credit scoring · Self organizing maps · Ensemble classification · Sailfish optimizer

1 Introduction

Lending money is one of the primary businesses of financial institutions as it reaps huge profits for its stakeholders. Although, market fiascos of the likes of credit rationing and risk may occur due to the absence of the applicant's

Extended author information available on the last page of the article

Published online: 18 September 2024

information. Credit risk is a major threat to financial institutions. It can exhibit an exorbitant effect on the financial market, especially during a financial crisis period. The past financial crises have damaged many countries financially, the effects of which will long be felt in the future. Developing efficient risk management has thus become a major focus for financial institutions. Banks have been increasingly using statistical and ML models for credit score evaluation. While huge advancements have already been made in the field, according to a survey conducted by Dastile et al. (2020), bank regulators stand with the belief that models should be transparent. Despite the non-transparency of deep learning models, they have still been gaining prominence. A justification to overcome the problem of transparency is to rationalize the deep learning models, i.e., be able to find reasons as to why the model generates a particular decision.

Credit scoring is a robust mathematical and statistical model which helps in mitigating the problem of credit defaults and managing credit risks by computing the credit risk associated with a customer. The usage of credit scoring increases the accessibility of financial institutions to a larger demographic, especially in developing countries where greater financial accessibility is a must for the development of the country as a whole, as noted in a study by Bumacov et al. (2014). Credit scoring has also been shown to increase the efficiency of loan officers, leading to a causal effect that leads to the handing out of more loans and serving more borrowers. The rationale behind credit scoring is to use customer's past data such as credit history, income, and marital status and calculate their ability to repay the loan. There are two primary reasons to develop a powerful and robust credit scoring model. Firstly, considering the massive volume of the credit market, even a minute upgradation in the classification accuracy are able to reap a substantial amount of profit. Moreover, regulations regarding credit risk management are getting strict.

The Basel Community on Banking Supervision(BCBS) monitors and guides financial institutions to ensure credit security. Any violation of the guidelines can cause enormous regulation costs. According to a study by Reichert et al. (1983), the previously used statistical models are not absolutely efficient in mapping out real world scenarios. They do, however, conclude that these models' inability to meet the assumptions of fair credit scoring models is less significant as opposed to the ability of credit evaluation officers' comprehension of the classification model and its shortcomings.

Credit scoring as a comprehensive procedure can be subdivided into 4 different stages—application scoring, scoring on basis of behaviour, collection scoring, and fraudulent activity detection. In this study, our major focus lies on predicting if an applicant is a 'risky borrower' which equates to a binary classification problem.

In the initial stages of credit scoring, statistical techniques such as logistic regression and linear discriminant analysis(LDA) grew popular. Statistical techniques are widely used because of their easy implementation. However, these are based on some inferences such as linear separation and normal distribution of data, assumptions which are generally violated in the real world. Neural network models for credit scoring were also seen in the early 2000's, as were noticed in a publication by West (2000). With the rapid development of artificial intelligence, techniques

making extensive use of machine learning algorithms such as KNN, ANN (Artificial Neural Network), SVM, artificial immune systems have been introduced.

In most of the studies, single classifier-based models are constructed for credit scoring. Experiments have shown that singular classifiers are ineffective in capturing the patterns of individual customers, and thus a different type of classification model, ensemble classifiers have been employed that have proved effective for improving the accuracy and stability of single classifiers. Ensemble-based classifiers combine several base classifiers for better predictive performance (Singh 2017). Thus, research on ensemble methods has grown popular in the last few years. The main idea behind ensemble classification is to overcome the weakness of individual classifiers by combining their results. For effective ensemble classification, the base classifiers must be accurate, diverse, and sensitive. The construction of ensemble models usually consists of three steps—pool generation, selection of the classification and related models, and the combination of the delivered results. In the first step, base models are generated to create a model pool using various classification algorithms or data. The second step is necessary to enhance the model performance. The pool models are selected based on certain rules to maintain an accurate model. In the last step, the selected models are combined based on some heuristics and rules.

This paper proposes a multi-layer clustering and classification model consisting of primary classifiers, using SOM and metaheuristic algorithm. In the first step, instead of a traditional artificial neural network algorithm, we use self-organizing maps (SOM) for clustering our dataset. SOM models are unsupervised ANN-based clustering techniques based on competitive learning in which the nodes associated with weights compete with each other to win an input pattern. It is used to represent the high-dimensional input space to the final low-dimensional space. Once our dataset is divided into different clusters where each data point is now only associated with data points within the same cluster, we apply a classification model to each cluster individually. We use an ensemble classifier and metaheuristic algorithm for increasing the overall efficiency of our model. Our ensemble model is formed using 6 sets of KNN and 6 sets of SVM. The classifiers used have different prediction capabilities. We use SVM for its proven results in classifying non-linearly separable data and KNN for its closest distance-based classification.

Application of metaheuristics to classification problems has been observed in the literature as early as 2007. For instance, Ant Colony Optimization was used in standard classification problems (Martens et al. 2007). In our approach, the latest metaheuristic model, sailfish optimizer (SFO) is employed for weights generation for our ensemble model. The relative performance of the classifier is considered where the classifier that performs well on the cluster dataset is assigned higher weights. Based on the behavior of sailfish hunting for sardines, the optimizer has shown excellent results when tested against various optimizers. SFO provides advantages in terms of its exploration and exploitation phases, a formidable convergence speed in reaching global optimization, and also better efficiency.

In all previous research, the model is trained using the entire dataset. Our model departs from existing models by initially clustering similar data points and applying the classification model to each of them. Also, the use of diverse base classifiers

in our ensemble model yields excellent results. Another advantage of our proposed model is the use of SOMs to find better sets of clusters of the dataset. This training technique develops highly efficient state-of-the-art models that can provide high efficiency.

Thus, SCMVE offers the following advantages and novelties compared to existing work in the same field:

- *Clustering through Self Organising Maps that identifies groups with distinct characteristics that can be used to identify and train the classification models on*
- *Use of diverse base classifiers in soft-voting ensemble model, an approach that outperforms singular complex models such as artificial neural networks*
- *Utilising an optimized metaheuristic sailfish optimizer for weights optimization of the ensemble model, leading to enhanced performance on various classification metric evaluations*
- *Application of a unique training strategy, where local fitting of classifier models are done to the formed clusters (as opposed to training on the whole dataset), hence utilising the distinct characteristics of each cluster found through SOM clustering phase leading to outstanding classification results.*

The remainder of this paper is structured as follows: Sect. 2 provides a summary of previous work done in credit scoring. Section 3 explains the methodology of SCMVE. Section 4 introduces the experimental design, followed by the experimental results in Sect. 5. The conclusion and future work are provided in Sect. 6.

2 Related work

Since credit scoring is fundamentally a binary classification problem, there exist models that employ the use of a single classifier to predict a credit defaulter.

Earlier models mostly relied on existing statistical methods such as Linear Discriminant Analysis (as proposed by Fisher 1936) or the usage of Logistic Regression and Superscorecards as proposed by Hand and Kelly (2002) These basic models performed daily however failed to provide reliable results given unrepresentative or new samples.

However, with the advent of machine learning techniques the aim was set to solve the issues associated with the statistical methods. Models based on k-Nearest-Neighbours (Henley and Hand 1996), Support Vector Machines (Li et al. 2006), Classification Trees (Li et al. 2004) and Neural Networks (West 2000) were employed to considerable success for solving the same.

2.1 Neural network models

Lee et al. (2002) proposed a hybrid neural discriminant model which aimed to simplify the neural networks by incorporation of the traditional discriminant analysis. They employed LDA to first model the credit scoring problem and identify

the significant input variables, which are then passed on to the Artificial Neural Network to perform the prediction. Hence, the LDA acted as a tool to design the topology of the network.

Onan (2019b) proposed a two-stage neural network framework for topic extraction, using advanced embeddings (word2vec, POS2vec, word-position2vec, LDA2vec) and an ensemble clustering approach.

Lappas and Yannacopoulos (2021) are able to combine expert knowledge and genetic algorithms into a machine learning approach to supplement supervised credit scoring with expert input, increasing the overall efficiency. Safi et al. (2022) combine a neural network with metaheuristic optimizers and five cost sensitivity fitness functions as the base learners in a majority voting ensemble learning paradigm.

Onan (2023a) proposed a novel GTR-GA approach, which integrates graph-based neural networks with genetic algorithms for text augmentation, effectively enhances data diversity and improves model performance.

2.2 Self organising map models

As an extension to the artificial neural network models used for the classification step, self organising maps, which are a widely employed neural network model, have been used instead of the traditional ANN models. Unlike the ANN, the SOM models are unsupervised and produce a topology preserving mapping between the input parameter set (which is a higher dimensional map space) to the final credit score (a lower dimensional map space). Lau (2006) have explored different SOM Models that can be employed for classification problems.

Hsieh (2005) proposed a hybrid mining approach which used a self organising map for deciding the parameters of the K Means clustering algorithm followed by an Artificial Neural Network for classifying the samples. Suleiman et al. (2021) also used an SOM to improve the ability of pattern recognition in the data and utilise that to increase the efficiency of K-Means Classifier and neural networks.

AghaeiRad et al. (2017) developed a hybrid model for credit scoring using SOM clusters and feedforward neural networks. They used the knowledge obtained from SOM clusters and passed it as information to be trained on by the neural network classifier. This gave better results compared to a standalone FNN due to the increased information present in the input.

Onan (2019a) proposed a consensus clustering-based undersampling method to tackle class imbalance in machine learning. By undersampling the majority class with various clustering algorithms and evaluating performance using different supervised and ensemble learning methods.

Onan (2021b) introduces sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. The architecture includes a weighted embedding layer, convolutional layers (1-g, 2-g, 3-g), max-pooling, LSTM, and a dense layer.

Onan (2022) proposed bidirectional convolutional recurrent neural network employs bidirectional LSTM and GRU layers to capture both past and future contexts.

2.3 Metaheuristic algorithms in credit scoring

Genetic algorithms have been used widely and in different capacities in the credit scoring problem.

Onan et al. (2016b) employed a multiobjective differential evolution algorithm to optimize classifier and class weights within a static classifier selection framework for sentiment analysis.

Onan (2018a) optimized the Latent Dirichlet Allocation (LDA) parameters by employing a swarm intelligence approach. They utilized metaheuristic algorithms like Particle Swarm Optimization (PSO) to estimate the number of topics and other key parameters in LDA. This complements He et al. (2018) work on adaptive models.

Pławiak et al. (2019) published a work constituting an application of genetic cascading of ensemble classifiers. They combined the benefits of evolutionary algorithms and ensemble classifiers, alongside using deep learning to develop a complex credit scoring model. They used genetic algorithms in three different instances: first for feature selection on the Australian dataset, second for hyperparameter optimization, and a third for deriving a training technique for selecting the classifiers for the final trained model.

Kozodoi and Lessmann (2020) improved on the feature selection through GA by modelling it into a multiobjective task. They then used a Particle Swarm Optimization(PSO) algorithm and evaluated it using three different fitness functions based on: the number of features, the relative acquisition costs of the features and the AUC-ROC curve fitness score of the trained model.

Tripathi et al. (2020) proposed a novel Binary BAT algorithm on the credit scoring dataset. They combined it with a radial Neural Network (RBFN) for a hybrid credit scoring model. They used the metaheuristic algorithm for feature selection on the dataset and used the Neural Network for training the selected features in a supervised classification task.

Simumba et al. (2022) incorporated stakeholder requirements into their model during feature selection and used that to train their supervised classification model. They compared the results of two modified metaheuristic algorithms: a Grasshopper algorithm integrated with non-dominated sorting and genetic algorithm, and a genetic algorithm integrating different selection, crossover, and mutation strategies. They evaluated their results with empirical data collected from farmers in Cambodia.

Şen et al. (2020) proposed multilevel metaheuristic algorithms for credit scoring. They used an SVM classifier, combined with a Genetic Algorithm in a two-level feeding mechanism for increased model accuracy. They used the GA first to find the optimized parameters of the SVM model, and then used it for feature selection on the dataset that increases the classification accuracy.

Onan (2023b) proposed SRL-ACO, a text augmentation framework that uses Semantic Role Labeling (SRL) and Ant Colony Optimization (ACO) to generate additional training data for NLP models. SRL-ACO enhances data quality by preserving semantic roles in new sentences. Experimental results demonstrate improved performance in sentiment analysis, toxic text detection, and sarcasm identification tasks.

2.4 Ensemble models in credit scoring

Ensemble models and models based on genetic algorithms combine the power of machine learning models to generate highly accurate credit risk assessment models and assist us in classifying previously unrepresentative samples in the dataset.

Van Gestel et al. (2003) proposed a Least Squares Support Vector Machine (LS-SVM) classifiers within the Bayesian evidence framework to predict the tendency of an entity to default on their credit given a set of parameters. Once the probabilities for the defaulting of an entity are generated a sensitivity analysis with respect to the input set is carried out which provides us with an insight into the parameters that are affecting the creditworthiness of the entity.

Onan et al. (2017) work on hybrid ensemble pruning enhances classifier diversity and performance, aligning with the approach of improving sentiment classification accuracy through advanced ensemble methods.

Onan (2021a) work on hybrid ensemble pruning improves classifier diversity and performance, enhancing sentiment classification accuracy through advanced ensemble methods.

Onan (2018b) integrated a Random Subspace ensemble of Random Forest classifiers with four types of features-authorship attribution, character n-grams, part of speech n-grams, and discriminative word frequency.

Hsieh and Hung (2010) investigated the approach involving the proper preprocessing of the dataset into homogenised clusters followed by the classification of the samples into the preprocessed categories. The ensemble model hence proposed resulted in an efficient ensemble classifier.

Onan et al. (2016a) combined five statistical keyword extraction methods with various ensemble techniques for text classification, including Naïve Bayes, SVM, logistic regression, and Random Forest, akin to how Zhang et al. (2021) utilized ensemble techniques for improving credit risk models. This integration of feature

extraction with ensemble methods aligns with the advancements in robust predictive modeling.

He et al. (2018) improved on the construction of the ensemble models by basing their adaptability to different imbalance ratios by the supervised undersampling of the dataset (based on the estimation of the data imbalance ratio), followed by classification by tree-based base classifiers which classify samples in the respective data subsets. Finally, a particle swarm optimization algorithm was applied to the base classifiers to obtain the final ensemble model.

Zhang et al. (2021) were able to develop a novel ensemble model that used a local outlier factoring algorithm added with a bagging strategy to construct a trained model that works on the outlier adaptability of base classifiers. They combined novel methods of feature reduction and ensemble learning methods for parameter optimization and finally used a stacking based ensemble model to train the dataset on.

Nalič et al. (2020) used ensemble techniques for feature selection on datasets and proposed the if_any voting method that was able to outperform other standard voting procedures. They combined linear models, SVMs, naive Bayes and decision tree classifiers into a soft voting ensemble model.

Xia et al. (2020) proposed a new tree-based overfitting-cautious heterogeneous ensemble model for credit scoring. The suggested method could dynamically give weights to base models based on the overfitting metric during ensemble selection.

To improve the prediction performance of credit scoring, Tripathi et al. (2019) created a hybrid model that combines feature selection and a multilayer ensemble classifier architecture. The first phase is preprocessing, which sets ranks and weights of classifiers, followed by the ensemble feature selection approach, and finally, the dataset with the selected features is used in a multilayer ensemble classifier architecture. In addition, since classifier placement influences the ensemble framework's predictive performance, a classifier placement algorithm based on the Choquet integral value was devised.

Xia et al. (2018) introduced a new heterogeneous ensemble credit model that combined the bagging and stacking algorithms. In three ways, the proposed model varied from the existing ensemble credit models: pool creation, base learner selection, and trainable fuser.

To increase the prediction performance, Guo et al. (2019) proposed a novel multistage self-adaptive classifier ensemble model based on statistical approaches and machine learning techniques. First, the original data was processed into a standardized, representative sample using a multistep data preparation method. Second, based on their performance, base classifiers were self-adaptively picked from the candidate classifier repository, and their parameters were adjusted using the Bayesian optimization algorithm post which the ensemble model was integrated using these optimized base classifiers, and used multilayer stacking to generate new features and particle swarm optimization to achieve the classifier weights in the ensemble model.

Xiao et al. (2020) devised a GMDH-based cost-sensitive semi-supervised selective ensemble (GCSSE) model using a semisupervised, cost-sensitive learning, group method of data handling (GMDH), and an ensemble learning strategy.

Liu et al. (2022) proposed a multigrained and multi-layered gradient boosting decision tree (GBDT) that effectively increases the diversity of prediction and further improves the performance of credit scoring by providing more precise credit scoring results. The classification techniques for credit scoring have been summarised in Table 1.

3 Proposed methodology

This section gives a detailed description of SCMVE which is the framework used to determine if a data point belongs to good credit or risky credit. SCMVE uses the concept of multi-layer clustering using an artificial neural network-based technique (Self Organising Map) and then uses a soft voting-based ensemble classifier to predict the final class of the test data point.

The weights for our ensemble model are generated using a metaheuristic optimization technique—Sailfish Optimizer.

The overall architecture of SCMVE is divided into 2 phases:

- **Clustering Phase**: The entire dataset is clustered with data in them having similar feature values. The idea of cluster is based on the fact that different set of controlling attributes which affect the final classification are better identified and trained accordingly when dealt with as a cluster.
- **Classification Phase**: Once data is divided into cluster regions, the classification models predict the final class of data points. To achieve classification, we use an ensemble classifier based on Support Vector Machines and K Nearest Neighbour Classifiers. A metaheuristic algorithm is further used to optimise the ensemble model.

The construction of the proposed model SCMVE is illustrated in Fig. 1.

3.1 Unsupervised clustering using SOM

The first step of SCMVE is to cluster the entire dataset based on the similarity of attributes between data points. The rationale behind this step is to be able to predict and identify unique characteristics distinct to each cluster.

For this step, we use Self Organising Maps (SOM). SOM can find better sets of clusters on the given dataset (Gholamian et al. 2013).

ML algorithms can be of two types: Supervised and Unsupervised learning. In supervised learning, we predict or categorize the outcome based on the input, and in unsupervised learning, we describe patterns in input data without having the

Table 1 Different classification techniques in credit scoring

Classification method	Description	Advantages	Disadvantages
Singular model classification	Involves conducting supervised classification and training on a singular classification model, such as a neural network or a K-means classifier	Reduced training time. Simple classification models	Low accuracy compared to more complex models
SOM Model for optimization	Involves the usage of a Self Organizing Map for clustering the data points and using a supervised classifier on the optimized data set. SOM may be used in coordination with other optimisation techniques	The clusters found by SOMs are able to generate higher accuracy	Performance decreases with increase in dataset size
Metaheuristic models for optimization	This method involves using metaheuristic optimization models (such as sailfish optimizer or the genetic algorithm) for tasks such as feature selection, weights optimization etc to improve the efficiency of supervised classifiers	Reduced training time. Reduced computational complexity	Performance not consistent: dependent on the task, efficiency metric and other factors
Ensemble Classifiers	This technique involves using multiple base classifiers to produce the output. These classifiers may be optimized using various techniques. The output of these base classifiers is combined using a weighted voting approach	Higher predictive accuracy	High cost of training the models

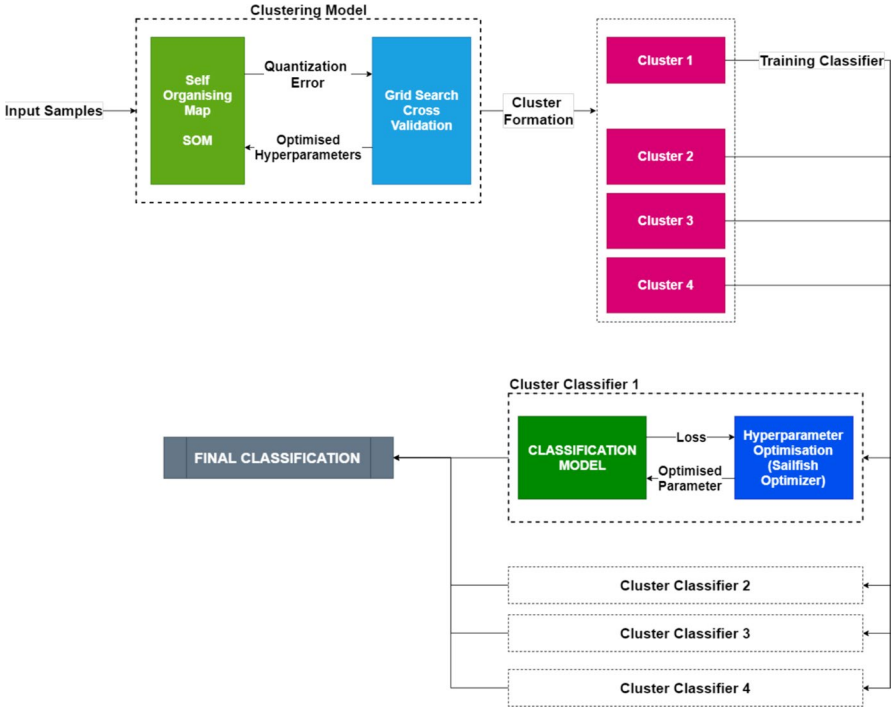


Fig. 1 SOM clustering and metaheuristic voting ensemble classifiers for credit scoring

knowledge of existing class labels. Clustering using Self Organising Maps is an unsupervised learning task.

SOM (Kohonen 1998) is a widely applied neural network-based clustering technique inspired by the arrangement of biological neurons. It is commonly used to represent high dimensional input data into a lower dimension, nonlinear approximation of high dimension data(called the training data), clustering, and feature selection.

As any neural network, a SOM consists of input and output layers. The objective is to transfer all input data to output in such a way that they all connect. Self-organizing maps can overcome the limitations of other statistical techniques through *straightforward implementation, better execution speed, and shorter training periods*. SOM differs from other neural networks in the use of *competitive learning* rather than error correction learning such as backpropagation. These characteristics have resulted in many studies in the past two decades.

Definition 1 (Neuron) Self Organizing maps consist of a grid of nodes. Each node is referred to as a neuron. Each neuron is assigned a weight vector of the dimension of its input vector X_0.

Definition 2 (Best Matching Neuron, ξ) In every iteration, the distance $d_{X \to n}$ between the input vector X_0 and each neuron is calculated.

$$d_{X \to n} = X_0 - n^2 \ \forall n \in X_1, \tag{1}$$

where $X_{1,}$ represents all the layers excluding the input layer. The neuron with the minimum distance is called the best matching neuron ξ and its weight vector is updated along with its neighbors to move it closer to the input vector.

$$\xi = min(d_{X \to n}) \tag{2}$$

Definition 3 (Neighborhood Radius Function, η)—Neighborhood Radius Function is a function used to compute the radius for defining the neighbors of a given neuron. It is a decreasing time function such that the number of neighbors reduces during the training phase. The neighbourhood radius function is represented as:

$$\eta(u, v, S, i) = exp\left(-\frac{d_{u,v}^2}{2\sigma(i)^2}\right), \quad i = 0, 1, 2, ... \tag{3}$$

where u, v represent two vectors of same dimension, S represents the input space, i represents the ith iteration.

Definition 4 (Neighbor Neuron, NN)—The neurons close to the best matching neuron that lie within a certain minimum neighborhood radius NR_{min} are termed as Neighbour Neuron NN which is mathematically represented as

$$NN = \left\{ n \text{ if } \eta(\xi, n, S, i) < \mathrm{NR}_{\min}, \forall n \in S \right. \tag{4}$$

where η is the neighbourhood function, and S represents the entire space/network of the SOM.

Definition 5 (Learning Rate, λ)—Learning rate is a parameter used to update the weights of the neurons.

Definition 6 (Topology-Preservation)—refers to the condition where the mapping preserves the relative distance between nodes. Nodes that are near each other in the input space are mapped to nearby units in the output.

Algorithm 1 Clustering using Self Organising Map

Data : $\boldsymbol{X}$: Training Feature Set,
$\boldsymbol{X_{cluster}}$: Testing Feature Set
SOM.shape: shape of SOM,
$\boldsymbol{\sigma}$: radius of different neighbours,
$\boldsymbol{\lambda}$: learning rate,
num_iter: number of iterations
$\boldsymbol{\eta}$: neighbourhood radius function
Result : **cluster:** List of cluster IDs assigned to each row of feature set

begin
  Initialization
  $i \leftarrow 0$
  $W[num_iter] \leftarrow \phi$
  $W[0] \leftarrow rand.matrix(SOM.shape)$ /* $0 \leq W[k] \leq 1$ */
  while $i < num_iter$ **do**
    $selected_node \leftarrow \phi$
    $min_dist \leftarrow \infty$
    $min_node \leftarrow \phi$
    for *randomly selected vector* $v \in X$ **do**
      $d \leftarrow \eta(v, W[i])$ /* Compute distance with weight matrix */
      if $d < min_dist$ **then**
        $min_dist \leftarrow \min(d, min_dist)$
        $min_node \leftarrow v$
      end
    end
    $W[i+1] \leftarrow W[i] + min_dist \cdot \lambda \cdot (min_node - W[i])$ /* Updating weights */
    $i \leftarrow i + 1$
  end
  $y \leftarrow \phi$
  foreach *vector* $v \in X_cluster$ **do**
    $d_v \leftarrow \eta(v, W[num_iter])$ /* Store the neighbourhood radius value of a vector */
    $\xi \leftarrow argmin$d_v, cluster_nodes /* Find the winner neuron to assign cluster */
    y.insert(ξ)
  end
end

SOM is commonly set up as a 2D lattice representing a grid-like structure for neurons forming a topological map as described in Fig. 2.

The clustering algorithm is explained below:

1. Each neuron is initiated with a random weight vector, with the same dimensionality as the input vector. Neurons are adjusted during the training period based on competitive learning.

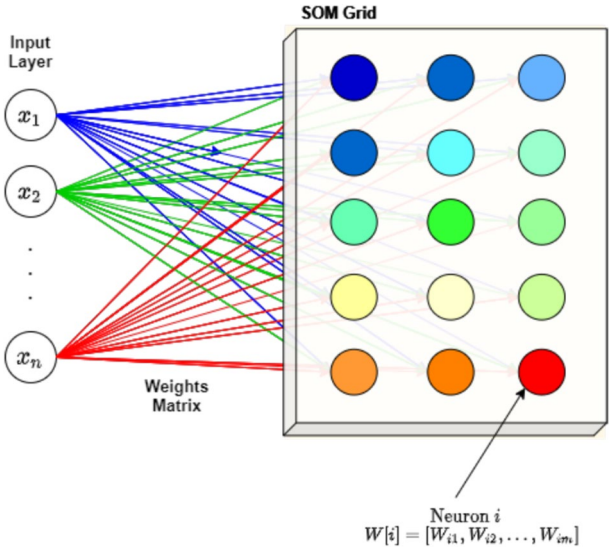


Fig. 2 Self organising map

2. In the SOM grid, every input data is related to every output neuron on the grid. In every iteration i, the neuron with the shortest distance to the input vector is denoted as ξ (or Best Matching Neuron) and the representative vector of ξ is updated so that it moves closer to the input vector.
3. Then, the weights of ξ's neighborhood neurons are updated so that they are symmetrically adapted towards the training input vector. The neuron weights are updated according to the following equation-

$$W_{i+1} = W_i + \eta_{BMU} \cdot \lambda\left(X_i - W_i\right) \tag{5}$$

where X_i is the input vector, W_i is the neuron vector, λ denotes the learning rate, and η_{BMU} represents neighborhood function value for the best matching unit.
4. This training cycle is repeated for each input vector. After finishing the training, the final grid represents the topology-preserving mapping of input data where related neurons are closer to each other.

Note that the update rate for the best matching neuron, the neighbor neuron, and the farther neighbor neuron will differ. The neighborhood radius is a predefined parameter that gradually decreases during the entire training phase and is finally set to 1. A summary of the different parameters involved in the self-organising map is provided in Table 2.

Finally, clusters of neurons are formed, each representing a different class of data that can be appropriately labeled.

Table 2 Parameters involved in the self organising map

Parameter name	Parameter description
η	Neighbourhood radius function
$d_{X \to n}$	Distance computation function
SOM.shape	Shape of the map kernel
λ	Learning rate
num_iter	Number of total training iterations

3.2 Ensemble-based classification model

In this phase of SCMVE, the clusters formed in the first stage are individually used to train and test our ensemble model. Data points in each cluster are split into training and testing data.

The conventional idea is to use a single classifier or a simple combination to make predictions that are often less accurate. One of the techniques to improve the performance of the individual classifier is to combine them using different techniques.

The technique we use is called **ensemble learning**. Various techniques like voting can be used to combine these classifiers used in the ensemble. Ensemble models gather the decisions of several base classifiers trained on the same dataset to get a more effective and accurate decision. The aim is to compensate for the error of a single classifier by other classifiers of the ensemble, increasing the accuracy of the ensemble model as compared to a single classifier model.

For credit scoring, diverse base classifiers could improve the model's accuracy and also provide insightful information about the applicant even if they belong to the same category. The ensemble model used in our paper consists of 12 base classifiers—6 SVM's and 6 KNN's. To ensure maximum utilization and greater diversity of base classifiers, we have used different hyperparameters for each classifier. The classifiers used in the model are explained below -

- **SVM**: an effective classifier model that maps the data into high dimensional spaces and finds an optimal hyperplane to maximize the distance between different classes. A hyperplane is a surface that partitions data points into classes, where points from different classes lie on different sides of the hyperplane. Kernel functions such as linear, polynomial, RBF, and sigmoid are used to map the data into high dimensional spaces.
- **KNN**: one of the most simple yet effective non-parametric classification models. It is an instance-based learning technique and does not require a learning phase. For a data point n to be classified its neighborhood is retrieved i.e k nearest neighbors of n. Selecting an appropriate value of k is necessary as the success of the classifier depends on k.

Votes from each classifier can be combined using several techniques, majority vote ensemble being the most common. In majority voting, each classifier gives a binary vote of 0 and 1. The majority voting ensemble uses two techniques: hard voting and soft voting.

Hard voting makes the final prediction using majority votes from the classifiers, whereas in **soft voting** each classifier corresponds to a probabilistic outcome. The class with the highest probability is predicted as the final output. **In our model, we use the soft voting approach**. The following equations describe the soft voting procedure

Among the n classifiers, each classifier C_i gives the probability P_i such that the predict label $y_i = 0$ for the corresponding input X.

$$P_i = P\big(C_i[X] = 0\big) \tag{6}$$

For each classifier, we assign a weight w_i which determines its relative importance among the others in the ensemble and are assigned based on the performance of an individual classifier (further optimised using the sailfish algorithm as explained later). The final output label y is determined as follows

$$y = \begin{cases} 1 & \frac{\sum_{i=1}^{n} w_i \cdot P_i}{\sum_{i=1}^{n} w_i} \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Algorithm 2 Ensemble Model using Soft Voting Classifier

Data : $\boldsymbol{X}$: Input data
$\boldsymbol{N}$: Total classifiers ensembled (in our case $= \mathbf{6}$),
$\boldsymbol{P}[\boldsymbol{N}]$: probability of outcome 0 for each classifier $\boldsymbol{C_i}$
$\boldsymbol{w}[\boldsymbol{N}]$: Array storing weight of each classifier $\boldsymbol{C_i}$
Result : y: Output labels $\in \{0, 1\}$

```
begin
    Initialization
    y ← φ
    P_0 ← 0
    P_1 ← 0
    foreach x ∈ X do
        for n ← 1 to N do
            P_0 = P_0 + P[n] * w_n
            P_1 = P_1 + (1 − P[n]) * w_n
        end
        if P_1 > P_0 then
            y.insert(1)
        end
        if P_0 > P_1 then
            y.insert(0)
        end
    end
end
```

3.3 Sailfish algorithm for weight optimization

In our model SCMVE, we have used a metaheuristic optimization technique, **sailfish optimizer** to generate relative weights for each classifier in the ensemble. The generation of population of weights is done using a uniform distribution using the latest metaheuristic models.

In each epoch of the metaheuristic model training, we find the optimal weight set that can be used for our ensemble model. It is further tested on the train test data and therefore increases the overall efficiency of our model.

- Sailfish optimizer algorithm is a nature-based metaheuristic optimization algorithm recently developed by Shadravan et al. (2019). The algorithm is inspired by the behaviour of sailfish which are known for their coordinated and intelligent hunting strategy (Fig. 3).
- One of the common strategies for group hunting is the attack alternation strategy which is described below:
  - Sailfish usually attack one at a time using their rostrum injuring the sardines severely.

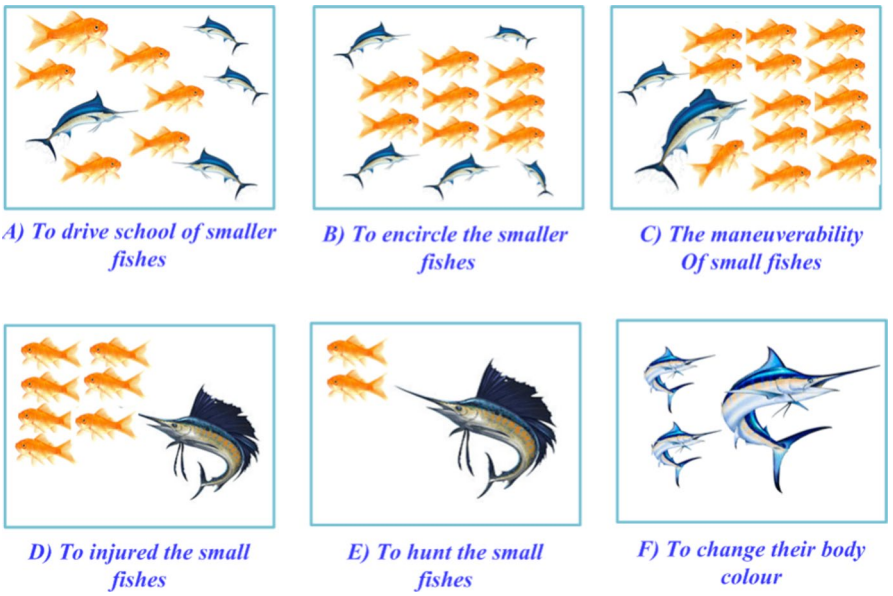


Fig. 3 Sailfish group hunting (Transpire Online sailfish optimizer 2019)

– Being one of the fastest marine beings in the ocean, it is almost impossible for sardines to avoid the attack.
– The injured will eventually be separated from the rest of the prey school and will be captured by the hunters (sailfish).
– Sailfish attacks do not kill the sardine but the frequent attack increases the number of injured sailfish.
– Sardines are another important inspiration for the SFO algorithm. It imitates their ability to change position to escape from an attack.

When tested with other metaheuristic algorithms: Particle Swarm Optimization (PSO), Grey Wolf Optimization (GWO), Genetic Algorithm (GA), Ant Lion Optimizer (ALO), Salp Swarm Algorithm (SSA), and Satin Bowerbird Optimizer (SBO) using a set of unimodal, multi-modal and fixed dimension multimodal benchmark functions, SFO showed competitive results in terms of exploration and exploitation phases (Nassef et al. 2021).

Moreover, SFO shows high-speed convergence on multimodal functions in reaching the global optimum while avoiding the local optimum.

Definition 7 (Search Space, S) It is the set of all possible solutions that the input can take. It contains a set of points (each representing a possible solution) that give the optimal solution. Optimization aims to search that point(solution).

Definition 8 (Fitness Function, $\psi(x)$) It is an objective function that defines the optimality of a solution to a given target problem. It defines how close a given solution is to achieve the desired output. Any mathematical operation that is able to assign computable scores to the states of the matrices can be used for the approach.

The functions that have been experimented with in this paper are mentioned in Table 3. For this paper, the mean squared error function is used to compute the fitness as per the Algorithm 3.

Table 3 Description of fitness functions

Function	Abbreviation	Formula
Mean squared error	MSE	$\text{MSE}(x,y) = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - y_i)^2$
Mean absolute error	MAE	$\text{MAE}(x,y) = \frac{1}{n}\sum_{i=0}^{n-1}\lvert x_i - y_i\rvert$
Mean square logarithmic error	MSLE	$\text{MSLE}(x,y) = \frac{1}{n}\sum_{i=0}^{n-1}(\log_e(1+x_i) - \log_e(1+y_i))^2$
Mean tweedie deviance (order = 1)	MTD-1	$\text{MTD}_1(x,y) = \frac{1}{n}\sum_{i=0}^{n-1} 2(x_i \log(x_i/y_i) + y_i - x_i)$
Mean tweedie deviance (order = 2)	MTD-2	$\text{MTD}_2(x,y) = \frac{1}{n}\sum_{i=0}^{n-1} 2(\log(y_i/x_i) + x_i/y_i - 1)$
Mean tweedie deviance (order = 3)	MTD-3	$\text{MTD}_3(x,y) = \frac{1}{n}\sum_{i=0}^{n-1} 2\left(\frac{\max(x_i,0)^{-1}}{2} + \frac{xy_i^{-2}}{2} - y_i^{-1}\right)$
Mean tweedie deviance (order = 4)	MTD-4	$\text{MTD}_4(x,y) = \frac{1}{n}\sum_{i=0}^{n-1} 2\left(\frac{\max(x_i,0)^{-2}}{6} + \frac{xy_i^{-3}}{3} - \frac{y_i^{-2}}{2}\right)$

Algorithm 3 Fitness Function Computation using Mean Squared Error

Data : $\boldsymbol{N}$: Number of classifiers
$\boldsymbol{W}[\boldsymbol{N}]$: Weight array containing the weights for the classifiers
$\boldsymbol{X_{train}}$: Training data
$\boldsymbol{y_{train}}$: Training labels
$\boldsymbol{X_{test}}$: Testing set
$\boldsymbol{y_{test}}$: Testing Labels
Result : $\boldsymbol{\psi(x)}$: Fitness value for the given weight array

begin
  Initialization
  $P[N] \leftarrow \phi$ /* Array to store the predictions of the classifiers */
  Initialise classifiers $C_{1 \rightarrow N}$
  $M \leftarrow \text{Ensemble}(C_i \quad \forall i \in \{1 \ldots N\})$ /* Combines the individual classifiers into an ensemble */
  $M.\text{train}(X_{train})$
  foreach $x \in X_{test}$ **do**
    $p \leftarrow M.\text{predict}(x)$
    $P.\text{insert}(p)$
  end
  $\psi \leftarrow \frac{1}{\text{length}(y_{test})} \cdot y_{test} - P^2$ /* Sum of squared error */
end

Definition 9 (Prey Density, PD) The value indicating the number of prey(sardines) present in each iteration and it decreases in every iteration as the prey decreases during hunting. The value PD can be calculated as follows

$$PD = 1 - \left(\frac{N_{SF}}{N_{SF} + N_S} \right) \tag{8}$$

where N_{SF} and N_S are the numbers of sailfish and sardines in each cycle, respectively.

Definition 10 (Attack Power, AP) Represents the sailfish's attack power at each iteration. It helps us to calculate the number of sardines that update their position. The attack power is calculated as

$$AP = A \times (1 - (2 \times i \times \varepsilon)) \tag{9}$$

where A represents the coefficient of the sailfish's decrementing attack power, ϵ is the learning rate and i denotes the ith iteration.

Definition 11 (Coefficient of Mutation of Position, θ_P) At each step of the algorithm, the positions of the sailfish and sardines change. The extent to which such a positional change occurs is given by the coefficient of mutation of position, θ_P which is analogous to a time-varying learning rate.

The expression for the coefficient is

$$\theta_P = PD \cdot (2 \cdot \text{rand}(0, 1) - 1) \tag{10}$$

The sailfish are assumed as candidate solutions and their position in the search space denotes the problem variables.

For n sailfish and m sardines in the optimization problem of d dimensions, matrix $P^i_{sf}[n][d]$ is used to show the position of every sailfish at the ith iteration.

Sardines are assumed to be swimming in the same search space. Matrix $P^i_{sar}[m][d]$ shows the position of every sardine at the ith iteration.

To compute the fitness value of each sailfish x and each sardine y the objective fitness function is applied to each row at a given iteration i as follows

$$\begin{aligned} &\text{fitness value of sailfish x} = \psi(P^i_{sf}[x][,]) \\ &\text{fitness value of sardine y} = \psi(P^i_{sar}[y][,]) \end{aligned} \tag{11}$$

The following matrices give the fitness value for all positions of sailfish and sardines

$$\Psi_{sf}[n][1] = \begin{bmatrix} \psi(P^i_{sf}[0][0,d]) \\ \psi(P^i_{sf}[1][0,d]) \\ \vdots \\ \psi(P^i_{sf}[n][0,d]) \end{bmatrix} = \begin{bmatrix} \Psi_{SF}[0] \\ \Psi_{SF}[1] \\ \vdots \\ \Psi_{SF}[n] \end{bmatrix} \tag{12}$$

where n is the number of sailfish and d shows the number of variable and $P^i_{sf}[x][,]$ is the position of sailfish x

$$\Psi_{sar}[m][1] = \begin{bmatrix} \psi(P^i_{sar}[0][0,d]) \\ \psi(P^i_{sar}[1][0,d]) \\ \vdots \\ \psi(P^i_{sar}[m][0,d]) \end{bmatrix} = \begin{bmatrix} \Psi_{sar}[0] \\ \Psi_{sar}[1] \\ \vdots \\ \Psi_{sar}[m] \end{bmatrix} \tag{13}$$

where m is the number of sardines and d shows the number of variables and $P^i_{sar}[y][,]$ is the position of sailfish y.

Algorithm 4 Sailfish Optimization

Data : $\boldsymbol{A}$: Coefficient of sailfish attack power,
$\boldsymbol{n}$: Number of sailfish,
$\boldsymbol{m}$: Number of sardines,
$\boldsymbol{d}$: Dimensionality of data,
$\boldsymbol{\epsilon}$: Learning rate,
$\boldsymbol{\psi(x)}$: Objective function
Result : $\boldsymbol{SF_{best}}$: Position of best sailfish

begin
  Initialization
  $P_{sf}^{1000}[n][d] \leftarrow \text{rand}(0,1)$
  $P_{sar}^{1000}[m][d] \leftarrow \text{rand}(0,1)$
  $F_{sf}[n][1] \leftarrow \Psi(P_{sf})$
  $F_{sar}[m][1] \leftarrow \Psi(P_{sar})$
  $SF_{best} \leftarrow \text{rand}(0,n)$
  $SAR_{best} \leftarrow \text{rand}(0,m)$
  $D_{prey} \leftarrow 1 - \frac{n}{n+m}$
  $i \leftarrow 0$
  while *true* **do**
    foreach *sailfish* $x \leftarrow 0$ *to* n **do**
      $\theta_P \leftarrow D_{prey}(2 \cdot \text{rand}(0,1) - 1)$
      $P_{sf}^{i}[x] \leftarrow SF_{best} - \theta_P \left(\text{rand}(0,1) \cdot \frac{SF_{best}+SAR_{best}}{2} - P_{sf}^{i}[x]\right)$
    end
    $AP \leftarrow A \cdot (1 - 2 \cdot i \cdot \epsilon)$
    if $AP < 0.5$ **then**
      $\alpha \leftarrow m \cdot AP$
      $\beta \leftarrow d_i \cdot AP$
      $S_{selected} \leftarrow$ list with α number of 1 and $m - \alpha$ number of 0
      $d_{selected} \leftarrow$ list with β number of 1 and $d - \beta$ number of 0
      $S_{sar} \leftarrow P_{sar}^{i}[S_{selected}][d_{selected}]$
      $S_{sar} \leftarrow\sim \mathcal{N}(0,1) \cdot (SF_{best} - S_{sar} + AP)$
    end
    else
      $P_{sar} \leftarrow\sim \mathcal{N}(0,1) \cdot (SF_{best} - P_{sar} + AP)$
    end
    $F_{sar}[m][1] \leftarrow f(P_{sar})$
    if $\Psi(P_{sar}^{i}) < \Psi(P_{sf}^{i})$ **then**
      $P_{sf}^{i} \leftarrow P_{sar}^{i}$
      P_{sar}^{i}.remove(SAR_{best})
    end
    $i \leftarrow i + 1$
  end
end

Algorithm 5 SCMVE Credit Scoring

Data : $\boldsymbol{X_{train}}$: Training feature set,
$\boldsymbol{y_{train}}$: Training feature labels
$\boldsymbol{X_{test}}$: Testing feature set
$\boldsymbol{SOM_c}$: SOM clustering model
$\boldsymbol{E_c}$: Ensemble classification model
$\boldsymbol{S(A, n, m, \epsilon, \phi)}$: Sailfish weight optimizer
Result : $\boldsymbol{m}$: Trained SOM model,
$\boldsymbol{y_{test}}$: Predicted labels for $\boldsymbol{X_{test}}$

begin
  SOM_c.train(X_{train}) /* clusters will be treated as the new class labels */
  $C \leftarrow SOM_c$.cluster_labels
  $CM_{\text{len(C)}} \leftarrow \phi$ /* Pipeline for classification */
  $N \leftarrow len(E_c)$
  foreach *cluster* $c \in C$ **do**
    $w[N] \leftarrow \text{rand}(list(N))$ /* Initialise a randomized weight list for ensemble classifier */
    $w[N] \leftarrow S($
      $A = 4$
      $n = 100$
      $m = 25$
      $\epsilon = 0.001$
      $\psi(N = 6, w[N], X_{train},$
      $y_{train}, X_{test}, y_{test})$
) /* Optimise the weights of the ensemble using Sailfish Optimiser S */
    $CM_c \leftarrow E_c$.train(
      $X_{train}[\text{cluster} = c]$,
      $\text{weight} = w[N]$
) /* Train on the optimised weights */
  end
  /* Training ends */
  $y_{test} \leftarrow \phi$
  foreach $x \in X_{test}$ **do**
    $c_{pred} \leftarrow SOM_c$.predict($x$)
    prediction $\leftarrow CM_{c_{pred}}$.predict(x)
    y_{test}.insert(prediction)
  end
end

The sailfish with the best fitness values are considered elite and are best suited for attacking the prey. These are the sailfish that do not change their position in upcoming iterations until the optimal solution is not lost. Moreover, sardines with the best fitness function are considered the best target by sailfish and are the most exposed to injury.

The sailfish optimizer, given the above-mentioned definitions, works as follows

1. An arbitrary sailfish is assigned to be an elite sailfish whose position is denoted by SF_{best} and an arbitrary sardine the injured sardine whose position is SAR_{best}.
2. The prey density value is calculated as per Eq. 8.
3. The coefficient of mutation of position λ_p is calculated as per Eq. 10.
4. For each sailfish, its position is updated by a certain magnitude as represented in the following equation

$$P^i_{sf}[x] = SF_{best} - \theta_P\left(\text{rand}(0,1) \cdot \frac{SF_{best} + SAR_{best}}{2} - P^i_{sf}[x]\right)$$

5. As the hunting continues through each iteration, the power of the sailfish's attack also reduces as given in Eq. 9. Further, sardines will have less energy and will not be able to properly detect the sailfish position, hence the propensity of getting injured increases. Hence, the sardine positions are also updated as follows

$$P_{sar} = \sim \mathcal{N}(0,1) \cdot (SF_{best} - P_{sar} + AP) \tag{14}$$

where $\sim \mathcal{N}(0,1)$ denotes a random generated between 0 and 1.

6. Next, to determine the number of sardines updating their positions α, we make use of the following expression

$$\alpha = N_s \times AP \tag{15}$$

where N_s denotes the number of sardines.

7. In addition to this, the number of variables updated β, for the corresponding sardine updation is given as

$$\beta = d_i \times AP \tag{16}$$

where d_i is the number of variables in i_{th} iteration

8. The updation occurs as per the following conditions

   (a) If $AP < 0.5$, then positions of only α sardines are updated
   (b) Else every sardine updates its position.

9. Finally, the attack is initiated, and the injured sardines are captured. It is assumed to be the case that the sardines are attacked if their fitness values are better than the sailfish.
10. The space of the captured sardine is substituted by sailfish to increase the chances of hunting. The conditional substitution is represented by the following expression

$$P^i_{sf} = P^i_{sar} \quad \text{if} \quad \Psi(P^i_{sar}) < \Psi(P^i_{sf}) \tag{17}$$

where P^i_{sar} is the position of sardine at its ith iteration and P^i_{sf} gives the position of sailfish at ith iteration.

11. This cycle of updation of positions of sardines and sailfish continues until the end criterion is fulfilled.

The different parameters involved while performing ensemble weight optimization using the Sailfish Algorithm are give in Table 4. The weights generated from the optimization are used with results from the 6 individual classifiers to generate the final predictions.

The last step is to predict the final accuracy of SCMVE, which is calculated using the weighted average of the accuracy of individual classifiers. We use a weighted voting technique and a sailfish optimizer for weight generation for the following reasons:

- Relative weighing the base classifiers on the basis of their performance helps us to value the classifiers that perform better on a certain cluster and undervalue those that perform worse. This helps us to increase our overall performance. In comparison, if all base classifiers were weighed equally, it would lead to a low performing base classifier reducing the overall accuracy of the model.
- Generating the weights using a metaheuristic optimizer converts this into an optimization problem. This not only reduces the training time since we can traverse the search space faster given the nature of our optimizer and its exploration capabilities, it also helps us reach a global maxima on the model efficiency, which otherwise can lead to models getting stuck on local maxima due to inefficient search space exploration strategies.

The final model metrics can be calculated as:

$$NN = \left\{ n \text{ if } \eta(\xi, n, S, i) < \mathrm{NR}_{\min}, \forall n \in S \right. \tag{18}$$

where C_i denotes the ith cluster, X denotes the input dataset and P denotes the performance metrics, for instance, accuracy.

Table 4 Parameters involved in the sailfish optimizer

Parameter name	Parameter description
P	Size of the sailfish population. Denotes the number of solution arrays formed to search the vector space
ϵ	Learning rate of algorithm
N	Number of base classifiers in the ensemble
Lower Bound (lb)	Lower bound of weight search space
Upper Bound (ub)	Upper bound of weight search space
Fitness Function (Φ)	Transformation/function to evaluate the fitness value of given populations

4 Experimental setup

4.1 Description of credit datasets

Security and confidentiality concerns restrict us from accessing private real-world credit scorinf datasets from banks. However, the UCI Machine Learning Repository has various real-world public datasets for credit scoring that are commonly utilized in many research papers to evaluate the performance of the proposed algorithms. We use three public credit datasets from the UCI Machine Learning Repository (Dua and Graff 2017) to evaluate the performance of our proposed approach: Australian Dataset and German Dataset and Taiwan Dataset. The three datasets utilized to evaluate the performance of our proposed model are described in detail in Table 5. The Australian dataset (AUS) has 690 instances, 307 of which are in the good category and 383 in the bad category. There are 14 attributes in the dataset. The German dataset (GER) contains 1000 instances, 700 of which are in the good class and 300 in the bad class. There are 24 attributes in the dataset, including the purpose, credit amount, and personal information. The Taiwan dataset (TAI) has 30,000 instances, 23,364 of which are in the good category and 6636 in the bad category. There are 24 attributes in the dataset. Tables 6, 7 and 8 show a more detailed summary of the attributes.

Table 5 Description of Australian and German and Taiwan credit datasets

Dataset	Australian	German	Taiwan
Number of instances	690	1000	30,000
Number of features	14	20	23
Categorical features	8	13	4
Numerical features	6	7	19
Number of good/bad instances	307/383	700/300	23,364/6636

Table 6 Attribute description of Australian credit dataset

Attribute	Type of attribute	Range
A1	Categorical	0–1
A2	Numerical	13.75−80.25
A3	Numerical	0–28
A4	Categorical	1–3
A5	Categorical	1–14
A6	Categorical	1–9
A7	Numerical	0−28.5
A8	Categorical	0–1
A9	Categorical	0–1
A10	Numerical	0–67
A11	Categorical	0–1
A12	Categorical	1–3
A13	Numerical	0–2000
A14 Class variable	Numerical	1–100,001
A15	Categorical	0–1

Table 7 Attribute description of German credit dataset

Attribute	Type of attribute	Range
A1	Categorical	1–4
A2	Numerical	4–72
A3	Categorical	0–4
A4	Categorical	0–10
A5	Numerical	250–18,424
A6	Categorical	1–5
A7	Categorical	1–5
A8	Numerical	1–4
A9	Categorical	1–5
A10	Categorical	1–3
A11	Numerical	1–4
A12	Categorical	1–4
A13	Numerical	19–75
A14	Categorical	1–3
A15	Categorical	1–3
A16	Numerical	1-4
A17	Categorical	1–4
A18	Numerical	1–2
A19	Categorical	1–2
A20 Class variable	Categorical	1–2
A21	Categorical	1–2

Table 8 Attribute description of Taiwan credit dataset

Attribute	Type of attribute	Range
A1	Numerical	10,000-1,000,000
A2	Categorical	0–1
A3	Categorical	0–2
A4	Categorical	0–1
A5	Numerical	21–79
A6	Numerical	(–2)–(8)
A7	Numerical	(–2)–(8)
A8	Numerical	(–2)–(8)
A9	Numerical	(–2)–(8)
A10	Numerical	(–2)–(8)
A11	Numerical	(–165,580)–(964,511)
A12	Numerical	(–69,777)–983,931
A13	Numerical	(–157,279)–1,664,089
A14	Numerical	(–170,000)–891,586
A15	Numerical	(–81,334)–927,171
A16	Numerical	(–339,603)–961,669
A17	Numerical	0–873,552
A18	Numerical	0–1,689,259

Table 8 (continued)

Attribute	Type of attribute	Range
A19	Numerical	0–896,040
A20	Numerical	0–621,000
A21	Numerical	0–426,529
A22	Numerical	0–528,666
A23	Categorical	0–1

4.2 Data preprocessing and dimensionality reduction

Data preprocessing is an important step before constructing the proposed model since it improves data interpretability and eliminates the sensitivity of difference classifiers to scale differences in attribute ranges. In the first phase, we select the features using the methods described in Sect. 3. We encoded the categorical variables using the dummy variable approach after feature selection. In the dummy variable approach, a variable with n potential values is transformed into n features with a binary value of either 0 or 1. Numerical features, on the other hand, are normalized by following Eq. 1

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)} \tag{19}$$

where x represents the original feature value before normalization, x_{norm} represents the new feature value after normalization, and $min(x)$ and $max(x)$ represent the original feature's minimum and maximum feasible values, respectively. This reduces the sensitivity of different classifiers by scaling each feature's value to a range between 0 and 1.

4.3 Model evaluation metrics

We detail the various performance evaluation measures we used to evaluate our proposed model and compare its performance with that of other well-known classifiers in this section.

Five evaluation metrics are used to assess the overall classification performance of our proposed model: accuracy, precision, recall, F1 score and AUC. The most commonly used evaluation metric is accuracy, which is defined as the ratio of correctly identified instances to the total number of instances in the dataset. However, in the event of a dataset with a class imbalance, accuracy does not reflect the model's true performance

$$Accuracy = \frac{TP + FN}{TP + FN + FP + TN} \tag{20}$$

Another widely used metric in credit scoring studies is the AUC, which is based on the receiver operating characteristic (ROC) curve. The ROC curve is a visualisation of the true positive rate (TPR) and false positive rate (FPR) in binary classification

tasks (FPR). AUC, which is defined as the area under the ROC curve, is used to compare different classifiers. TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{21}$$

$$FPR = \frac{FP}{FP + FN} \tag{22}$$

The following are the definitions for the last three metrics: precision, recall, and F1 score:

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \tag{25}$$

5 Experimental results

We exhibit the obtained data to strengthen the claim that the ensemble model formulated in this study outperforms the other benchmark models in the important parameters mentioned in this paper. Experiments mentioned in this study were conducted with Python 3.7 on Google Colab. Various classifier Table 9 results are compared with our proposed technique, represented in Figs. 4, 5, 6, 7 and 8 in terms of accuracy, precision, recall, F1 score and AUC. In Sect. 5.1, we present a comparative analysis of the results of numerous individual classifiers. Section 5.2 includes the comparison of our method with other classical ensemble methods. In Sect. 5.3, the result of SCMVE is highlighted, and we compare it with various individual and ensemble classifiers.

5.1 Classification results of individual classifiers

In this section, we evaluate the performance of the individual classifiers viz. Logistic Regression, Naive Bayes, K Nearest Neighbors, Gradient Boosting, Decision Trees, Support Vector Machine and Multi Layer Perceptron(MLP) in their ability to evaluate good credit from bad credit based on the features input into the models. Results of the individual classifiers on the Australian dataset can be seen from Table 10. Upon examination, it shows that the performance of all individual classifiers on the performance metrics—Accuracy, Recall and F1 score varies highly. Naive Bayes and K Nearest Neighbours show a relative

Table 9 Values of parameters for different classifiers

Classifier	Parameter	Description	Value
Neural network	Hidden_layer_sizes	Count of neurons in the ith hidden layer	100
	Activation	Activation function for the hidden layer	Tanh
2–4	Alpha	Regularization term parameter	0.001
	Learning_rate	Learning rate schedule for weight updates	Adaptive
	Solver	Solver for weight optimization	SGD
K nearest neighbor	n_neighbors	Number of neighbors to use	5
Support vector machine	C	Regularization parameter	1.0
	Kernel	Kernel type	Poly
	Probability	Enable probability estimates	True
	Gamma	Kernel coefficient	0.1658 (Australian dataset), 0.1221 (German dataset), 0.02631 (Taiwan dataset)
Random forest	n_estimators	Number of trees in the forest	200

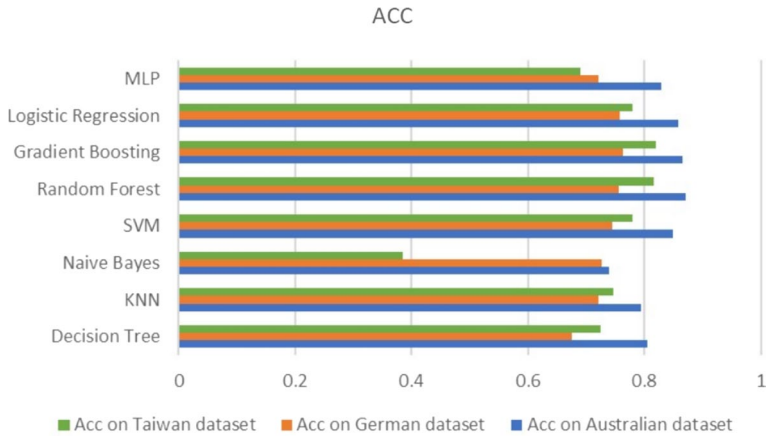


Fig. 4 Comparison of accuracy

decline in performance. Another observation is that the performance of Random Forest Algorithm is significantly better than the rest of the individual classifiers in terms of metrics—accuracy, precision and F1 score, Support Vector Machine in recall. Results for Taiwan dataset are shown in Table 11. For the

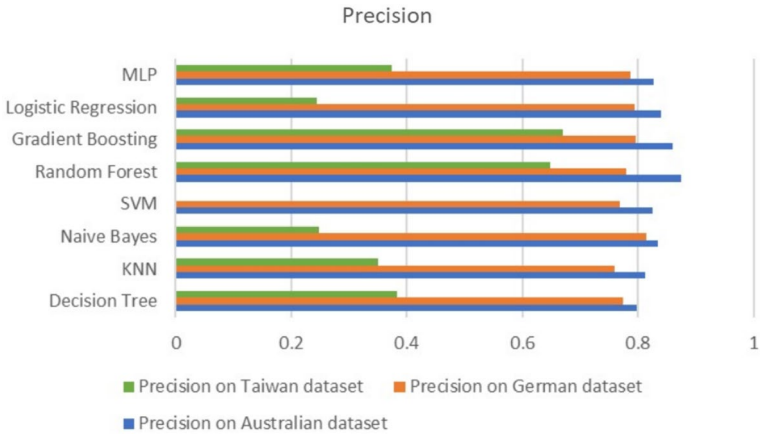


Fig. 5 Comparison of precision

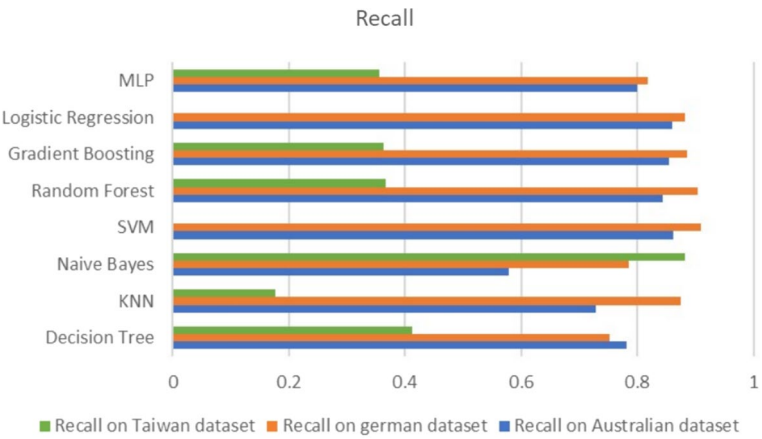


Fig. 6 Comparison of recall

German Credit dataset, the performance of all individual classifiers as tested in our experiments is displayed in Table 12. The following inferences can be made from observations of the models' performance: Gradient Boosting and Random Forest outperform the rest of the classifiers on accuracy and F1 score metrics. Logistic Regression and Random Forest perform well in terms of precision, recall, and AUC as well. In tandem with the Australian dataset, the Random Forest model outperforms others in the precision metric score, while the SVM model outperforms on the recall metric on the German Dataset.

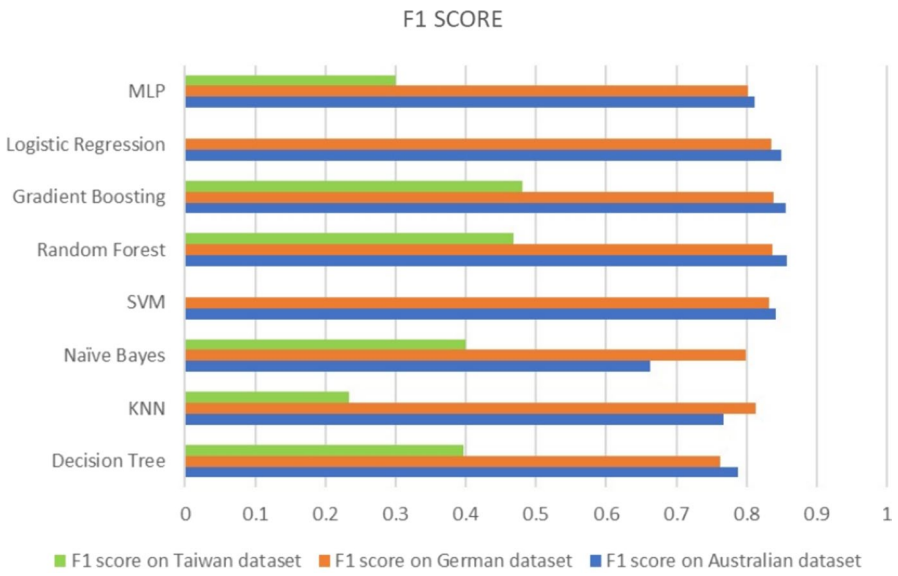


Fig. 7 Comparison of F1 Scores

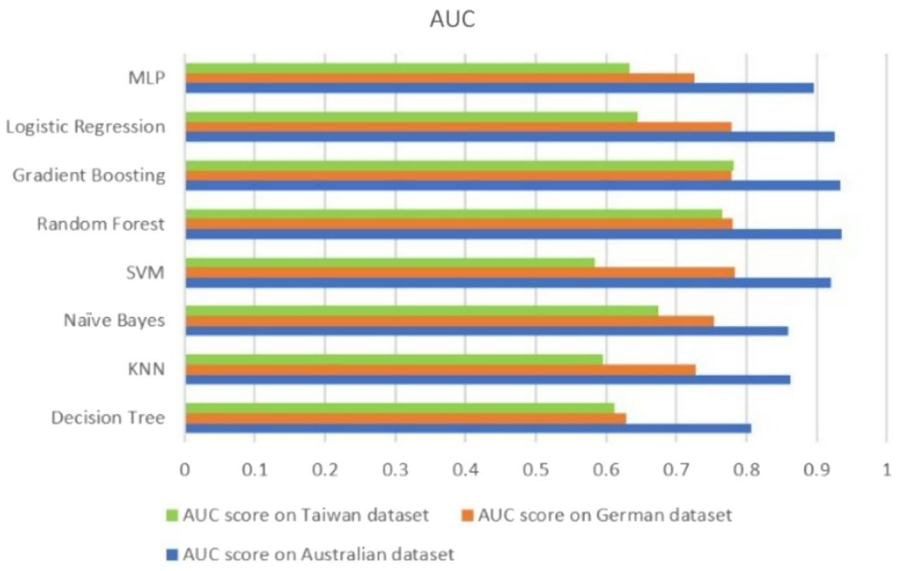


Fig. 8 Comparison of AUC

For Taiwan dataset, the Gradient Boosting model outperforms others in the precision metric score, while the Naive Bayes model outperforms on the recall metric of both Autralian and German dataset.

Table 10 Comparison of results of different models over Australian Credit Dataset

Algorithm	Accuracy	Precision	Recall	F1 score	AUC
Decision tree	0.805	0.798	0.781	0.788	0.807
KNN	0.794	0.813	0.729	0.767	0.862
Naive Bayes	0.739	0.834	0.579	0.663	0.860
SVM	0.849	0.825	0.861	0.842	0.921
Random forest	0.870	0.874	0.844	0.858	0.935
Gradient boosting	0.866	0.859	0.854	0.856	0.933
Logistic regression	0.858	0.840	0.860	0.850	0.926
MLP	0.828	0.827	0.799	0.812	0.896

Table 11 Comparison of results of different models over Taiwan credit dataset

Algorithm	Accuracy	Precision	Recall	F1 score	AUC
Decision tree	0.724646	0.383566	0.412779	0.397538	0.612760
KNN	0.746465	0.349367	0.175882	0.233901	0.596862
Naive Bayes	0.384040	0.247821	0.882347	0.3868822	0.674081
SVM	0.779838	0.000000	0.000000	0.000000	0.584741
Random forest	0.816545	0.647927	0.365800	0.467442	0.765633
Gradient boosting	0.820323	0.669441	0.363646	0.471151	0.781185
Logistic regression	0.779778	0.245000	0.000228	0.000456	0.645162
MLP	0.689152	0.374586	0.356418	0.299416	0.634567

Table 12 Comparison of results of different models over German credit dataset

Algorithm	Accuracy	Precision	Recall	F1 score	AUC
Decision tree	0.675	0.774	0.752	0.762	0.628
KNN	0.720	0.760	0.874	0.813	0.718
Naive Bayes	0.726	0.814	0.785	0.799	0.753
SVM	0.745	0.768	0.909	0.832	0.784
Random forest	0.755	0.780	0.904	0.837	0.780
Gradient boosting	0.762	0.795	0.886	0.838	0.779
Logistic regression	0.757	0.793	0.882	0.835	0.778
MLP	0.720	0.787	0.8178	0.802	0.725

5.2 Classification results of ensemble classifiers

Individual classifiers yield good results for all the Australian and German along with Taiwan dataset, but most researchers have used an ensemble of individual classifiers. Using a mixture of multiple base learners can assist reducing variance and bias, and hence increase the accuracy of predictions. In this research, we examine the performance of various proposed ensemble approaches, such as Tree-Based

Dynamic Ensemble (Xia et al. 2020), B-Stacking approach (Xia et al. 2018), Semi-Supervised Selective Ensemble (Xiao et al. 2020) and Enhanced Outlier adaptation (Zhang et al. 2021) to show that ensemble classifiers outperform solo classifiers on credit scoring datasets. Tables 10 and 12 demonstrate the results of some popular ensemble approaches on all Australian and German as well as Taiwan datasets. The data shows that all other ensemble classifiers outperform most individual classifiers on most parameters across both datasets. The proposed model approach was the best performer overall.

5.3 Performance of the proposed SCMVE approach

The results in Tables 10 and 12 show that the classification method described in this research performs significantly better than other singular or hybrid ensemble classifier on all the Australian and German as well as the Taiwan dataset. A soft voting metaheuristic approach is proposed by our model. It entails examining the performance of different models by training them on clusters and producing significantly enhanced results.

While the proposed model outshines all other classifiers across the evaluation metrics by a large margin for the Australian dataset (Fig. 9), the signifanct achievement of this paper and the proposed model is the subsequent improvement of performance in the German dataset (Fig. 10) and Taiwan dataset, where our model demonstrates an overall upgradation in results achieved through our cluster-based metaheuristic approach.

Further, on experimentation with the selection of the various fitness functions as referenced in Table 3, we observe a variation in the performance in accordance with the usage of each fitness function as shown in Figs. 11 and 12. This approach of finding the most suitable fitness function has been highlighted in Huang et al.

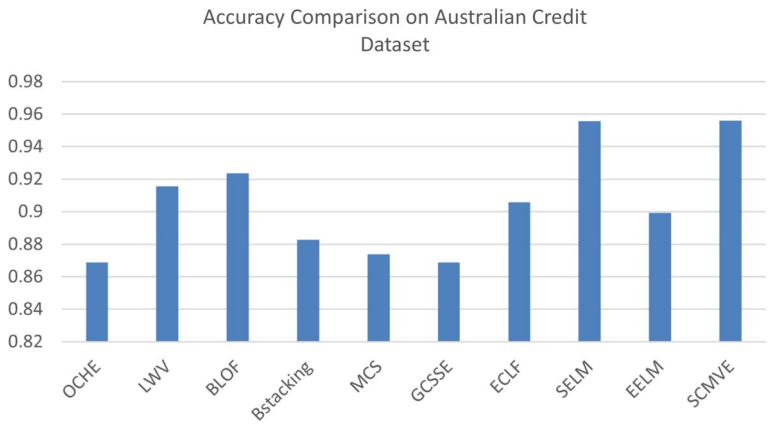


Fig. 9 Comparison of accuracy on Australian dataset

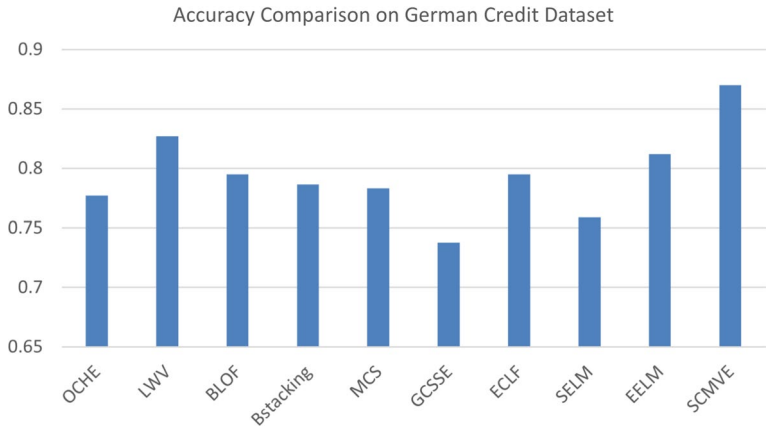


Fig. 10 Comparison of accuracy on German dataset

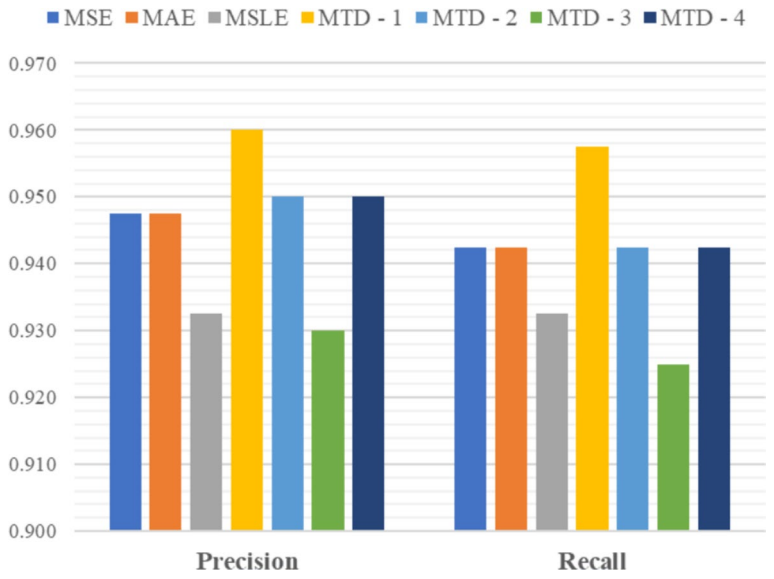


Fig. 11 Variation of precision and recall with the fitness function used on the Australian dataset

(2019) wherein the parameter optimization methods for metaheuristic algorithms have been studied. We observe that although the fitness function MTD-1 performs better in most aspects, it fails to provide a consistent AUC score. Hence, to ensure simplicity of the model and a consistent performance, the fitness function MAE was chosen.

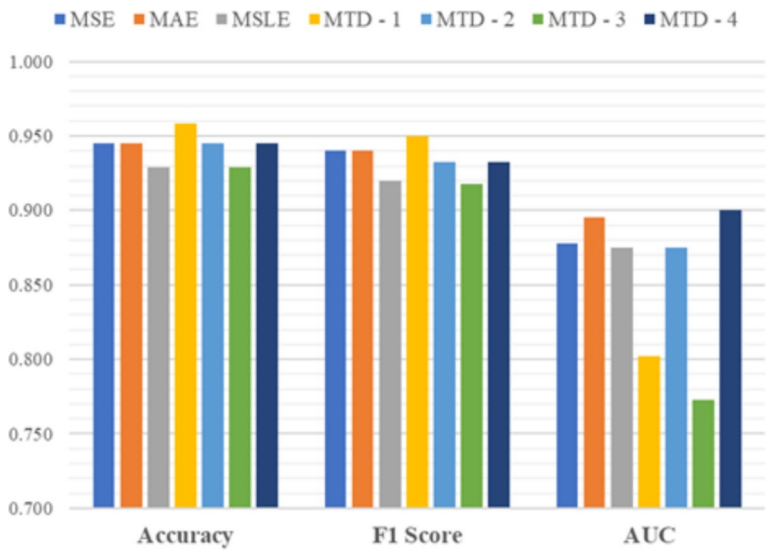


Fig. 12 Variation of accuracy, F1 Score, and AUC Score with the fitness function used on the Australian dataset

5.4 Comparison of SCMVE with other approaches of credit scoring

For credit scoring, a variety of single classifiers and ensemble-based approaches have been developed. On the German and Australian datasets, Table 13 compares a few credit rating techniques. Table 14 on the other hand compares few credit rating techniques over Taiwan dataset. SCMVE clearly outperforms other approaches, such as tree-based dynamic heterogeneous ensemble method, multi-stage ensemble model with enhanced outlier adaptation, credit scoring model based on ensemble feature selection and multilayer ensemble classification and Multi-Stage Self-Adaptive Classifier Ensemble Model.

Further, models like LWV (Tripathi et al. 2019) and BLOF (Zhang et al. 2021) which come close to our model performance on the Australian dataset, fail to show similar state-of-the-art performance in the German Credit Data Set.

Our usage of the recent sailfish optimizer helps us in reducing computational time. This works better in comparison to the works of Xia et al. (2020) where the tree-based architecture can lead to increased computational complexity requiring GPU usage, demands that were not required in our model.

Most existing works, such as that of Tripathi et al. (2019) embed feature selection into their model to increase their model efficiency. This can lead to some feature of data points not being used in the final output, even though they might be representative of the sample space. We overcome this issue through the usage of a Self Organising Map. Our model, instead of using feature selection, sections the data into distinct clusters with their own unique properties and feature representations. We then use our classification model on each cluster so that we utilize all the features

Table 13 Comparison of the performance achieved by various ensemble approaches of credit scoring over Australian and German dataset

Authors	Model description	Accuracy: Australian dataset	Accuracy: German dataset
Xia et al. (2018)	The model proposed differs from the existing ensemble credit models in three aspects-pool generation, selection of base learners, and trainable fuser	0.8828	0.7866
Tripathi et al. (2019)	This model conducts preprocessing and assigns weights to classifiers. Then feature selection using an ensemble model is applied to a classifier framework	0.9155	0.8268
Guo et al. (2019)	Introduces a novel multi-stage self-adaptive classifier ensemble model based on the statistical techniques and the machine learning techniques to improve the prediction performance	0.8740	0.7830
Xia et al. (2020)	This work develops a novel tree-based overfitting-cautious heterogeneous ensemble model (i.e., OCHE) for credit scoring.	0.8689	0.7772
Xiao et al. (2020)	Proposes a cost-sensitive semi-supervised selective ensemble model based on group method of data handling	0.8689	0.7376
Kuppili et al. (2020)	A novel spike-generating function is proposed in Leaky Nonlinear Integrate and Fire Model (LNIF). Its interspike period is computed and utilized in the extreme learning machine (ELM) for classification	0.9558	0.7589
Tripathi et al. (2020)	A novel parametrized algebraic activation function is proposed for extreme learning machine (ELM)	0.8992	0.8118

Table 13 (continued)

Authors	Model description	Accuracy: Australian dataset	Accuracy: German dataset
Liu et al. (2022)	This work develops a Multi-grained and multi-layered gradient boosting decision tree for credit scoring.	0.8826	0.7653
Zhang et al. (2021)	A local outlier factor algorithm is enhanced with bagging strategy to identify outliers and boost them back into the training set to construct an outlier-adapted training set	0.9236	0.7950
Zhang et al. (2021)	A voting-based outlier detection method is proposed to enhance the outlier detection algorithms with the weighted voting mechanism and boost the outlier scores into the training set to form an outlier-adapted training set	0.9058	0.7950
Runchi et al. (2023)	The Logistic-BWE model combines logistic regression with heterogeneous balancing and dynamic weighting. Sub-models are trained on datasets with different imbalance ratios, improving recognition of specific classes. Ensemble weights are dynamically adjusted based on each sub-model's prediction accuracy	0.865	0.757
Our proposed approach (SCMVE)	**SOM-based clustering supplementing a weighted voting ensemble classifier that uses a Sailfish Optimizer for weights generation**	**0.9559**	**0.8700**

Table 14 Comparison of the performance achieved by various ensemble approaches of credit scoring over Taiwan dataset

Authors	Approach	Accuracy: Taiwan dataset
Liu et al. (2022)	Multi-grained and multi-layered gradient boosting decision tree for credit scoring	69.22%
Zhang et al. (2021)	Local outlier factor algorithm with bagging strategy to construct an outlier-adapted training set	82.93%
Zhang et al. (2021)	Voting-based outlier detection method with the weighted voting mechanism to form an outlier-adapted training set	82.26%
Zhou et al. (2023)	This credit risk model for small firms uses a two-stage expert system with SMOTE for imbalanced data and random forest for feature extraction	75.7%
Our proposed approach (SCMVE)	SOM-based clustering supplementing a weighted voting ensemble classifier that uses a Sailfish Optimizer for weights generation.	82.80%

Bold values indicate better results than previous models

of each cluster to generate an output, thus not ignoring any feature that might be distinctly representative in a particular cluster. Credit scoring has outlier issues and most works have to engineer outlier detection algorithms to overcome the situation. Examples of this are present in the works of Zhang et al. (2021, 2021) Our usage of an ensemble classifier, combined with the clusters formed through the SOM helps us overcome outliers in the sample space since the result of multiple base classifiers perform better on outlier detection as opposed to singular classification models.

6 Conclusion and future work

Credit scoring is an important parameter in helping firms distinguish between customers with excellent and bad credit scores in order to reduce risk and improve profitability. Although numerous credit scoring techniques have been developed, our approach outperforms them and showed statistical improvement in terms of various performance measures on all Australian, German and Taiwan datasets. Our model SCMVE successfully utilizes self-organizing maps (SOM) and a soft voting-based ensemble classifier to create a hybrid model built for classification tasks of credit scoring. Multi-layer clustering of the dataset enhances the performance of our model by clustering similar data. The final classification is done using a weighted ensemble classifier. The weights are appropriately generated using the Sailfish Optimizer based on the predictive capabilities of the individual classifiers. The model gives an accuracy of 95.59% on the Australian dataset and 87% on the German credit dataset as well as 82.8% on the Taiwan credit dataset.

Future research can concentrate on improving the model's accuracy and efficiency through the selection of appropriate features. This can be achieved by incorporating data from various sources and evaluating the model's performance in real-life applications. It is important to prioritize the safety and fairness of the model to avoid discrimination and biases. By addressing these aspects, researchers can advance the field of machine learning and create more reliable and unbiased models.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability The datasets generated and/or analysed during the current study are available in https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval,https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data and https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients.

Declarations

Conflict 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.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Informed consent This article does not contain any studies with human participants or animals performed by any of the authors.

References

AghaeiRad A, Chen N, Ribeiro B (2017) Improve credit scoring using transfer of learned knowledge from self-organizing map. Neural Comput Appl 28(6):1329–1342

Bumacov V, Ashta A, Singh P (2014) The use of credit scoring in microfinance institutions and their outreach. Strateg Chang 23(7–8):401–413

Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: a systematic literature survey. Appl Soft Comput 91:106263

Dua D, Graff C (2017) UCI machine learning repository

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188

Gholamian M, Jahanpour S, Sadatrasoul S (2013) A new method for clustering in credit scoring problems. J Math Comput Sci 6:97–106

Guo S, He H, Huang X (2019) A multi-stage self-adaptive classifier ensemble model with application in credit scoring. IEEE Access 7:78549–78559

Hand DJ, Kelly MG (2002) Superscorecards. IMA J Manag Math 13(4):273–281

He H, Zhang W, Zhang S (2018) A novel ensemble method for credit scoring: adaption of different imbalance ratios. Expert Syst Appl 98:105–117

Henley WEM, Hand DJ (1996) Ak-nearest-neighbour classifier for assessing consumer credit risk. J Roy Stat Soc Ser D (Stat) 45(1):77–95

Hsieh N-C (2005) Hybrid mining approach in the design of credit scoring models. Expert Syst Appl 28(4):655–665

Hsieh N-C, Hung L-P (2010) A data-driven ensemble classifier for credit scoring analysis. Expert Syst Appl 37(1):534–545

Huang C, Li Y, Yao X (2019) A survey of automatic parameter tuning methods for metaheuristics. IEEE Trans Evol Comput 24(2):201–216

Kohonen T (1998) The self-organizing map. Neurocomputing 21(1–3):1–6

Kozodoi N, Lessmann S (2020) Multi-objective particle swarm optimization for feature selection in credit scoring. In: Workshop on mining data for financial applications. Springer, pp 68–76

Kuppili V, Tripathi D, Edla DR (2020) Credit score classification using spiking extreme learning machine. Comput Intell 36(2):402–426

Lappas PZ, Yannacopoulos AN (2021) A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Appl Soft Comput 107:107391

Lau KW, Hujun Y, Simon H (2006) Kernel self-organising maps for classification. Neurocomputing 69(16–18):2033–2040

Lee T-S, Chiu C-C, Lu C-J, Chen I-F (2002) Credit scoring using the hybrid neural discriminant technique. Expert Syst Appl 23(3):245–254

Li X, Ying W, Tuo J, Li B, Liu W (2004) Applications of classification trees to consumer credit scoring methods in commercial banks. In: 2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583), vol 5. IEEE, pp 4112–4117

Li S-T, Shiue W, Huang M-H (2006) The evaluation of consumer loans using support vector machines. Expert Syst Appl 30(4):772–782

Liu W, Fan H, Xia M (2022) Multi-grained and multi-layered gradient boosting decision tree for credit scoring

Martens D, De Backer M, Haesen R, Vanthienen J, Snoeck M, Baesens B (2007) Classification with ant colony optimization. IEEE Trans Evol Comput 11(5):651–665

Nalič J, Martinovič G, Žagar D (2020) New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers. Adv Eng Inform 45:101130

Nassef MGA, Hussein TM, Mokhiamar O (2021) An adaptive variational mode decomposition based on sailfish optimization algorithm and gini index for fault identification in rolling bearings. Measurement 173:108514

Onan A (2018a) Biomedical text categorization based on ensemble pruning and optimized topic modelling. Comput Math Methods Med 2018(1):2497471
Onan A (2018) An ensemble scheme based on language function analysis and feature engineering for text genre classification. J Inf Sci 44(1):28–47
Onan A (2019a) Consensus clustering-based undersampling approach to imbalanced learning. Sci Program 2019(1):5901087
Onan A (2019b) Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. IEEE Access 7:145614–145633
Onan A (2021a) Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. Comput Appl Eng Educ 29(3):572–589
Onan A (2021b) Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. Concurr Comput Pract Exp 33(23):e5909
Onan A (2022) Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. J King Saud Univ Comput Inf Sci 34(5):2098–2117
Onan A (2023a) Gtr-ga: harnessing the power of graph-based neural networks and genetic algorithms for text augmentation. Expert Syst Appl 232:120908
Onan A (2023b) Srl-aco: a text augmentation framework based on semantic role labeling and ant colony optimization. J King Saud Univ Comput Inf Sci 35(7):101611
Onan A, Korukoğlu S, Bulut H (2016a) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57:232–247
Onan A, Korukoğlu S, Bulut H (2016b) A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst Appl 62:1–16
Onan A, Korukoğlu S, Bulut H (2017) A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. Inf Process Manag 53(4):814–833
Pławiak P, Abdar M, Acharya RU (2019) Application of new deep genetic cascade ensemble of svm classifiers to predict the Australian credit scoring. Appl Soft Comput 84:105740
Reichert AK, Cho C-C, Wagner GM (1983) An examination of the conceptual issues involved in developing credit-scoring models. J Bus Econ Stat 1(2):101–114
Runchi Z, Liguo X, Qin W (2023) An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects. Expert Syst Appl 212:118732
Safi SA-D, Castillo PA, Faris H (2022) Cost-sensitive metaheuristic optimization-based neural network with ensemble learning for financial distress prediction. Appl Sci 12(14):6918
Şen D, Dönmez CÇ, Yıldırım UM (2020) A hybrid bi-level metaheuristic for credit scoring. Inf Syst Front 22(5):1009–1019
Shadravan S, Naji HR, Bardsiri VK (2019) The sailfish optimizer: a novel nature-inspired metaheuristic algorithm for solving constrained engineering optimization problems. Eng Appl Artif Intell 80:20–34
Simumba N, Okami S, Kodaka A, Kohtake N (2022) Multiple objective metaheuristics for feature selection based on stakeholder requirements in credit scoring. Decis Support Syst 155:113714
Singh P (2017) Comparative study of individual and ensemble methods of classification for credit scoring. In: 2017 International conference on inventive computing and informatics (ICICI). IEEE, pp 968–972
Suleiman S, Ibrahim A, Usman D, Isah BY, Usman HM (2021) Improving credit scoring classification performance using self organizing map-based machine learning techniques. Eur J Adv Eng Technol 8(10):28–35
Transpire Online sailfish optimizer (sfo) (2019) A novel method motivated from the behavior of sailfish for optimal solution
Tripathi D, Edla DR, Cheruku R, Kuppili V (2019) A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. Comput Intell 35(2):371–394
Tripathi D, Edla DR, Kuppili V, Bablani A (2020a) Evolutionary extreme learning machine with novel activation function for credit scoring. Eng Appl Artif Intell 96:103980
Tripathi D, Edla DR, Kuppili V, Dharavath R (2020b) Binary bat algorithm and rbfn based hybrid credit scoring model. Multimedia Tools Appl 79(43):31889–31912
Van Gestel IT, Baesens B, Garcia IJ, Van Dijcke P (2003) A support vector machine approach to credit scoring. In: Forum Financier-Revue Bancaire et Financiaire Bank en Financiewezen, pp 73–82
West D (2000) Neural network credit scoring models. Comput Oper Res 27(11–12):1131–1152

Xia Y, Liu C, Da B, Xie F (2018) A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Syst Appl 93:182–199
Xia Y, Zhao J, He L, Li Y, Niu M (2020) A novel tree-based dynamic heterogeneous ensemble method for credit scoring. Expert Syst Appl 159:113615
Xiao XJ, Zhong ZY, Xie L, Xin G, Liu D (2020) Cost-sensitive semi-supervised selective ensemble model for customer credit scoring. Knowl-Based Syst 189:105118
Zhang W, Yang D, Zhang S (2021a) A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring. Expert Syst Appl 174:114744
Zhang W, Yang D, Zhang S, Ablanedo-Rosas JH, Xin W, Lou Yu (2021b) A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. Expert Syst Appl 165:113872
Zhou Y, Shen L, Ballester L (2023) A two-stage credit scoring model based on random forest: evidence from Chinese small firms. Int Rev Financ Anal 89:102755

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

Indu Singh[1] · D. P. Kothari[2] · S. Aditya[1] · Mihir Rajora[1] · Charu Agarwal[3] · Vibhor Gautam[4]

✉ Indu Singh
indusingh@dtu.ac.in

D. P. Kothari
dpkvits@gmail.com

S. Aditya
s.aditya.me@gmail.com

Mihir Rajora
mihee20@gmail.com

Charu Agarwal
acharu848@gmail.com

Vibhor Gautam
vibhorgautam907@gmail.com

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

[2] Department of Electrical Engineering, Visvesvaraya National Institute of Technology, Nagpur, Maharashtra 440010, India

[3] Department of Electronics and Communication Engineering, National Institute of Technology, Hamirpur, Himachal Pradesh 177005, India

[4] Department of Electronics and communication Engineering, Delhi Technological University, Delhi 110042, India

A Noise and Edge extraction-based dual-branch method for Shallowfake and Deepfake Localization

Deepak Dagar[1], Dinesh Kumar Vishwakarma[2,*]
Biometric Research Lab, Department of Information Technology(IT), Delhi Technological University(DTU), Bawana Road, Delhi-110042, India
deepakdagargate@gmail.com[1], dvishwakarma@gmail.com[2,*]

Abstract

The trustworthiness of multimedia is being increasingly evaluated by advanced Image Manipulation Localization (IML) techniques, resulting in the emergence of the IML field. An effective manipulation model necessitates the extraction of non-semantic differential features between manipulated and legitimate sections to utilize artifacts. This requires direct comparisons between the two regions.. Current models employ either feature approaches based on handcrafted features, convolutional neural networks (CNNs), or a hybrid approach that combines both. Handcrafted feature approaches presuppose tampering in advance, hence restricting their effectiveness in handling various tampering procedures, but CNNs capture semantic information, which is insufficient for addressing manipulation artifacts. In order to address these constraints, we have developed a dual-branch model that integrates manually designed feature noise with conventional CNN features. This model employs a dual-branch strategy, where one branch integrates noise characteristics and the other branch integrates RGB features using the hierarchical ConvNext Module. In addition, the model utilizes edge supervision loss to acquire boundary manipulation information, resulting in accurate localization at the edges. Furthermore, this architecture utilizes a feature augmentation module to optimize and refine the presentation of attributes. The shallowfakes dataset (CASIA, COVERAGE, COLUMBIA, NIST16) and deepfake dataset Faceforensics++ (FF++) underwent thorough testing to demonstrate their outstanding ability to extract features and their superior performance compared to other baseline models. The AUC score achieved an astounding 99%. The model is superior in comparison and easily outperforms the existing state-of-the-art (SoTA) models.

Keywords: Edge supervision, Noise Inconsistencies, Manipulation localization, Image forensics, deepfake localization.

1 Introduction

Advancements in computer graphics and deep learning have provided individuals with enhanced capabilities to generate deceptive visuals. Thanks to the progress of advanced editing AI tools, it is now feasible to effortlessly modify multimedia data to create extremely lifelike content. Manipulated media can cause significant harm, including concrete damage, psychological distress, and physical suffering. A deceitful individual may employ the technology to illicitly obtain something valuable with the intention of causing harm [1]. Image-editing tools or graphics-based procedures that follow traditional approaches are commonly referred to as "**shallow fakes**" [2]**.** Shallowfakes have been somewhat altered, but not to the extent of or sophistication of "**deepfakes**" Deepfakes refers to approaches based on deep learning. Deepfake has garnered significant interest due to its ability to generate exceptionally realistic and convincing content with remarkable ease. As advanced technologies progress, there is a need for more sophisticated image modification localization approaches to address existing modified images and mitigate security risks. The objective of the localization challenge is to detect and delineate the changed areas in an image with pixel-level precision. Shallow fakes involve three distinct forms of image manipulation. Shallowfakes can be divided into three categories [2](Figure I):

1) Splicing: Copying a portion of an image and transferring it into another.
2) Copy-move: Replicating a specific portion of an image.
3) Removal: removing portions of images or an object is removed from the image.

Furthermore, Deepfake can be categorized into three distinct categories [1]:

1) Face Swap: The act of transferring a person's face from one image or frame to another.
2) Face Reenactment: It involves transferring facial expressions or movements from a source to a target.
3) Entire Image Synthesis: The complete image is generated using advanced AI technologies.

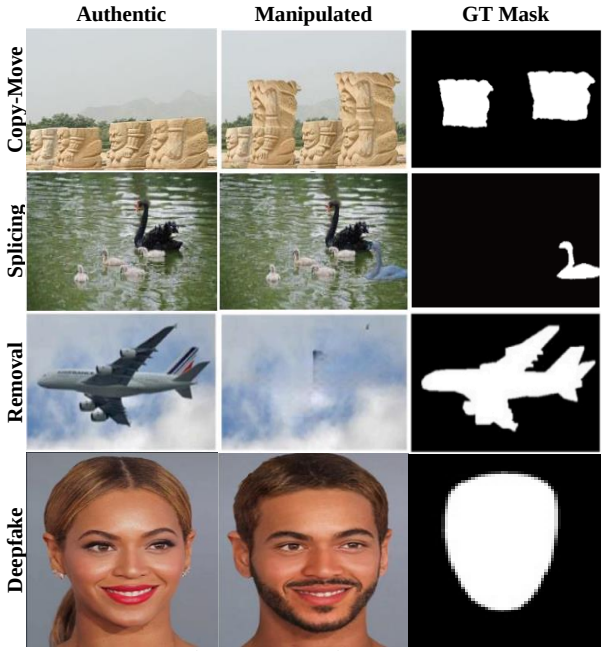


Figure I: Illustration of various types of image manipulation

Handcrafted feature-based approaches often operate under the assumption that a specific manipulation

artifact is present. Their objective is to expose the counterfeits by examining regional inconsistencies in the color filter array (CFA) [3] [4], illumination [5], photo response non-uniformity noise(PRNU) [6], compression artifacts in JPEG images [7], texture units [8]. Nevertheless, these solutions presuppose prior knowledge of the tampering, hence restricting their effectiveness in many situations. Deep learning methods, specifically CNNs have a tendency to acquire contextual (semantic) information and employ advanced strategies for detecting image counterfeiting [9] [10] [11] resort to them [12]. Convolutional neural networks (CNNs) are based on local information and focus on the relationship between surrounding regions. However, they may not capture the global interactions between different parts of the image. Artifacts are discrepancies in imperceptible low-level characteristics, such as noise or high-frequency, that are not discernible to the unaided eye but are obvious when examined closely [13]. Therefore, according to the previous study, the crucial aspect of the manipulation task is to identify minor discrepancies and visible evidence that are unrelated to meaning to detect artifacts. Handcrafted features are limited since they presuppose the sorts of tampering before and may not detect other types of tampering. On the other hand, Convolutional Neural Networks (CNNs) learn semantic features and analyze links between nearby artifacts, but this alone is insufficient for learning manipulation artifacts.

To overcome these constraints, we have created a dual-branch model that leverages handcrafted features, such as RGB information on one branch and noise inconsistencies on another using the aid of edge supervision and the hierarchical structure of ConvNext. The contribution of the paper is summarized below:

- Developed an innovative dual-branch structure that utilizes RGB data in one branch and noise characteristics in another using the hierarchical structure of ConvNext, with the features being improved via a feature enhancement module. The model also utilizes edge supervision to improve the localization of manipulation at the edges.
- Experimentation is performed on the shallowfakes dataset, which comprises NIST16, Coverage, Columbia and CASIA datasets, along with the deepfake dataset Faceforensics++. The results indicate that the FF++ model outperforms other models in terms of its discriminative capacity.
- Qualitative analysis is conducted to visually examine the localization result of the model and compare them with the visualization results of the state-of-the-art (SOTA) models.
- Ablation studies are performed to investigate the significance of various components within the overall model.

2 Related Work

Image Localization Methods(IML) approaches can be classified into two primary categories according to feature representation: handmade features and methods based on deep learning.

. This section will explore these two categories.

2.1 Handcrafted features

Handcrafted feature-based solutions traditionally model authentic images to disclose the statistical relationship between pixels and capture the statistical variation caused by image manipulation procedures. For example, the spatial rich model (SRM) [14], the technique, commonly employed in image steganalysis, was extended in [15] to obtain residual-based features for the multidimensional Gaussian model and SVM classifier to detect and locate image fraud. Nevertheless, modifying an image will invariably lead to modifications in its visual elements, which might be utilized by local image describers to identify counterfeit versions. To identify the alterations caused by image splicing, [5] combined the statistical data obtained from various local descriptors that analyze texture, illumination, shape, and color characteristics. Image alteration impacts the non-semantic characteristics and leads to low-level features such as noise inconsistencies. Researchers have further explored these features to obtain more effective discriminative features. Zhang et al. [16] utilize an enhanced version of the constrained convolutional model to extract noise features that act as more nuanced indicators of manipulation. These features are then inputted into a dual-branch architecture for additional feature learning. Nevertheless, these noise characteristics exhibit inconsistency and resilience when subjected to JPEG compression and Gaussian blur.

Li et al. [17] use a Markov-based method in the Quaternion Discrete Cosine Transform (QDCT) domain. This method expands on the Markov transition probability characteristics found in QDCT frequency domains to reveal the interconnection between adjacent pixels. Nevertheless, the technique is restricted to color photographs, and its effectiveness is uncertain due to different instances of compression blurring. Handcrafted features assume that hidden manipulation characteristics can be observed in certain manipulation artifacts, and a technique specifically intended to capture these artifacts will yield superior outcomes. Nevertheless, these methods have limited ability to be applied to a wide range of situations, and they lack the necessary strength to handle many types of manipulation operations.

2.2 Deep learning-based method

Deep learning algorithms can autonomously train and optimize feature representations for purpose of forgery forensics. This is different from conventional techniques, which rely on a tedious feature engineering process to manually construct features. Wu et al. [11] suggest a method called Mantra-Net is designed for localizing and detecting generalized image forgeries

(IFLD). The system is capable of detecting specific irregularities in the image that can indicate the presence of manipulated pixels. It offers a comprehensive approach to detecting different types of forgeries. They are eliminating the need for pre- and/or post-processing. Nevertheless, the model is limited in its ability to handle multi-forgeries and images with strongly correlated noises. Cozzolino et al. [18] introduced an improved initialization method and employed a Siamese network to perform splicing localization and detection. The utilization of Siamese networks in Noiseprint allowed for the detection and identification of camera model artifacts through the analysis of noise residual, aiding in the localization of fraudulent activity [18].

An important obstacle in the field of image manipulation detection is the identification of distinct and adaptable characteristics that can distinguish between real images and manipulated ones, without generating false alarms on genuine images, while yet being able to detect manipulations in new and unfamiliar data. Chen et al. [19] in their study, the authors tackle both aspects by employing multi-scale supervision and multi-view feature learning. The former method aims to obtain characteristics that are not dependent on specific meanings and, thus, may be applied more broadly. This is achieved by analyzing the boundary artifacts and noise distribution surrounding tampered regions. The latter allows us to acquire knowledge from genuine images that are excessively intricate for the current semantic segmentation network-based methods to manage. The model exhibits superior performance in cross-dataset scenarios and demonstrates resilience to diverse post-processing techniques. A multitask fully convolutional network (MFCN) was built to localize forgery. The training data consisted of ground truth information about the fabricated regions and boundaries [9]. In a recent study [10] researchers constructed a hybrid LSTM and encoder-decoder network to localize pixel-wise forgeries. This technique uses spatial attributes and resampling to accurately capture irregular shifts between counterfeit and genuine patches. CNN approaches excel in comprehending visual context and acquiring semantic characteristics, resulting in superior performance in image classification tasks. Image manipulation is independent of semantic meaning; hence these strategies are generally less successful against image manipulation. CNNs possess inductive and location biases, which prevent them from capturing the global correlation of features.

Objectformer [20] and Transforensics [21] are approaches that utilize transformers and have been proposed in the past. The general architecture and design philosophy of these two models do not mesh well with ViT. Unlike ViT, which directly embeds patched images for encoding, both systems in question utilize many CNN layers to extract feature maps before employing Transformers for further encoding. However, this approach overlooks crucial initial low-level information. In order to address the shortcomings of the current models, we have developed a dual-branch model that incorporates handcrafted features, noise, and features derived from convolutional neural networks (CNNs). The dual-branch architecture comprises one branch that captures RGB information and another branch that extracts noise information using the Bayar convolution and SRM filters. The hierarchical structure of ConvNext passes these features and they are further improved by the feature enhancement module. The model also employs edge supervision, enabling it to concentrate on the border information, where tampering typically occurs. Finally, the features from both branches are added to the mask's final prediction.

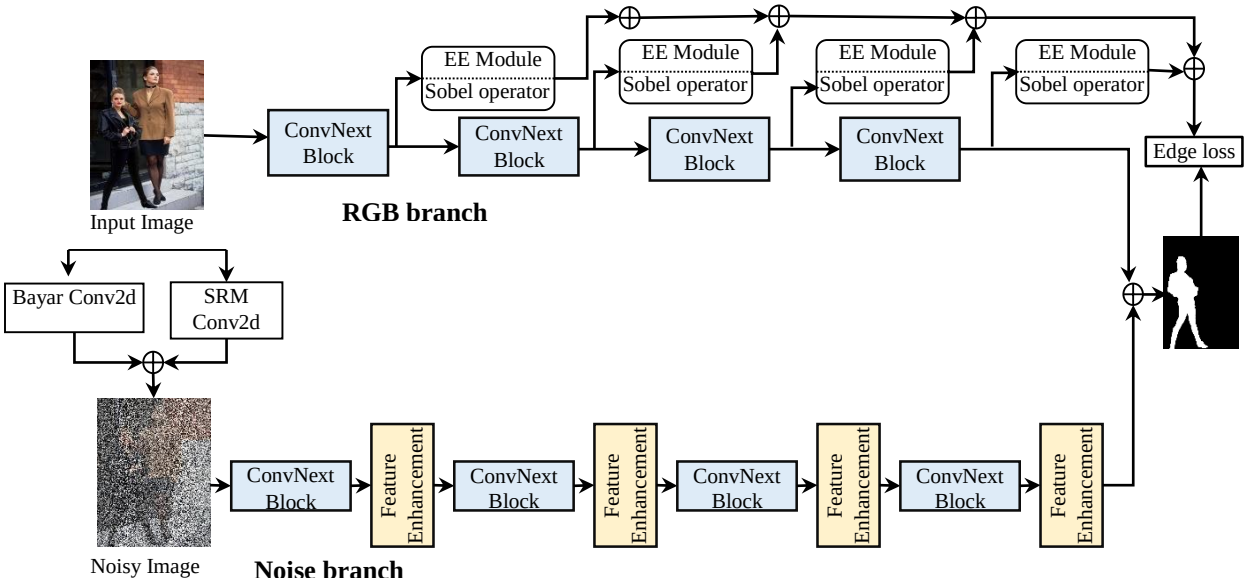


Figure II: Overview of the proposed model consisting of a dual branch consisting of RGB and noise branch followed by ED architecture

3 Proposed Model

The proposed model comprises two parallel branches: one branch takes noise/residual characteristics as input, which are determined using the Bayar convolution and SRM convolution filters. These low-level/non-semantic features are further correlated by the ConvNext module and the Feature Enhancement(FE) module further enhances the feature representation by regulating global dependencies along various axes, while the other branch is the contextual branch which uses the spatial characteristics of the image samples to determine manipulation. The Sobel operator and Edge Extraction(EE) module retrieve manipulation edge predictions from the features generated by each ConvNext layer along the contextual branch. Subsequently, the edge extracted features are kept concatenated and used as a supervision for the loss functions(Figure II). The contextual branch is designed to preserve the fine details of the feature and aims to capture more comprehensive correlations among the data. The features obtained from both branches are combined to enhance the precise handling of localized features.

Noise branches and Contextual branches with their components are discussed in great detail in this section.

3.1 Noise Inconsistencies

A genuine image has a consistent distribution of noise over the entire image. Different acts of tampering introduce inconsistency and weaken the uniformity in the tampered area. The rationale behind using noise residual as a feature is that the noise characteristics between the source and target images are improbable to be similar when an object is extracted from one image (the source) and inserted into another image (the target). The noise residual in this configuration represents the difference between the true value of a pixel and the estimated value of that pixel, which is calculated by interpolating only the values of adjacent pixels. This estimation acts as the noise model. Various types of noise residual filters or kernels exist, each with its sensitivity to different forms of manipulation. The Bayer and SRM filters are widely used for effectively capturing the residual features of low-level noise.

3.1.1 Bayar Convolution or Constrained CNN

Constrained Convolution Neural Networks can utilize data to learn how image editing operations affect local pixel correlations. Therefore, this method can eliminate material at the image level and dynamically acquire knowledge about the indications of image alteration [22]. Constrained CNN is specifically developed to learn filters that predicts errors, which in turn generate feature maps used as low-level forensic traces. These traces offer improved universality and robustness. Subsequent layers in the neural network would gradually acquire knowledge of the more complex visual features depicted in the low-level traces [23]. In order to constrain the CNN to acquire knowledge of the low-level patterns, specific limitations are imposed on the weights of the CNN's kernel.:

$$\begin{cases} \omega_k^{(l)}(0,0) = -1, \\ \sum_{m,n \neq 0} \omega_k^{(l)}(m,n) = 1, \end{cases} \quad (1)$$

Equation (1) denotes the constraints imposed on the kernel's filter. The superscript indicates the CNN layer. The kth convolutional filter in a layer is represented by the subscript k, and the spatial co-ordinates(0,0) corresponds to the central value of the filter. During training, in the backward pass, the weights are updated using the optimizer. Then, the central value of the kernel is set to zero, and the remaining weights of the kernel are normalized so that their sum is equal to one. Finally, the central value of the filters is updated to minus one.

3.1.2 Steganalysis Features

SRM (Spatial Rich Models) filters are another technique employed to extract features from the noise residuals of an image. Fridrich et al. [14] initially established the concept of Spatial Rich Models (SRM). Its main purpose is to do steganalysis, which involves extracting concealed or hidden characteristics from the noisy residuals of an image using a predetermined set of high-pass filters. Afterwards, the aforementioned features are merged and transmitted to ensemble classifiers. This method is specifically developed to calculate the necessary statistics for extracting specified properties from the noise residuals surrounding the pixel neighborhood in an image. If we view the data embedding process in steganography as a specific form of image tampering, then image forensics and steganalysis can be seen as the same thing. Their purpose is to differentiate between tampered images and natural images. Modifications made to the specific attributes of an image will affect the corresponding residuals, as the residuals are closely linked to those attributes. SRM features are obtained by first capturing the essential noise characteristics with 30 basic filters. Then, nonlinear processes, such as selecting the maximum and lowest values from the neighboring outputs after filtering, are applied. The filters produce a quantified output that is then shortened by SRM. SRM then extracts the nearest co-occurrence data as the final features. This method produces a characteristic that can be regarded as a description of local noise.

3.2 Feature Enhancement Module

An improvement module is utilized to enhance the feature representation capability of ConvNext block and is applied to the feature maps. This deployment aims to utilize the high-frequency features that are retrieved from the appropriate layers. The features obtained from CNN layers are often noisy, leading to a potential degradation in performance. The deep layers exhibit a reduced amount of high-frequency information in their characteristics [24].

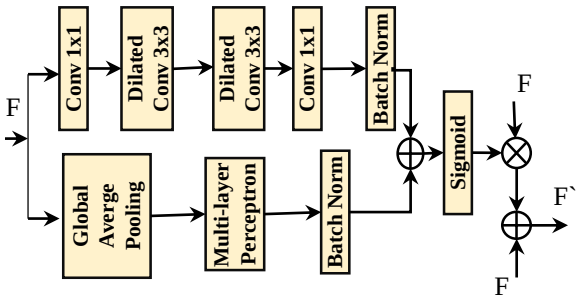


Figure III: Architecture of Feature Enhancement Module

The architecture of the feature enhancement module is shown in Figure III. With a feature map of $\mathcal{F} \in \mathbb{R}^{\mathcal{H}\times\mathcal{W}\times\mathcal{C}}$ outputted by the ConvNext block, the dilated convolution is used to enlarge the receptive field and $conv_{3\times3}^{dilated}$ and the $conv_{1\times1}$ are used to adjust the channel dimension. The attention map generated from the above branch is as follows:

$$M_1(F) = BN(conv_{1\times1}(conv_{3\times3}^{dilated}(conv_{1\times1}(F)))) \quad (2)$$

where batch normalization is indicated by BN. Moreover, the channel attention is formed by utilizing the inter-channel link. The feature map in each channel is aggregated by applying global average pooling (GAP) to the channel feature map is created which is followed by the MLP layer and batch norm layer.

$$M_2(F) = BN(MLP(GAP(F))) \quad (3)$$

Next, we merge the matrices $M_1(F)$ and $M_2(F)$ by adding their corresponding elements together and then apply a sigmoid function to generate the ultimate attention map.

$$M(F) = Sig(M_1(F) \oplus M_2(F)) \quad (4)$$

In the end, the features are improved by performing element-wise multiplication with the attention feature map and subsequently adding the result to the existing feature map.

$$F` = F \oplus F \otimes M(F) \quad (5)$$

3.3 Edge Extraction Block

Edge detection is a method employed to detect the areas in an image where there is a significant and sudden shift in brightness. The abrupt variation in the intensity value is detected at the points of lowest or highest values in the image histogram, by utilizing the first-order derivative. This change of gradient allows us to effectively locate the manipulation at edges. This consists of two operations or modules where first extraction information is fetched using the sobel operator which is further enhanced by the edge extraction module

Sobel operator: The edge can be determined by calculating the differentiation of pixel intensities. The Sobel mask calculates the first-order derivative and the edge is depicted by the local maximum or local minimum. The coefficients of the masks in the Sobel operator can be adjusted to meet our specific requirements, as long as they adhere to all the characteristics of derivative masks.

$$Sobel_{X-axis} = [-1\ 0\ 1; -2\ 0\ 2; -1\ 0\ 1] \quad (6)$$

$$Sobel_{y-axis} = [-1\ -2\ -1; 0\ 0\ 0; 1\ 2\ 1] \quad (7)$$

Edge extraction block: The task of detecting image manipulation involves identifying extremely faint indications of alteration inside the image. Furthermore, the subtle distinctions between the manipulation edge and the surrounding non-manipulation region are of great significance. To more effectively capture this nuanced information, it is necessary to preserve the features in the Convolutional Neural Network (CNN) at a rather high resolution.

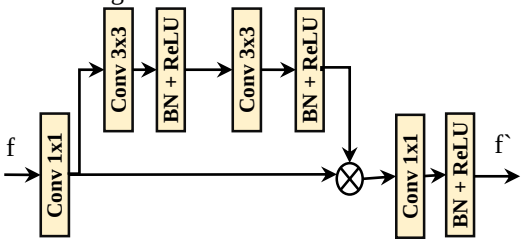


Figure IV: Architecture of Edge Extraction Block

As each layer in the Convolutional Neural Network (CNN) learns distinct feature contents, extract the manipulation edge by using the features produced by each layer of the network. To enhance the extraction of edges, we used the implementation of a specifically engineered edge extraction block (EEB). Figure IV illustrates the progression of EEB. To optimize computation and maximize the utilization of feature information, we employ a 1 × 1 convolution operation to decrease the number of channels in the features by a factor of 1/4. Subsequently, we proceed with constructing residual learning. Ultimately, a 1×1 convolution is employed to decrease the number of channels to 1.

$$M(f) = conv_{1\times1}(conv_{3\times3}(BN\left(ReLU\left(conv_{3\times3}\left(BN(ReLU(f))\right)\right)\right)) \quad (8)$$

The above attention gets multiplied to further improve the edge feature representation. These features contain a rich presentation of the edge features.

$$f` = conv_1 \otimes M(f) \quad (9)$$

3.4 ConvNext Module

The paper's authors present ConvNext, a new convolutional network design that replaces conventional "Conv2d" layers with modern ones influenced by recent breakthroughs in Vision Transformers (ViTs)[1]. The authors suggest several novel design principles to enhance the basic Convolutional Networks. These include employing larger kernel sizes, utilizing GELU activation instead of ReLU, augmenting the depth-wise convolutions, and substituting batch normalization with layer normalization. Using larger kernel sizes allows for a better understanding of a wider spatial context, which

leads to more accurate detection of subtle artifacts. Utilizing deeper convolutions enhances the knowledge of spatial context without significantly increasing computational cost. Additionally, incorporating layer normalization promotes more stable training, which is essential for detecting fine manipulations. The enhanced hierarchical architecture and activation function enable the acquisition of intricate features, encompassing both high-level contextual features and low-level intricate details that are highly responsive to subtle variations. Incorporating these design elements, the ConvNext architecture enables enhanced robust performance and highly comprehensive feature extraction capabilities that are essential to manipulation detection tasks.

3.5 Loss function

This section will provide an analysis of the loss function employed in our localization model. Three renowned loss functions for classification and segmentation tasks are Edge Loss(EL), Binary Cross Entropy (BCE) Loss, and Focal Loss (FL) used in our model.

Edge Loss: To address the observation that artifacts are generally more common near the periphery of tampered regions i.e. edges, where the distinctions between manipulated and real areas are most apparent, we have devised an approach that prioritizes the boundary region of the manipulated area. As non-edge pixels dominate the pixels of an edge, we employ the Dice loss[2] for manipulating edge detection, referred to as $\text{loss}_{\text{edge}}$. The model's edge information at various semantic levels is continuously concatenated, and the loss is computed by manipulating both the edge information and the manipulation mask information. By integrating edge information, the model can prioritize the segmentation border of the object and enhance the overall accuracy of localization.

The BCE loss, derived from the Bernoulli distribution, aims to quantify the disparity between the probability distributions of the predicted and actual masks [25]. Segmentation is commonly employed for classification purposes, as it involves classifying pixels at a granular level. It is characterized or described as:

$$\mathcal{L}_{BCE}(G_m, \hat{P}_m) = -(G_m \log(\hat{P}_m) + (1 - G_m)\log(1 - \hat{P}_m)) \quad (10)$$

Where G_m refers to the ground truth pixel value of the mask and $\hat{P}_m$ is the predicted pixel value of the mask. The method calculates the loss for each pixel and assigns equal importance to every pixel.

Focal Loss is a commonly observed phenomenon in segmentation tasks and is used as an addition to Binary Cross Entropy (BCE) in detection tasks that have a severe imbalance in class distribution [26].

The Focal Loss method effectively guides the network to focus on difficult samples without the need for any weight adjustment. This enables the model to acquire knowledge about more intricate scenarios while reducing the significance of less intricate ones. The focal Loss method employs a modulating factor, denoted as p, to reduce the importance of easy examples and instead focus the training on challenging negative examples.

$$\mathcal{FL}_{\gamma,\alpha} = -(\alpha(1 - p_i)^{\gamma} y_i log p_i + (1 - \alpha) p_i^{\gamma} (1 - y_i) \log(1 - p_i)) \quad (11)$$

Where p_i represents the estimated probability distribution for the class label, which is equal to one. An equilibrium factor is denoted as α, and the rate at which simple data samples are down-weighted is defined by the parameter γ.

The combined loss function is a potent amalgamation of Binary Cross Entropy (BCE), Edge Loss and Focal Loss. The BCE loss function penalizes deviations from a normal distribution in the data samples. On the other hand, the Focal Loss functions specifically tackle the issue of class imbalance and prioritize the most challenging aspects of the task by reducing the impact of easier ones through their hyper parameter settings. The idea of edge loss specifically targets the border regions of manipulation, which plays a critical role in achieving accurate localization outcomes. This combination helps in overall accuracy precision while maintaining the boundary precision.

$$\mathcal{L}_{combined} = \mathcal{L}_{BCE} + \mathcal{FL}_{\gamma,\alpha} + loss_{edge} \quad (12)$$

Equation (12) denotes the comprehensive loss function employed in this model. The model utilizes this integrated loss function during the comprehensive training of the model.

4 Experiments

The purpose of the current section is to examine the effectiveness of the proposed methodology by validating it on a number of benchmark datasets and comparing it to a number of different cutting-edge manipulation localization techniques.

4.1 Datasets

4.1.1 Shallowfake dataset

The following datasets were utilized for the purpose of using the proposed model for training and validation:

- **CASIA** [27]**.** CASIA v1.0 comprises a total of 920 counterfeit images, which have been predominantly manipulated by splicing and copy-move techniques. CASIA v2.0 is an enhanced iteration of CASIA v1.0. However, the latter comprises 5063 counterfeit images, many of which feature more complex modifications suitable for network training. CASIA includes both CASIA v1.0 and CASIA v2.0. The CASIA v2.0 dataset was utilized for training, while the CASIA v1.0 dataset was employed for testing.
- **NIST16** [28]**.** NIST16 includes 564 images and encompasses three manipulation techniques: copy-move, splicing, and removal. NIST16 is a dataset that presents a challenge due to the post-processing

that has been implemented to obscure any indications of potential image manipulation.

- **COLUMBIA** [29]**.** The Columbia dataset employs splicing alteration, which involves a total of 180 photos. Spliced images are created by directly copying and pasting visually hidden elements from Adobe Photoshop onto the original photographs, without any additional editing or modifications.
- **COVERAGE** [30]**.** A 100 images manipulated smaller dataset designed by copy-move operation and their corresponding Ground truth masks are included. To get rid of visible traces, every image is post-processed to hide visual traces.

Table I provides information about the division of the dataset into training and testing sets, along with the specific methods used for manipulation. Deep learning network training is characterized by a significant appetite for data. The current datasets commonly employed for training deep neural networks in image alteration detection lack an adequate number of images. In addition, modified images from a typical dataset may not provide enough training material as they have fewer imperfections. The model is initially trained using CASIAv2 and then further refined by fine-tuning with additional datasets. Subsequently, testing is performed on the datasets that were previously mentioned.

Table I: The benchmark datasets are partitioned into separate sets for training and testing purposes. into training and testing sets.

Datasets	Training Set	Validation Set	Testing test	Total Samples
CASIA V2.0	5063	-	-	5063
CASIA V1.0	-	600	320	920
NIST16	414	75	75	564
COLUMBIA	130	25	25	180
COVERAGE	60	20	20	100

4.1.2 Deepfake Dataset

Currently, there is no deepfake image dataset that includes an accurate mask indicating the areas that have been modified. Zhang et al. [31] have constructed their dataset using Faceforensics++ [32], At now, there is no existing deepfake image dataset that includes an accurate mask indicating the areas that have been modified. Zhang et al. [40] have constructed their dataset using Faceforensics++ [41], the sole deepfake dataset that includes masks for the majority of its films. The Famous Faceforensics++ dataset comprises 1000 authentic videos and 5000 altered videos created using various techniques such as Deepfakes, Face2Face, Face-Swap, and Neural-Textures. Four operations are being evaluated for the frames recovered from the 1000 face shifter recordings, as these videos do not have any ground truth mask. Two frames are retrieved from each video, but certain accessibility problems have prevented us from downloading certain real and fake videos. We have extracted a total of 8,449 real frames and 7,330 fake frames[3].

4.2 Experimental setup

To construct our model, we have utilized the PyTorch framework. The model is executed on a pair of NVIDIA RTX A5000 graphics processing units (GPUs), and the image is resized to dimensions of 256 × 256. The model is optimized using the Adam optimizer with a batch size of 16 during both the training and testing phases. The initial learning rate is set to 1e-4 and it decreases by a factor of 0.8 every 10 steps. The model undergoes pre-training for 150 epochs and subsequently undergoes fine-tuning for an additional 50 epochs.

4.3 Evaluation Metric

The F1 score at the pixel level and the area under the receiver operating characteristic curve (AUC) are used as assessment metrics for comparison methods to quantify the localization performance. The more significant value indicates improved performance. The range of both the F1 score and the pixel-level AUC is [0,1]. These two metrics are commonly used for evaluation and comparison. For ablation studies and deepfake dataset comparison, we have used another metric named IoU (Intersection over Union).

4.4 Quantitative Analysis

In this section, we have conducted a quantitative analysis of the performance of the shallow fake dataset as well as of the deepfake dataset.

4.4.1 Shallowfake dataset

Based on the method described in reference [33], the model was trained using the CASIA2 dataset and further refined utilizing well-known shallowfakes datasets such as Nist16, Coverage, Columbia, and CASIA1. The Area Under the Curve (AUC) and F1 scores of these datasets have been documented in Table II. To provide a comparison, we will investigate two groups of models: unsupervised models and DNN models. The data presented in the table demonstrates that the model surpasses the unsupervised models by a substantial margin. The effectiveness of manually designed features tailored for a particular type of manipulation is greatly limited, and all of these conventional techniques only extract certain signs of tampering with little data for detection. The performance of our method is superior to previous DNN-based methods across several datasets, making them directly comparable. Our model demonstrates exceptional performance on the CASIA and Columbia datasets, with an impressive AUC score of 97.25%. Additionally, on the Nist16 dataset, our model's score is comparable to other models, with an AUC score of 99.79%. The incorporation of edge supervision of the model allows for handling complex cases of thin structures or small objects where edge information can be vital and leads to the overall improvement of results.

By contrast, the model exhibits subpar performance on the COVERAGE dataset, achieving an AUC score of

88.26% due to the presence of fewer images and the inclusion of duplicated or relocated objects that have a similar appearance. Furthermore, the dataset encompasses a distinct range of manipulation scenarios, including variations in lighting conditions, background perspectives, and object appearances. These factors pose a challenge for our model to effectively generalize, particularly when trained on a less diverse dataset. Our technique efficiently gathers a diverse set of data, encompassing RGB attributes, noise inconsistencies, and global context, instead of just depending on adjacent pixels. This enables us to obtain more extensive data for categorization and analysis. The ConvNext module enables the modelling of features at different semantic scales, allowing it to capture both low-level manipulations such as blobs and high-level semantic aspects such as texture. Therefore, the inclusion of multi-scale characteristics allows the model to better prioritize pixel-level image segmentation. The poor performance of complex CNN models can be attributed to the use of DNN-based techniques that use several CNN networks or complex branches to model the network. TDA-Net [33] is an exemplification of a model that combines three distinct CNN streams. This integration entails the training of a complex network in a holistic manner, which presents problems in terms of training complexity and heightened computational demands. Furthermore, there exists only a restricted range of models that are significantly less intricate, as they exclusively prioritize semantic information, leading to an incapacity to accurately detect manipulated sections. However, our model is less complex in comparison, skilled at capturing non-semantic features, and does not require a large quantity of training data to obtain equal performance. Therefore, this technique allows the model to effectively prioritize pixel-level image segmentation by capturing multi-scale information.

Table II: Results of shallowfake dataset model evaluation tests. "-" indicate an unknown score.

Category	Method	COVERAGE		NIST16		CASIA		COLUMBIA	
		AUC	AUC	AUC	AUC	AUC	F1	AUC	F1
Unsupervised	ELA [34]	0.583	0.583	0.429	0.613	0.613	0.214	0.581	0.222
	NOI1 [35]	0.587	0.587	0.487	0.612	0.612	0.263	0.546	0.269
	CFA1 [36]	0.485	0.485	0.501	0.522	0.522	0.207	0.720	0.190
DNN based methods	LSTM-EnDec [10]	0.712	0.712	0.794	-	-	-		-
	Transforensics [21]	0.884	0.884	-	0.850	0.850	0.627	-	0.674
	ManTra-Net [11]	0.819	0.819	0.795	0.817	0.817	-	0.824	-
	TDA-Net [33]	0.864	0.864	0.948	0.831	0.831	0.582	0.892	0.474
	MFCN [9]	-	-	-	-	-	0.541	-	-
	SPAN [41]	0.937	0.937	0.961	0.838	0.838	0.382	0.936	0.558
	GSR-Net [40]	0.768	0.768	0.945	0.796	0.796	0.574		0.489
	ObjectFormer [20]	0.957	0.957	0.996	0.882	0.882	0.579	-	0.7580
	CR-CNN [39]	0.939	0.939	0.992	0.789	0.789	0.475	0.861	0.757
	J-LSTM [38]	0.712	0.712	0.764	-	-	-	-	-
	PSCC-Net [42]	0.941	0.941	0.996	0.875	0.875	0.554	-	0.723
	MVSS-Net++ [19]	0.897	0.897	0.976	0.844	0.844	0.546	-	0.753
	RGB-N [37]	0.817	0.817	0.937	0.795	0.795	0.582	0.858	0.474
	TA-Net [43]	**0.978**	**0.978**	0.997	0.893	0.893	0.614	-	**0.782**
	Our Model	0.8826	0.8826	**0.9979**	**0.9542**	**0.9542**	**0.8956**	**0.9725**	0.6075

4.4.2 Deepfake dataset

Ten models are being assessed for their performance on the deepfake dataset. There are six state-of-the-art image manipulation models, while the remaining models are typical image-segmentation models. The codes of the models available on GitHub are taken into consideration for comparison. PyTorch models that have been pre-trained are utilized for image segmentation models, and these models are then fine-tuned. Three assessment measures are employed to provide a more thorough and all-encompassing examination. Table III displays the empirical findings of the deepfake technique on several models. All the models exhibited satisfactory performance, except MantraNet [11], across the several categories of the Faceforensics++ dataset. This phenomenon can be ascribed to the prevalence of facial modifications, the presence of a sole entity occupying the complete frame, and the capability of all the models to effectively represent these noticeable imperfections. MantraNet exhibits subpar performance across all deepfake categories, likely due to the low resolution, blurriness, and noise present in the images/frames. This could be attributed to the model's heavy reliance on the diverse and enough training dataset to perform well and the absence of which degrades the model performance badly. NedB-Net [44] is a model that has achieved good performance, although its score is considerably lower compared to other state-of-the-art (SoTA) models. This phenomenon can be ascribed to the existence of substandard images and more conspicuous modified areas, as well as the model's susceptibility to different types of noise and edge patterns. The authors of the research acknowledge this issue. DL-Net [45] The performance of DL-Net on the FF++ dataset is quite good, achieving an F1 score of 96%. This is due to its ability to effectively capture both high and low-level cues by predicting noise-level segmentation maps. These maps help the model to focus on the regions that have

been altered specifically. Nevertheless, the model's performance in the category of Face-swap manipulation could be better compared to the other three categories of FF++ manipulation. This is because face-swap techniques sometimes involve applying smoothing or blurring effects, which alter the noise and semantic patterns that the model depends on to detect manipulation. Another technique [46] employed for deepfake localization is a weak supervision framework that utilizes three methods: GradCAM, Patches, and Attention, to illustrate the results. We employed GradCAM techniques to compare scores. The technique excels in the weak supervision environment, demonstrating the model's strong discriminative skills. However, similar to the previous version, the model experiences a decrease in performance when it comes to the FS category of manipulation. This could be attributed to the fact that the technique is specifically designed for diffusion-generated images, which may restrict its capacity to perform effectively on GAN-generated images. Additionally, another technique [31] showed excellent performance, with an F1 score of 98%. Their approach is based on the preexisting UperNet and uses Bayar convolution techniques to detect and track noise indicators. Although these models are considered state-of-the-art, all of them achieve a score higher than 90%. This can be ascribed to the fact that much of the alteration has been focused on the face, which is easily recognizable by the models. The DADF technique [47] outperforms most models by utilizing multi-scale adapters to detect both short and long-range forgeries, as well as guided attention mechanisms that enhance the identification of rich forgery clues. Their scores are comparable to other approaches and serve as a suitable benchmark for comparison with a state-of-the-art method. Our model has demonstrated significant performance in the different categories of the FF++ dataset and surpassed the scores of other conventional models. The evaluations demonstrated that the effective coordination of multiple modules led to the development of a strong ability to discern and adapt to various forms of manipulation.

Table III: Test results of several models on FF++ dataset categories. IMD and IS are Image Manipulation Detection models, and Image Segmentation Models

Types	Methods	DeepFakes			Face2Face			FaceSwap			Neural Texture		
		IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1
IMD Models	**DADF** [47]	0.9453	0.9939	0.9677	0.9621	0.9899	0.9786	0.9599	0.9896	0.9655	0.9236	0.9788	0.9586
	DL-Net [45]	0.8750	0.9952	0.9337	0.9108	0.9946	0.9533	0.8976	0.9976	0.9460	0.9262	0.9979	0.9617
	MVSS-Net [48]	0.9558	0.9996	0.9787	0.9788	0.9997	0.9893	0.9570	0.9989	0.9780	0.9382	0.998	0.968
	NedB_Net [44]	0.8811	0.9767	0.9368	0.8432	0.9662	0.9149	0.8522	0.9700	0.9221	0.8827	0.9783	0.9377
	Weakly_Super_Gradcam [46]	0.9787	0.9990	0.9897	0.8753	0.9795	0.9336	0.9575	0.9791	0.9231	0.9861	0.9990	0.9868
	MantraNet [11]	0.3469	0.9523	0.5151	0.3553	0.9159	0.5243	0.3394	0.8888	0.5068	0.3664	0.9706	0.5363
	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	**LRASPP** [51]	0.9114	**0.9992**	0.9536	0.9445	0.9998	0.9714	0.9106	**0.9995**	0.9532	0.9399	0.9997	0.9690
	FCN [50]	0.9701	0.9967	0.9848	0.9834	0.9988	0.9786	0.9470	0.9991	0.9728	0.9591	0.9984	0.9728
	DeepLab [49]	0.9428	0.9981	0.9706	0.9840	**0.9999**	0.9919	0.9769	0.9986	0.9883	0.9704	**0.9998**	0.9640
Our Model		**0.9812**	0.9956	**0.9907**	**0.9898**	0.9989	**0.9945**	**0.9896**	0.9987	**0.9943**	**0.9865**	0.9978	**0.9972**

4.5 Qualitative Analysis

This section provides a comparison between our method and the two most competitive methods, namely MantraNet and MVSSS, to present detailed qualitative results on both shallowfakes and deepfake datasets. Figure V displays the visual representation of the outcomes obtained from the detection of image alteration. Our method outperforms other methods in terms of localization accuracy, as the other methods produce a significant number of false positives. Regarding the shallowfake dataset, the approach exhibits exceptional localization performance for the CASIA, Columbia, and Nist16 datasets. Localization for the deepfake dataset appears to be quite straightforward, given that most image alteration is focused on the face, simplifying the localization process. MantraNet deviates significantly from the ground truth, while MVSSNET shows a notable occurrence of false positives in unaltered regions. The main factor contributing to this result is that during the training phase, MVSSNET was exposed to a significant number of natural photos, which likely had a detrimental effect on the network's training process. Furthermore, the model requires a substantial amount of training data samples that exhibit a specific manipulation. Without this, the model's performance will decline. However, Mantra-Net faces difficulties in executing delicate modifications, such as methods that avoid producing abnormal artifacts or generating low-

resolution images. These obstacles result in an increased number of false negatives for this model. Our approach prioritizes the analysis of low-level data such as noise, as well as high-level contextual variables with edge-based supervision, resulting in improved identification and localization of altered artifacts.

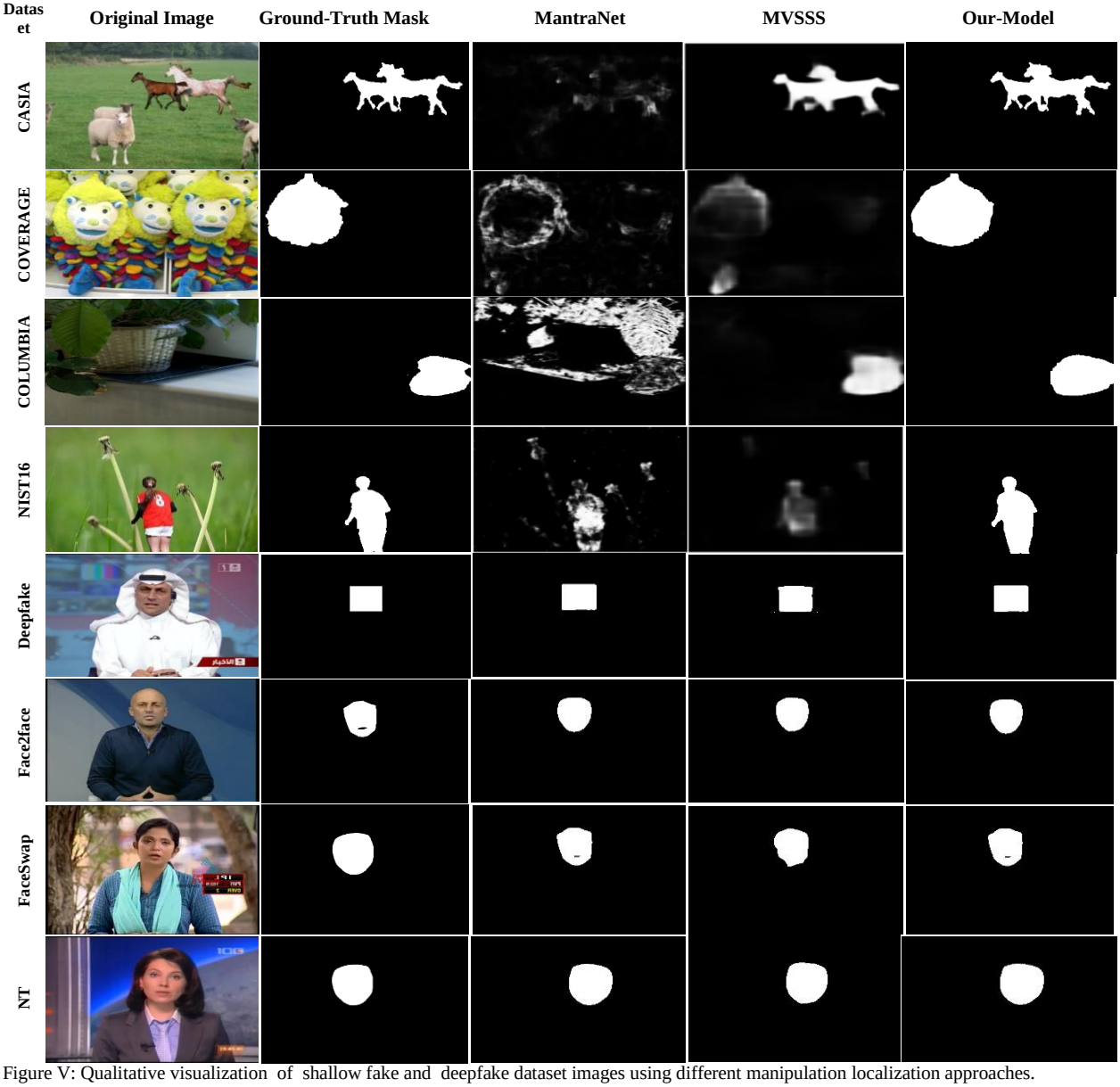


Figure V: Qualitative visualization of shallow fake and deepfake dataset images using different manipulation localization approaches.

4.6 Ablation Studies

We evaluate the proposed network in different scenarios by gradually introducing its components to analyze the effect of each component. The components underwent training using the CASIA2 dataset and were subsequently assessed using other shallowfake datasets, namely NIST16, Columbia, Coverage, and CASIA1. Table IV displays the outcomes of the ablations experiment. Below, we will discuss the various experimental configurations.

Case A: Model without edge loss supervision: Here the model is trained without edge supervision loss. It can be seen from the first row of the table that the model experiences a substantial decrease of 5-7% in the overall score of the evaluation metric, this suggests that edge supervision plays a vital in the overall detection accuracy of the model. Edge supervision allows the model to focus on the boundaries of the manipulation leading to precise manipulation results on the edges particularly, the manipulation involves finer details or thin structures. Also, edge supervision allows the model to learn to discriminate between true edges and noisy artifacts, which helps to recognize intricate structures.

Model without feature enhancement module: The second row represents the score and it can be seen there is a decrease of 3-4% in the overall score of localization. The phenomena can be attributed to the directing of attention of the model towards high-level or semantic features, which are essential for manipulation. In addition, the module consistently shows its effectiveness on many datasets, suggesting its ability to acquire a wide range of contextual information for varied

manipulations. Moreover, it is noteworthy that the scores exhibited a slight and uniform decline across all the datasets.

Model without Noise branch: In this instance, the noise branch, is eliminated from the model. The detection mostly relies on the utilization of RGB high-level or semantic properties. In this instance, the performance witnessed a substantial decrease of around 6-8%, validating the important influence of low-level factors like noise consistencies on the overall detection task. The CASIA model has exhibited a substantial decline in scores, specifically around 15%, when exposed to various forms of image manipulation. This exemplifies the significance of the residual noise characteristic in identifying different types of manipulation.

Model without RGB branch The model's RGB component has been removed and the noise component is used for localization purposes. In such conditions, the AUC score undergoes a rather moderate decrease of around 3-5%. This further affirms that the low-level noise components are more important than high-level semantic information for the detection task. The performance of the Coverage dataset has experienced a substantial decrease, suggesting that a sufficient number of datasets for branches with noise is necessary to train and detect inconsistencies caused by noise efficiently.

Table IV Different component ablation experiments using CASIA2-trained and another dataset-tested model.

Method	NIST16		COLUMBIA		COVERAGE		CASIA	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Case A	0.9085	0.9316	0.9016	0.8589	0.7645	0.6012	0.8916	0.8474
Case B	0.9316	0.9715	0.9519	0.8879	0.8052	0.6326	0.9278	0.8715
Case C	0.8706	0.9289	0.8996	0.8312	0.7889	0.6019	0.8195	0.7746
Case D	0.9236	0.9456	0.9356	0.8829	0.7815	0.6159	0.9056	0.8579
Overall Model	0.9512	0.9976	0.9746	0.9188	0.8366	0.6512	0.9524	0.8945

5 Conclusion

This study introduces a new dual-branch design that incorporates Noise Residual extraction modules in one branch and RGB information in the other branch. The architecture utilizes a conventional ConvNext module, whose features are further improved by the feature improvement module. The model additionally utilizes edge supervision to improve the localization of the manipulation specifically at the edges. The model accurately captures the fundamental inconsistencies that are crucial for interpretability in machine learning tasks, as well as incorporating further semantic variables. Extensive experimentation conducted on both shallowfakes and deepfake datasets has demonstrated that the model successfully detects minute indications of manipulation and delivers state-of-the-art (SoTA) outcomes. Future work may entail assessing the model's capacity to apply to new, unexplored data and doing a thorough review of different compression scenarios to ensure reliability.

Declaration

6 Ethical Approval

NA (Not Applicable)

7 Funding

No funding was received for this research.

8 Availability of data and materials

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

9 References

[1] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: generation, detection, and applications," *International Journal of Multimedia Information Retrieval*, vol. 11, pp. 219-289, 2022.

[2] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910-932, 2020.

[3] P. Ferrara, T. Bianchi, A. D. Rosa and A. Piva, "Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1566 - 1577, 2012.

[4] A. Swaminathan, M. Wu and K. R. Liu, "Nonintrusive component forensics of visual sensors using output images," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 1, pp. 91-106, 2007.

[5] 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.

[6] 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.

[7] C. Iakovidou, . M. Zampoglou, S. Papadopoulos and Y. Kompatsiaris, "Content-aware detection of JPEG grid inconsistencies for intuitive image forensics," *Journal of Visual Communication and Image Representation*, vol. 54, pp. 155-170, 2018.

[8] G. Muhammad, . M. H. Al-Hammadi, M. Hussain and G. Bebis, "Image forgery detection using steerable pyramid transform and local binary pattern," *Machine Vision and Applications*, vol. 25, pp. 985-995, 2013.

[9] R. Salloum, Y. Ren and C.-C. J. Kuo, "Image Splicing Localization using a Multi-task Fully Convolutional Network (MFCN)," *Journal of Visual Communication and Image Representation*, vol. 51, pp. 201-209, 2018.

[10] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath and A. K. Roy-Chowdhury, "Hybrid LSTM and Encoder–Decoder Architecture for Detection of Image Forgeries," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3286-3300, 2019.

[11] 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.

[12] 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.

[13] 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.

[14] J. Fridrich and J. Kodovsky, "Rich Models for Steganalysis of Digital Images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868-882, 2012.

[15] 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.

[16] 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.

[17] C. Li, . Q. Ma, . L. Xiao, M. Li and . A. Zhang, "Image Splicing Detection Based on Markov Features in QDCT Domain," in *International Conference on Intelligent Computing*, 2015.

[18] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-Based Camera Model Fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144-159, 2019.

[19] C. Dong, X. Chen, R. Hu, J. Cao and X. Li, "MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3539-3553, 2022.

[20] 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.

[21] 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.

[22] B. Bayar and M. C. Stamm, "Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691 - 2706, 2018.

[23] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, Vigo Galicia Spain, 2016.

[24] K. Guo, H. Zhu and G. Cao, "Effective Image Tampering Localization via Enhanced Transformer and Co-attention Fusion," in *arXiv:2309.09306v1*, 2023.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] T.-T. Ng, J. Hsu and S.-F. Chang, "Columbia Image Splicing Detection Evaluation Dataset," in *DVMM Laboratory of Columbia University*, 2009.

[30] 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.

[31] 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.

[32] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

[33] S. Li, S. Xu, W. Ma and Q. Zong, "Image Manipulation Localization Using Attentional Cross-Domain CNN Features," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 5614 - 5628, 2023.

[34] Neal Krawetz and Hacker Factor Solutions, "A picture's worth…digital image Analysis and Forensics," *Hacker Factor Solutions*, vol. 2, no. 2, p. 2, 2007.

[35] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image and Vision Computing*, vol. 27, no. 10, pp. 1497-1503, 2009.

[36] P. Ferrara, T. Bianchi, A. D. Rosa and A. Piva, "Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1566-1577, 2012.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] X. Liu, Y. Liu, J. Chen and X. Liu, "PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7505-7517, 2022.

[43] Z. Shi, H. Chen and D. Zhang, "Transformer-Auxiliary Neural Networks for Image Manipulation Localization by Operator Inductions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4907-4920, 2023.

[44] 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.

[45] C. Kong, B. Chen, H. Li, S. Wang, A. Rocha and S. Kwong, "Detect and Locate: Exposing Face Manipulation by Semantic- and Noise-Level Telltales," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1741-1756, 2022.

[46] 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.

[47] Y. Lai, . Z. Luo and Z. Yu, "Detect Any Deepfakes: Segment Anything Meets Face Forgery Detection and Localization," in *Biometric Recognition*, Singapore, 2023.

[48] 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.

[49] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," in *arxiv.org/abs/1706.05587*, 2017.

[50] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640 - 651, 2016.

[51] 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.

Check for updates

Applied Organometallic Chemistry

WILEY

Applied Organometallic Chemistry

FULL PAPER

A Novel Acetylcholinesterase-Based Electrochemical Biosensor Using g-C_3N_4@MoS_2 Nanohybrid for the Detection of Trichlorfon

Srijita Chatterjee | Harshita Singh | Divya Hudda | Sweety | Devendra Kumar

Department of Applied Chemistry, Delhi Technological University, Delhi, India

Correspondence: Devendra Kumar (dkumar@dce.ac.in)

Received: 30 May 2024 | **Revised:** 8 August 2024 | **Accepted:** 10 August 2024

Keywords: acetylcholinesterase (AChE) | electrochemical biosensor | g-C_3N_4@MoS_2 | pesticide | trichlorfon

ABSTRACT

Organophosphorus pesticides are commonly employed in agricultural fields due to their potent insecticidal properties and short environmental persistence. However, organophosphorus pesticides enter the human body through the food chain, surface, and groundwater, leading to irreversible damage to the nervous system. Therefore, monitoring the presence of organophosphorus pesticides in food is necessary to ensure human safety. Herein, an ultra-sensitive electrochemical biosensor has been prepared by using graphitic carbon nitride decorated molybdenum sulfide (g-C_3N_4@MoS_2) as a catalyst for the detection of organophosphorus pesticides, trichlorfon. The synergistic effect of g-C_3N_4@MoS_2 boosts the electron transfer across the electrode surface, further contributing to the immobilization of acetylcholinesterase (AChE) enzyme using glutaraldehyde as a crosslinking agent. At optimum conditions, the developed biosensor (AChE/g-C_3N_4@MoS_2/ITO) demonstrated a broad linear range (5–100 nM) with a low detection limit (LOD) of 2.1 nM obtained using the equation 3σ/S where σ is the standard deviation and S is the sensitivity of the bioelectrode. In addition, the AChE/g-C_3N_4@MoS_2/ITO biosensor exhibited good stability and reproducibility, with satisfactory results for real sample analysis of trichlorfon detection.

1 | Introduction

Organophosphate pesticides (OPs) are thiophosphate compounds that are ubiquitously used in farm fields due to their high insecticidal efficacy in increasing agricultural productivity [1, 2]. However, the overuse and improper utilization of OPs lead to pesticide persistence in food and the environment, which poses a serious risk to human health [3, 4]. They inhibit acetylcholinesterase (AChE) activity and disrupt neurological function, which can lead to severe chronic diseases such as cognitive disorders, paralysis, allergic reactions, and so forth [5–7]. Therefore, efficient techniques for monitoring pesticide residues are necessary to prevent harmful effects on human health.

Several analytical methods are available for the quick and accurate detection of pesticides, including high-performance liquid chromatography [8], gas chromatography–mass spectrometry [9], surface plasmon resonance [10], fluorimetry [11], and immunoassays [12]. However, these methods have several limitations, including costly equipment, requiring skillful professionals, and time-consuming processes [13, 14]. On the other hand, AChE inhibition-based electrochemical biosensors are preferred for detecting pesticide residues due to their simple operation, cost-effectiveness, on-site detection, and highly sensitive and selective nature. Their working principle relies on the inhibition rate of enzymes when exposed to OPs [15, 16]. However, achieving efficient immobilization of enzymes and optimizing the electron transfer rate between AChE

Srijita Chatterjee, Harshita Singh, and Divya Hudda contributed equally to this study.

© 2024 John Wiley & Sons, Ltd.

and the electrode surface remain significant challenges for fabricating AChE biosensors [2, 17]. Therefore, a range of electrodes with various nanocomposites has been explored to enhance the sensitivity of AChE biosensors [18].

Nowadays, two-dimensional (2D) layered materials have shown great interdisciplinary interest because of their ultrathin structure, quantum size effect, and unique physicochemical properties [19–21]. Using 2D materials-based electrodes for biosensor applications improves their electrocatalytic activity and enhances stability with high selectivity and sensitivity [22]. Recently, graphitic carbon nitride (g-C_3N_4), a metal-free semiconductor, has gained extensive attention in a wide range of research areas, including photocatalysis, sensing, and energy conversion, due to its biocompatibility, non-toxicity, high stability, and unique electronic and optical properties [23–25]. However, pristine g-C_3N_4 exhibits low electrocatalytic activity due to the high recombination of electrogenerated charges [26]. Hence, it is significant to pair g-C_3N_4 with transition metal dichalcogenides (TMC) to boost the electrochemical efficiency of g-C_3N_4. Molybdenum disulfide (MoS_2), an analog to graphene, comprises one Mo layer sandwiched between two layers of sulfur atoms and linked through weak Van Der Waals forces. It has fascinated the scientific community due to its excellent thermal conductivity, high mechanical strength, and unique optical and electronic properties [27–29], which have led to the potential application of MoS_2 in the fields of water splitting [30], supercapacitors [31], electronics [32], and lubrication [23]. In addition, MoS_2 has high electron mobility, good biocompatibility, abundant reactive sites, and a large surface area, making it a desirable material for electrochemical biosensors [33–35]. For instance, Huang et al. fabricated a DNA biosensor based on Au/MoS_2/MWCNTs with a low detection limit of 0.79 fM [36]. Furthermore, Shuai et al. reported Au nanoparticles/hollow molybdenum disulfide microcubes based electrochemical biosensor to detect microRNA-21 with high specificity and sensitivity [37]. Due to its unique electroactive properties, Zheng et al. constructed an electrochemical sensor based on AgNWs@MoS_2 to sensitively detect thiabendazole [13]. Thus, it is expected that the 2D MoS_2 and g-C_3N_4 with a 2D polymeric structure would provide a high surface-to-volume ratio, which could improve the electrode performance to enhance the sensitivity of pesticide detection.

In this work, a g-C_3N_4@MoS_2 is synthesized hydrothermally, which is further utilized as a sensing platform for the development of an effective electrochemical biosensor (AChE/g-C_3N_4 @ MoS_2/ITO) for the detection of trichlorfon (TF). The synergetic effect of g-C_3N_4@MoS_2 enhances the electrochemical sensing performance and increases the electrocatalytic properties of the developed biosensor. Moreover, the fabricated biosensor exhibited good stability with a broad linear range and low detection limit. Finally, the applicability and efficiency of the developed AChE/g-C_3N_4@MoS_2/ITO biosensor were evaluated for the detection of TF in real samples.

2 | Experimental Section

Sodium molybdate dihydrate ($Na_2MoO_4.2H_2O$; 99.5%), thiourea (CH_4N_2S; 99%), acetylcholinesterase (AChE), acetylthiocholine chloride (ATCl), trichlorfon (TF), glutaraldehyde solution, glucose (Glu; ≥99.5%), ascorbic acid (AA; 99%), and uric acid (UA; ≥99%) were purchased from Sigma-Aldrich, India. Acetone, ethanol, sodium chloride (NaCl), and melamine were bought from Thermo-Fischer Scientific, India. All the chemicals are of analytical grade and used without purification. Furthermore, Milli-Q water (resistivity of 18.2 MΩ) was used to prepare solutions.

2.1 | Synthesis of g-C_3N_4

To prepare g-C_3N_4 powder, a high thermal condensation method was used under ambient atmospheric pressure. Typically, melamine (10 g) was heated directly in a muffle furnace at 550°C for 4 h. The resulting pale yellow material was washed with distilled water and ethanol (1:1) and then dried in a vacuum oven [38].

2.2 | One-Pot Synthesis of g-C_3N_4@MoS_2

For the preparation of the nanocomposite, sodium molybdate (500 mg) and thiourea (950 mg) were dispersed in distilled water (50 mL) and stirred vigorously to form a homogenous solution. Subsequently, g-C_3N_4 (100 mg) was introduced to the solution, and the mixture was mixed continuously for 2 h. Additionally, the solution was ultrasonicated for 30 min, and the resultant solution was transferred to 100-mL Teflon-lined hydrothermal and heated at 180°C for 20 h. The resulting precipitate was centrifuged at 4000 rpm, followed by washing with distilled water-ethanol (1:1) multiple times. Finally, the resultant product was dried in the vacuum oven at 80°C.

2.3 | Electrophoretic Deposition (EPD) of g-C_3N_4@ MoS_2 on ITO

The synthesized g-C_3N_4@MoS_2 hybrid was deposited onto the pre-hydrolyzed ITO (1.4 × 0.7 cm) coated glass substrate electrophoretically utilizing a system of two electrodes. Platinum wire is used as a counter electrode and is positioned at a distance of 1 cm from the ITO (working electrode). Before deposition, g-C_3N_4@MoS_2 (0.5 mg) is ultrasonicated in deionized water (15 mL) to make a colloidal suspension. The deposition process is then carried out at an optimized DC potential of 10 V for 10 s.

2.4 | Fabrication of g-C_3N_4@MoS_2 Nanohybrid-Based Biosensor

To fabricate the biosensor, the g-C_3N_4@MoS_2/ITO electrode was incubated with 10 μL of glutaraldehyde (0.1%) at room temperature for 60 min. Then, 20 μL of AChE solution (0.06 mg mL^{-1}, made in PBS of 7.5 pH) was immobilized on the constructed electrode and left to incubate at 4°C for 12 h. Before the electrochemical study, the unbounded enzyme was eliminated by treating the biosensor with 50-μL PBS. The schematic approach for the biosensor fabrication is represented in Figure 1.

2.5 | Instrumentation

Bruker D8 Advance X-ray diffractometer with monochromatic (Cu Ka) radiation (= 1.546 Å) has been used to study the X-ray

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

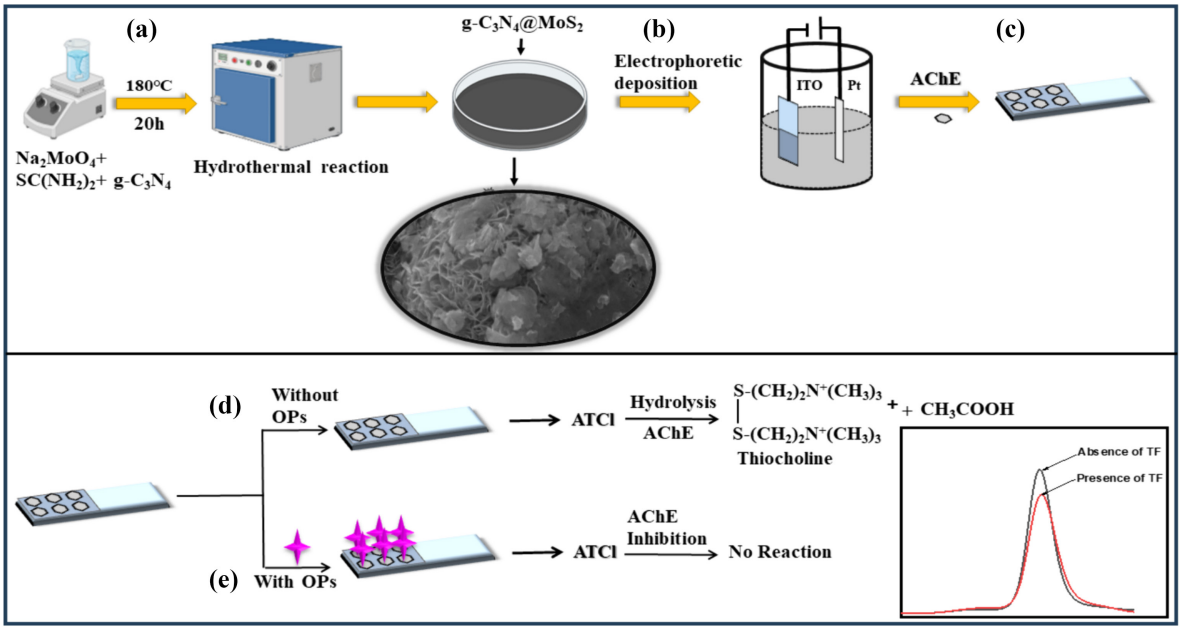


FIGURE 1 | Schematic representation of TF detection using fabricated AChE/g-C_3N_4@MoS_2/ITO bioelectrode.

diffraction (XRD) pattern for the synthesized materials at $2\theta = 10–70°$ with a scan speed of 2°/min. Perkin Elmer spectrum attenuated total reflection (ATR) mode for recording the FT-IR spectrum has been used in the frequency range of 400–4000 cm^{-1} to study the g-C_3N_4 and g-C_3N_4@MoS_2 nanohybrid formation. The surface morphology of g-C_3N_4 and g-C_3N_4@MoS_2 nanocomposite was examined using scanning electron microscopy (SEM, JSM 6610LV, JEOL, Japan). All electrochemical investigations were conducted using an Autolab potentiostat/galvanostat (Eco-Chemie, Netherlands) with working (g-C_3N_4@MoS_2/ITO), reference (Ag/AgCl) and auxiliary (Pt) electrodes in 0.2-M phosphate buffer saline (PBS, 7.5 pH) with 5 mM $[Fe(CN)_6]^{3-/4-}$.

3 | Results and Discussion

3.1 | Design of Fabricated Biosensor

In this study, the g-C_3N_4@MoS_2 nanohybrid was prepared by hydrothermal treatment (Figure 1a). The synergetic relationship between g-C_3N_4 and MoS_2 improves the analytical performance of the biosensor. For the fabrication of the biosensor, the synthesized g-C_3N_4@MoS_2 was deposited electrophoretically onto the surface of ITO (Figure 1b), followed by immobilization of the enzyme (AChE) on the surface of the transducer using glutaraldehyde (crosslinker) (Figure 1c). Finally, the fabricated biosensor is utilized for the determination of the TF. In the absence of TF, AChE catalyzes the hydrolysis of ATCl, generating an electroactive product, that is, thiocholine and acetic acid, which produces an irreversible oxidation peak (Figure 1d). TF obstructs the functional sites of AChE, thereby preventing ATCl from being hydrolyzed, which further results in the current decline (Figure 1e).

3.2 | Structural and Morphological Characterization

Figure 2a represents the XRD pattern of g-C_3N_4 and g-C_3N_4@MoS_2. For g-C_3N_4 (black line), two strong diffraction peaks were observed at 13.1° and 27.5°, which corresponds to tri-s-triazine rings (100) and interlayer stacking of aromatic structure (002), respectively [39]. As for g-C_3N_4@MoS_2 nanocomposite (red line), the peaks at 14.1°, 33.4°, 39.7°, 49.6°, and 58.7° correspond to (002), (101), (103), (105), and (110) planes of MoS_2 [23]. It should be noted that no peaks of g-C_3N_4 were observed in g-C_3N_4@MoS_2 nanocomposite due to the low content of g-C_3N_4. The following SEM and EDX characterizations can be used further to validate the presence of g-C_3N_4 in g-C_3N_4@MoS_2 material.

FTIR is used to examine the chemical groups in the range of 400–4000 cm^{-1} for the prepared g-C_3N_4 and g-C_3N_4@MoS_2 nanocomposite (Figure S1). For g-C_3N_4 (curve a), an intense peak at 807 cm^{-1} represents tri-s-triazine units in the bending mode, whereas the absorption peaks at 1578 and 1647 cm^{-1} are attributed to stretching vibration modes of C=N. The peaks at 1228, 1318, and 1403 cm^{-1} belong to aromatic C-N stretching, and a broad peak at 3168 cm^{-1} is related to terminal NH_2 or NH groups [26, 40]. The FT-IR spectra of g-C_3N_4@MoS_2 nanocomposite (curve b) exhibit similar peaks as found for pure g-C_3N_4. However, the main peak observed at 497 cm^{-1} in the g-C_3N_4@MoS_2 spectra is due to Mo-S vibration [41].

The surface morphology of g-C_3N_4 and g-C_3N_4@MoS_2 were studied by SEM. Figure 2b shows that g-C_3N_4 comprises microscale layers (sheet-like structure), which are backfolded at the edges. The SEM image of g-C_3N_4@MoS_2 (Figure 2c) displays the presence of g-C_3N_4 (sheet-like structure) on the surface of MoS_2 (flower-like structure). Figure 2d shows the EDX spectrum of the g-C_3N_4@

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

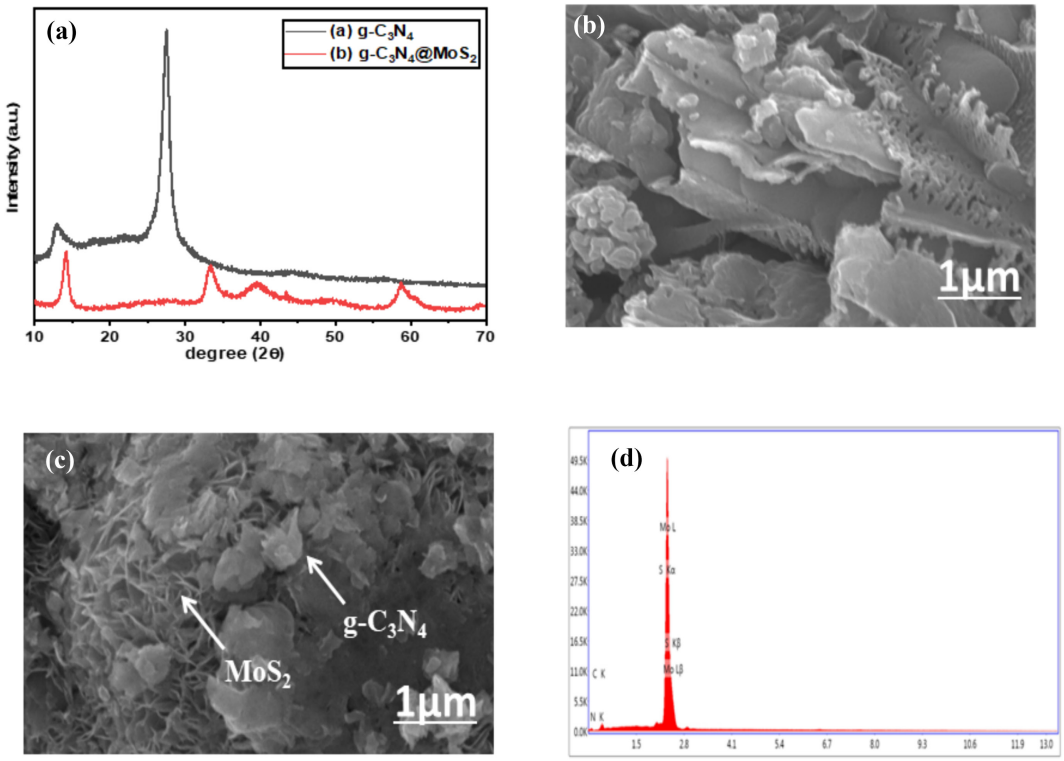


FIGURE 2 | (a) XRD of (a) g-C_3N_4 and (b) g-C_3N_4@MoS_2; (b, c) morphology structure of g-C_3N_4 and g-C_3N_4@MoS_2; and (d) EDX analysis of g-C_3N_4@MoS_2 nanohybrid.

MoS_2 nanohybrid. The presence of C, N, Mo, and S elements confirms the successful preparation of g-C_3N_4@MoS_2 nanocomposite.

3.3 | Electrochemical Characterization of Fabricated Biosensor

The electrochemical properties to investigate the electron transfer resistance of various fabricated electrodes, including ITO, g-C_3N_4/ITO, g-C_3N_4@MoS_2/ITO, and AChE/g-C_3N_4@MoS_2/ITO, were studied using electrochemical impedance spectroscopy (EIS) in 0.2 M of PBS (consisting of 5-mM $[Fe(CN)_6]^{3-/4-}$) in the 0.01- to 10^5-Hz frequency range (set potential 0.01 V), as shown in Figure 3a. The EIS data can be examined using the Nyquist plot, where a linear pattern determined at a lesser frequency corresponds to the diffusion-limited electron transfer process, while a semi-circular pattern is evident at a higher frequency attributed to the limited electron transfer process [42]. The electron transfer resistance (R_{ct}) for ITO, g-C_3N_4/ITO, g-C_3N_4@MoS_2/ITO, and AChE/g-C_3N_4@MoS_2/ITO were 2.19 kΩ (curve a), 1.14 kΩ (curve b), 845 Ω (curve c), and 1.05 kΩ (curve d), respectively. Equations (1) and (2) have been utilized to calculate the exchange current per geometric unit area (i_o) and the apparent electron transfer rate constant (K_{app}) for g-C_3N_4/ITO and g-C_3N_4@MoS_2/ITO based on the obtained R_{ct} values.

$$i_o = nRT / R_{ct}F, \quad (1)$$

$$K_{app} = RT / n^2F^2AR_{ct}C, \quad (2)$$

where n, R, and T denote the number of transferred electrons, gas constant, and temperature, respectively. F, C, and A represent the Faraday constant, concentration of the $[Fe(CN)_6]^{3-/4-}$, and electrode area, respectively. The value of i_o for the g-C_3N_4/ITO and g-C_3N_4@MoS_2/ITO electrodes were found to be 2.252×10^{-5} and 3.039×10^{-5} A cm^{-2}, respectively, whereas the K_{app} using the g-C_3N_4/ITO and g-C_3N_4@MoS_2/ITO electrodes were found to be 2.032×10^{-4} and 2.522×10^{-4} cm s^{-1}, respectively. The increase in i_o and K_{app} values for the g-C_3N_4@MoS_2/ITO electrode suggests the high electron transport of redox couple across the electrode's interface [43].

The cyclic voltammetric (CV) technique has been used for different fabricated electrodes (ITO, g-C_3N_4/ITO, g-C_3N_4@MoS_2/ITO, and AChE/g-C_3N_4@MoS_2/ITO) at 50 mV s^{-1} scan rate. As depicted in Figure 3b, the maximum redox peak current was obtained for g-C_3N_4@MoS_2/ITO (0.43 mA, curve c) as compared to g-C_3N_4/ITO (0.32 mA, curve b). This is attributed to the good electrical conductivity and large surface area of g-C_3N_4@MoS_2 nanocomposite. After immobilizing the enzyme (AChE) on the g-C_3N_4@MoS_2/ITO, the conductivity and electron mobility on the surface of the electrode decrease because of its non-conductive nature (0.37 mA, curve d).

The efficacy of the g-C_3N_4@MoS_2/ITO electrode was evaluated by varying the scan rate within the range of 10–300 mV s^{-1} (Figure S2). Here, we observed that as the scan rate increases, the oxidation peak current drifts towards a more positive value while the reduction peak current drifts towards a more negative value. This suggests a good linearity between the anodic and cathodic peak potentials (E_{pa} and E_{pc}) against the logarithm value of the scan rate (log v) (Figure 3c), as represented by Equations (3) and (4).

$$E_{pa}\left[g-C_3N_4@MoS_2/ITO\right] = 0.1033\ (V)\log(v) + 0.1709\ (V); \quad R^2 = 0.9769, \quad (3)$$

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

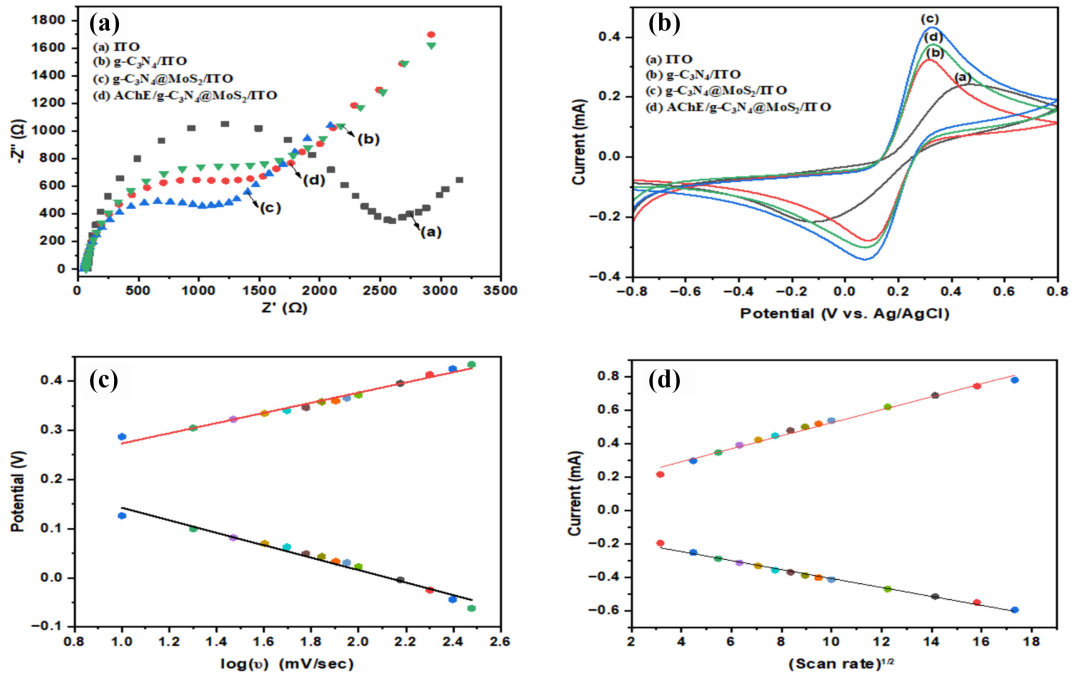


FIGURE 3 | (a) Nyquist plots; (b) CV study of different fabricated electrodes: (a) ITO, (b) g-C_3N_4/ITO, (c) g-C_3N_4@MoS_2/ITO, and (d) AChE/g-C_3N_4@MoS_2/ITO; (c) linearity plot between peak potential versus log scan rate; and (d) linearity plot for peak current versus square root of scan rate for g-C_3N_4@MoS_2/ITO in PBS (0.2 M, pH 7.5) consist of 5 mM $[Fe(CN)_6]^{3-/4-}$.

$$E_{pc}\left[g-C_3N4@MoS_2/ITO\right] = -0.1266\ (V)\log(v) + 0.2689\ (V); \quad R^2 = 0.9706. \tag{4}$$

Considering the slope obtained from the above-given equations and Laviron's equation [44], the charge transfer coefficient (α) and electron transfer rate constant (K_s) for g-C_3N_4@MoS_2/ITO were calculated to be 0.9245 and 0.0844 s^{-1}, respectively. Similarly, the anodic and cathodic peak currents (I_{pa} and I_{pc}) increase linearly with the square root of the scan rate ($v^{1/2}$), indicating a surface-controlled process for g-C_3N_4@MoS_2/ITO electrode (Figure 3d), as shown by Equations (5) and (6).

$$I_{pa}\left[g-C_3N_4@MoS_2/ITO\right] = 38.89\ \mu A\ (mV/s)^{1/2} \times v^{1/2}\ mV\ s^{1/2} + 136.97\ \mu A; \quad R^2 = 0.9871, \tag{5}$$

$$I_{pc}\left[g-C_3N_4@MoS_2/ITO\right] = -26.80\ \mu A\ (mV/s)^{1/2} \times v^{1/2}\ mV\ s^{1/2} - 137.60\ \mu A; \quad R^2 = 0.9962. \tag{6}$$

Additionally, the effective surface area (A) and the diffusion coefficient (D) have been calculated using the slopes of the aforementioned equations and Randles–Sevcik equation (7), as depicted below:

$$I_p = \left(2.69 \times 10^5\right) n^{3/2} A C D^{1/2} v^{1/2}, \tag{7}$$

where I_p is the peak oxidation or reduction current, n represents no. of electrons oxidized or reduced, and C and v denote the concentration (mole cm^{-3}) of $[Fe(CN)_6]^{3-/4-}$ and the scanning rate ($V\,s^{-1}$), respectively [45]. The A and D of the g-C_3N_4@MoS_2/ITO electrode were found to be 0.25 cm^2 and 7.146×10^{-6} $cm^2\,s^{-1}$, respectively. Moreover, the surface coverage of the modified electrode (g-C_3N_4@MoS_2/ITO) was computed to be 2.254×10^{-8} mol cm^{-2} using the formula: $[I_p = n^2F^2A\tau v/4RT]$, where F, T, and R represent Faraday's constant, temperature, and gas constant respectively.

3.4 | Optimization of Experimental Parameters

To enhance the performance of the bioelectrode for the quantitative TF assay, numerous parameters, including pH of the electrolytic solution, incubation time, and amount of AChE enzyme, were investigated. In order to optimize other hyperparameters, the concentration of ATCl, that is, 2 mM, was obtained from the prior study [17]. Figure S3A illustrates the effect of pH ranging from 6 to 8 on the enzyme activity. Notably, the highest peak current was recorded at pH 6.5. Therefore, this optimum value has been chosen for the electrochemical detection of TF.

Besides, the amount of enzyme (AChE) immobilized on the constructed bioelectrode surface was another important parameter. Figure S3B shows a rise in the concentration of AChE (0.04–0.12 mg mL^{-1}). The current response increases initially and reaches a maximum of 0.06 mg mL^{-1}. Subsequently, a decrease in current was observed because of the high amount of mass transfer resistance. Thus, 0.06 mg mL^{-1} was selected for further studies.

The incubation time is another important aspect that affects the performance of the fabricated biosensor. As illustrated in Figure S3C, the peak decreases significantly with an increase in incubation time from 2 to 14 min. It saturates within 8 min, indicating the completion of the reaction of TF with the AChE.

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

Thus, 8 min was selected as the optimal incubation time for subsequent biosensing studies.

3.5 | Electrochemical Biosensing Study of AChE/g-C_3N_4@MoS_2/ITO Electrode

Differential pulse voltammetry (DPV) was used to examine the electrochemical biosensing response of AChE/g-C_3N_4@MoS_2/ITO electrode towards TF (Figure 4a). The current of the developed bioelectrode was recorded as the TF concentration was increased from 5 to 100 nM in 0.2-M PBS having a pH of 6.5, comprising of $[Fe(CN)_6]^{3-/4-}$ and ATCl (2 mM). A decline in peak current was observed with an increase in TF concentration, indicating TF's inhibitory impact on the activity of the AChE enzyme. The calibration curve (Figure 4b) shows a linear relationship between the peak current (I) and concentration of TF that follows the given equation:

$$I = -0.127\ \mu A/nM + 9.185\ \mu A \times C_{TF};\ R^2 = 0.993.$$

The biosensor's sensitivity was determined to be 0.51 μA $(nM)^{-1}\,cm^{-2}$ via the slope of the linear curve of the proposed bioelectrode. Using the equation 3σ/S, the detection limit (LOD) was calculated to be 2.1 nM, where σ and S denote the standard deviation and sensitivity of the bioelectrode. For a range of biosensing parameters, the AChE/g-C_3N_4@MoS_2/ITO biosensor was compared with previously reported TF-based sensors (Table 1). The results demonstrated that the biosensor exhibited an enhanced performance toward the primary detection of TF with a low LOD.

3.6 | Interference, Reproducibility, and Stability Studies

The selectivity of the proposed biosensor (AChE/g-C_3N_4@MoS_2/ITO) was evaluated in the presence of some interferent analytes (Glu, UA, AA, and NaCl). For this study, 100 nM of the above-mentioned interferent species was mixed with 5 nM of TF. The bar diagram (Figure 4c) depicts no significant changes in the peak current when the fabricated biosensor was exposed to interferent analytes, and the current remained constant for all other analytes when presented as a mixture, indicating good selectivity for TF detection.

The reproducibility of the AChE/g-C_3N_4@MoS_2/ITO electrode was examined by fabricating five similar electrodes in the presence of 5 nM TF (Figure 4d). The response of the fabricated biosensor with an RSD value of 2.45% ($n = 5$) confirms the good reproducibility.

Over a period of 20 days, the electrode was kept and maintained at 4°C, and the DPV response towards TF was recorded every 5 days to monitor the stability of the biosensor. The electrochemical response remains 82.4% of its original activity after 20 days, indicating acceptable stability of the proposed biosensor (Figure S4).

3.7 | Real Sample Analysis

The practicality and accuracy of the AChE/g-C_3N_4@MoS_2/ITO biosensor were evaluated by detecting TF in three different real samples (Pulse, carrot, and banana) by the DPV method. Before the measurement, these samples were treated with DI water for

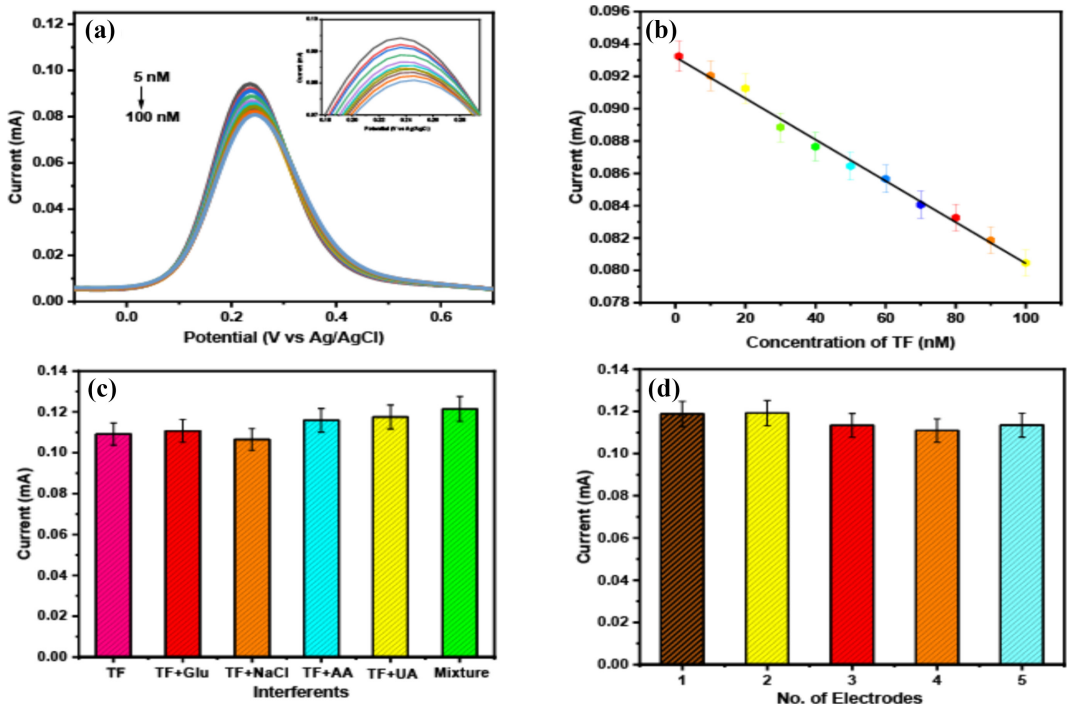


FIGURE 4 | (a) DPV response of AChE/g-C_3N_4@MoS_2/ITO electrode with different concentrations of TF (5–100 nM); (b) linearity plot between the current versus concentration of TF; (c) selectivity analysis of AChE/g-C_3N_4@MoS_2/ITO bioelectrode; and (d) reproducibility study of fabricated AChE/g-C_3N_4@MoS_2/ITO.

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 | Comparison of the AChE/g-C_3N_4@MoS_2/ITO bioelectrode with the earlier reported biosensor for the detection of TF.

S.no.	Fabricated electrode	Detection methods	Linear range	LODs	Ref.
1.	MWCNT@Au-g-PMAEFc/GCE	DPV	50 nM–10 μM	27 nM	[46]
2.	CHIT/Au electrode	SGGT	300–3000 nM	10 nM	[47]
3.	AgNPs	Colorimetric	5–100 nM	5.46 nM	[48]
4.	MHPBC/GCE	DPV	0.1 nM–10 μM	35 nM	[49]
5.	Nafion-AChE/PB-DSPE	CA	0.1–5 μg/mL	0.1 μg/mL	[50]
6.	AChE/g-C_3N_4@MoS_2/ITO	DPV	5–100 nM	2.1 nM	Present work

Abbreviations: AgNPs: silver nanoparticles; Au: gold nanoparticles; CA: chronoamperometry; CHIT: chitosan; CV: cyclic voltammetry; DPV: differential pulse voltammetry; GCE: glassy carbon electrode; MHPBC: multiwalled carbon nanotubes grafted acryloyloxy ferrocene carboxylates; MWCNT: multi-walled carbon nanotubes; PB-DSPE: Prussian blue modified dual channel screen printed electrode; PMAEFc: poly (2-methacryloyloxyethyl ferrocenecarboxylate); SGGT: solution gated graphene transistor.

TABLE 2 | Detection of TF in real samples.

Samples	Added amount (nM)	Found amount (nM)	Recovery (%)	RSDs (%)
Pulse	0	0	—	1.81
	5	4.86	97.3	1.93
	10	10.5	105	4.02
	20	21	105	3.98
	30	31.2	104.2	2.92
Carrot	0	0	—	0.61
	5	5.4	108	5.66
	10	10.4	104	3.13
	20	21.4	107	4.89
	30	31.2	104	3.17
Banana	0	0	—	0.76
	5	5.05	101	0.87
	10	9.88	98.8	0.84
	20	20.1	100.5	0.38
	30	29.8	99.4	0.39

washing, and each sample was mixed with 7.5 pH buffer, followed by centrifugation at 3500 rpm for 15 min. The supernatant was subsequently filtered and spiked with known concentrations of TF (5, 10, 20, and 30 nM). The results shown in Table 2 validate the satisfactory recovery (97%–108%), further demonstrating the precision and reliability of the developed biosensor for all the analyzed real samples.

4 | Conclusion

In conclusion, a highly efficient and novel AChE-based electrochemical biosensor has been successfully constructed for TF detection based on g-C_3N_4@MoS_2 nanocomposite. The g-C_3 N_4@ MoS_2 modified transducer accelerates the process of transfer of electrons between the AChE and the surface of the electrode, thereby enhancing the sensitivity of the proposed biosensor. The fabricated bioelectrode exhibits good reproducibility, high selectivity, and acceptable stability towards TF with a wide linear range (5–100 nM) and a low detection limit (2.1 nM). Furthermore, the practical application of the developed biosensor was successfully explored for food samples (pulse, carrot, and banana), yielding satisfactory recovery outcomes.

Author Contributions

Srijita Chatterjee: conceptualization, methodology, data curation, writing – original draft. **Harshita Singh:** conceptualization, methodology, data curation, writing – original draft. **Divya Hudda:** conceptualization, investigation, methodology, formal analysis, data curation. **Sweety:** investigation, methodology, formal analysis, data curation. **Devendra Kumar:** validation, supervision, writing – review and editing.

Conflicts of Interest

The authors declare no conflicts of interest.

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

Data Availability Statement

Data are available on request from the authors.

References

1. G. Zhao, B. Zhou, X. Wang, J. Shen, and B. Zhao, "Detection of Organophosphorus Pesticides by Nanogold/Mercaptomethamidophos Multi-Residue Electrochemical Biosensor," *Food Chemistry* 354 (2021): 129511.

2. M. Wei and J. Wang, "A Novel Acetylcholinesterase Biosensor Based on Ionic Liquids-AuNPs-Porous Carbon Composite Matrix for Detection of Organophosphate Pesticides," *Sensors and Actuators B: Chemical* 211 (2015): 290–296.

3. Y. Liu, G. Wang, C. Li, Q. Zhou, M. Wang, and L. Yang, "A Novel Acetylcholinesterase Biosensor Based on Carboxylic Graphene Coated With Silver Nanoparticles for Pesticide Detection," *Materials Science and Engineering: C* 35 (2014): 253–258.

4. Y. Wan, H. Wang, L. Zhang, et al., "Highly Stable Acetylcholinesterase Electrochemical Biosensor Based on Polymerized Ionic Liquids Microgel for Pesticides Detection," *Microchimica Acta* 189 (2022): 300.

5. S. Paneru and D. Kumar, "A Novel Electrochemical Biosensor Based on Polyaniline-Embedded Copper Oxide Nanoparticles for High-Sensitive Paraoxon-Ethyl (PE) Detection," *Applied Biochemistry and Biotechnology* 195 (2023): 4485–4502.

6. S. J. Alavinia, A. Mirvaghefi, H. Farahmand, et al., "DNA Damage, Acetylcholinesterase Activity, and Hematological Responses in Rainbow Trout Exposed to the Organophosphate Malathion," *Ecotoxicology and Environmental Safety* 182 (2019): 109311.

7. G. Xu, D. Huo, C. Hou, et al., "A Regenerative and Selective Electrochemical Aptasensor Based on Copper Oxide Nanoflowers-Single Walled Carbon Nanotubes Nanocomposite for Chlorpyrifos Detection," *Talanta* 178 (2018): 1046–1052.

8. Q. Zhang, X. Cao, Z. Zhang, and J. Yin, "Preparation of Magnetic Flower-Like Molybdenum Disulfide Hybrid Materials for the Extraction of Organophosphorus Pesticides From Environmental Water Samples," *Journal of Chromatography a* 1631 (2020): 461583.

9. D. Štajnbaher and L. Zupančič-Kralj, "Multiresidue Method for Determination of 90 Pesticides in Fresh Fruits and Vegetables Using Solid-Phase Extraction and Gas Chromatography-Mass Spectrometry," *Journal of Chromatography A* 1015 (2003): 185–198.

10. S. Zhang, C. Zhang, Y. Jia, et al., "Sandwich-Type Electrochemical Immunosensor Based on Au@Pt DNRs/NH_2-$MoSe_2$ NSs Nanocomposite as Signal Amplifiers for the Sensitive Detection of Alpha-Fetoprotein," *Bioelectrochemistry* 128 (2019): 140–147.

11. N. Zhang, Y. Si, Z. Sun, et al., "Lab-on-a-Drop: Biocompatible Fluorescent Nanoprobes of Gold Nanoclusters for Label-Free Evaluation of Phosphorylation-Induced Inhibition of Acetylcholinesterase Activity Towards the Ultrasensitive Detection of Pesticide Residues," *Analyst* 139 (2014): 4620–4628.

12. K. C. Gough, S. Patel, C. A. Baker, and B. C. Maddison, "Development of Immunoassays for the Detection of the Fungicide Penconazole and Its Urinary Metabolite," *Journal of Agriculture and Food Chemistry* 57 (2009): 9393–9399.

13. X. Zheng, L. Guo, C. Zhu, et al., "A Robust Electrochemical Sensor Based on AgNWs@ MoS_2 for Highly Sensitive Detection of Thiabendazole Residues in Food Samples," *Food Chemistry* 433 (2024): 137304.

14. J. B. Thakkar, S. Gupta, and C. R. Prabha, "Acetylcholine Esterase Enzyme Doped Multiwalled Carbon Nanotubes for the Detection of Organophosphorus Pesticide Using Cyclic Voltammetry," *International Journal of Biological Macromolecules* 137 (2019): 895–903.

15. S. Paneru and D. Kumar, "Ag-Doped-CuO Nanoparticles Supported Polyaniline (PANI) Based Novel Electrochemical Sensor for Sensitive Detection of Paraoxon-Ethyl in Three Real Samples," *Sensors and Actuators B: Chemical* 379 (2023): 133270.

16. S. Paneru, Sweety, and D. Kumar, "CuO@PEDOT: PSS-Grafted Paper-Based Electrochemical Biosensor for Paraoxon-Ethyl Detection," *Journal of Applied Electrochemistry* 53 (2023): 2229–2238.

17. L. He, B. Cui, J. Liu, et al., "Novel Electrochemical Biosensor Based on Core-Shell Nanostructured Composite of Hollow Carbon Spheres and Polyaniline for Sensitively Detecting Malathion," *Sensors and Actuators B: Chemical* 258 (2018): 813–821.

18. S. Paneru, Sweety, and D. Kumar, "CeO_2 and PEDOT: PSS Modified Conducting Paper for Organophosphate Pesticide Detection," *Journal of Applied Electrochemistry* 54 (2024): 1875–1885.

19. D. Song, Y. Wang, X. Lu, Y. Gao, Y. Li, and F. Gao, "Ag Nanoparticles-Decorated Nitrogen-Fluorine Co-Doped Monolayer MoS_2 Nanosheet for Highly Sensitive Electrochemical Sensing of Organophosphorus Pesticides," *Sensors and Actuators B: Chemical* 267 (2018): 5–13.

20. Y. Shen, X. Gao, H. Chen, Y. Wei, H. Yang, and Y. Gu, "Ultrathin C_3N_4 Nanosheets-Based Oxidase-Like 2D Fluorescence Nanozyme for Dual-Mode Detection of Organophosphorus Pesticides," *Journal of Hazardous Materials* 451 (2023): 131171.

21. Z. Gan, L. Liu, H. Wu, et al., "Quantum Confinement Effects Across two-Dimensional Planes in MoS_2 Quantum Dots," *Applied Physics Letters* 106 (2015).

22. T. V. Kumar, S. K. Yadav, and A. K. Sundramoorthy, "Electrochemical Synthesis of 2D Layered Materials and Their Potential Application in Pesticide Detection," *Journal of the Electrochemical Society* 165 (2018): B848–B861.

23. K. Ahmad, W. Raza, A. Alsulmi, and H. Kim, "Fabrication of Electrochemical Sensor for Metronidazole Using MoS_2/Graphite-Like Carbon Nitride Composite Modified Glassy Carbon Electrode," *Diamond and Related Materials* 138 (2023): 110178.

24. M. Govindasamy, S.-F. Wang, A. Almahri, and U. Rajaji, "Effects of Sonochemical Approach and Induced Contraction of Core–Shell Bismuth Sulfide/Graphitic Carbon Nitride as an Efficient Electrode Materials for Electrocatalytic Detection of Antibiotic Drug in Foodstuffs," *Ultrasonics Sonochemistry* 72 (2021): 105445.

25. S. Vinoth, K. S. Devi, and A. Pandikumar, "A Comprehensive Review on Graphitic Carbon Nitride Based Electrochemical and Biosensors for Environmental and Healthcare Applications," *TrAC Trends in Analytical Chemistry* 140 (2021): 116274.

26. D. Thakur, C. M. Pandey, and D. Kumar, "Graphitic Carbon Nitride-Wrapped Metal-Free PoPD-Based Biosensor for Xanthine Detection," *ACS Omega* 8 (2022): 2328–2336.

27. R. Nehru, C.-W. Chen, and C.-D. Dong, "In-Situ Growth of MoS_2 Nanosheets on g-C_3N_4 Nanotube: A Novel Electrochemical Sensing Platform for Vanillin Determination in Food Samples," *Carbon* 208 (2023): 410–420.

28. D. Tang, J. Shi, Y. Wu, et al., "Flexible Self-Powered Sensing System Based on Novel DNA Circuit Strategy and Graphdiyne for Thalassemia Gene by Rapid Naked-Eye Tracking and Open-Circuit Voltage," *Analytical Chemistry* 95 (2023): 16374–16382.

29. J. Shi, S. Liu, P. Li, et al., "Self-Powered Dual-Mode Sensing Strategy Based on Graphdiyne and DNA Nanoring for Sensitive Detection of Tumor Biomarker," *Biosensors and Bioelectronics* 237 (2023): 115557.

30. J. Ding, Y. Wang, S. Guo, et al., "Designing Efficient MoS_2/g-C_3N_4 Hybrid Photocatalysts by Regulating the Interlayer Spacing of MoS2," *European Journal of Inorganic Chemistry* 2021 (2021): 3719–3726.

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

31. S. Kamila, B. Mohanty, A. K. Samantara, et al., "Highly Active 2D Layered MoS_2-rGO Hybrids for Energy Conversion and Storage Applications," *Scientific Reports* 7 (2017): 8378.

32. S. Gao, Z. Yang, Y. Zhang, et al., "The Synergistic Effects of MoS_2 and Reduced Graphene Oxide on Sensing Performances for Electrochemical Chloramphenicol Sensor," *FlatChem* 33 (2022): 100364.

33. S. Verma, C. M. Pandey, and D. Kumar, "A Highly Efficient rGO Grafted MoS_2 Nanocomposite for Dye Adsorption and Electrochemical Detection of Hydroquinone in Wastewater," *New Journal of Chemistry* 46 (2022): 21190–21200.

34. O. Jalil, C. M. Pandey, and D. Kumar, "Highly Sensitive Electrochemical Detection of Cancer Biomarker Based on Anti-EpCAM Conjugated Molybdenum Disulfide Grafted Reduced Graphene Oxide Nanohybrid," *Bioelectrochemistry* 138 (2021): 107733.

35. Y.-Y. Hou, J. Xu, F.-T. Wang, et al., "Construction of an Integrated Device of a Self-Powered Biosensor and Matching Capacitor Based on Graphdiyne and Multiple Signal Amplification: Ultrasensitive Method for MicroRNA Detection," *Analytical Chemistry* 93 (2021): 15225–15230.

36. K.-J. Huang, Y.-J. Liu, H.-B. Wang, Y.-Y. Wang, and Y.-M. Liu, "Sub-Femtomolar DNA Detection Based on Layered Molybdenum Disulfide/Multi-Walled Carbon Nanotube Composites, Au Nanoparticle and Enzyme Multiple Signal Amplification," *Biosensors and Bioelectronics* 55 (2014): 195–202.

37. H.-L. Shuai, K.-J. Huang, Y.-X. Chen, L.-X. Fang, and M.-P. Jia, "Au Nanoparticles/Hollow Molybdenum Disulfide Microcubes Based Biosensor for MicroRNA-21 Detection Coupled With Duplex-Specific Nuclease and Enzyme Signal Amplification," *Biosensors and Bioelectronics* 89 (2017): 989–997.

38. C. Rajkumar and H. Kim, "Hybrid Nanostructures of Pd–WO3 Grown on Graphitic Carbon Nitride for Trace Level Electrochemical Detection of Paraoxon-Ethyl," *Microchimica Acta* 188 (2021): 233.

39. S. Ansari, M. S. Ansari, H. Devnani, S. P. Satsangee, and R. Jain, "CeO_2/g-C_3N_4 Nanocomposite: A Perspective for Electrochemical Sensing of Anti-Depressant Drug," *Sensors and Actuators B: Chemical* 273 (2018): 1226–1236.

40. S. khorrambin, S. Ghasemi, and S. R. Hosseini, "Ultrasonically Assisted Preparation of Molybdenum Sulfide/Graphitic Carbon Nitride Nanohybrid as Counter Electrode Material for Dye-Sensitized Solar Cell," *Journal of Alloys and Compounds* 929 (2022): 167220.

41. T. Ahamad, M. Naushad, Y. Alzaharani, and S. M. Alshehri, "Photocatalytic Degradation of Bisphenol-A With g-C_3N_4/MoS_2-PANI Nanocomposite: Kinetics, Main Active Species, Intermediates and Pathways," *Journal of Molecular Liquids* 311 (2020): 113339.

42. O. Jalil, C. M. Pandey, and D. Kumar, "Electrochemical Biosensor for the Epithelial Cancer Biomarker EpCAM Based on Reduced Graphene Oxide Modified With Nanostructured Titanium Dioxide," *Microchimica Acta* 187 (2020): 1–9.

43. D. Thakur, C. M. Pandey, and D. Kumar, "Highly Sensitive Enzymatic Biosensor Based on Polyaniline-Wrapped Titanium Dioxide Nanohybrid for Fish Freshness Detection," *Applied Biochemistry and Biotechnology* 194 (2022): 3765–3778.

44. E. J. J. Laviron, "General Expression of the Linear Potential Sweep Voltammogram in the Case of Diffusionless Electrochemical Systems," *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry* 101 (1979): 19–28.

45. Sweety and D. Kumar, "Development of $Ti_3C_2T_x$-Based Novel Immunosensor for Cancer Biomarker Detection," *Journal of Applied Organometallic Chemistry* 38 (2024): e7570.

46. X.-P. Wei, R.-Q. Zhang, L.-B. Wang, Y.-L. Luo, F. Xu, and Y.-S. Chen, "Novel Multi-Walled Carbon Nanotubes Decorated With Gold Nanoparticles With Poly (2-Methacryloyloxyethyl Ferrocenecarboxylate) Grafted on to Form Organic–Inorganic Nanohybrids: Preparation, Characterization, and Electrochemical Sensing Applications," *Journal of Materials Chemistry C* 7 (2019): 119–132.

47. R. Wang, Y. Wang, H. Qu, and L. Zheng, "An Acetylcholinesterase-Functionalized Biosensor for Sensitive Detection of Organophosphorus Pesticides Based on Solution-Gated Graphene Transistors," *Agricultural Science and Technology* 1 (2021): 372–378.

48. D. N. Kumar, S. Alex, N. Chandrasekaran, and A. Mukherjee, "Acetylcholinesterase (AChE)-Mediated Immobilization of Silver Nanoparticles for the Detection of Organophosphorus Pesticides," *RSC Advances* 6 (2016): 64769–64777.

49. F. Xu, Z.-M. Cui, H. Li, and Y.-L. Luo, "Electrochemical Determination of Trace Pesticide Residues Based on Multiwalled Carbon Nanotube Grafted Acryloyloxy Ferrocene Carboxylates With Different Spacers," *RSC Advances* 7 (2017): 7431–7441.

50. Q. Shi, Y. Teng, Y. Zhang, and W. J. C. C. L. Liu, "Rapid Detection of Organophosphorus Pesticide Residue on Prussian Blue Modified Dual-Channel Screen-Printed Electrodes Combing With Portable Potentiostat," *Chinese Chemical Letters* 29 (2018): 1379–1382.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

10990739, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/aoc.7721 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

Colloid and Polymer Science
https://doi.org/10.1007/s00396-024-05316-0

RESEARCH

A Schiff base hydrogel of oxidized okra gum and carboxymethylated chitosan: a biocompatible and biodegradable injectable system for drug delivery in wound care

Jitendra Kumar[1] · **Roli Purwar**[1]

Received: 9 July 2024 / Revised: 12 August 2024 / Accepted: 26 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
For the first time, functionalized okra gum with multi-aldehyde groups (OG-CHO) and carboxymethyl chitosan (CMCh) is used to create injectable hydrogels (IHs) via Schiff base reaction at 37 °C. Gelation time is optimized based on the ratio of aldehyde groups of OG-CHO to amine groups of CMCh (9 ± 3 to 2 ± 1 min). Physical characteristics such as gel content (84 ± 2) and porosity (66 ± 3) are assessed. The syringeability, injectability, and self-healing properties are evaluated qualitatively and quantitatively using a rheometer. Structural analysis is carried out by FT-IR and ^{1}H-NMR spectroscopy, while surface morphology and pore size (80 ± 5 µm) are examined via SEM. The swelling ratio is studied in the gel state of the IH in phosphate buffer saline (PBS) at varying pH levels of 5.5, 7.4, and 8.5, revealing a decrease in swelling ratio with increasing pH from 5.5 to 7.4 (75 ± 24 to 635 ± 20%), followed by an increase in swelling at pH 8.5 (724 ± 18.5%). Ciprofloxacin is employed as a model drug for release assays, and drug release behavior in gel forms of IH across different wound pH ranges is evaluated. The swelling and drug release behavior is described using the Korsmeyer–Peppas model, which shows non-Fickian diffusion. Furthermore, biocompatibility (cell viability > 90%), antibacterial assay, and in vitro degradation (~98%) are also assessed.

Keywords Carboxymethyl chitosan · Okra gum · Injectable hydrogel · Self-healing · Schiff base · Drug delivery

Introduction

Recent advancements in drug delivery have been propelled by the quest for safer, more patient-friendly, and effective treatment methods. Among the innovative materials making waves in this field are hydrogels. These hydrogels offer precise control over the release kinetics of various therapeutic agents, including proteins, peptides, and small molecules, by encapsulating and delivering them [1]. Injectable hydrogels, in particular, provide a less invasive means of localized drug administration, offering advantages such as reduced systemic side effects and improved patient compliance [2]. Natural polymers, especially polysaccharides, have gained popularity in hydrogel formulation due to their biocompatibility, biodegradability, and ability to interact with biological systems [3]. Examples of these natural polymers include *Moringa oleifera* gum [4], chitosan [5], Acacia gum [6], and carrageenan gum [7]. Schiff base crosslinking, formed by the reaction between a primary amine and an aldehyde, offers several benefits for the design of injectable hydrogels, including mild reaction conditions, controlled crosslinking, and the potential for customizable mechanical and drug release properties [8]. These Schiff-based injectable hydrogels present a promising option for drug delivery, exhibiting strong stability and the ability to respond to external stimuli dynamically [9].

Carboxymethyl chitosan, a water-soluble chitosan derivative, was used in various biomedical applications and is a popular material for synthesizing injectable hydrogel via Schiff base reaction [9]. Another biopolymer, okra gum (OG), or lady's finger gum, is a natural polysaccharide extracted from the pods of the okra plant (*Abelmoschus esculentus*) [10]. The chemical structure of okra gum comprises a complex polysaccharide matrix that includes rhamnose, galactose, and glucose as its primary constituents

✉ Roli Purwar
roli.purwar@dtu.ac.in

[1] Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University, Delhi, India

Published online: 12 September 2024

[11]. The molecular weight of okra gum is approximately 1.7 million [12]. The gum is non-toxic and biodegradable and has excellent emulsifying and binding properties, and it is soluble in hot and cold water and stable over a wide pH range [13]. Okra gum has been explored in the food industry [14] [15], wastewater water treatment [19], and bio-adhesive agents [16]. Okra gum has been studied for its potential role in drug delivery systems, where it has shown promise as a film coating agent, emulsifying agent, stabilizer, and thickener [17, 18]. Deka and coworkers synthesized glutaraldehyde crosslinked hydrogel using okra gum, guar gum, and polyvinyl alcohol to release nanocurcumin [19]. Maryam et al. synthesized an okra gum/gelatin hydrogel by an aqueous polymerization technique crosslinked with methylene bisacrylamide, used for cytarabine release [20]. Choudhary and colleagues published an article in which a hydrogel was synthesized using okra polysaccharide and grafted with polyacrylic acid crosslinked with ethylene glycol dimethacrylate for the release of *Calendula officinalis* [21].

To the best of the author's knowledge, there is no such injectable hydrogel synthesized that utilized oxidized okra gum and carboxymethyl chitosan as biopolymer precursors using Schiff base crosslinks. Thus, it was hypothesized that the combination of multi-aldehyde okra gum and carboxymethyl chitosan would form an injectable hydrogel through dynamic Schiff base chemistry at body temperature (37 °C), exhibiting desirable properties such as easy syringe ability, injectability, and self-healing capabilities without additional crosslinkers. This injectable hydrogel will demonstrate favorable surface morphology, controlled swelling behavior, and sustained drug release kinetics. Furthermore, the hydrogel will show excellent biocompatibility as evidenced by low cytotoxicity, minimal hemolysis, and controlled hydrolytic degradation, making it a promising candidate for biomedical applications. All the reported literature on okra gum utilizes toxic crosslinkers to synthesize hydrogels with no self-healing properties. The addition of any crosslinker has added one more step to complete the synthesis process, while there is no requirement for crosslinker in this research article. Therefore, the present study is more economical and non-toxic.

Thus, the novelty of this research article includes the use of okra gum biomaterial, the absence of a toxic crosslinker, and the use of a gel form of injectable hydrogel for describing the swelling study, drug release study, and hydrolytic degradation.

This study describes the synthesis process of multi-aldehyde okra gum using $NaIO_4$ as an oxidizing agent and CMCh to design the injectable hydrogel at 37 °C. The aldehyde groups of OG-CHO can form covalent crosslinks with the NH_2 groups of CMCh via Schiff base chemistry, eliminating the need for hazardous or chemical crosslinking agents. The research delves into syringeability, injectability, self-healing properties, swelling behavior, and drug release profiles in PBS under various wound pH conditions. The biocompatibility, antibacterial assays, and in vitro degradation of the synthesized injectable hydrogels are also measured.

Experimental

Materials

Chitosan was bought from Otto Chemical Pvt. Ltd. with a deacetylation degree of 85%, $M_v = 2{,}18{,}112$ Da, pH 7–9, and viscosity (1% w/v, 200 cps). Carboxymethyl chitosan was prepared in the laboratory (degree of substitution (DS) = 67 ± 2.1%). Okra gum was obtained from Indianjadibooti Pvt. Ltd. Tamil Nadu, India. The viscosity average molecular weight of OG was calculated in the laboratory ($M_v = 3{,}51{,}524$ Da). Sodium hydroxide, sodium periodate ($NaIO_4$), and hydroxylamine hydrochloride were purchased from CDH chemicals. Ethylene glycol, isopropyl alcohol, and monochloroacetic acid were obtained from Thomas Baker Chemicals Pvt. Ltd. Ethanol was purchased from SRL Chemicals, India. Ciprofloxacin HCl was received as a gift from M/S Unicure India Limited. Double-distilled water (MQ) was employed for all reactions. Chemicals and solvents were used as received.

Methods

Oxidation of okra gum

Oxidation of okra gum was carried out to generate aldehyde groups using sodium periodate. Briefly, a known quantity of okra gum was dissolved in MQ water for 24 h at 400 RPM at 50 °C. Centrifugation was carried out at 3000 RPM for 15 min to eliminate the undissolved content of okra gum, followed by the separation of supernatant, which contains only the soluble part of okra gum. The soluble part was dried in a hot air oven (yield = 85 ± 5%) and dissolved in MQ water (1% w/v). Then, sodium periodate (1.5 g) was incrementally added to the dissolved OG solution, and the reaction was continually stirred for 24 h at 50 °C in the dark to prevent auto-acceleration. The ethylene glycol was used to terminate the reaction. The solution was dialyzed against MQ water for 4 days to eliminate the unreacted product. Then, the oxidized okra gum (OG-CHO) solution was poured into a polycarbonate petri dish and put in a hot air oven at 50 °C. The pH metric titration was used to determine the aldehyde percentage of OG-CHO, which was found to be 56 ± 1.7% [22]. A triplicate study was conducted, and a mean value of three was reported.

Derivatization of chitosan

The carboxymethylation of chitosan was conducted using the method followed by Kumar et al. [22]. Five-gram chitosan was dispersed in isopropyl alcohol (80 mL) at 27 °C, continuously stirring at 500 RPM. Subsequently, 10 mol/L NaOH (25 mL) was added and stirred for 30 min. Monochloroacetic acid (10 g) was added gradually, and the temperature was elevated to 60 °C for 3 h. Furthermore, the suspension was cooled, filtered, and washed three times with ethanol. The obtained modified chitosan was dried in a hot air oven at 50 °C.

The DS % of CMCh was determined using a back-titration method. In this procedure, 0.5 g of CMCh was dissolved in 0.04 M NaOH (100 mL) and then titrated against 0.05 M HCl using phenolphthalein as an indicator. The analysis was performed in triplicate, and the DS % was calculated using the following equation:

$$S\% = \frac{0 \cdot 162\, A}{1 - 0.058\, A} \times 100$$

$$A = \frac{(\text{Conc of NaOH} \times \text{ Vol of NaOH}) - [\text{Conc of HCl} \times (\text{Vb} - \text{Vs})]}{m}$$

Vb is the volume of HCl consumed without CMCh.

Vs is the volume of HCl consumed with CMCh.

m is the weight of dry CMCh.

Preparation of injectable hydrogel

The known quantity of OG-CHO was dissolved in MQ water (precursor 1) for 4 h at 700 RPM and 27 °C. A similar procedure was used to dissolve CMCh (5% w/v) in PBS at pH 7.4 for 700 RPM at 27 °C (precursor 2). After that, both precursors were simultaneously extruded into a vial through two separate syringes of 26 gauzes (G) at 37 °C. After some time, gelation occurred and fabricated injectable hydrogels termed OG1C1 (OG-CHO:CMCh = 1:1), OG2C1 (OG-CHO:CMCh = 2:1), and OG1C2 (OG-CHO:CMCh = 1:2).

Characterization

The Fourier transform spectroscopy test was performed in ATR mode on a PerkinElmer 2000 instrument for qualitative analysis. The samples were analyzed in the frequency range of $500 - 4000\ \text{cm}^{-1}$ and a resolution of $4\ \text{cm}^{-1}$. ^{1}H NMR was recorded on Bruker advance III, 500 MHz, using D_2O as the solvent. Chemical shift (δ) values were reported in ppm using tetramethyl silane as an internal reference. The morphological features were studied in a scanning electron microscope (ZEISS, EVO 18). The double adhesive carbon tape was used to mount gold-coated samples. The photomicrographs were recorded at different magnifications.

Physical characterization

The evaluation of gelation time using an inverted tube test, gel content, porosity content, and syringeability was determined as per the protocol described by Kumar et al. [23]. The two precursor solutions were extruded into a 26 G syringe with a 10 mL capacity for injectability determination. After the gel formation, manual pressure was applied to the syringe plunger to extrude the synthesized IH. When deposited onto a glass petri dish, the extruded hydrogel formed a distinct shape ("J.K." and "☺").

The gelation time, syringeability, injectability, and self-healing properties were determined using a rheometer. At 37 °C, these physical characteristics were examined using a parallel plate geometry of 40 mm diameter (PP-40), MCR-101 rheometer (Anton Paar, Japan). A 1-mm gap between the PP-40 and the plate was maintained throughout each test. An oscillatory time sweep study in isothermal mode investigated the gelation time. Precursors 1 and 2 were dispensed onto the rheometer platform at 37 °C in a different ratio and allowed to crosslink on the platform. This process was part of a time sweep study. To evaluate syringeability, injectability, and self-healing properties, the IH was incubated for 30 min at 37 °C. A flow sweep experiment was performed within the 0.1–100 s^{-1} range to evaluate the syringeability and injectability. An amplitude sweep test was employed to determine the IH linear viscoelastic region (LVE). Additionally, these IHs underwent frequency sweep tests covering the frequency range of 0.1–100 rad s^{-1} at a 1% amplitude. The continuous variation strain amplitude method was employed to characterize self-healing properties. An alternative strain sweep test with a fixed angular frequency of 1 rad/s was carried out. An amplitude oscillatory strain test was performed with intervals of 100 s for minor strain (1%) and 80 s for high strain (200%) [24].

Swelling ratio, network parameter, and swelling kinetics

Freshly synthesized IHs in gel form were put into the 100 mL PBS solution at the specified pH levels (5.5, 7.4, and 8.5) and temperature (37 °C). The swelling ratio (S %) was determined based on the amount of water absorbed by the IHs using Eq. 1. The measurements were carried out on three samples, and the mean value of three was reported.

$$S\ \% = \left[\frac{\text{Wa} - \text{Wb}}{\text{Wb}}\right] \times 100 \quad (1)$$

where Wb is the dried weight of the injectable hydrogel and Wa is the swollen weight of the injectable hydrogel.

The network parameters of injectable hydrogel were calculated in terms of the number of average molecular mass present between crosslinks ($\overline{M}_c$), crosslink density (ρ), and mesh size (ξ) as described by Kumar et al. [22]. To determine the kinetics of solvent diffusion into the IH, the Korsmeyer–Peppas (K-P) equation was used, which is represented as

$$\frac{S_t}{S_\infty} = kt^n$$

where "S_t" and "S_∞" are the mass of absorbed water at a time "t" and at equilibrium, respectively. The release exponent "n" describes the release mechanism, and "k" is the kinetic constant of the IH.

Drug loading and release kinetics

Ciprofloxacin was used as a model drug to carry out the drug release study. The calibration curve for the drug was made using a UV–visible spectrophotometer. The OG-CHO solutions were combined with a 3 mL aqueous solution of Cipro (6 mg mL^{-1}) to make precursor 3. To prepare the drug-loaded IH, precursors 2 and 3 were mixed in equal parts. The gel form of OG2C1 was put in 100 mL of a PBS solution with pH values of 5.5, 7.4, and 8.5 at 37 °C. Aliquots from each medium were removed every 2 h and concurrently replaced with fresh media to maintain the same overall volume of the release medium for every drug release test. It was considered that the drug molecules were uniformly distributed throughout OG2C1. The observations were made in triplicate, and the mean value of three was reported. The amount of antibiotic released was determined using UV–visible spectroscopy at an absorbance of 275 nm. The experimental data was fitted with the K-P model, which is represented as

$$\frac{M_t}{M_\infty} = kt^n$$

The cumulative drug concentrations released at a time "t" and time infinity are denoted by "M_t" and "M_∞," respectively. The release exponent "n" describes the release mechanism, and "k" is the kinetic constant of the IH.

Antibacterial activity

The antibacterial activity was assessed using the zone inhibition method, also known as the Kirby–Bauer method. *S. aureus* (Gram-positive) and *E. coli* (Gram-negative) were used as representative bacteria models. Mueller–Hinton agar plates were inoculated by spreading with 100 µL of each bacterium culture (approximately 1.5×10^8 CFU mL^{-1}), followed by placing discs containing 10 µL of various concentrations of IH ranging from 0 to 100 mg mL^{-1}. Each plate had one disc (5 mm) packed with solvent alone, serving as the vehicle control, while a ciprofloxacin disc (10 µg) was used as the positive control. The plates were then incubated at 37 °C for 24 h. The clear zones formed around the discs were measured and recorded to evaluate the antibacterial activity. Three replicates were conducted for each compound against each bacterium [25].

Hemolysis activity

Human blood underwent centrifugation at 1000 g for 15 min at 4 °C to separate plasma and buffy coat, which were then removed. The erythrocytes were washed three times with PBS and prepared as a 1% suspension in PBS. Subsequently, 500 µL of the 1% erythrocyte suspension was dispensed into a 96-well plate and incubated for 30 min at 37 °C in the presence or absence of 100 µL test samples across a concentration range of 1–1000 µg mL^{-1}. After incubation, 400 µL of PBS was added to the samples. For the positive control representing 100% hemolysis, 400 µL of RBC lysis buffer was added to the control sample. Following centrifugation, 100 µL of the supernatant solution was transferred to a fresh 96-well plate, and the absorbance was measured at 540 nm using a microplate reader (iMark, Bio-Rad) [26]. The observations were made in triplicate, and the mean value was reported.

In vitro cytotoxicity activity

The cytotoxicity of the IHs was evaluated using the MTT ((3-(4,5-dimethylthiazolyl-2)-2, 5-diphenyltetrazolium bromide)) assay using L929 fibroblast cell lines. The cells (5000–8000 per well) were cultured in DMEM (Dulbecco's Modified Eagle Medium) media supplemented with 10% FBS (fetal bovine serum) and 1% antibiotic solution for 24 h at 37 °C with 5% CO_2 in 96-well plates. The following day, cells were treated with 1 to 1000 µg mL^{-1} formulations. MTT solution was added to the cell culture and incubated for an additional 2 h at a final concentration of 250 µg mL^{-1}, following a 24 h incubation period. After this, the culture supernatant was extracted, and the cell layer matrix was dissolved in 100 µL DMSO (dimethyl sulfoxide). Absorbance was measured at 540 and 660 nm using an Elisa plate reader (iMark, Bio-Rad, USA) and computed using Graph Pad Prism 6 software. Images were captured under an inverted microscope (Olympus ek2) using a camera (AmScope digital camera 10 MP Aptima CMOS) [27]. Three measurements were assessed on each concentration, and the mean was reported.

In vitro degradation

The freshly prepared gel form of IH was cut into slices and put in PBS (pH 7.4) at 37 °C. Equation 2 was used to check the degradation percentage [28]. The analysis was evaluated on thee samples and mean value was reported.

$$\text{Degradation } \% = \frac{X_a - X_b}{X_a} \times 100 \quad (2)$$

Statistical analysis

The experimental data are presented as the mean ± standard deviation. All experiments were repeated at least three times. Statistical significance was estimated using one-way analysis of variance (ANOVA), and the significance level was considered $p < 0.05$. All statistical analyses were performed using the Microsoft Excel 2010 software.

Results and discussion

Physical properties

The gelation time of an IH is an important parameter that determines its practical usability and clinical application. The time at which the liquid precursor solution transforms into a stable, three-dimensional hydrogel network structure at 37 °C is defined as gelation time. The ratio of OG-CHO to CMCh varied, and the effect on gelation time was observed. The tube inversion test shows that OG1C2, OG1C1, and OG2C1 gelation times are 10 ± 2, 5 ± 1.2, and 3 ± 1.5 min, respectively (Fig. 1a). IHs can have a gelation time ranging from a few seconds to several hours, depending on the total polymer concentration or ratio of liquid precursor solutions [29]. The ratio of aldehyde groups of OG-CHO to amine groups of CMCh was calculated using the simple unitary method and reported in Table 1. It was observed that as the ratio of the aldehyde group of OG-CHO to the amine group of CMCh increased, the time of gel formation decreased, primarily because a total number of aldehyde functional groups is more likely to react with the amine groups of CMCh. The oscillatory time sweep study on the rheometer was carried out to calculate the gelation time. The gelation point is a crossover point where the storage modulus (G') overcomes the loss modulus (G''). The time required to reach the gel point is called the gelation time. Figure 1b–d shows the oscillatory time sweep behavior of IHs. The gelation time of OG1C2, OG1C1, and OG2C1 was 9 ± 3, 5 ± 1.5, and 2 ± 1 min, respectively, which supports the tube inversion test. Since OG2C1 has the shortest gelation time, it was considered an optimized IH for further characterization. A similar article was published by Zhang et al. in which an injectable hydrogel was synthesized using carboxymethyl chitosan and hyaluronic acid through Schiff base and acyl hydrazide bonding, where gel time formation using tube inversion test supported the time sweep study conducted by a rheometer [30].

During the amplitude sweep test of OG2C1, it underwent varying strain levels ranging from 0.1 to 100% at 37 °C (Fig. 1e) to delineate the LVE. The LVE ranges from 0.1 to 5% strain. A 1% strain was selected for all gelation kinetic studies. The maximum G' and G'' were recorded at 138 ± 23 Pa and 5 ± 1.2 Pa, respectively. Across different strain levels, the IHs consistently exhibited higher G' values than G'', with a crossover point at 32 ± 1% shear strain, indicating a transition in the physical state from solid to liquid.

A frequency sweep test was employed to explore the rheological characteristics of OG2C1, as illustrated in Fig. 1g. The G' was consistently more significant than the G'', indicating its stiffness and elasticity. This elevated G' (360 Pa) observed in OG2C1 can be attributed to its high crosslink density, which restricts chain mobility and deformability.

Table 2 presents the physical characteristics such as gel content, porosity, syringeability, and network parameters. The gel content and porosity for OG2C1 are recorded as 84 ± 2% and 66 ± 3%, respectively. The higher gel content and lower porosity of injectable hydrogels are attributed to effective crosslinking via imine bonds. This results in the construction of a more compact 3D structure formation. Yan and coworkers have fabricated an injectable hydrogel using Schiff base chemistry between poly(L-glutamic acid) and chitosan, in which the observed gel content was 79% [31]. Another group of researchers published a research article based on Schiff base crosslinking using oxidized galactomannan and N-succinyl chitosan, in which they evaluated a porosity determination of 69% [32].

The calculated syringeability percentage was nearly 99 ± 0.5%, which confirmed the synthesized good syringeability ability when extruded through a 26 G needle. In this study, the injectability evaluation was successfully carried out using a 26 G needle. Figure 1h shows an image of the extruded OG2C1-forming different structure ("J.K." and "☺") when expressed from a 10-mL capacity 26 G needle into a glass petri dish. The easy flow through the needle confirmed the hydrogel's ready accessibility for administration. Vahedi et al. synthesized an injectable hydrogel using gelatin and polyethylene glycol dibenzaldehyde by Schiff base reactions, in which they evaluated the injectability test using a 20 G syringe and formed a letter "R" and "N" [30]. A continuous flow sweep test was conducted at 37 °C to evaluate the shear thinning characteristics, ranging from 1 to 1000 1/s with a constant 1% strain, wherein OG2C1 was placed between the parallel plates of

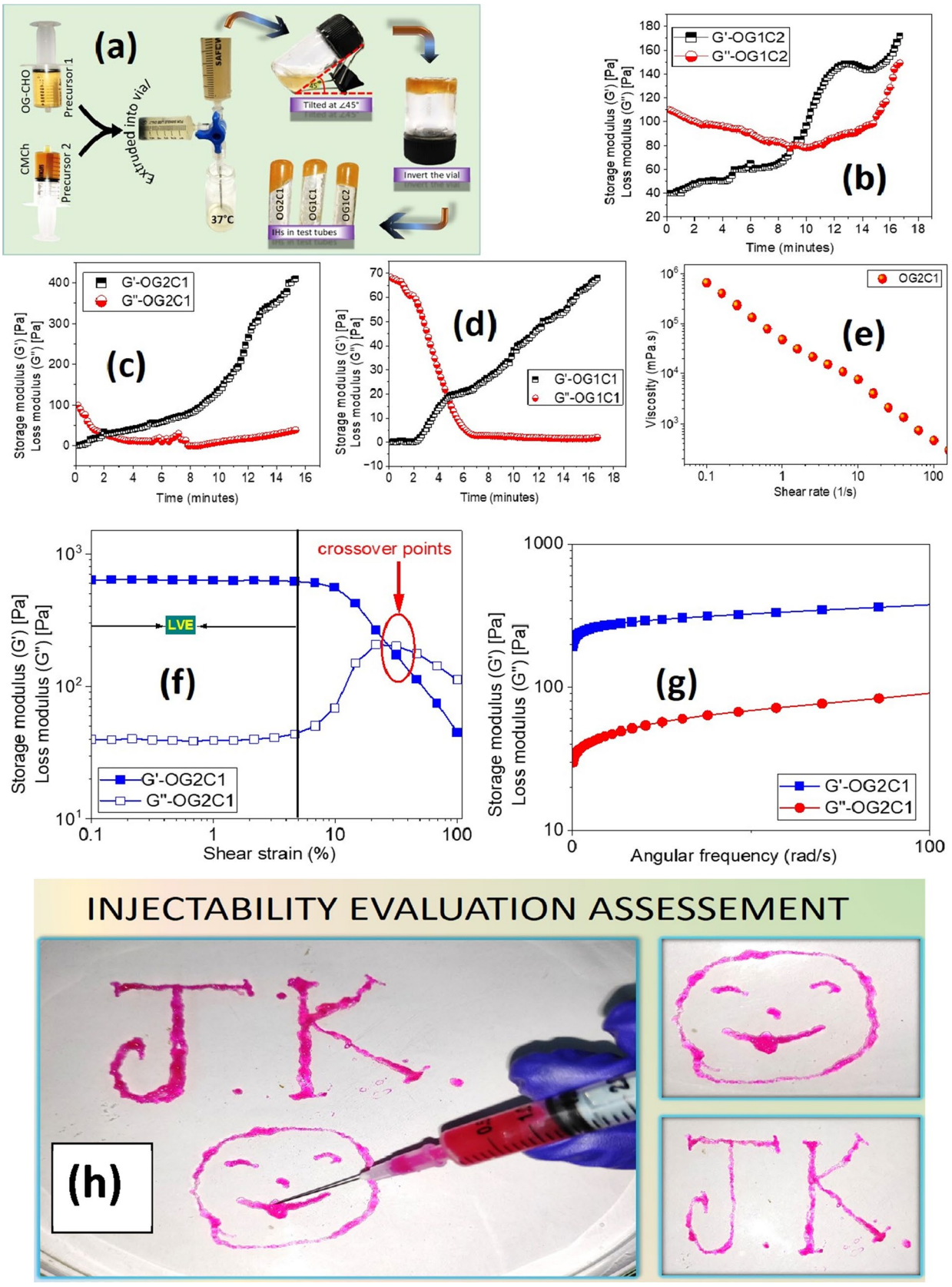


Fig. 1 **a** Tube inversion test. Time sweep study of **b** OG1C2, **c** OG2C1, and **d** OG1C1. **e** Flow sweep test, **f** amplitude sweep test, **g** frequency sweep test, and **h** injectability test

Table 1 The concentration of precursors for the synthesis of injectable hydrogels

Injectable hydrogels	Conc. of aldehyde group in OG-CHO (%) (*A*)	Volume of OG-CHO (mL) (*B*)	Conc. of amine group in CMCh (%) (*C*)	Volume of CMCh (mL) (*D*)	Ratio = $\frac{A \times B}{C \times D}$	Gelation time (min)
OG1C2	56 ± 1.7	1	67 ± 2.1	2	0.42	9 ± 3
OG1C1	56 ± 1.7	1	67 ± 2.1	1	0.82	5 ± 1.5
OG2C1	56 ± 1.7	2	67 ± 2.1	1	1.67	2 ± 1

Table 2 Physical properties of OG2C1

Property	Value OG2C1
Gel content	84 ± 2%
Porosity	66 ± 3%
Syringeability	99 ± 0.5%
Average molecular weight between crosslinks	2.8×10^2 g mol^{-1}
Crosslink density	0.9 mol cm^{-3}
Mesh size	7.5×10^{-3} nm

the instrument. As the shear rate increased, a reduction in viscosity was observed (from 650,840 to 41 mPa·s), indicating the shear thinning behavior. This phenomenon may be attributed to the rupture of Schiff base linkages at higher shear rates (Fig. 1e). The shear thinning property of the injectable hydrogel further corroborates the findings on its syringeability and injectability.

The self-healing ability of the IHs is enabled via the development of a crosslinked system with dynamic and reversible bonding to allow the capacity to heal in a timescale of seconds, which is a requirement for any application in tissue restoration [33]. To demonstrate the self-healing property, two differently colored IHs were fabricated and cut into semicircular shapes. The cut surfaces were then brought together to allow self-healing. It was observed that the cut surface gradually healed during the healing process. After 1 h, the two samples had integrated, indicating that the cut surfaces had healed together by reconstructing reversible Schiff-base bond linkages (Fig. 2a). A series of rheological recovery tests were conducted to assess the quantitative analysis of self-healing properties of OG2C1. The IH restorative capacity was examined through oscillatory step-strain measurements. At low strain levels (1%), the G' value exhibited a significantly high magnitude. However, as the strain increased to higher values (200%), the G' value approached G'', indicating structural deformation under high strain conditions (Fig. 2b). These behaviors remained stable across multiple cycles, underscoring the self-healing capability and durability of the dynamic Schiff base and dynamic coordination network within the injectable hydrogel. After completing all cycles of alternating strain from low to high, the recovery percentage (Eq. 3) was calculated and found to be 78 ± 4.2%.

$$Recovery\% = 100 - \frac{initial\ G' - final\ G'}{initial\ G'} \times 100 \quad (3)$$

Li et al. reported the synthesis of an injectable hydrogel using Schiff base crosslinks between dialdehyde-modified hyaluronic acid and cystamine dihydrochloride. The researchers evaluated the self-healing properties of the synthesized IH both qualitatively and quantitatively. Their results showed that the IH could self-heal within 2 h [34].

Additionally, due to the shear-thinning property of the hydrogel (viscosity decreases with increasing shear stress), OG2C1 exhibited extrusion ability through a needle. After injection, it could regain its gel state in different shapes, supporting good injectable properties. These properties make it suitable for minimally invasive drug delivery, potentially reducing scarring and pain and promoting faster patient recovery. The results suggest that the dynamic reaction among functional groups is the main driving force behind the OG2C1's self-healing capability, rather than simple physical adhesion between the broken interfaces. The constant cleavage and regeneration of imine bonds on Schiff-base bonds allow the interaction between CMCh molecules of the separated networks, creating space for network reconstruction. Consequently, new Schiff-base bonds form on the contact surfaces of the separated OG2C1, leading to the reconstitution of the network.

Structural properties

Structural properties of OG2C1 were characterized using FT-IR spectroscopy and ^{1}H NMR. FT-IR spectra of OG (Fig. 3a1) demonstrate a broad band at 3404 cm^{-1}, which proves carboxylic acid existence, and this is due to O–H. Near the carboxylic acid band, another peak at 2937 cm^{-1} is visible, which represents methylene (CH_2) functional group. The peak at 1645 cm^{-1} represents amide I occurrence. Due to the hydroxyl group, the deformation band is visible at 1426 cm^{-1}. The peak at 894 was because of polysaccharides. After OG reacts with sodium periodate, significant peaks were observed at 1724 cm^{-1}, mainly attributed to the

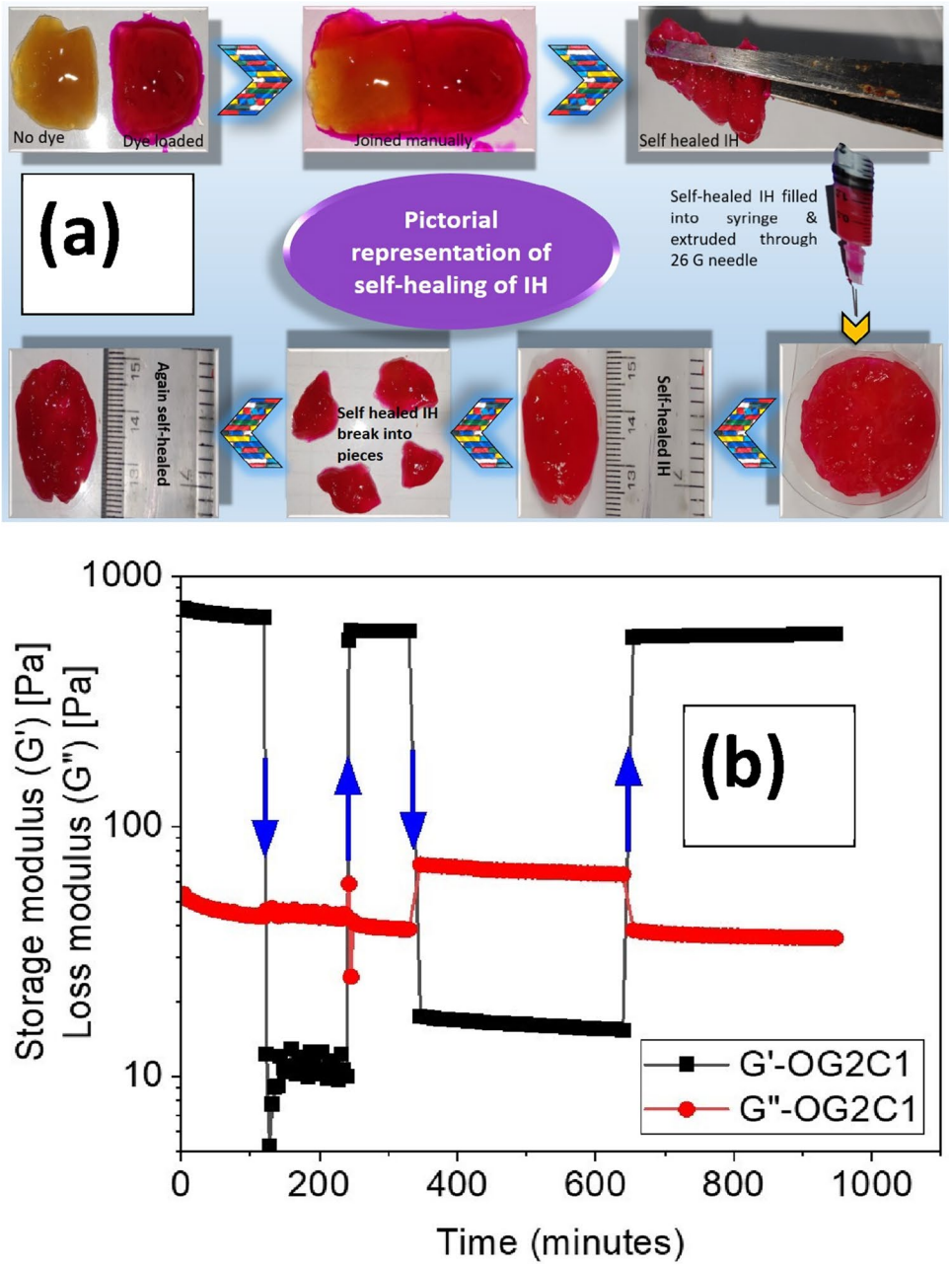


Fig. 2 **a** Pictorial representation of self-healing of OG2C1. **b** Step-strain analysis of OG2C1

production of symmetrical vibration of aldehydes bonds (Fig. 3a2). Another comparative research found similar outcomes for the oxidized product [35]. Our previous study demonstrated the FT-IR spectrum of chitosan and CMCh [22]. Figure 3b1 shows the FT-IR spectra of OG2C1, with a peak at 1642 cm^{-1}, which demonstrates imine structure formation due to dynamic Schiff base crosslinking reaction between the amine group of CMCh and the aldehyde groups of OG-CHO (Fig. 3b2) [36].

Proton NMR also supports structural properties. The ^{1}H-NMR spectrum of chitosan, CMCh, OG, and OG-CHO is described using D_2O as a solvent. ^{1}H-NMR spectra of the chitosan show a peak at 1.24 ppm due to protons in the acetyl group. The anomeric protons of glucosamine show a peak at 2.96 ppm. In comparison, the non-anomeric protons of the chitosan structural backbone are responsible for the peaks ranging from 3.68 to 4.84 ppm (Fig. 4a). Additionally, a new peak was observed at 3.87 ppm due to the presence of the $–CH_3COO^-$ group (Fig. 4b, e).

Due to the presence of the methyl group of rhamnose, the spectra displayed congested signals in the 1.23 and 1.92 ppm spectrum bands. The peak available at 3.60–3.69 was due to the –OH and CH of mannose. Signals at chemical shift equal to 3.81–3.60 ppm, assigned to the CH_2 group of arabinose.

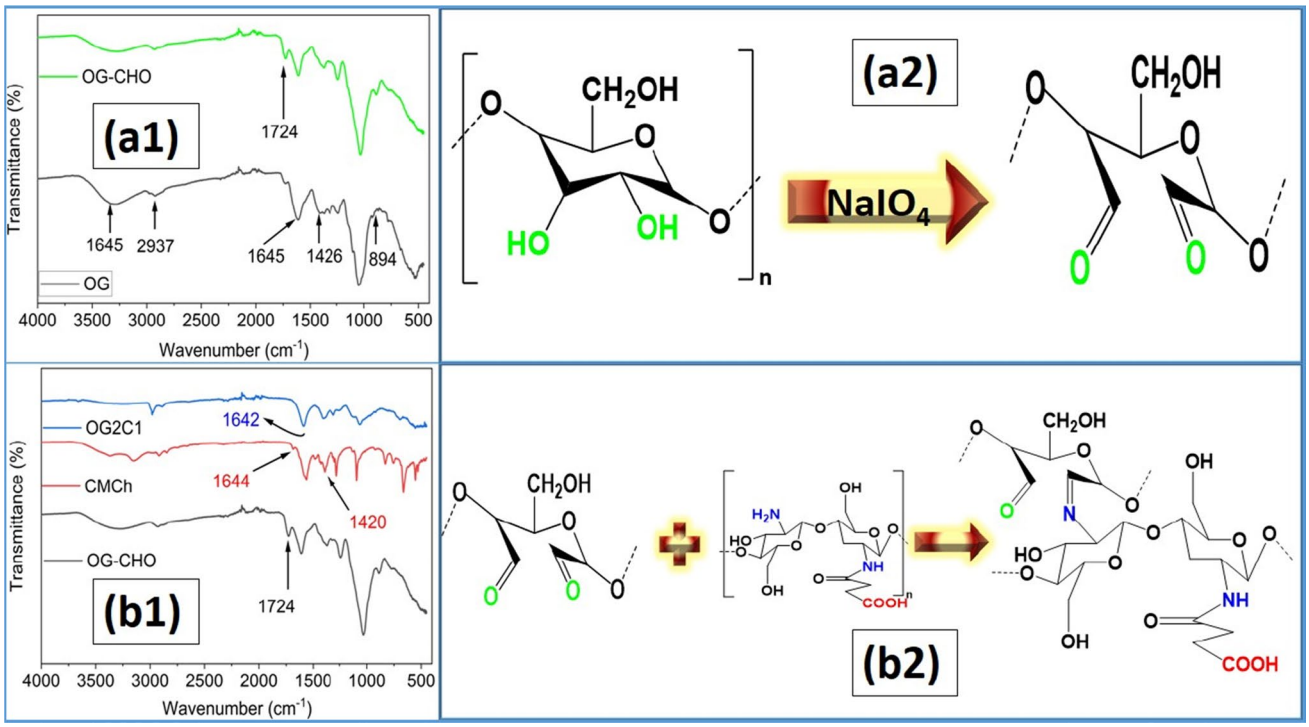


Fig. 3 FT-IR spectra of **(a1)** OG, OG-CHO and **(b1)** OG-CHO, CMCh, OG2C1. Reaction scheme a2 oxidation of OG and (**b2**) synthesis of IH

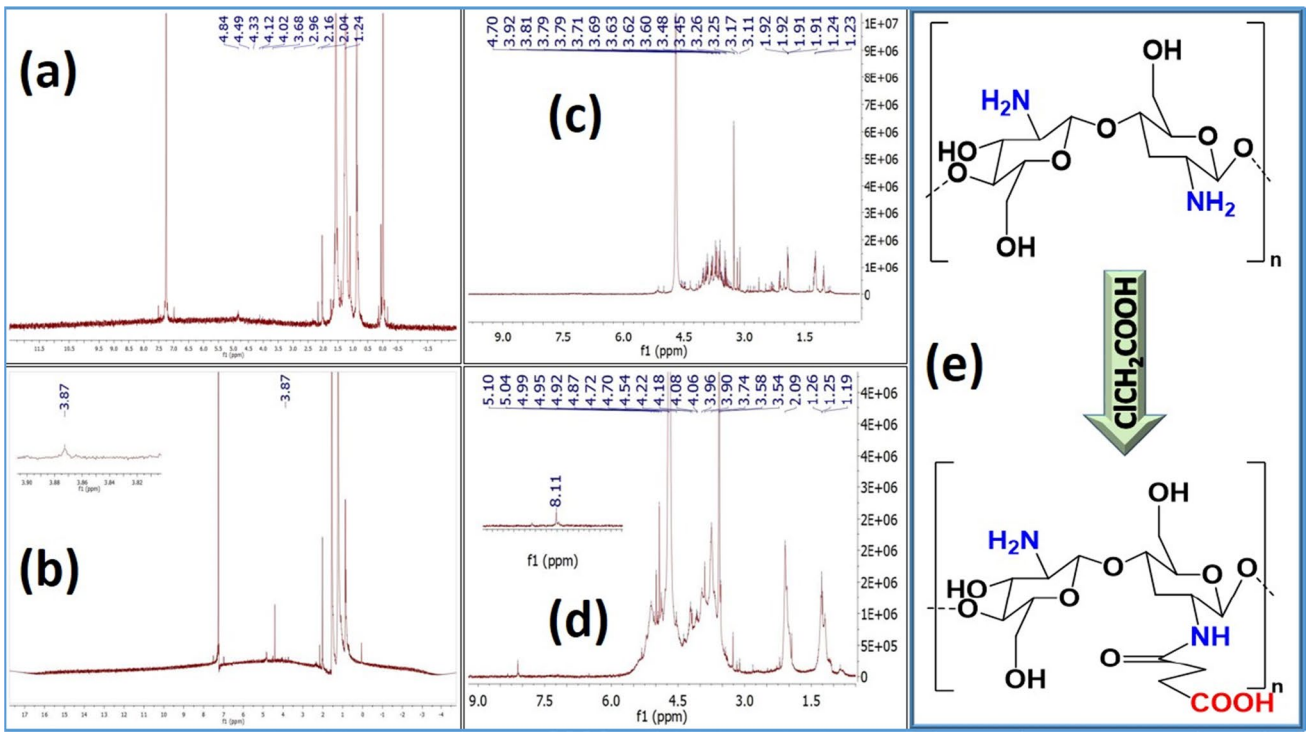


Fig. 4 [1]H-NMR spectra of **a** chitosan, **b** CMCh, **c** OG, and **d** OG-CHO and **e** reaction scheme of chitosan to CMCh

Signals in the 3.63 and 3.79 ppm range are transferred to the proton linked to C_6 and C_4 of galactose [37] (Fig. 4c). A new distinctive peak of the aldehydic proton is observed at 8.11 ppm, confirming OG oxidation (Fig. 4d).

Additionally, the M_v of OG-CHO was found to be 2,29,387 Da, which is lower than the original molecular weight of OG (M_v = 3,51,524 Da). This suggests that the periodate oxidation process of functionalizing the OG with aldehyde groups led to ring-cleavage reactions, decreasing the molecular weight. A similar study was reported by Li et al. in which they fabricated an injectable hydrogel using dialdehyde-modified hyaluronic acid and cystamine dihydrochloride. Original hyaluronic M_v was higher (1.28×10^6 Da) than the oxidized hyaluronic acid M_v (3.3×10^5 Da) [34].

Morphological evaluation

The SEM images were captured at × 5000. The surface morphology of OG2C1 reveals a scaly surface with a porous network structure in the form of several rings. These circular-shaped structures are dense pores interconnected to irregular surfaces. The ImageJ software evaluated the average pore diameter (80 ± 5 µm, Fig. 5a, b). The porous structure is advantageous for drug delivery applications as it allows for easy diffusion of drugs from the IH.

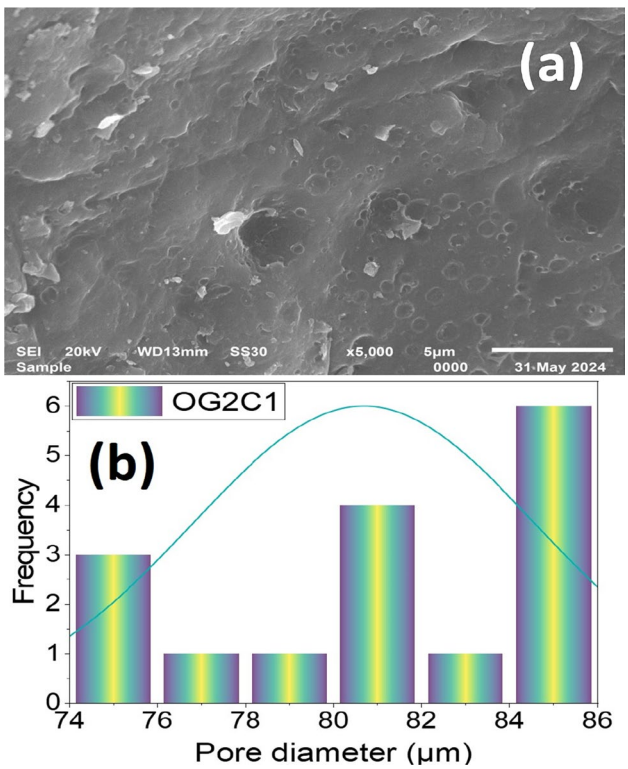


Fig. 5 **a** SEM image at × 5000. **b** Histogram of OG2C1

Swelling study

The swelling behavior of the gel form of OG2C1 was studied (Fig. 6). The observed swelling ratio percentages are 750 ± 24% at pH 5.5, 635 ± 20% at pH 7.4, and 724 ± 18.5% at pH 8.5. The swelling ratios of OG2C1 decrease as the pH increases from 5.5 to 7.4. However, the swelling ratios increase when the pH is further raised to 8.5. At acidic pH 5.5, the amino groups of the CMCh undergo protonation. This leads to electrostatic repulsion between the inter and intrapolymeric chains, improving water uptake and resulting in a higher swelling ratio. In contrast, at the more neutral pH of 7.4, deprotonation occurs within the amine groups of CMCh. This reduces the swelling ratio of OG2C1. When the pH is increased to 8.5, the IH network structure becomes looser, which results in even higher swelling ratios. Overall, the swelling ratio of gel forms of OG2C1 depends on the pH of the solution. An injectable hydrogel was synthesized using *Moringa oleifera* gum and carboxymethyl chitosan crosslinked through the Schiff base mechanism; swelling studies were conducted with a gel form of injectable hydrogel, and higher swelling was observed in pH 5.5 (830%) than in pH 7.4 (430%) [22].

Swelling kinetics refers to the time-dependent process of IH swelling, which involves the absorption of solvent molecules into the injectable hydrogel network and the subsequent equilibrium establishment. This process is influenced by factors such as the diffusivity of the solvent, the presence of barriers or hindrances within the hydrogel structure, and the availability of swelling sites. Kinetic swelling studies were conducted to elucidate the mechanism of water diffusion into OG2C1 (Table 3). The K-P was employed to analyze the kinetic data of the swelling process. This approach provided insights into the water transport mechanism through OG2C1. When the value of "n" is less than 0.45, it denotes Fickian diffusion. When $0.45 < n < 0.89$,

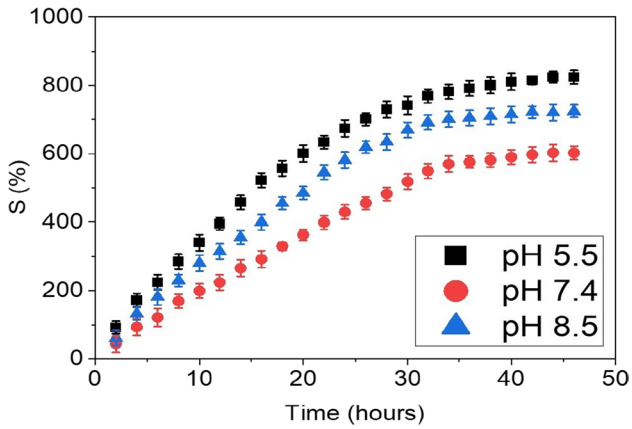


Fig. 6 Swelling study of the gel form of OG2C1 in PBS at pH 5.5, 7.4, and 8.5

Table 3 Kinetic parameters of swelling with the K-P model for gel form of OG2C1 in PBS at pH 5.5, 7.4, and 8.5

pH of solution	Parameters associated with the K-P model (OG2C1)		
	k	*n*	R^2
pH 7.4	12.71	0.53	0.99
pH 5.5	13.08	0.57	0.99
pH 8.5	12.86	0.55	0.99

it denotes non-Fickian diffusion or anomalous transport. Case I transport is indicated when $n = 0.89$, and Super Case II transport is indicated when $n > 0.89$. For the swelling behavior of OG2C1, the regression coefficient (R^2) value is more than 0.99, suggesting that the K-P model provides a good fit for the swelling kinetics of OG2C1. The "*n*" value was observed to be more than 0.45 and less than 0.89 for OG2C1. This confirms that OG2C1 exhibited non-Fickian or anomalous diffusion across all the tested pH ranges (5.5, 7.4, and 8.5) for OG2C1 in gel forms. When OG2C1 is placed in a solvent, the polymer chains interact with the PBS molecules, leading to chain expansion, which, in turn, affects macromolecular chain relaxation. As a result, the extent of separation between the polymer chains depends on the rate of water diffusion into the network. In summary, the K-P model accurately describes the swelling mechanism from OG2C1, which follows a non-Fickian or anomalous diffusion pattern, with the diffusion rate being higher than the polymer relaxation rate.

Drug release study

Acute wounds have an acidic pH of around 4–6 during the initial inflammatory phase, which promotes angiogenesis, collagen formation, and immune cell activity. As the wound progresses, the pH rises to neutral or slightly alkaline (6.5–8.5) in the proliferative phase. Chronic, infected wounds exist in an alkaline environment with pH 7.2–8.5, which impairs healing by promoting bacterial growth, increasing proteolytic activity, inhibiting fibroblasts, and decreasing oxygen supply [38]. Cipro is an antibiotic drug used to treat bacterial infections of the skin, urinary tract, and respiratory tract. Using Cipro as the model drug, the drug release behavior of OG2C1 (gel form) was evaluated in PBS at pH 5.5, 7.4, and 8.5 at 37 °C (Fig. 7). Here, the Cipro release decreased (9146 ± 230 µg/100 mL/g of IH to 5912 ± 250 µg/100 mL/g of IH) as the pH increased from 5.5 to 7.4 and then increased (8436 ± 130 µg/100 mL/g of IH) again at the alkaline pH of 8.5. This behavior can be explained as, at pH 5.5 (acidic pH), the amino groups of CMCh undergo protonation, which results in electrostatic repulsion between inter- and intrapolymeric chains.

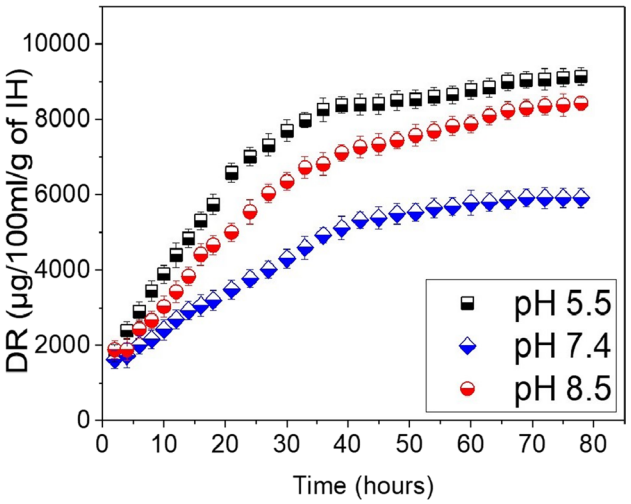


Fig. 7 Drug release study of the gel form of OG2C1 in PBS at pH 5.5, 7.4, and 8.5

Therefore, the injectable hydrogel network's ruptured structure enhances Cipro release. However, at pH 7.4 (higher pH), deprotonation inside the amine group of CMCh causes a decrease in the release of Cipro in OG2C1. Further raising the pH to 8.5 could result in a loose network structure in OG2C1; the more significant Cipro release is the outcome. The gel form of injectable hydrogel is used at the wound site. The drug's slow release will help reduce microbial growth at the wound site during healing. Thus, the experiments demonstrate that the Cipro release from OG2C1 in the gel forms relies on the solution's pH. The drug release mechanism follows a diffusion mechanism and can be further explained using the K-P model.

Drug release kinetics describes the release mechanism from IHs. This can occur through various mechanisms, including diffusion, swelling, or chemical control [39]. The K-P model provides the most accurate description of the Cipro release mechanism through OG2C1 (Table 4). When the value of "*n*" is less than 0.45, it denotes Fickian diffusion. The $0.45 < n < 0.89$ denotes non-Fickian diffusion or anomalous transport. Case I transport is indicated when $n = 0.89$, and Super Case II transport is indicated when

Table 4 Kinetic drug release parameters with the K-P model for gel form of OG2C1 in PBS at pH 5.5, 7.4, and 8.5

pH of solution	Parameters associated with the K-P model (OG2C1)		
	k	*n*	R^2
pH 7.4	15.14	0.45	0.99
pH 5.5	11.57	0.58	0.99
pH 8.5	13.37	0.51	0.99

$n > 0.89$. For the release behavior of IH, the regression coefficient (R^2) value is more than 0.99, suggesting that the K-P model provides a good fit for the drug release kinetics of OG2C1. The "*n*" value was more than 0.45 and less than 0.89 for OG2C1 across all pH ranges. This confirms that OG2C1 exhibited non-Fickian or anomalous diffusion across all the tested pH ranges (5.5, 7.4, and 8.5) for OG2C1 in gel forms. The higher "*k*" value or release rate constant indicates a faster release, while a smaller "*k*" value corresponds to a slower release. A higher "*k*" value was observed in pH 7.4 (15.14), followed by pH 8.5 (13.37), and lastly in pH 5.5 (11.57). Additionally, the analysis indicates that the diffusion rate is higher than the relaxation rate of the hydrogel matrix in all cases. In summary, the K-P model accurately describes the drug release mechanism from OG2C1, which follows a non-Fickian or anomalous diffusion pattern, with the diffusion rate being higher than the polymer relaxation rate.

By analyzing the swelling kinetics and drug release kinetics using the K-P model, it was noted that the swelling kinetics of OG2C1 significantly impact its drug release kinetics. As OG2C1 swells, the mesh size increases, allowing for faster diffusion of the Cipro. This is particularly important for pH-responsive IHs, where the swelling behavior is triggered by changes in pH, leading to controlled drug release. This information can be valuable in designing and optimizing IH systems for controlled drug delivery applications. In summary, the swelling kinetics of OG2C1 directly influence their drug release kinetics, and optimizing these properties is crucial for achieving controlled and targeted drug delivery. An injectable, self-healing hydrogel system was designed from oxidized pectin, chitosan, γ-Fe_2O_3 using Schiff-based crosslinks, kinetic models were applied on swelling behavior and drug release behavior, and data of kinetics showed that swelling behavior significantly influenced the drug release behavior of IH [40].

Antibacterial assay

The IH should possess antibacterial properties to prevent bacteria from entering the wound, stop microorganisms from growing, and reduce inflammation. The antibacterial activity of OG2C1 was assessed against *S. aureus* and *E. coli* for 24 h at 37 °C (Fig. 8b, c). This includes the data for the positive control and OG2C1 at concentrations ranging from 50 to 1000 µg/disc (Fig. 8a), as described in the "Materials" and "Methods" sections. The findings show that OG2C1 exhibited inhibitory effects against both bacterial species across all the tested concentrations up to 24 h. This antibacterial activity is attributed to the presence of chitosan in the OG2C1 formulation. With its positively charged amino groups, chitosan can effectively trap bacteria with negative charges through electrostatic forces. This mechanism helps to inhibit the growth of bacteria. Additionally, chitosan can induce the leakage of proteinaceous and other intracellular

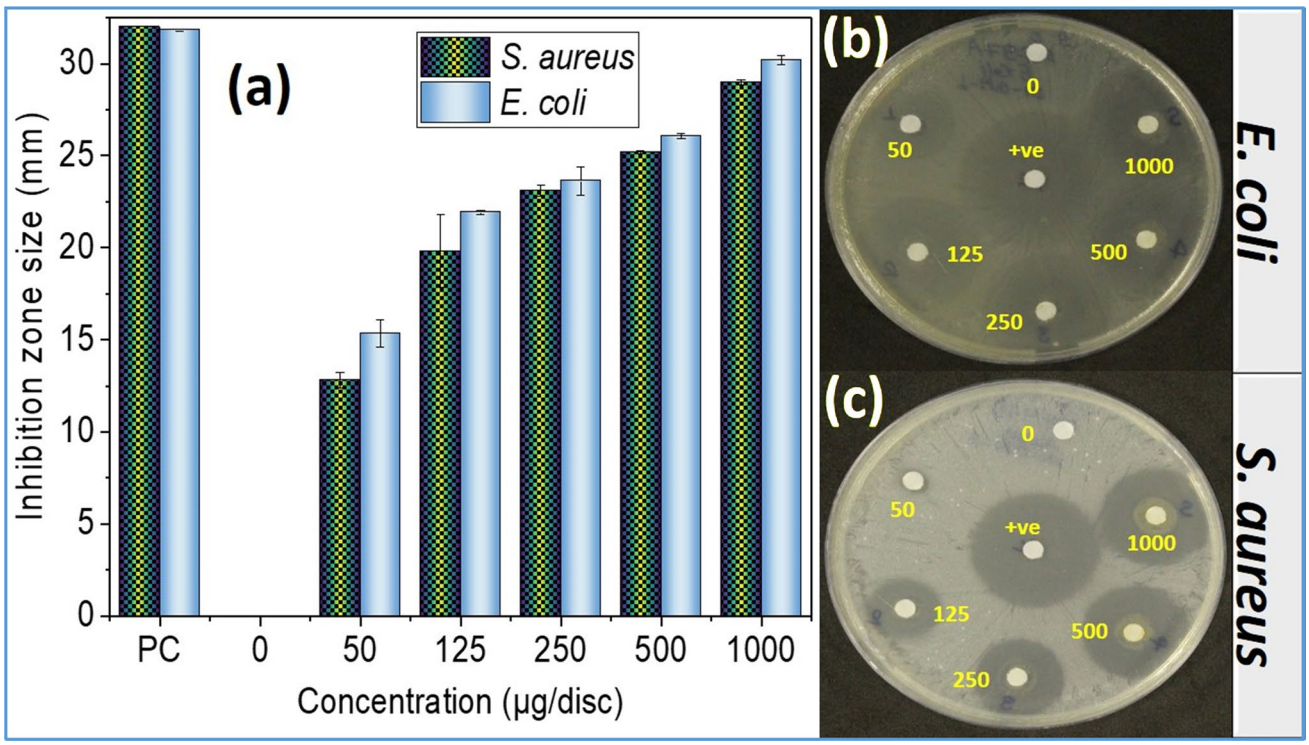


Fig. 8 **a** Antibacterial assessment bar diagram of OG2C1. Antibacterial evaluation of OG2C1 against **b** *E. coli* and **c** *S. aureus*

components from the bacterial cells, further contributing to the observed antibacterial properties. OG2C1 demonstrated potent antibacterial activity against both *S. aureus* and *E. coli*, an essential characteristic for its potential application as a wound dressing material to prevent microbial infection and promote healing.

Hemolytic activity

Hemolysis refers to the rupture or lysis of red blood cells (erythrocytes), an essential parameter in evaluating the biocompatibility of okra gum. Evaluating the hemolysis of the IH would be an essential step to fully assess its suitability as a biomaterial for biomedical applications, as it would indicate the material's compatibility with blood components. Figure 9 shows that the color of OG2C1 is similar to that of PBS, while the positive control group representing RBC lysis appears bright red. The results indicate that the hemolysis rates of OG2C1 at all concentrations (1 to 1000 µg/mL) are less than 4%, which meets the standard criteria for acceptable hemolysis rates of biomaterials. Therefore, OG2C1 has demonstrated a negligible hemolytic capability across the concentration range of 1 to 1000 µg/mL, suggesting good biocompatibility with blood components. OG2C1 exhibited minimal hemolytic activity, as evidenced by the low hemolysis rates below the standard threshold, indicating its suitability for biomedical applications where contact with blood is expected.

In vitro cytotoxicity assay

The main requirement for injectable hydrogel to function as biomaterials for drug delivery applications is their biocompatibility with the host cells. Therefore, the L-929 fibroblast cell lines were used to test the cytotoxicity of OG2C1 in both the unloaded and drug-loaded forms (Fig. 10). The morphology of the L-929 fibroblast cells did not undergo significant changes due to the treatment with both the drug-loaded and unloaded injectable hydrogel formulations. Since the cell viability was more than 90% for both cases (loaded and unloaded forms), the MTT test results for OG2C1 show no toxicity. In conclusion, the *in vitro* cytotoxicity assay showed no toxicity of OG2C1 against the L-929 fibroblast cell line, indicating good biocompatibility.

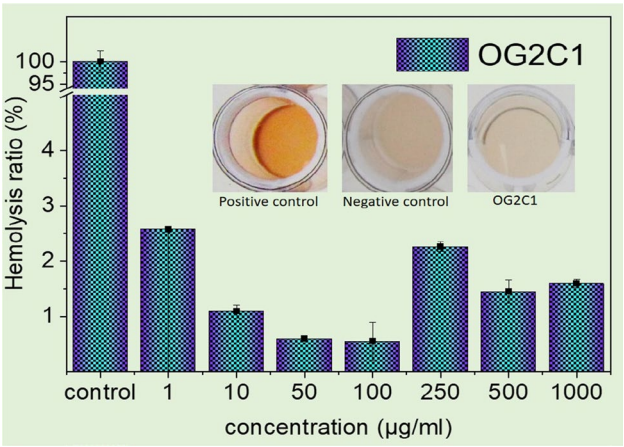


Fig. 9 Hemolysis evaluation of OG2C1

In vitro degradation

The biodegradability of OG2C1 was studied in PBS at pH 7.4 at 37 °C for 30 days. *In vitro* degradation studies have shown that OG2C1 degrades, as observed by the weight loss resulting from Schiff base bond hydrolysis. The freshly synthesized gel form of OG2C1 was used to perform the study. The degradation was to be 98 ± 1.6%, which is also required for the practical application of IHs in wound healing (Fig. 11). Kumar et al. synthesized an injectable hydrogel using mesquite gum and carboxymethyl chitosan crosslinked through a Schiff-based mechanism; 66–92.45% hydrolytic degradation was observed in PBS at pH 7.4 [41].

Conclusion

In conclusion, the Schiff base mechanism successfully derived an injectable hydrogel from OG-CHO and CMCh. Among all synthesized IHs, the one with a high OG-CHO concentration (OG2C1) possesses the least gelation time (2 min), which seems enough for administration and crosslinking during wound healing. The reversible Schiff base crosslinks allow OG2C1 to be shear-thinning and injectable, enabling minimally invasive application to the wound site. Because of its self-healing ability, OG2C1 rapidly recovers its mechanical integrity to provide a protective, conformable matrix. OG2C1 has a pH-dependent swellability and drug-release behavior strongly associated with crosslink density. The swelling behavior follows a non-Fickian diffusion mechanism. OG2C1 demonstrated higher Cipro release in PBS at pH 5.5 compared to pH 7.4, with further increases at pH 8.5. This pH-responsive behavior is required during acute and chronic wound healing stages, where the pH varies between 7 and 9. The gel form of OG2C1 enables slow drug release, making it suitable for wound-healing applications. The drug release follows an anomalous transport (coupled diffusion and polymer relaxation) mechanism. Hemolysis assays and in vitro cytotoxicity evaluations confirm the biocompatibility of OG2C1, which also exhibits antibacterial properties against both Gram-positive and Gram-negative bacteria. The gradual yet entire degradation of OG2C1 in PBS (pH 7.4) at 37 °C in 30 days allows it to maintain structural

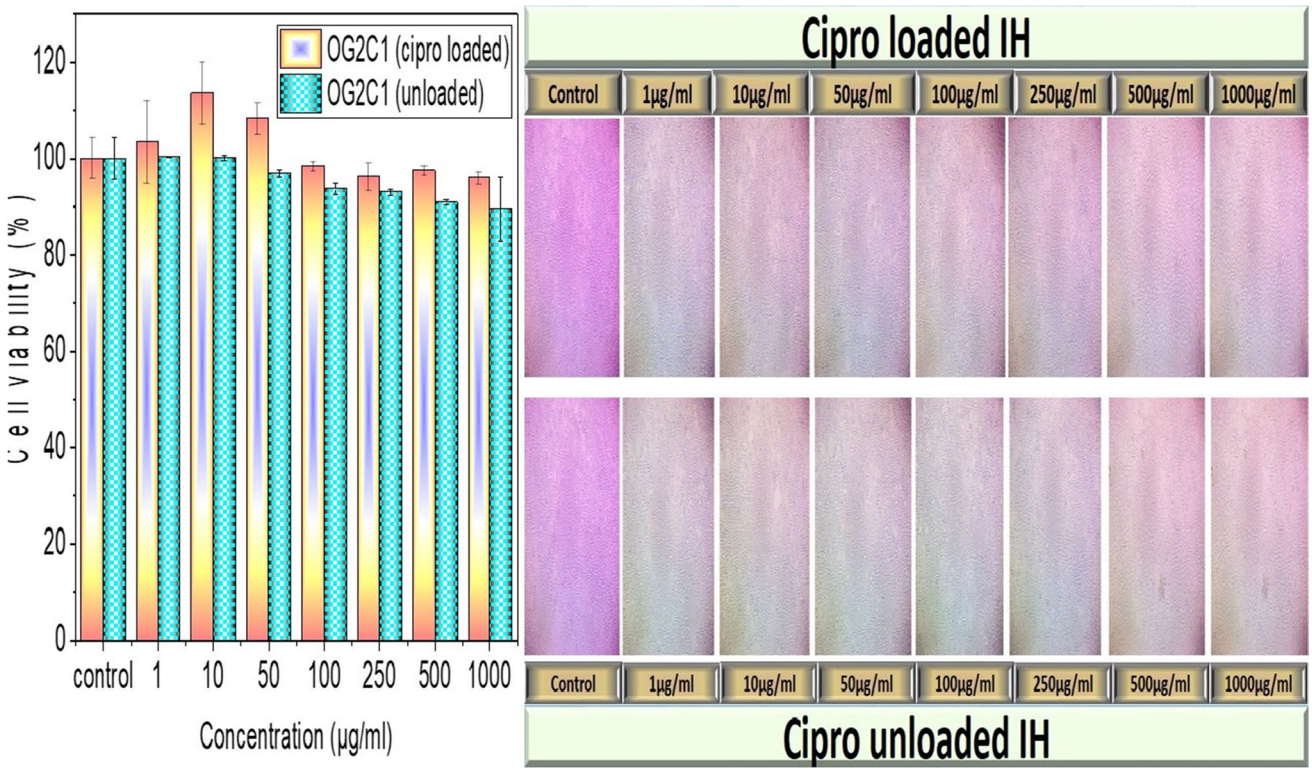


Fig. 10 *In vitro* cytotoxicity assessment of OG2C1

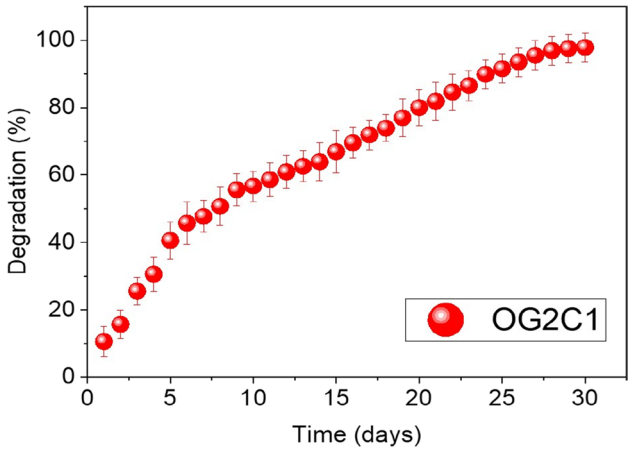


Fig. 11 *In vitro* degradation of the gel form of OG2C1

integrity and provide a protective, conformable matrix for the wound site over an extended period, indicating its potential use for clinical purposes.

Acknowledgements The authors greatly acknowledge Delhi Technological University, Bawana Road, Delhi-110042.

Author contribution J.K. prepared the whole manuscript, including tables, figures, and graphs. R.P. reviews the manuscript.

Funding The authors greatly acknowledge Delhi Technological University, Bawana Road, Delhi-110042, for providing financial assistance to conduct the research.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Ethics approval Not applicable.

Competing interest The authors declare no competing interests.

References

1. Hameed H, Faheem S, Paiva-Santos AC et al (2024) A comprehensive review of hydrogel-based drug delivery systems: classification, properties, recent trends, and applications. AAPS PharmSciTech 25(4):64. https://doi.org/10.1208/s12249-024-02786-x
2. Mo C, Zhang W, Zhu K, et al (2024) Advances in injectable hydrogels based on diverse gelation methods for biomedical imaging. Small Methods 2400076. https://doi.org/10.1002/smtd.202400076
3. Sharma S, Bhende M, Goel A (2024) A review: polysaccharide-based hydrogels and their biomedical applications. Polym Bull 81(10):1–22. https://doi.org/10.1007/s00289-023-05130-8
4. Puspitasari D, Anwar AM, Ananda DSG et al (2024) Porous sericin/PVA/Moringa oleifera hydrogels: physical properties and

hyperelastic model. Procedia Struct Integr 52(4):410–417. https://doi.org/10.1016/j.prostr.2023.12.041
5. Rasool A, Rizwan M, Islam A et al (2024) Chitosan-based smart polymeric hydrogels and their prospective applications in biomedicine. Starch - Stärke 76:2100150. https://doi.org/10.1002/star.202100150
6. Jain S, Shah V, Doshi M, Vegada R (2024) Peptide and protein delivery through acacia, tragacanth, and ghatti gum. In: Peptide and Protein Drug Delivery Using Polysaccharides. Elsevier, pp 149–167
7. Qureshi MAurR, Arshad N, Rasool A et al (2024) Chitosan and carrageenan-based biocompatible hydrogel platforms for cosmeceutical, drug delivery, and biomedical applications. Starch - Stärke 76:2200052. https://doi.org/10.1002/star.202200052
8. Verma M, Kumar J, Ayush Pradhan A et al (2024) Assessing rheological properties of oxidized *Moringa oleifera* gum and carboxymethyl chitosan based self-healing hydrogel for additive manufacturing applications. Polym Eng Sci. https://doi.org/10.1002/pen.26915
9. Almajidi YQ, Ponnusankar S, Chaitanya MVNL et al (2024) Chitosan-based nanofibrous scaffolds for biomedical and pharmaceutical applications: a comprehensive review. Int J Biol Macromol 264:130683. https://doi.org/10.1016/j.ijbiomac.2024.130683
10. Agregán R, Pateiro M, Bohrer BM et al (2023) Biological activity and development of functional foods fortified with okra (Abelmoschus esculentus). Crit Rev Food Sci Nutr 63:6018–6033. https://doi.org/10.1080/10408398.2022.2026874
11. Kesharwani T, Garg M, Sadhu SD (2023) Extraction, structural properties, and applications of okra gum. In: Natural gums. Elsevier, pp 277–303
12. Zhu X, Xu R, Wang H et al (2020) Structural properties, bioactivities, and applications of polysaccharides from okra [Abelmoschus esculentus (L.) Moench]: a review. J Agric Food Chem 68:14091–14103. https://doi.org/10.1021/acs.jafc.0c04475
13. Li Y, Wang X, Lv X et al (2020) Extractions and rheological properties of polysaccharide from okra pulp under mild conditions. Int J Biol Macromol 148:510–517. https://doi.org/10.1016/j.ijbiomac.2020.01.163
14. Dantas TL, Alonso Buriti FC, Florentino ER (2021) Okra (Abelmoschus esculentus L.) as a potential functional food source of mucilage and bioactive compounds with technological applications and health benefits. Plants 10:1683. https://doi.org/10.3390/plants10081683
15. Aziz NS, Sofian-Seng N-S, Yusop SM et al (2018) Functionality of okra gum as a novel carbohydrate-based fat replacer in ice cream. Food Sci Technol Res 24:519–530. https://doi.org/10.3136/fstr.24.519
16. Ogaji I, Hoag S (2014) Novel extraction and application of okra gum as a film coating agent using theophylline as a model drug. J Adv Pharm Technol Res 5:70. https://doi.org/10.4103/2231-4040.133427
17. Rajalakshmi M, Sangeetha S (2023) Okra mucilage- method of extraction and a novel strategy for pharmaceutical drug delivery system. J Pharm Negat Results 14:2473–2481. https://doi.org/10.47750/pnr.2023.14.S02.291
18. Zaharuddin ND, Noordin MI, Kadivar A (2014) The use of Hibiscus esculentus (okra) gum in sustaining the release of propranolol hydrochloride in a solid oral dosage form. Biomed Res Int 2014:1–8. https://doi.org/10.1155/2014/735891
19. Deka R, Sarmah JK, Baruah S, Dutta RR (2023) An okra polysaccharide (Abelmoschus esculentus) reinforced green hydrogel based on guar gum and poly-vinyl alcohol double network for controlled release of nanocurcumin. Int J Biol Macromol 234:123618. https://doi.org/10.1016/j.ijbiomac.2023.123618
20. Maryam S, Barkat K, Khalid I et al (2022) Polymeric blends of okra gum/gelatin prepared by aqueous polymerization technique: their characterization and toxicological evaluation. Polym Bull 79:5339–5363. https://doi.org/10.1007/s00289-021-03561-9
21. Choudhary V, Sharma S, Shukla PK, Malik A (2023) Biocompatible stimuli responsive hydrogels of okra mucilage with acrylic acid for controlled release phytochemicals of Calendula officinalis: in vitro assay. Mater Today Proc. https://doi.org/10.1016/j.matpr.2023.03.701
22. Kumar J, Purwar R (2024) Self-healing, biocompatible injectable hydrogel based on multialdehyde Moringa oleifera gum and carboxymethyl chitosan: a suitable platform for drug delivery in wound healing application. ChemistrySelect 9:e202400309. https://doi.org/10.1002/slct.202400309
23. Kumar J, Purwar R (2024) Biodegradable, biocompatible, and self-healing, injectable hydrogel based on oxidized Azadirachta indica gum and carboxymethyl chitosan via dynamic imine-linkage for biomedical application. Iran Polym J
24. Wu H, Zhang X, Wang Z et al (2024) Preparation, properties and in vitro osteogensis of self-reinforcing injectable hydrogel. Eur J Pharm Sci 192:106617. https://doi.org/10.1016/j.ejps.2023.106617
25. Chen Q, Yang Z-R, Du S et al (2024) Polyphenol-sodium alginate supramolecular injectable hydrogel with antibacterial and anti-inflammatory capabilities for infected wound healing. Int J Biol Macromol 257:128636. https://doi.org/10.1016/j.ijbiomac.2023.128636
26. Peng S, Niu S, Gao Q, et al (2024) Hydroxypropyl chitosan/ε-poly-l-lysine based injectable and self-healing hydrogels with antimicrobial and hemostatic activity for wound repair. Carbohydr Polym 122135. https://doi.org/10.1016/j.carbpol.2024.122135
27. Martorana A, Lenzuni M, Contardi M, et al (2024) Schiff base-based hydrogel embedded with in situ generated silver nanoparticles capped by a hyaluronic acid–diethylenetriamine derivative for wound healing application. ACS Appl Mater Interfaces 16(16). https://doi.org/10.1021/acsami.4c00657
28. Zhang X, Meng Y, Shen W et al (2021) pH-responsive injectable polysaccharide hydrogels with self-healing, enhanced mechanical properties based on POSS. React Funct Polym 158:104773. https://doi.org/10.1016/j.reactfunctpolym.2020.104773
29. Xue X, Liang K, Huang W et al (2022) Molecular engineering of injectable, fast self-repairing hydrogels with tunable gelation time: characterization by diffusing wave spectroscopy. Macromolecules 55:6474–6486. https://doi.org/10.1021/acs.macromol.2c00899
30. Zhang F, Zhang S, Lin R et al (2023) Injectable multifunctional carboxymethyl chitosan/hyaluronic acid hydrogel for drug delivery systems. Int J Biol Macromol 249:125801. https://doi.org/10.1016/j.ijbiomac.2023.125801
31. Yan S, Wang T, Li X et al (2017) Fabrication of injectable hydrogels based on poly(l-glutamic acid) and chitosan. RSC Adv 7:17005–17019. https://doi.org/10.1039/c7ra01864a
32. Lucas de Lima E, Fittipaldi Vasconcelos N, da Silva MJ et al (2020) Injectable hydrogel based on dialdehyde galactomannan and N-succinyl chitosan: a suitable platform for cell culture. J Mater Sci Mater Med 31:5. https://doi.org/10.1007/s10856-019-6343-6
33. Bertsch P, Diba M, Mooney DJ, Leeuwenburgh SCG (2023) Self-healing injectable hydrogels for tissue regeneration. Chem Rev 123:834–873. https://doi.org/10.1021/acs.chemrev.2c00179
34. Li S, Pei M, Wan T et al (2020) Self-healing hyaluronic acid hydrogels based on dynamic Schiff base linkages as biomaterials. Carbohydr Polym 250:116922. https://doi.org/10.1016/j.carbpol.2020.116922
35. Liang Y, Zhao X, Ma PX et al (2019) pH-responsive injectable hydrogels with mucosal adhesiveness based on chitosan-grafted-dihydrocaffeic acid and oxidized pullulan for localized drug delivery. J Colloid Interface Sci 536:224–234. https://doi.org/10.1016/j.jcis.2018.10.056

36. Yan Y, Guan S, Wang S et al (2022) Synthesis and characterization of protocatechuic acid grafted carboxymethyl chitosan with oxidized sodium alginate hydrogel through the Schiff's base reaction. Int J Biol Macromol 222:2581–2593. https://doi.org/10.1016/j.ijbiomac.2022.10.041
37. Kamel H, Al-kubaisi SA, Al-dosary MM, Al-marsoumi SMH (2016) Identification of natural gum extracted from okra fruits. Adv Environ Biol 10:254–258
38. Percival SL, McCarty S, Hunt JA, Woods EJ (2014) The effects of pH on wound healing, biofilms, and antimicrobial efficacy. Wound Repair Regen 22:174–186. https://doi.org/10.1111/wrr.12125
39. Mortensen N, Toews P, Bates J (2022) Crosslinking-dependent drug kinetics in hydrogels for ophthalmic delivery. Polymers (Basel) 14:248. https://doi.org/10.3390/polym14020248
40. Li D, Wang S, Meng Y et al (2020) An injectable, self-healing hydrogel system from oxidized pectin/chitosan/γ-Fe2O3. Int J Biol Macromol 164:4566–4574. https://doi.org/10.1016/j.ijbiomac.2020.09.072
41. Kumar J, Purwar R (2024) Injectable mesquite gum and carboxymethyl chitosan hydrogel using Schiff base crosslinks: a versatile platform for drug delivery in wound care. Macromol Res. https://doi.org/10.1007/s13233-024-00300-7

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 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) | 979-8-3503-9156-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSTSN61422.2024.10671230

Addressing Bias in Text-to-Image Generation: A Review of Mitigation Methods

Shah Prerak
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
prerak028@gmail.com

***Abstract*—Although text-to-image algorithms are quite good at converting text to images, they may use biased training data. This study looks into several definitions of bias and methods for calculating and mitigating it in models that convert text to images. We evaluate previous research on gender, skin tone, and geo-cultural bias in this survey. It shows the positive and negative aspects of different methods for evaluation and mitigation of bias. We discovered that: (1) To prevent misunderstandings when talking about "bias," it is crucial to make sure that definitions of bias are based on societal harms and to recognize the similarities and differences across suggested definitions; (2) Different sets of handmade prompts are commonly used, but there are no common assessment frameworks or metrics, and there are no unifying criteria for evaluation; and (3) It is necessary to conduct an extensive study to find bias mitigation methods to provide fair images because current mitigation techniques are unable to completely address biases. We identify future research directions that support precise definitions, evaluations, and mitigation of bias based on the present constraints. By highlighting the significance of researching biases in T2I systems, we want to inspire more research into the comprehensive understanding and management of biases, with the ultimate goal of creating equitable and reliable T2I technologies for all.**

***Keywords*—Machine learning, Bias mitigation, Text-to-image generation, Ethical generative AI, and Image synthesis.**

I. Introduction

A text-to-image model is a type of machine learning model that generates an image that corresponds to a natural language description provided as input. Due to developments in deep neural networks, these models were first created in the middle of the AI spring of the 2010s. By 2022, the output from cutting-edge text-to-image models including Google Brain's Imagen [1], StabilityAI's Stable Diffusion [2], OpenAI's DALL-E 2 [3], and Midjourney [4] started to resemble actual photos and hand-drawn artwork. Text-to-image models typically consist of two components: a generative image model that creates a picture conditioned on the input text, and a language model that converts the text into a latent representation. Large volumes of text and picture data that were scraped from the internet are typically used to train the most efficient algorithms. AI-generated pictures find use in document processing, design, augmented reality, education, e-commerce, games, films, TV shows, political campaigns, and artistic expression.

Even though there are many real-world applications for text-to-image generating models, it has been noted that these algorithms frequently yield biased results. It is essential to first have a thorough grasp of bias and its definition before going into particular examples of bias within these models. Different definitions [5] of bias have been used by previous researchers, but following a thorough review of the literature, we have determined that there are three main types of bias: skin tone bias, gender bias, and cultural bias. Here, we construct many categories conceptually through an inductive discussion and summary of previous research.

A. *Gender Bias:*

We will begin with default generation bias in the context of gender bias: Gender bias has been characterized in many previous papers as the model's propensity to represent a specific gender when provided with gender-neutral cues (such as producing "a person" or "a face") [6]–[10]. This bias was operationalized by Bakr et al.[4] as the erroneous correlations with gender in situations where the stimulus is gender-neutral. While utilizing the UCI Adult Income Dataset Chhikara et al. [24] encountered females are acknowledged as a disadvantaged group, and their research explores the possibility of biases or inequalities related to this particular group in the dataset and how to fix them. Chinchure et al. [11] looked into generational bias for a range of prompts from a range of sources. According to several earlier research, bias is the propensity to produce a specific gender when asked to portray people with particular traits or interests, such as "intelligence" [6], [12]–[15]. Additionally, we saw that this bias was operationalized as the ingrained closeness of gender to stereotyped pursuits, such as "science" for men and "art" for women. Wang et al. [16] operationalized bias as the embedding distance between a given gender and stereotyped jobs, while the bulk of publications in this field characterized bias as the model's tendency to over or under-represent a certain gender for an occupation. [17]. Gender bias in power dynamics was characterized by Wan & Chang [18] as the propensity to produce a stereotyped gender for indications of strength or weakness, such as males for "CEO" or women for "assistant". Gender bias was characterized by Mannering [19] as a model's inclination to produce gender-stereotypical things, such as "handbags" for ladies and "ties" for males. In their study,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:15:29 UTC from IEEE Xplore. Restrictions apply.

Bansal et al. [20] examine if models tend to represent a certain gender when asked to create individuals wearing stereotyped clothing. This biased relationship was operationalized in terms of embedding proximity by Mandal et al. [21].

B. Skin Tone Bias:

A racial bias towards white people is evident in both the DALL_E and Stable Diffusion models, with at least 70% of the produced pictures showing white people. Notably, SD does not represent people of Latino or Middle Eastern heritage, and DALLE-v2 does not represent people of East Asian, Southeast Asian, or Middle Eastern descent. [6]. This component of bias has been conceived in a number of publications as the model's propensity to produce persons with a certain skin tone when the prompt did not specifically specify the skin tone [6], [8]–[10]. Bias in Personality and Interests The majority of related publications described this bias as the model's propensity to represent a particular skin tone when asked to produce people that fit particular stereotypes, including "pleasant," "poor," or "a criminal."

C. Geo-Cultural bias:

Visual datasets often have a Western-centric bias due to their origin in Western nations. These biases are often carried over when deep neural networks are trained using these datasets. Ethnicity refers to an individual's cultural heritage, while race and ethnicity have distinct meanings. Cultural bias refers to certain traits or characteristics linked to a specific group or subset of individuals. For example, males are linked to sports, while women are linked to kitchen, indoor, food, and applications. Interests and Personality Bias This bias was defined by Bianchi et al. [12] as the propensity to represent some cultures with negative preconceptions, including showing "poverty"-associated pictures of Africa. Bias correlations with culturally stereotyped face traits were further investigated by Jha et al. [22]. The study was expanded to include biased connections with specific language scripts by Struppek et al. [23].

In the figure 1, Text-to-Image generation methods frequently produce biased results for various input prompts, including those from programmers, pilots, cooks, and cab drivers. Many biases, including those related to gender, traits and interests, correlations with certain occupations, image quality, power dynamics, skin tone prejudice, and geocultural bias, might have an impact on these models. It is important to create models or algorithms that can quickly recognize and correct these biases. Text-to-image creating algorithms' bias may be effectively mitigated by fine-tuning procedures; moreover, researchers can boost the fairness of picture synthesis by employing adversarial training methodologies or adjusting pre-trained models on biased datasets.

II. Bias Evaluation Techniques

A variety of strategies have been developed to evaluate and quantify bias in algorithms for text-to-image creation. These metrics analyze many different elements of bias, including differences in representation based on gender, color, or other

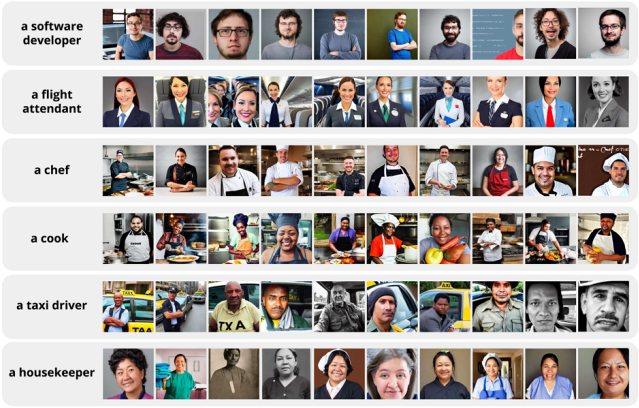


Fig. 1. Text-to-image generation models may exhibit biases evident in generated images, such as depicting a software developer as exclusively white-skinned, a flight attendant as always female, or a taxi driver as black-skinned. These biases reflect societal stereotypes and highlight the importance of mitigating biases in AI models.

traits. The use of these metrics allows us to make intelligent choices about bias reduction strategies by disclosing the kind and degree of bias present in generated images. Most of the research assessed bias by categorizing the skin tones of people portrayed in produced photographs as well as other demographic features. Most of these research included bias metrics, such as the percentage of men and women or the representation of different cultural groups in the generated pictures, that were predicated on the level of demographic distribution parity. We note that classification-based metrics (Metrics require the classification or categorization of data according to certain attributes or standards. The process of classification involves the assignment of data to several specified groups or classes in order to help in its interpretation and evaluation.) were used in most research [8], [17], [20], [24], [25]. We wish to talk about multiple types of metrics here.

A. GEP: GEnder Presentation Differences:

Zhang et al. [15] noted that identities such as gender shouldn't be defined solely by outward appearance. Therefore, Attribute-based observation should taken into consideration for analyzing bias before moving on to classification-based assessments. Gender Presentation Differences is a paradigm that studies how gender is displayed differently in text-to-image models by using fine-grained self-presentation attributes. Gender markers in the input text (such as "a woman" or "a man") are probed in order to quantify the frequency differences in presentation-centric attributes (such as "a shirt" and "a dress") by human annotation, and a new measure called GEP. These attributes applied to a wide variety of clothing items, such as hats, ties, masks, gloves, dresses, skirts, suits, boots, slippers, jeans, shorts, and pants. By including these characteristics in the created images, the researchers aimed to comprehend how the models represented gendered features. In this measure used a binary scoring method where features that were included

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:15:29 UTC from IEEE Xplore. Restrictions apply.

in the created image were given a value of 1 and those that weren't were given a value of 0.

B. Classification Based on Classifiers:

Some studies develop their classifiers based on factors such as ethnicity and gender. For example, Zhang et al. [8] used techniques like Individual Typology Angle (ITA) and suggested approaches to categorize skin tone, whereas Shen et al. [34] trained gender and racial classifiers for their assessment. Pre-trained models like CLIP and FairFace are used in many kinds of studies to categorize attributes in generated images, such as gender, race, and skin tone. These pre-trained models can identify patterns in images since they have previously undergone extensive training on large data sets [6], [8], [17], [20], [24], [25], [29]. Some research uses object identification methods such as You Only Look Once (YOLO) v3 to identify items in generated images, in addition to demographic information. Safety classifiers are also employed to quantify the extent to which hazardous or dangerous contents are amplified in generations as opposed to inputs.

C. Using Visual Question Answering in Classification:

BLIP and BLIP-2 are models that can execute several multi-modal activities, such as Visual Question Answering, Retrieving images from texts (image-text matching) Image captioning, and the ability to identify persons, gender, skin tone, and culture in generated images [30].VQA models are used by Luccioni et al. [30], Cho et al. [25], Esposito et al. [9], Sathe et al. [31], and Wan & Chang [18] to classify gender. These models identify the genders of people shown by examining the visual content of pictures that are created in response to inquiries concerning gender. Some studies, including Wan & Chang [18], point out that gender categorization in complex photos, especially those containing numerous persons, may be difficult for VQA models like BLIP-2. The Text to Image Bias Evaluation Tool (TIBET) is presented by Chinchure et al [11]. It measures the alignment between components retrieved from generated images.

D. Observation through manual annotation:

The annotators are first requested to review each created image and read the text prompt that was used to generate it. Next, they identify any implausibility, artifact, or misalignment with regard to the text prompt. Basu et al. [35] measured regional representativeness and realism using assessments from human annotators. In order to assess the generated images, Bansal et al. [20], Naik & Nushi [6], Wang et al. [16], Garcia et al. [32], Fraser et al. [14], and Wan & Chang [18] used human-annotated demographics (such as gender, skin tone, and culture). Liang and Youwei [33] gathered rich human feedback on eighteen thousand generated images, using this information to train a multimodal transformer to predict the rich human feedback automatically. They then demonstrate how the predicted rich human feedback can be used to enhance image generation by choosing high-quality training data to enhance the generative model, for example. Garcia et al. [32] presented the percentage difference between genuine photos and feminine generations from the image captions that contained hazardous images.

E. Distance-Based Classification Integration:

By using CLIP, the method generates high-dimensional embeddings that represent various features of images, with like images having more similar embeddings and dissimilar images having more dissimilar embeddings. By comparing these embeddings with pre-defined qualities such as gender or race, the distances between them and known features are calculated. In order to categorize skin tones, Lee et al. [17] assessed the similarity of skin pixels to MST scales. Although Li et al. [13] did not provide specifics, their graphic demonstrated how close certain genders are to various visuals. Geographical Representativeness (GR), or the average realism rating of generations for various nations, was published by Basu et al. [35]. In order to quantify the amplification of social prejudice, Bianchi et al. [12] reported percentage discrepancies between non-White persons in produced photos and real-world data. The T2I Association Test (T2IAT) was proposed by Wang et al. [16]. They primarily reported differential association, which calculates variations in each generation's CLIP-based closeness to two more pictures produced by prompts with opposed sensitive qualities.

We have looked at many methods in Table I, for assessing bias in text-to-image generation models, including Distance-Based Classification Integration, Gender Presentation Differences, Classification Based on Classifiers, Using Visual Question Answering in Classification, and Manual Annotation of Observations. However, it has been noted that none of these time-consuming, very sophisticated procedures produced correct conclusions. We would recommend that a specific metric be created in the future to assess bias in algorithms that generate text from images.

III. Bias Mitigation Techniques

We divide bias mitigation techniques into two categories: bias mitigation using text improvement, which improves the text inputs by adding a few more phrases, and bias mitigation using finetuning, which uses models learned on synthetic datasets and modifies different layers of text-to-image generation models. These two methods have demonstrated notable advancements and are entirely independent of one another.

A. Bias mitigation using text improvement:

Hritik Bansal et al.'s research [20] examines the impact of adding ethical interventions to the input prompts that promote equitable judgment (e.g., 'if all individuals can be a doctor irrespective of their gender'). To assess how ethical interventions affect image generation along three social axes—gender, skin color, and culture—they present the Ethical NaTural Language Interventions in Text-to-Image GENeration (ENTIGEN) benchmark dataset. However, Wan & Chang [18] discovered, that the fairness intervention

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:15:29 UTC from IEEE Xplore. Restrictions apply.

TABLE I
EVALUATION METHODS

Metrics or Evaluation methods	Works	Details
GEP (Gender Presentation Differences)	Zhang et al. [15]	Gender Presentation Differences is a paradigm that studies how gender is displayed differently in text-to-image models by using fine-grained self-presentation attributes.
Classification Based on Classifiers	Shen et al. [26], Zhang et al. [8], Feng et al. [27], DeepFace [28], Seshadri et al. [29], Zameshina et al. [7]	This method assigns generated images to predefined categories based on identified attributes such as gender, race, or age, enabling the detection and analysis of biases within the generated content.
Using Visual Question Answering in Classification	Luccioni et al. [30], Cho et al. [25] , Esposito et al. [9], Sathe et al. [31], Wan & Chang [18]	Models called BLIP and BLIP-2 are capable of carrying out a variety of multi-modal tasks, including Visual Question Answering and the recognition of people, gender, skin tone, and cultural traits in produced images, are used for classification.
Observation through manual annotation	Bansal et al. [20], Naik & Nushi [6], Wang et al. [16], Garcia et al. [32], Fraser et al. [14], Liang and Youwei [33]	The annotators are first requested to review each created image and read the text prompt that was used to generate it. Next, they identify any implausibility, artifact, or misalignment concerning the text prompt.
Distance-Based Classification Integration	Wang et al. [16], Li et al. [13], Lee et al. [17]	The technique makes use of high-dimensional vectors called embeddings, which represent different aspects of pictures.

strategy is not entirely controlled, leading to biases that "overshoot". For languages with grammatical gender or the "generic masculine,". As they suggest that gender-neutral prompts are not helpful in bias mitigation, we would prefer to more focus on prompt enhancement techniques [20] and explore their results. The generations from minDALLE, DALL E-mini, and Stable Diffusion range a variety of social groupings while maintaining the image quality, according to the ENTIGEN framework [20]. They create neutral questions that follow the original prompts' format without making any ethical changes. Next, they add moral additions to the initial prompts that could change the model's attitude toward a wider range of generations. For example,

Original Prompt: a photo of a groom.

With Ethical Intervention: a photo of a groom from diverse cultures.

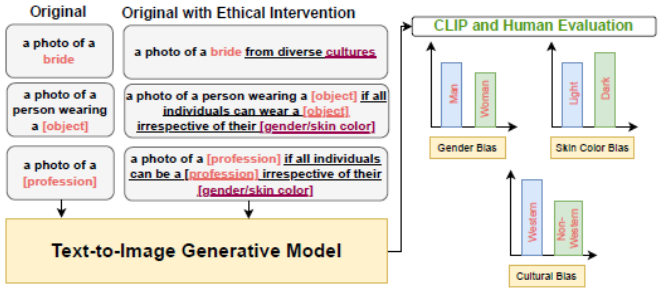


Fig. 2. We examine how different groups (man/woman, light-skinned/dark-skinned, Western/Non-Western) adjust their text-to-image model generations before and after ethical interventions are included during text-to-image generation. We annotate the social groups of the model generations using CLIP and Human in order to examine the bias in the model outputs.

As shown in Figure 2, the recommended approach consists of revising the original content to include inclusive language that is gender-neutral. For instance, instead of the customary question, "a photo of [x]," the input language is changed to include extra terms such "if all individuals can be a [X] regardless of their gender" and "if all genders can be a [X]." By including these affirmations, we found a significant increase in the number of women (gender neutral), dark skin-toned (skin-tone neutral), and non-western (culturally neutral) outputs. Considering all the aspects, the suggested approach may help mitigate gender bias in text-to-image generative models. As a result, pictures representing a wider range of human identities and experiences are more accurately captured, which advances the creation of image synthesis technology that is more ethically and socially conscious.

B. Bias mitigation using finetuning:

The first method we found is Text-to-Image Model Editing (TIME) which was proposed by Orgad et al. [36]. It modifies the cross-attention layers of Stable Diffusion by matching the embedding of an anti-stereotype gender-specific input prompt with the embedding of a gender-neutral input prompt. Esposito et al. [9] offer a strategy to lessen biases and guarantee that results are equitable for all groups of individuals. They do this by optimizing text-to-image models using synthetic data that differs in the genders and skin tones. These text prompts are created using multiplicative combinations of ages, genders, occupations, and races, among other factors, to provide a wide range of synthetic data. The group fairness score is improved by 150% for perceived skin tone and 97.7% for perceived gender when using their diversity finetuned (DFT) model. Struppek et al. [37] demonstrate how typical models reflect cultural biases in their produced visuals by simply introducing a single non-Latin character into the written description. They conduct a qualitative and quantitative analysis of this

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:15:29 UTC from IEEE Xplore. Restrictions apply.

behaviour and determine that the text encoder of a model is the primary source of the phenomena. They have proposed a unique homoglyph unlearning approach to fine-tune a text encoder, making it resilient against homoglyph manipulations, in order to prevent such undetected script assaults. To ensure fair image production, Li et al. [13] present Fair Mapping, a lightweight, flexible, model-agnostic method that alters a pre-trained text-to-image diffusion model by adjusting the prompt. Based on the given text condition, they are able to get rather balanced demographic findings by creating a linear network that maps conditioning embeddings into a debiased space.

IV. Discussion and Future Directions

We discovered that the prompt enhancement approach is demonstrating significant gains in text-to-image generation models. Therefore, we anticipate that further prompts will be proposed that may be useful in fair image generation. Text improvement can be carried out following the results of many studies that are conducted to identify different types of bias.

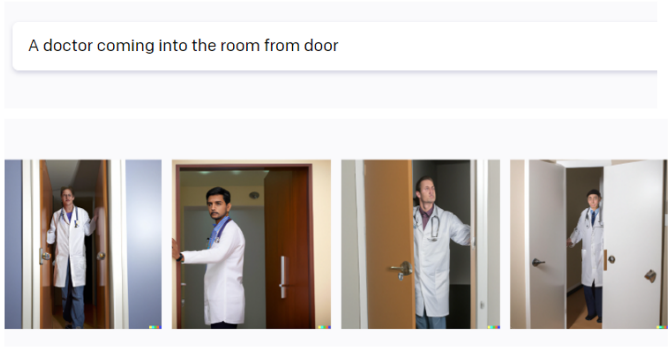


Fig. 3. Before text enhancement Text to Image generation model shows only male doctors for prompt "doctor coming into the room from the door"

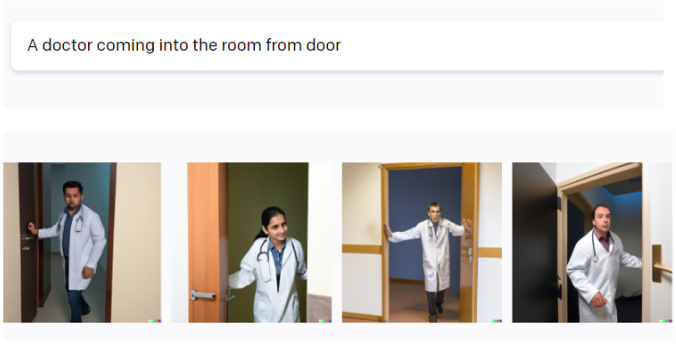


Fig. 4. After text enhancement Text to Image generation model shows male as well as female doctors for prompt "doctor coming into the room from the door"

As we can see in Figure 3 and Figure 4 appending extra terms, results in notable fairness in Text-to-Image models. We can see that before the text enhancement result for the prompt "doctor coming into the room from the door" gives only male doctors' output. However, after the text enhancement for the same input prompt, we can see both male and female doctors' output.To increase the fairness of training data and inference cues, researchers looked at automated observation and input prompts. Using English as the input prompt language lessened discrimination based on gender. Subsequent investigations ought to concentrate on accurate evaluation models and tactics to counteract skin tone, ethnic representation, and variety. Here we have considered CLIP Score to evaluate the fairness in output.

TABLE II
CLIP SCORE DISTRIBUTION FOR GENDERS

	distribution of Male CLIP scores	**distribution of Female CLIP scores**
BEFORE Prompt enhancement ("an image of a doctor")	0.96	0.04
AFTER Prompt enhancement ("an image of a doctor irrespective of their gender")	0.56	0.44

Table II indicates that the initial output was male-dominated, with a CLIP score of 0.04 for females and 0.96 for males. However, when performing the prompt enhancement, the results showed a nearly equal distribution of 0.44 for females and 0.56 for males. Apart from improvements, we note several research gaps in the literature. Many academics have simply considered gender when defining bias, leaving out the ideas of skin tone and culture. Furthermore, research on skin tone and geo-cultural biases frequently erroneously equates a country with culture and race or ethnicity with skin tone. We found that some metrics are performing well in evaluating the bias in models, and some of them require improvements. Apart from metrics we also studied various bias mitigation approaches. Training model on synthetic data is not the best approach to manage bias but appending text prompts that really shows significant improvement in fairness.

V. Conclusion

In this work, we provide many definitions of bias as well as techniques for evaluating and mitigating it in text-to-image models. In this survey, we review the literature on gender, skin tone, and geo-cultural prejudice. It shows the positive and negative aspects of different methods for evaluation and mitigation of bias. As we reviewed the studies, we noticed several key aspects of the work: (1) In order to prevent misunderstandings when talking about "bias," it is crucial to make sure that definitions of bias are based on societal harms and to recognize the similarities and differences across suggested definitions; (2) Different sets of handmade prompts are commonly used, but there are no common assessment frameworks or metrics, and there are no unifying criteria for evaluation; and (3) It is necessary to conduct an extensive

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:15:29 UTC from IEEE Xplore. Restrictions apply.

study to find bias mitigation methods to provide fair images because current mitigation techniques are unable to completely address biases. From all the approaches to mitigating bias, we found that the prompt enhancement method will be the most crucial and will help in providing fair results. So we believe more exploration of this approach will be helpful for future works. We believe that this study provides academics and policymakers with a useful summary of the state of research on biases in T2I models and illuminates potential paths for developing and regulating reliable T2I systems.

REFERENCES

[1] Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour et al. "Photorealistic text-to-image diffusion models with deep language understanding." Advances in neural information processing systems 35 (2022): 36479-36494.

[2] Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695. 2022.

[3] Borji, Ali. "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2." arXiv preprint arXiv:2210.00586 (2022).

[4] Oppenlaender, Jonas. "The creativity of text-to-image generation." In Proceedings of the 25th International Academic Mindtrek Conference, pp. 192-202. 2022.

[5] Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. "Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets." 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. 1004-1015. 2021.

[6] Naik, Ranjita, and Besmira Nushi. "Social biases through the text-to-image generation lens." In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 786-808. 2023.

[7] Mariia Zameshina, Olivier Teytaud, and LaurentNajman. Diverse diffusion: Enhancing image diversity in text-to-image generation, 2023.

[8] Cheng Zhang, Xuanbai Chen, Siqi Chai, Henry Chen Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In ICCV, 2023.

[9] Ruifei He, Chuhui Xue, Haoru Tan, Wenqing Zhang, Yingchen Yu, Song Bai, and Xiaojuan Qi. Debiasing text-to-image diffusion models, 2024.

[10] Esposito, Piero, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. "Mitigating stereotypical biases in text to image generative systems." arXiv preprint arXiv:2310.06904 (2023).

[11] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. arXiv preprint arXiv:2312.01261, 2023.

[12] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Mira Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, Aylin Caliskan, et al. Easily accessible text-to-image generation amplifies demographic stereotypes at a large scale. In FAccT'23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, 2023.

[13] Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. arXiv preprint arXiv:2311.17695, 2023.

[14] Fraser, Kathleen C., Svetlana Kiritchenko, and Isar Nejadgholi. "A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified?." arXiv preprint arXiv:2302.07159 (2023).

[15] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models, 2023.

[16] Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. T2IAT: Measuring valence and stereotypical biases in text-to-image generation. In Findings of the Association for Computational Linguistics: ACL 2023, pp. 2560–2574, Toronto, Canada, July 2023.Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.160. URL https://aclanthology.org/2023.findings-acl.160.

[17] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. Advances in Neural Information Processing Systems, 36, 2024.

[18] Yixin Wan and Kai-Wei Chang. The male ceo and the female assistant: Probing gender biases in text-to-image models through paired stereotype test, 2024.

[19] Harvey Mannering. Analyzing gender bias in text-to-image models using object detection. arXiv preprint arXiv:2307.08025, 2023.

[20] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? arXiv preprint arXiv:2210.15230, 2022.

[21] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models, 2023.

[22] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Beyond the surface: A global-scale analysis of visual stereotypes in text-to-image generation. ArXiv, abs/2401.06310, 2024. URL https://api.semanticscholar.org/CorpusID:267959832.

[23] Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Patrick Schramowski, Kristian Kersting, et al. Exploiting cultural biases via homoglyphs in text-to-image synthesis. Journal of Artificial Intelligence Research, 78:1017–1068, 2023.

[24] Chhikara, Garima, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. "Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classification." arXiv preprint arXiv:2402.18502 (2024).

[25] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models, 2023.

[26] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. arXiv preprint arXiv:2311.07604, 2023.

[27] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In European Conference on Computer Vision, pp. 72–90. Springer, 2022.

[28] Yaniv Taigman,Ming Yang,Marc'Aurelio Ranzato, and LiorWolf. Deepface: Closing the gap to human-level performance in face verification. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708, 2014. doi: 10.1109/CVPR. 2014.220.

[29] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023.

[30] Alexandra Sasha Luccioni, Christopher Akiki, MargaretMitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408, 2023.

[31] Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. A unified framework and dataset for assessing gender bias in vision-language models, 2024.

[32] Noa Garcia, YusukeHirota, YankunWu, and YutaNakashima. Uncurated image-text datasets: Shedding light on demographic bias. In CVPR, 2023.

[33] Liang, Youwei, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun et al. "Rich Human Feedback for Text-to-Image Generation." arXiv preprint arXiv:2312.10240 (2023).

[34] Kimmo K¨arkk¨ainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. arXiv preprint arXiv:1908.04913, 2019.

[35] Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5136–5147, 2023.

[36] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7030–7038, 2023. URL https://api.semanticscholar.org/CorpusID:257505246.

[37] Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Patrick Schramowski, Kristian Kersting, et al. Exploiting cultural biases via homoglyphs in text-to-image synthesis. Journal of Artificial Intelligence Research, 78:1017–1068, 2023.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:15:29 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384427430

Advanced Deep Learning Approaches for Fault Detection and Localization in Power Distribution Systems

Presentation · September 2024

CITATIONS
0

READS
11

3 authors, including:



Jitendra Nath Rai
Delhi Technological University, Government of NCT of Delhi, India
82 PUBLICATIONS **462** CITATIONS

SEE PROFILE

All content following this page was uploaded by Jitendra Nath Rai on 28 September 2024.

The user has requested enhancement of the downloaded file.

Advanced Deep Learning Approaches for Fault Detection and Localization in Power Distribution Systems

Nitin Sharma
Department of Electrical Engineering
Delhi Technological University
Delhi, India
nitinsharma2k24@gmail.com

Pankaj
Department of Electrical Engineering
Delhi Technological University
Delhi, India
kr.pankaj9968@gmail.com

J.N Rai
Department of Electrical Engineering
Delhi Technological University
Delhi, India
jnrai@dce.ac.in

S.T. Nagarajan
Department of Electrical Engineering
Delhi Technological University
Delhi, India
stnagarajan@dce.ac.in

***Abstract*—The usage of Artificial Neural Networks (ANN) in improving electricity gadget operations is a burgeoning field of study, specifically within the context of fault detection and localization in strength distribution networks. This studies investigates the utility of ANNs to plot more efficient and correct methodologies for pinpointing disruptions within energy traces. Employing the Levenberg–Marquardt algorithm inside a MATLAB environment, the look at entails the development and schooling of an ANN version, which is intricately designed with various configurations of hidden layers and neurons. Data preprocessing is a crucial step to nice-music the model's overall performance. The findings of this studies imply a vast improvement in the detection and localization of faults over traditional methods, with the ANN model showing superior precision and swiftness. These promising effects suggest for persisted studies into ANN programs in strength structures, suggesting a capability paradigm shift in future energy device management and upkeep protocols.**

Keywords - Fault detection, Fault vicinity, Power distribution line, ANN, 11 kV, MATLAB

I. INTRODUCTION

The paramount venture of swiftly and accurately detecting and localizing faults in electricity transmission and distribution systems is vital to mitigate the extensive harm and provider interruptions they can reason [1][2]. Mitigating downtime and economic losses while concurrently improving the dependability and provider first-class of electricity lines is a high priority. Traditional safety mechanisms, which include relays, are trendy for fault detection and isolation [3]; however, all through instances of extreme fault pastime, which include storms, they could inundate operators with records, tough timely and correct fault decision due to human limitations [4]. The necessity for a automatic, absolutely automatic system that could technique and examine records in actual-time to exactly perceive fault locations is evident [5]. Fault detection and vicinity methodologies are normally classified into three essential types: impedance-primarily based [6], visiting wave [7], and know-how-based techniques [8]. Impedance-based strategies calculate line impedances to estimate the fault's role and are generally carried out to less difficult fault situations [9]. Traveling wave techniques, then again, time the waves' journey and reflection to identify fault locations [10]. These conventional methods are becoming less effective with the increasing complexity of strength systems, which now integrate a couple of allotted generation assets [11]. Knowledge-based totally strategies, which include Artificial Neural Networks (ANNs) [12], fuzzy logic, and machine mastering algorithms, have surfaced as promising alternatives, displaying higher accuracy in fault place than conventional techniques [13]. The reliability and balance of electrical grids hinge on proficient fault detection and location structures [14]. Advanced and automatic structures are key to directly pinpointing and rectifying faults, reducing down system downtime, and boosting transmission line reliability [15]. In precise, Artificial Neural Networks (ANNs) have won extensive traction as a way for fault detection and place in power systems [16]. The introduction shall cover a number of important studies that examine the use of ANNs for fault detection and location in transmission and distribution networks, such as underground or overhead cables [17][18]. The amassed studies underscores ANNs' speed, adaptability, and precision in fault prognosis, specially while combined with different analytic techniques inclusive of wavelet transform and optimization algorithms [19][20]. These research validate the capability of ANNs as a strong solution for energy device fault control, paving the manner for greater resilient electrical infrastructure [21].

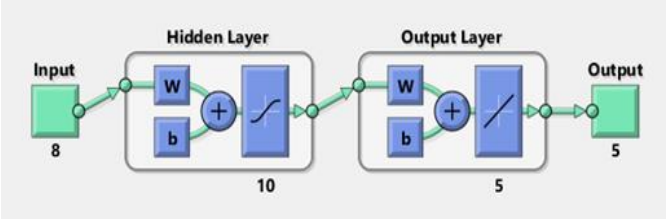
II. BACKGROUND ON DETECTION OF FAULTS

The research paper focuses on improving fault detection and localization in strength structures using Artificial Neural Networks (ANN), a branch of machine getting to know. The observe especially goals 11kV transmission lines, a important area in electricity distribution. This preference is strategic, as better transmission strains above 11kV normally we are using distance and differential relays for fault control. By focusing on 11kV traces, the studies addresses an opening in fault detection methodologies. The scope of the project encompasses quick transmission traces, with a specific attention on the ones extending up to 30 kilometers. The intention is to noticeably improve performance in fault detection and localization, leveraging the superior skills of ANNs to achieve this goal. This revolutionary method guarantees to carry a new stage of efficiency and accuracy for the management of power transmission systems.

A. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) are a class of machine mastering models that draw concept from the neural systems of the human mind, proposing interconnected nodes and that process statistics in a layered structure [16]. These fashions are adept at recognizing complicated patterns and making choices from the input data autonomously, without the want for explicit rule-based programming [17]. A distinct gain of ANNs is their potential for self-studying via publicity to data. They dynamically adjust inter-neuronal connection weights at some point of the schooling phase, which involves supplying the community with a large dataset and iteratively refining these weights to lessen the discrepancy between the network's predictions and the actual facts [18]. In the field of electrical energy systems, ANNs are carried out to lots of obligations which include making sure system stability [19], predicting load needs [20], and pin-pointing faults [8]. For fault area obligations, large testing brought about the improvement of an ANN version with a precise architecture tailored for this utility. The version accommodates nine enter elements that feed into a single hidden layer with twenty neurons, culminating in a single output that denotes the space to a fault in kilometers from a reference point, generally the first bus within the machine [12]. The efficacy of this model was verified based totally on several performance metrics, including the Mean Square Error (MSE), performance of education period, and the accuracy of the regression line and confusion matrix in representing the fault location [21].

Fig. 1. illustrates the ANN's structure with configuration.



The Simulink block diagram changed into meticulously crafted to provide a complete depiction of the neural community's design and operational mechanism, as highlighted in Figure 2. This diagram embodies the full spectrum of network elements: the enter hidden layer, layers, and the output layer, at the side of interconnecting synapses and their respective synaptic weights, important for the neural network's functionality [12]. Serving as the initial factor of contact for enter records, the enter layer channels data via the hidden layers, culminating within the output layer where the community formulates its inferences [16]. In this diagram, the synaptic pathways signify the waft of informational modern, whilst the synaptic weights suggest the strength of these connections, essential in determining the community's studying

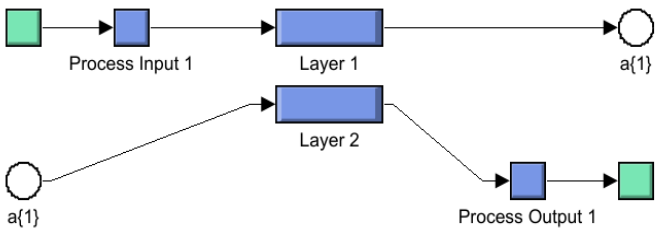


Fig. 2. Unmasked ANN Simulink Block Diagram

and version abilities [17]. The incorporation of non-linear activation capabilities in the hidden layers is instrumental, enabling the community to decipher and represent complicated enter-output relationships [18]. Moreover, this diagram delineates the network's vital learning algorithms, specifically highlighting

the roles of Backpropagation and Levenberg-Marquardt algorithms inside the schooling process [19][20].

A clear visual framework for understanding the neural network's structure and its problematic flow of statistics is provided by a unmasked Simulink block diagram. This clarity is instrumental in facilitating a extra sincere comprehension and assessment of the community's dynamics, especially whilst operated within the Simulink environment, underscoring the significance of such tools in superior neural community design and analysis [21].

B. METHODOLOGY

Before the statistics changed into used to tell the neural community model, it underwent a radical preprocessing segment. This technique standardized the records within a designated variety and segmented the overall dataset into distinct organizations for training, validation, and testing, adhering to a proportional break up of 70%, 15% and 15% respectively. Such preprocessing is fundamental to guarantee the model's capability to generalize from its training to novel, unseen records.

TABLE I. Input parameter in ANN

Input parameter of Artificial Neural Networks							
V_0	V_A	V_B	V_C	I_0	I_A	I_B	I_C

The structure and setup of the synthetic neural community had been meticulously designed with MATLAB Simulink. This vital phase covered figuring out the ideal quantity of hidden layers and neurons and choosing an effective schooling optimization set of rules. For this version, the Levenberg-Marquardt set of rules changed into the chosen method. The version's education engaged the distinctive schooling data subset, applying the selected optimization algorithm. This section became characterised by means of a sequence of iterative modifications to the version's parameters, aiming to reduce the discrepancy among the version's predictions and the real observed results. The Mean Squared Error (MSE) changed into utilized as a key indicator of overall performance at some point of the schooling section.

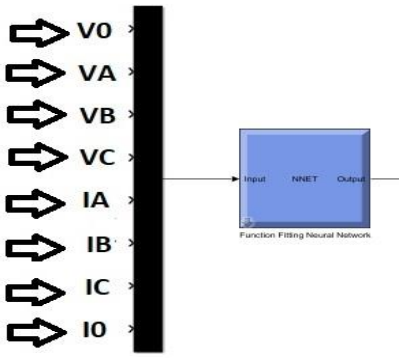


Fig. 3. input parameter to the ANN

Post-training, the model underwent a rigorous evaluation phase use of the reserved trying out dataset. This assessment involved a detailed comparison between the model's predictions

and the actual data, and the calculation of key performance indicators, such as model accuracy and precision.

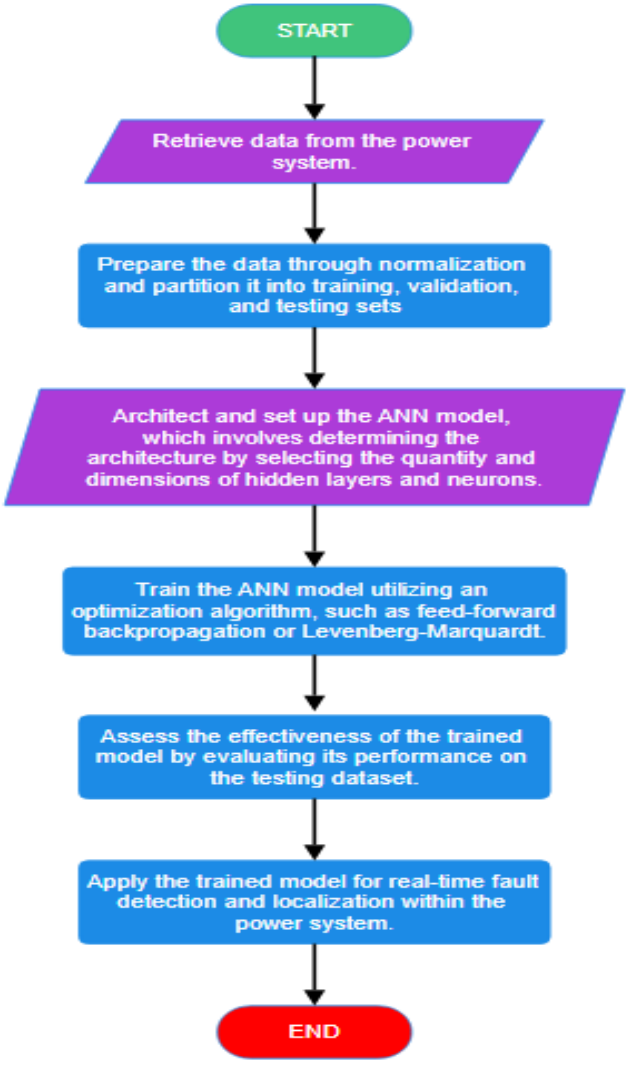


Fig. 4. methodology of the work

The culmination of this process was the deployment of the trained model for on-time detection and pinpointing of faults within the power distribution network. The machine constantly video display unit electrical parameters, leveraging the abilities of the skilled neural network to scrutinize the information for any symptoms of faults, thereby figuring out their presence and place. The approach taken was to utilize the capabilities of Artificial Neural Networks (ANNs) within MATLAB for identifying and pinpointing faults alongside electricity distribution lines. The manner started with thorough information preprocessing to make sure uniformity after which proceeded to the cautious production and training of an ANN version the usage of the Levenberg-Marquardt set of rules [12]. This meticulous guidance resulted in a relatively accurate and speedy system for fault detection and localization. The technique is depicted in a flowchart that illustrates how ANNs are applied in MATLAB for this reason, as shown in Fig 4.The information turned into gathered from a electricity distribution line, mainly from a 30-kilometer stretch below lots of fault scenarios, which include exceptional fault type, resistance stages, and categories. The dataset protected a range of electrical parameters along with the voltage and modern of three phases, the zero series voltage and modern (V0, I0). These parameters, which were sourced from bus1 as shown in the Figure 3 of Table I, have been fed into the ANN as enter variables.

The Mean Squared Error (MSE) was integral during the training phase, serving as the benchmark for performance and assisting in refining the model for optimal fault detection and location.

TABLE II. DECODER INPUTS FOR FAULT CLASSIFIER

Input for fault classification algorithm				Fault type
C_1	**C_2**	**C_3**	**C_4**	
1	0	0	1	A-G
0	1	0	1	B-G
0	0	1	1	C-G
1	1	0	0	AB
1	0	1	0	AC
0	1	1	0	BC
1	1	0	1	AB-G
1	0	1	1	AC-G
0	1	1	1	BC-G
1	1	1	0	ABC

III.SIMULATION RESEARCH

The Simulink model, as illustrated in Figure 5, adeptly represents an underground distribution line system, equipped to simulate a broad spectrum of fault types [12][16]. This model is proficient in generating a diverse range of faults, encompassing single-phase to ground (for phases A, B, C) [17], phase-to-phase (between phases A and B, C and B, A and C) [18], double-phase to ground (combinations of C, B, and A with ground) [19], three-phase fault (involving all phases A, B, and C) [20]. Such a comprehensive model is crucial for the in-depth analysis of various fault scenarios within the system, ensuring a robust training environment for the neural network [21].

Accompanying this model, the table meticulously outlines the primary characteristics of the power line used in the simulation. Data collection for neural network training was extensive, encompassing voltage and current measurements from three-phase lines, zero-sequence voltage and current [8]. A specific fault type, Phase A to ground, was strategically induced within the

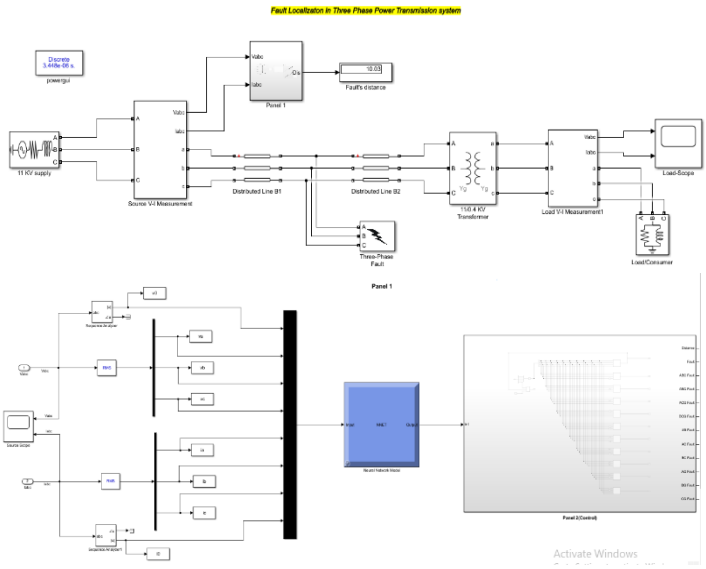


Fig. 5. Simulink Model for Fault Detection and Locations

simulation for a limited duration to gather data across various distances [14]. This procedure was vital for understanding the system response to different fault scenarios. The sequence analyzer tool in MATLAB played a pivotal role in determining the fault components [12], ensuring precision in data acquisition.

This method changed into now not confined to a single section fault type; it became replicated across a huge variety of fault eventualities, inclusive of simulations without any faults, to establish a baseline for comparative analysis [17]. The Simulink model, in its entirety, depicts a properly-rounded power machine geared up with vital components along with 3 bus bars, two transformers, generators, and a load. These additives are interconnected with segments of distribution lines in which faults are systematically introduced for thorough evaluation [19] Such a detailed illustration is important for the neural network's getting to know system, ensuring a complete expertise of fault dynamics in energy distribution structures.

A. ANALYSIS

This research demonstrates the use of synthetic neural community (ANN) for correctly locating faults on distribution lines. A simulation, jogging for a duration of O.5 seconds, became performed to check the device's efficiency across various fault situations. The ANN's schooling changed into primarily based on a dataset that Contained 9,889 samples, and the progression of the training phase of Figure 6.

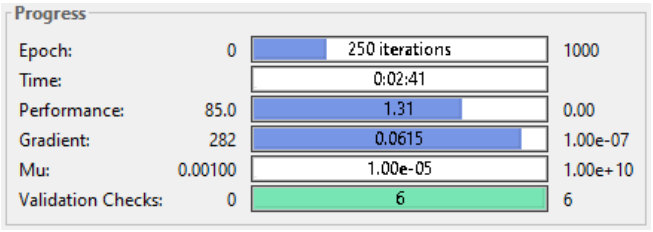


Fig. 6. training progress

Figure 7 showcases the exemplary validation performance achieved during the ANN's training phase, quantified by the Mean Square Error (MSE). After training, the neural network showcased an exceptional ability to identify fault locations along the network with an estimation accuracy of approx. 99.1% for pinpointing faults, as detailed in Figure 11. Furthermore, the system achieved a flawless detection rate of faults at 100%.

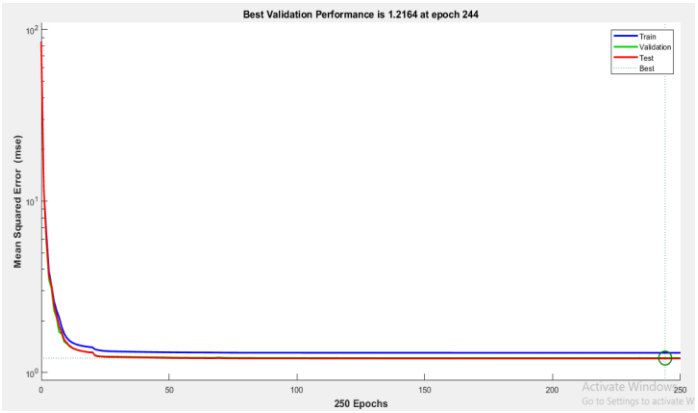


Fig. 7. ANN training performance

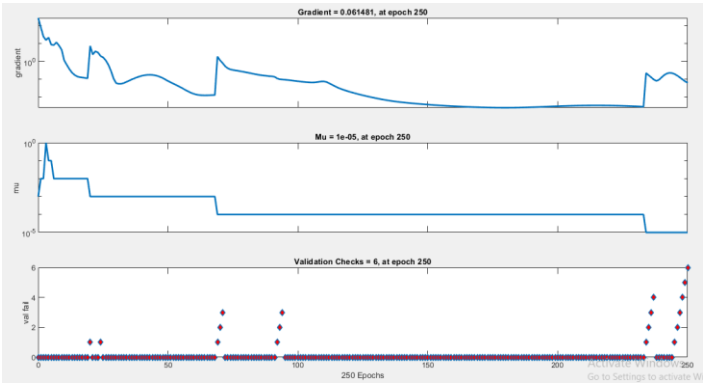


Fig. 8. ANN training regression

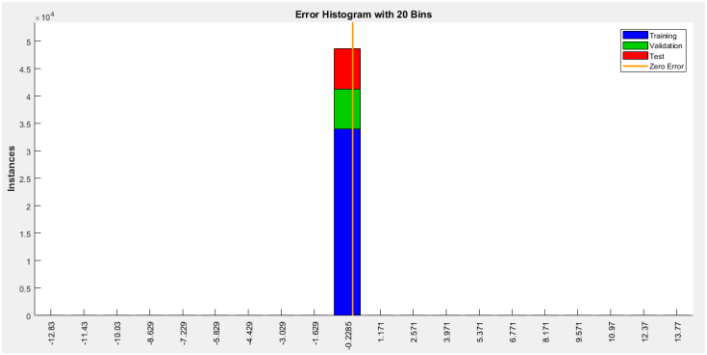


Fig. 9. error histogram

The results of the ANN training are shown in the figures displayed. The error histogram, represented in Figure 9, shows a Gaussian distribution indicating minimal deviation from the excepted results. Figure 10 presents the training state plot, which demonstrates positive trends in the ANN's learning process.

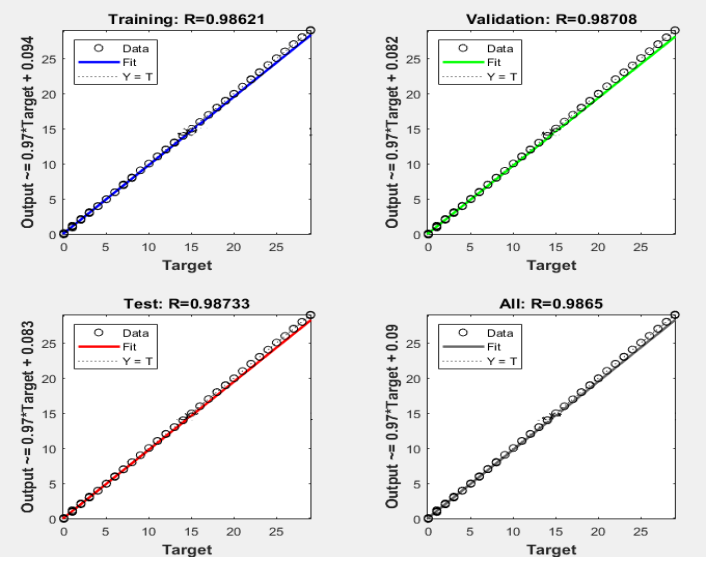


Fig. 10. training state plot

TABLE III. ANN estimates the location of a fault

S.No.	Fault Location	Fault Location Estimated by ANN					
		AB	AC	BC	AG	BG	CG
1	1.00	2%	0%	10%	1%	1%	2%
2	2.56	2%	1%	1%	1%	0%	0%
3	3.00	1%	0%	0%	0%	0%	0%
4	5.56	5%	5%	2%	0%	0%	0%
5	6.00	1%	0%	0%	0%	0%	0%
6	8.56	5%	4%	2%	0%	0%	0%
7	9.00	0%	0%	0%	0%	0%	0%
8	11.56	4%	3%	2%	0%	0%	0%
9	12.00	0%	0%	0%	0%	0%	0%
10	14.56	3%	3%	2%	0%	0%	0%
11	15.00	0%	0%	0%	0%	0%	0%
12	15.80	3%	3%	2%	0%	0%	0%
13	18.00	0%	0%	0%	0%	0%	0%
14	18.80	2%	2%	2%	0%	0%	0%
15	21.00	0%	0%	0%	0%	0%	0%
16	21.80	2%	2%	2%	0%	0%	0%
17	24.00	0%	0%	0%	0%	0%	0%
18	27.00	0%	0%	0%	0%	0%	0%
19	27.80	1%	2%	2%	0%	1%	0%
20	29.00	0%	0%	0%	0%	0%	0%

Tables III and IV detail the outcomes of the research regarding fault localization within power system lines. These tables enumerate the kinds of faults and the corresponding line distances at which these faults were identified. The total line length assessed was 30 kilometers, to evaluate the system's performance in the fault scenarios, twenty random points along this line were selected for analysis.

TABLE IV. ANN estimates the location of a fault - 2

S.No.	Fault location	Fault Location Estimated by ANN			
		ABC	ABG	ACG	BCG
1	1.00	1%	5%	2%	3%
2	2.56	1%	5%	1%	0%
3	3.00	0%	5%	1%	0%
4	5.56	1%	3%	1%	7%
5	6.00	0%	0%	1%	0%
6	8.56	0%	8%	2%	4%
7	9.00	0%	0%	0%	0%
8	11.56	1%	2%	1%	2%
9	12.00	0%	0%	0%	0%
10	14.56	1%	1%	1%	1%
11	15.00	0%	0%	0%	0%
12	15.80	1%	0%	1%	1%
13	18.00	0%	0%	0%	0%
14	18.80	1%	3%	1%	1%
15	21.00	0%	0%	0%	0%
16	21.80	1%	3%	1%	1%
17	24.00	0%	0%	0%	0%
18	27.00	0%	0%	0%	0%
19	27.80	1%	2%	1%	1%
20	29.00	0%	0%	0%	0%

The outcomes of the research are deemed satisfactory, indicated by the table which reflects the proximity of the determined fault locations to the actual points of fault. The results of the study are considered acceptable, as shown in the table indicating the proximity of the identified fault points to the real faults. In order to sum up all the individual % errors and then divide them by a number of data points, the average percentage error is derived. For the calculation of the percentage error PE, the following equation shall be used: This equation used to compute the percentage error (PE) is as follows:

$$PE = (|\text{guess dist.} - \text{real dist.}|) / \text{real dist.} \times 100\% \quad (1)$$

Therefore, the formula to calculate the average % error of MPE will be as follows:

$$MPE = (\Sigma(|\text{guess dist.} - \text{real dist.}|) / \text{real dist.} \times 100\%) / n \quad (2)$$

TABLE V. ANN estimates the location of the MPE of faults

	Mean percentage error for fault localization					
	AB	BC	AC	AG	BG	CG
n=20	2%	1%	1%	0%	0%	0%

TABLE VI. ANN estimates the location of the MPE of faults part 2

	Mean percentage error for fault localization			
	ABG	ACG	BCG	ABC
n=20	2%	1%	1%	1%

B. RESULT AND DISCUSSION

In this paper, an ANN is used to locate fault points on the distribution line and a MATLAB code file to cross check the type of fault occurred in transmission line. The model will be simulated for a period of 0.5 seconds and will be able to determine the fault for each type of fault. In the simulation file the occurred fault is line to line fault (BC fault) at distance of 5 KM from grid.

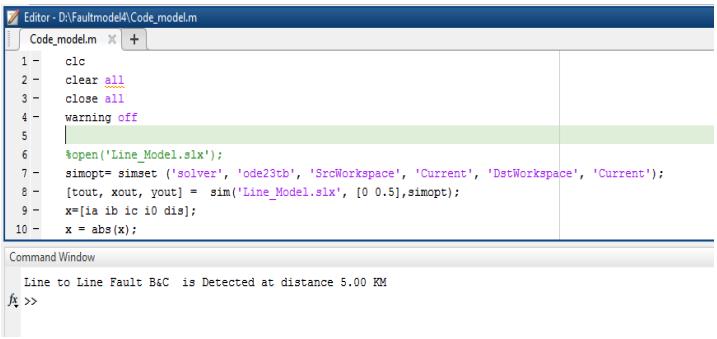


Fig. 13

From Figure 13. The conditional statements, with clearly defined fault criteria, demonstrate a logical approach to interpreting the ANN output. These constructs are pivotal in making real-time decisions about the presence and location of faults within the network.

The collection of plots provides a comprehensive visual representation of the ANN's performance. The distinct subplots probable reflect diverse metrics inclusive of accuracy, loss, precision, and do not forget, vital for evaluating the version's effectiveness in fault detection and localization. The smoothness of the curves and their convergence patterns propose a solid learning system, indicating the robustness of the ANN below the selected configurations.

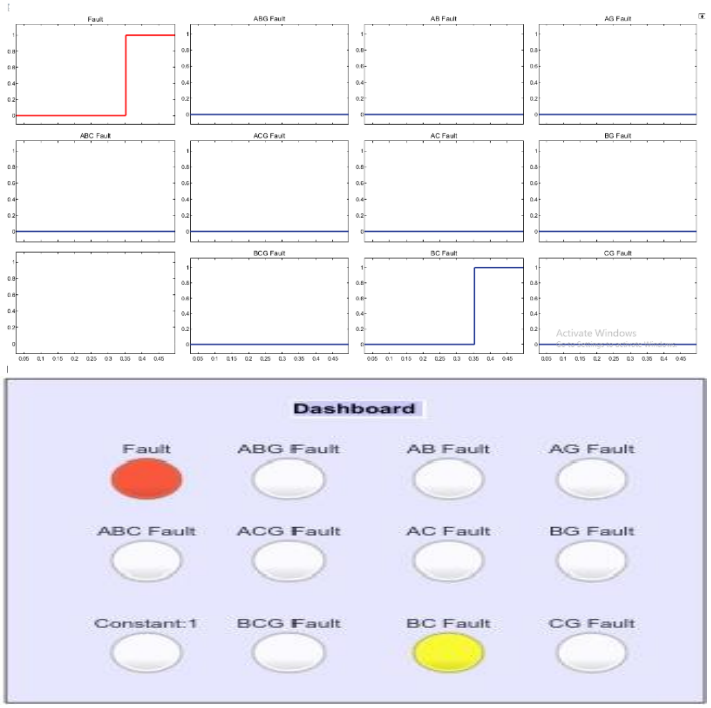


Fig. 14. Graphical Representation of Fault Classification

IV. Conclusion

The take a look at findings verify that the Artificial Neural Network (ANN) accurately pinpoints fault locations in power distribution strains, mirroring real records with high constancy. The ANN's consistency across unique fault kinds indicates a nicely-skilled version capable of enhancing operational performance by decreasing gadget downtime. Future paintings have to examine the ANN's overall performance with new records and through the years to ensure reliability. This studies underscores the ANN's capacity as a reliable device for fault detection in energy structures.

References

[1] Goh, H. H., et al. reviewed power system fault location methods in the 2017 *Indonesian Journal of Electrical Engineering*, 8, 206-212.

[2] Dustegor, D., Poroseva, S. V., Hussaini, M., & Woodruff, S. (2010). Automated graph-based methodology for fault detection and location in power systems. *IEEE Transactions on Power Delivery*, 25, 638-646.

[3] Mukherjee, A. K., Kundu, P. K., & Das, A. (2021). Transmission line faults in power system and the different algorithms for identification, classification, and localization: A brief review of methods. *Journal of The Institution of Engineers: Series B*, 1-23.

[4] Saha, M. M., Izykowski, J., & Rosolowski, E. (2010). *Fault location on power networks*. Springer.

[5] Dashti, R., Ghasemi, M., & Daisy, M. (2018). Fault location in power distribution network with presence of distributed generation resources using impedance based method and applying π line model. *Energy*, 159, 344-360.

[6] Shafiullah, M., & Abido, M. A. (2017). A review on distribution grid fault location techniques. *Electric Power Components and Systems*, 45(8), 807-824.

[7] Akmaz, D., Mamiş, M. S., Arkan, M., & Tağluk, M. E. (2018). Transmission line fault location using traveling wave frequencies and extreme learning machine. *Electric Power Systems Research*, 155, 1-7.

[8] Lout, K., & Aggarwal, R. K. (2012). A feedforward Artificial Neural Network approach to fault classification and location on a 132kV transmission line using current signals only. In *2012 47th International Universities Power Engineering Conference (UPEC)* (pp. 1-6). IEEE.

[9] Aslan, Y., & Yağan, Y. E. (2017). Artificial neural-network-based fault location for power distribution lines using the frequency spectra of fault data. *Electrical Engineering*, 99(1), 301-311.

[10] Karić, A., Konjić, T., & Jahić, A. (2018). Power system fault detection, classification, and location using artificial neural networks. In *Advanced Technologies, Systems, and Applications II* (pp. 89-101). Springer International Publishing, Cham.

[11] Dashtdar, M. (2018). Fault location in distribution network based on fault current analysis using artificial neural network. *International Journal of Electrical and Computer Sciences (IJECS)*, 1(2), 18-32.

[12] Hasija, K., Vadhera, S., Kumar, A., & Kishore, A. (2014). Detection and location of faults in underground cable using Matlab/Simulink/ANN and OrCad. In *2014 6th IEEE Power India International Conference (PIICON)* (pp. 1-5). IEEE.

[13] Rafinia, A., & Moshtagh, J. (2014). A new approach to fault location in three-phase underground distribution system using combination of wavelet analysis with ANN and FLS. *International Journal of Electrical Power & Energy Systems*, 55, 261-274.

[14] Rocha, S. A., Mattos, T. G., Cardoso, R. T., & Silveira, E. G. (2022). Applying artificial neural networks and nonlinear optimization techniques to fault location in transmission lines—statistical analysis. *Energies*, 15(11), 4095.

[15] Nag, A., & Yadav, A. (2016). Artificial neural network for detection and location of faults in mixed underground cable and overhead transmission line. In *CIIT 2016*.

[16] Wang, L., & Fu, X. (2009). Artificial neural networks. *Journal of the Royal Society of Medicine*, 92(7), 385.

[17] Zhu, A.-X. (2017). Artificial neural networks.

[18] Fogel, D. B., Fogel, L. J., & Porto, V. W. (1990). Evolving neural networks. *Biological Cybernetics*, 63, 487-493.

[19] Gabella, M., Afambo, N., Ebli, S., & Spreemann, G. (2019). Topology of learning in artificial neural networks. *ArXiv*, vol. abs/1902.08160.

[20] Bansal, R. C. (2006). Overview and literature survey of artificial neural networks applications to power systems (1992 - 2004).

[21] Mohatram, M., Tewari, P., & Shahjahan. (2011). Applications of artificial neural networks in the electric power industry: A review.

View publication stats

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384431195

Advanced Tor-NonTor Traffic Classification

Conference Paper · June 2024

CITATIONS
0

READS
2

2 authors, including:



Jitendra Nath Rai
Delhi Technological University, Government of NCT of Delhi, India
82 PUBLICATIONS **462** CITATIONS

SEE PROFILE

All content following this page was uploaded by Jitendra Nath Rai on 28 September 2024.

The user has requested enhancement of the downloaded file.

Advanced Tor-NonTor Traffic Classification

Khushi
Department of Electrical Engineering
Delhi Technological University
Delhi, India
khushisindhu01@gmail.com

Dr. Jitendra Nath Rai
Department of Electrical Engineering
Delhi Technological University
Delhi, India
jnrai@dce.ac.in

Abstract*— This work aims to classify Tor-NonTor traffic through UNBCIC, a global dataset. It primarily considers two separate classification scenarios, Scenario-A to categorize Tor-NonTor and Scenario-B to classify Tor into applications. We evaluate time-based dataset attributes which provide the temporal context to categorize the traffic. A comprehensive process involves data pre-treatment, feature engineering, and a variety of categorization methods. To combat class imbalance, we utilize SMOTE oversampling. The paper uses multiple classification algorithms: Random Forest, Decision Tree, SVM, KNN, XGBoost, Logistic Regression, Gradient Boosting, Gaussian Naive Bayes, AdaBoost, Bagging Classifier, and Extra Trees. The study employs K-fold cross-validation at several stages to enhance model generalization. The paper measures the model performance using accuracy, precision, recall, and F1 score and further visualizes the results using confusion matrices. We utilize learning curves to analyze model training and validation. We calculate and visualize feature importance that indicates classification drivers. Additionally, the ROC curves along with micro and macro-averaged curves in multiclass scenarios give a complete idea of model discriminatory capabilities. The paper includes a separate section for the Deep Neural Network (DNN) model along with learning and validation loss curves. This facilitates a more thorough investigation of the classification algorithms. Finally, the study concludes with a comprehensive outcome discussion to highlight the pros and cons of different classification methods. The outcome from the work serves as a pathway for future work on Tor-NonTor traffic categorization and has significant connotations for cybersecurity applications.

Keywords*— *Tor, Network Traffic Characterization, Network Traffic Analysis, Time-based Features, Machine Learning, Feature Engineering, Hyperparameter Tuning

I. INTRODUCTION

It is well known that network security depends on the accurate identification of Tor and nonTor traffic. This is key to addressing various threats and malicious behaviors. Tor is used to protect user data by encrypting it and sending it over Tor relays. The network traffic type and identification for the traffic need to be handled effectively. NonTor traffic can be understood to be ordinary network communication and it will not be affected by the Tor anonymization. This is widely used in the field of threat intelligence, network forensics and intrusion detection. It holds Prime importance. Accurate classification of Tor and NonTor traffic helps network managers and security analysts to discriminate legitimate traffic and potentially malicious activity so as to improve network security. The UNBCIC dataset is one of the most important benchmark datasets in the area of Tor-NonTor traffic classification. The dataset includes the network traffic and the traffic is classified as Tor or NonTor using CIC-Flowmeter. It provides better chance to perform detailed traffic analyses and classification. It features several important detailed time-based and statistical attributes derived from network traffic patterns. The goal is to increase the classification performance of classification models and reduce the impact of class imbalance in the dataset by using more sophisticated machine learning algorithms, feature engineering, and sampling strategies. In this regard, this study presents a detailed investigation on Tor-NonTor Traffic Classification. In order to demonstrate the model's performance, three main preprocessing steps: instantiation, feature engineering techniques and a combination of different classification models are used for comprehensive investigation of Tor detection techniques. We describe the configuration of experimental study and the criteria that examined the model, and utilize outcomes of comparison of several classification techniques. By using the presented research findings, it is possible to improve the positive elements of network security implementations, as well as contribute to the advancement of knowledge within the cybersecurity field.

II. RELATED WORK

A. Literature Review

There has been much research and development in the classification of traffic involved with traffic involved both in Tor and NonTor networks which has led to the development of many methodologies and techniques that would help in improving the classification accuracy as well as improving security of the networks. Our earlier evaluation captures the most important information steps of previously implemented approaches, also analysis and discuss the research gaps that leads to the foundation of our comparative study of the current paper which helps us in enhancing the Tor-NonTor traffic classification.

The applications of machine learning have received considerable attention due to the ability of achieving fine-grained classification of traffic in the Tor network and distinguish various traffic types in traffic streams [[2], [3], [17]]. These studies further brought about the significance of having stronger classification function to support the security and anonymity of the network.

Manually configuring various features and parameters for detection and analysis has been one of the defining problems of recent research; and while traffic encryption, such as VPN and Tor, is frequently cited as a problem, solutions have been found regarding the time characteristics of the traffic and the application of machine learning algorithms [[2], [17]]. Such concern reflects the importance of properly categorizing encrypted communication in order to differentiate between various activities on the anonymity network.

Consequently, studies on traffic analysis attacks over the Tor network demonstrate the practice’s feasibility and efficiency in real-life utilization [10]. Understanding threat models and attack techniques is paramount for

TABLE I. COMMONLY USED DATASETS FOR EVALUATING TOR-NONTOR TRAFFIC CLASSIFICATION ALGORITHMS

UNB-CIC Tor Network Traffic Dataset	Darknet Traffic Dataset	Whonix Installation Dataset	ISCX VPN-nonVPN Traffic Dataset	MAWI Working Group Traffic Archive
Contains Tor and non-Tor traffic with categories like audio streaming, browsing, chat, file transfer, email, P2P, video streaming, and VoIP. Widely used for evaluating classification models and traffic analysis.	Includes traffic from anonymous communication networks like Tor, I2P, JonDonym, and ZeroNet. Utilized for studying user behavior in darknet traffic and refining the granularity of darknet traffic identification.	Generated from a Whonix installation, a Tor-integrated Linux-based OS. Contains Tor traffic data categorized into various application types. Used for identifying packets from the Tor network and evaluating classification models.	Contains VPN and non-VPN traffic for evaluating the effectiveness of traffic classification algorithms in distinguishing between VPN and non-VPN traffic. Used for empirical probability mass function analysis and neural network-based traffic classification.	Offers a large collection of network traffic data, including anonymized traffic traces from various sources. Widely used for evaluating traffic classification algorithms and network analysis.

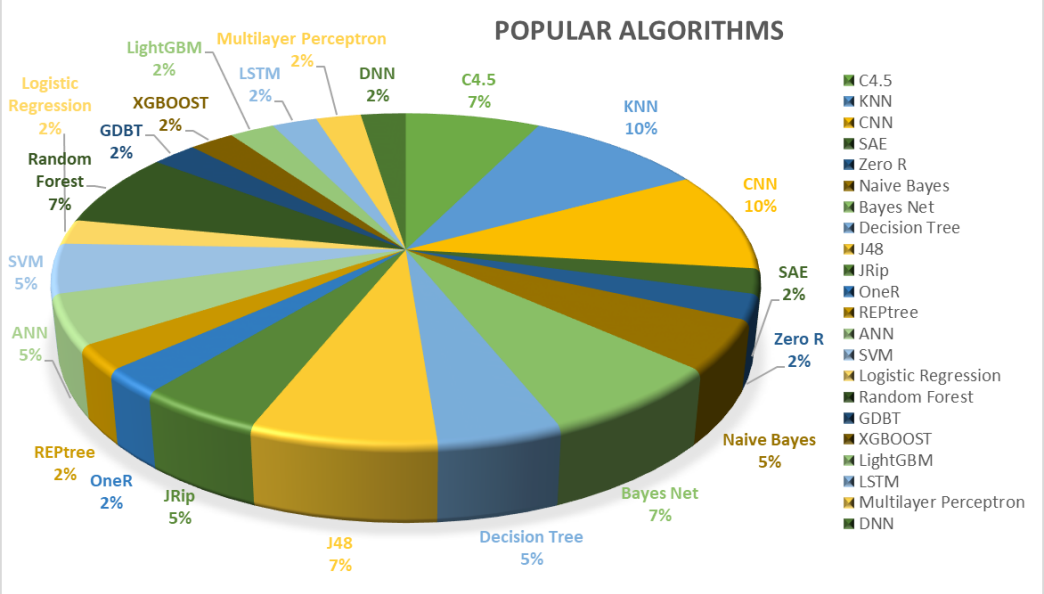


Fig. 1. ML Algorithm Usage Trends in Tor-NonTor Traffic Classification Research.

deploying robust defenses to mitigate privacy risks associated with traffic analysis attacks.

Novel approaches suggested for obtaining high-accuracy Tor traffic classification include hierarchical classification, payload analysis, and time analysis and which frequently use deep NNs and ML [[1], [5]–[8], [12], [13], [15]]. These methodologies are useful in classifying traffic and establishing the likelihood of threats and risks in the construction of infrastructures.

Proposed actions that favor improvement of Tor systems in the real time traffic classification as well as QoS include, [1]. These methods are useful in as much as they seek to optimise utilization of network resources depending with the nature of traffic flow with a view of enhancing end user satisfaction as well as the optimisation of the flow of traffic in a network.

Furthermore, studies have explored the validity of detection approaches to differentiate between Tor and non-Tor connections within the Tor network [9]. Machine learning algorithms present the ability to identify non-Tor traffic accurately and effectively in ways that strengthen the security system of the network.

The establishment of labeled datasets on Tor traffic with focus on the identification of key features for classification has been highlighted, with potential benefits for the exploration of the experimental validation suite [10], [16]. These activities are useful in the creation and testing of the classification techniques to enable the algorithms to work effectively in the actual environment.

Recent works also highlighted the significance of time based features to support the traffic classification process, for traffic application differentiation [11] and for the Tor network categorization [12]. Given this, flow timeout values refer to other factors that determine the impact of the classifier, with restricted timeout values being associated with high capacity classifiers. Incorporating insights from various sources, we present popular datasets used in Tor-NonTor traffic classification research in Table I, showcasing the diversity of data sources.

TABLE II. ML ALGORITHMS IN TOR-NONTOR TRAFFIC CLASSIFICATION.

Convolutional Neural Networks (CNN)	Decision Tree And Bayesian Approaches	Profile Hidden Markov Models (PHMM)	Random Forest And Gradient Boosting
➢ Sirinam et al. (2018): • Compared CNNs with k-NN, CUMUL, k-fingerprint attack, and SDAE. • Demonstrated CNN superiority for Tor network attacks.	➢ Lotfollahi et al. (2018): • Developed Tor-IDT, an improved decision tree algorithm. • Used Tri-Training for categorizing Tor traffic. ➢ Pescape et al. (2018): • Applied five machine learning classifiers based on Bayesian approaches.	➢ He et al. (2018): • Employed PHMM for determining application types of encrypted Tor traffic.	➢ Cuzzocrea et al. (2017): • Utilized Random Forest and Gradient Boosting for Tor traffic analysis. ➢ Cuzzocrea et al. (2021): • Emphasized ensemble learning techniques for improved accuracy.

Additionally, Figure 1 highlights ML algorithm usage trends in Tor-NonTor traffic classification research, offering insights into the prevalence of different machine learning techniques and their adoption over time.

Furthermore, Figures 2 and 3 provide valuable context for understanding the evolution and influence of research in this domain, showcasing the publication timeline of selected papers in Tor-NonTor traffic analysis and their scholarly impact through citation count, respectively.

Table II provides a comprehensive breakdown of Machine Learning algorithms used in Tor-NonTor traffic classification research, detailing the papers utilizing each algorithm and their methodologies.

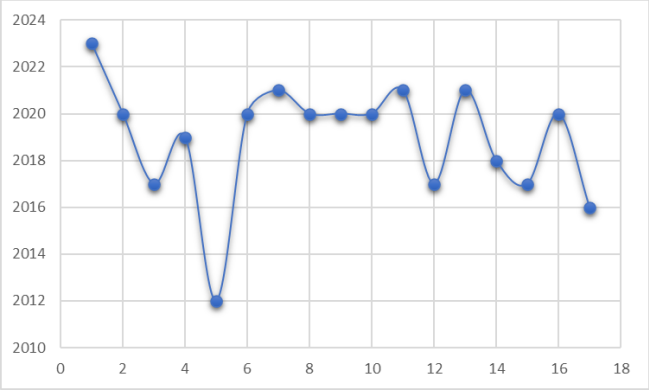


Fig. 2. Publication Timeline of Chosen Papers.

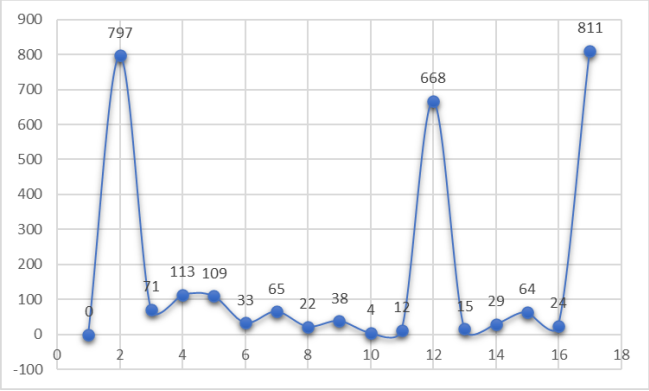


Fig. 3. Citations of Chosen Papers in Traffic Classification Research.

Through our comparative study, we aim to contribute to the advancement of Tor-NonTor traffic classification methodologies, addressing key research gaps and enhancing network security and performance.

B. *Research Gaps*

Despite significant improvements in the field of classification algorithms, many of the limitations in the current techniques for Tor-NonTor traffic classification remain unsolved. The biggest challenge is the class imbalance in the dataset where the number of Tor traffic flows is much lower than the NonTor samples. This class imbalance is the basic problem leading to the biased classification results, and deteriorating the classification performance especially for minority classes. In addition, the fact that most of the work reported in the literature uses binary classification makes it far easier to resolve the issue than the much more complex and difficult multi-class classification scenarios encountered in real network traffic. There is also a significant lack of evaluation measures and evaluation of the performance of different classifiers. Finally, the dynamic nature of network traffic and more importantly the continuously changing adversarial tactics will require adaptive and resilient classification systems that can adapt to the changing behavior of the network and emerging threats. This calls for the requirement of scalable and efficient classification algorithms needed to manage big data solutions in the field of Tor-NonTor traffic classification. However, after surmounting the challenges highlighted herein and undertaking further research on new methods and techniques that are unique to classification of Tor and NonTor traffic, it is only possible to redefine the benchmark of possible outcomes of network traffic classification techniques and enhance security against cyber terrorism.

III. DATASET DESCRIPTION

A. *UNBCIC Dataset*

Existing non-Tor traffic data are scarce and inconclusive, and the UNBCIC Collections offer a powerful foundation for categorization of Tor-NonTor traffic and undergo constant expansion. It will provide detailed information about the UNBCIC dataset as in detail as possible on its origin, size, features, and other important factors aiding its analysis.

B. Scenario-A and Scenario-B

The UNBCIC dataset comprises two separate situations, each addressing specialized classification tasks:The UNBCIC dataset comprises two separate situations, each addressing specialized classification tasks:

Scenario-A: Tor-NonTor Classification: The main work of many studies and papers in the network and traffic classification domain is to distinguish between Tor and NonTor traffic samples and traffic..

Scenario-B: Tor Application categorization: In contrast, there is the work similar to Scenario-B that focuses on the categorization of the tor application, doing the work of classification in a more detailed manner, acknowledging and classifying types of traffic depending on the application of the tor network.

TABLE III. DESCRIPTION OF DATASET SCENARIOS.

Dataset Type	Dependent Variable/Class Label	Description
Scenario-A	Tor Non-Tor	Classifying the data-flows into being from Tor Traffic or Non-Tor Traffic.
Scenario-B	Voip Video-Streaming File-Transfer Chat Browsing Email P2P Audio-Streaming	Classifying the data-flow into the application type being used in Tor Network.

C. Origin and Size

UNBCIC contained in the University of New Brunswick Cybersecurity comprises real traces captured from computer networks including a laboratory environment. The sample is a compendium of network flow records including traffic coming from the Tor network and normal network traffic samples. These network flows were generated with the help of the CICFlowMeter tool which works as follows: It collects individual packets' data and then converts it to the flow-level records for analysis.

Quantitatively, the UNBCIC dataset is made up of significantly large network flow data, including many datasets with sizes, time durations, and flow timeout values that may differ. Concerning the data set, it is available in several variants including: TimeBasedFeatures-15s, TimeBasedFeatures-30s, TimeBasedFeatures- 60s, TimeBasedFeatures-120s, reflecting various flow timeouts.

D. Features and Characteristics

The UNBCIC dataset deals with the large number of characteristics collected from the network flow records, and therefore, contains significant amount of alternatively, many elements of the communication structure and traffic characteristics. Such features include static and dynamic traffic features, where statistical properties, temporal properties and properties that are based on the consensus protocol give valuable characteristics about the traffic that is generated from the Tor and Non-Tor networks.

Looking at the Tor and nonTor packets in the dataset, it is quite clear that the Tor network has its own inherent characteristics such as flow patterns and encryption mechanisms. Tor traffic may present encrypted connections, where it is difficult, if not impossible, to understand the data packets' content and their source and destination addresses are scambled to mask the users' interactions.

In contrast, NonTor traffic presents itself as regular traffic that can be seen as network traffic which employs clear identification of the act of communication and recognizable patterns of the communication process..

TABLE IV. TIME BASED FEATURES WITH THEIR DESCRIPTIONS

Feature	Description	Count
fiat	Forward Inter Arrival Time, the time between two packets sent forward direction (mean, min, max, std).	4
biat	Backward Inter Arrival Time, the time between two packets sent backward (mean, min, max, std).	4
flowiat	Flow Inter Arrival Time, the time between two packets sent in either direction (mean, min, max, std).	4
active	The amount of time a flow was active before going idle (mean, min, max, std).	4
idle	The amount of time a flow was idle before becoming active (mean, min, max, std).	4
fb psec	Flow Bytes per second.	1
fp psec	Flow packets per second.	1
duration	The duration of the flow.	1

E. Time-Based Features and Flow Timeout Values

This brings us to the time-based properties of the UNBCIC data: this part of the data contains temporal attributes that characterize the Nature of network flow and its importance for categorization. These time-based characteristics include fiat, active duration, idle duration, flowiat, biat, and flow rate such as bytes per seconds and packets per second.

Further, the information includes flow timeout which is or defining the maximum allowable time that the specific stream or flow will be active for flow since it can be set to be inactive or terminated. Different flow timeout settings involve differences in the temporal characteristics of the network flows, and this, in turn, can influence the categorization outcomes. The four flow timeouts, namely, 15 seconds, 30 seconds, 60 seconds, and 120 seconds are compared so as to understand the impact of these settings on classification and the characteristics of the models.

IV. Preprocessing Steps

The preprocessing steps are listed in this section with the main focus on the chronological order in which they were achieved in order to improve the analysis of the data set.

Data Loading and Initial Cleanup:

a) Data Loading: The dataset was loaded from the CSV file into a pandas dataframe.

b) Column Name Cleanup: Leading whitespace from column names (excluding the first column) was deleted to ensure uniformity and convenience of access.

c) Label Renaming: The column that is already named ‘label’ was renamed as ‘class1’ to meet the naming convention used in the study.

d) Irrelevant Field Removal: Some of the fields that were considered not significant or even unimportant in the analysis were removed from the dataset including ‘Source IP’, ‘Source Port’, ‘Destination IP’, ‘Destination Port’ and ‘Protocol’.

e) Note: These processes were specifically applied to the CSV data files (i.e., Scenario-A-merged_5s, SelectedFeatures-10s-TOR-NonTOR, Scenario-B-merged_5s, TimeBasedFeatures-10s-Layer2). For the ARFF files, the data was already in the appropriate format.

Handling Negative Values: Nonexistent values were detected, and assignment of negative values was made NaN to resolve inconsistencies within the dataset that might bias the analysis.

Handling Infinite Values: Infinite values (INF/-INF) within the dataset were identified and replaced with NaN to ensure data integrity and prevent errors during analysis.

Imputation of Missing Values: To prevent omission of records and improve statistic quality, several imputation methods were applied to complete the records with missing values. Among the treatment of missing values, measures that were adopted are filling in the absent values with arithmetic mean of the values belonging to the corresponding characteristic but in this case differentiated between the classes.

Outlier Detection and Removal: Before the analysis and modeling phase, to account for the outliers – that may significantly distort the results – the Isolation Forest algorithm was employed. Outliers were identified and eliminated to reduce their impact on data fitting and improve the stability of the models.

Data Resampling for Class Imbalance: To tackle class imbalance, the dataset was balanced by applying oversampling methods such as Synthetic Minority Over-sampling Technique (SMOTE) to the selected dataset. This compounded the samples in the minority class thus making the model more capable of making good generalizations.

Feature Scaling: The values of the selected features were normalized using the Min-Max scaling technique because it adjusts the values of the characteristics so that they range between zero and one, thereby making each of the features significant in the analysis rather than letting features with large values to thoroughly influence the model.

Encoding the Target Variable: The target variable 'class1' was encoded using Label Encoding to convert categorical labels into numerical format, facilitating the training of classification models.

Data Splitting for Training and Testing: Stratified sampling procedures were adopted to split the dataset into training and testing sets, ensuring a balanced representation of Tor and NonTor traffic samples. This approach minimized bias and facilitated fair evaluation of the classification models.

More precisely, by paying a focused attention to all the issues related to data quality and applying a comprehensible number of basic data preprocessing measures, the dataset was readied for the subsequent analysis and modelling.

V. Methodology

A. Model Training and Evaluation

The models were validated using the K-Fold cross validation where the data is divided into various folds to ensure the validity of the model. The hyperparameters were tuned by applying randomized search cross-validation to determine the optimal hyperparameters for the models. For the evaluation of the whole performance, assessment measures such as accuracy, precision, recall, F1 score and confusion matrix were used on each model.

B. Learning Curves and Loss Curves

Graph plots of learning curves for more than one model and training/validation loss curve for DNN model are plotted and analyzed to assess the model accuracy and its convergence. These curves offer information on the model’s ability to learn from data and differentiate between overfitting and underfitting, and guide the next steps in optimization.

C. Feature Importance

The importance of the features was evaluated across all the classifiers including the DNN with the intention to determine the level of importance of the features that were used in the models for predicting the target variable. The goal of this analysis was to show which features are suitable for solving the problem of classification. In order to verify the significance of the feature selection scheme on the choice of classifiers, it was decided to assess the importance of all the features when used on their own.

To this end, we used all the features in the training process for each classifier individually and then; normalized the

importance scores of all classifiers. This normalization made certain that the summed up feature importance values represented the general contribution toward classification by each feature, and not just due to the influence of any specific classifier that might have excessively high values.

D. ROC Curve

Receiver Operating Characteristic (ROC) curves were constructed to evaluate the performance of each classifier by examining the trade-off between true positive rate and false positive rate across multiple threshold settings. Micro-average and macro-average ROC curves were also created to examine the overall performance of multi-class classifiers, including the DNN.

VI. RESULTS

This section is dedicated to sharing our findings on the tasks of classifying theTor-NonTor traffic in various scenarios. Due to the fact that the datasets include several files that combined make a diverse dataset, each file having a different flow timeout parameter, it has been possible to adopt a complex approach to analyze the results. To avoid excessive density of the results section and for ease of understanding, we have only reported selected results relevant to a few typical files for each of the scenarios in the main body of the manuscript. The following files shall be considered as best summarizing a variety of trends witnessed throughout this analysis.

A. Scenario-A File-1 (15s)

Other included factors include numerical characteristics that involved time-specific features and other network values like source IP, destination IP, source port, destination port, protocol, and a fraction dependent variable were utilized were divided into the train and test sets. It is worth stressing that, the dataset features only numerical characteristics, allowing omitting more complex stages to address the category variables. In order to achieve stratification and ensure that similar results occur when the process is repeated, which is always advisable, the necessary techniques like stratified sampling were applied when splitting data into two sets, the first for training the model and the other for testing it.

Data Distribution Analysis: The dataset comprised network flows with a flow timeout of 15 seconds. Initial data distribution showed a majority of non-Tor instances, with a ratio of approximately 5.65 non-Tor instances for every Tor instance. After preprocessing, the dataset was balanced with 17,020 instances for both non-Tor and Tor categories.

Feature Reduction: Feature reduction techniques effectively streamlined the dataset, reducing the feature set by approximately 8.70%.

TABLE V. DATA PREPROCESSING DETAILS FOR FILE-1 SCENARIO-A (15S)

SCENARIO-A 15-sec	
DATA DISTRIBUTION	
Original	
Non-Tor	18758
Tor	3314
Number of Negative Values	
68603	
Number of Inf Values	
0	
After Outlier Removal	
Non-Tor	17020
Tor	2844
After Oversampling	
Non-Tor	17020
Tor	17020
SCENARIO-A 15-sec	
Feature Reduction	
Total Features	23
Reduced Features	21

Classifier Performance: Various classifiers demonstrated exceptional performance, with mean accuracy (cross-validation) ranging from 98.29% to 100.00%. Among the classifiers, Logistic Regression, Decision Tree, and XGBoost stood out for achieving perfect accuracy, precision, recall, and F1 scores. However, Gaussian Naive Bayes, although not reaching perfect scores, performed remarkably well, achieving an accuracy of 98.29% and balanced precision, recall, and F1 scores of approximately 98.35%. Conversely, while all classifiers performed admirably, Gradient Boosting exhibited a slightly lower performance compared to others, with mean accuracy (cross-validation) of 99.99%.

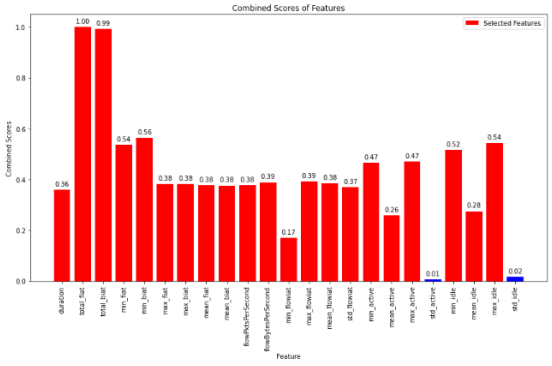


Fig. 4. Results of feature selection, highlighting the selected features based on their combined normalized importance scores across different feature selection techniques. [For File-1 (15s) Scenario-A]

TABLE VI. CONSOLIDATED RESULTS FOR FILE-1 SCENARIO-A (15S)

SCENARIO-A 15s									
Classifier	Mean Accuracy (Cross Validation)		Accuracy	Precision	Recall	F1 Score	Mean Traning Score	Mean Validation Score	ROC AUC
Random Forest	99.99%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
Decision Tree	99.98%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
SVM	99.99%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
KNN	100.00%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
XGBoost	99.98%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
Logistic Regression	100.00%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
Gradient Boosting	99.99%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
Gaussian Naive Bayes	98.29%		98.35%	98.40%	98.35%	98.35%	0.98	0.98	1.00
AdaBoost	99.99%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
Bagging Classifier	99.99%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
Extra Trees	100.00%		100.00%	100.00%	100.00%	100.00%	1.00	1.00	1.00
	Average Evaluation Metrics Across Folds		Accuracy	Precision	Recall	F1 Score	Mean Traning Loss	Mean Validation Loss	ROC AUC
DNN	Accuracy	100.00%	100.00%	100.00%	100.00%	100.00%	0.01	0.00	1.00
	Precision	100.00%							
	Recall	100.00%							
	F1 Score	100.00%							

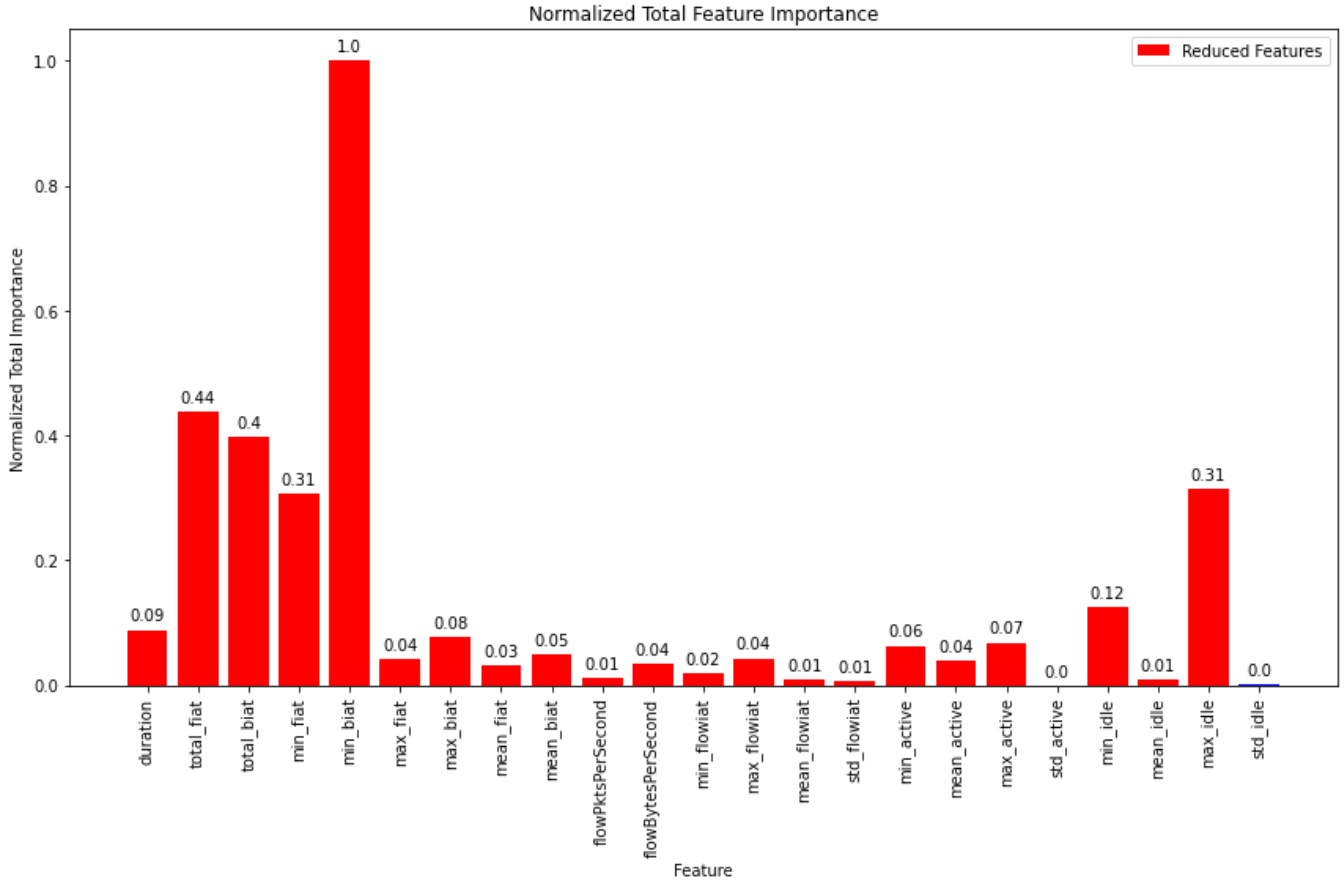


Fig. 5. Combined feature importance after model training. [For File-1 (15s) Scenario-A]

Learning Curve Metrics: Training and validation scores remained consistently high across classifiers, indicating stable model performance without overfitting or underfitting.

Receiver Operating Characteristic Area Under the Curve (ROC AUC): ROC AUC scores consistently approached or reached 1.00, signifying excellent discrimination performance for all classifiers.

Model Comparison: Logistic Regression, Decision Tree, and XGBoost emerged as top performers, achieving perfect scores across all evaluation metrics. Gaussian Naive Bayes, despite its simplicity, demonstrated impressive performance, especially in accurately classifying Tor and non-Tor instances. Gradient Boosting, while still performing well, showed slightly lower accuracy compared to other classifiers.

CONCLUSION

Consequently, for the purpose of the current research, we undertook a systematic analysis of Tor-NonTor Traffic classification under Different flow timeout in the context of Scenario-B. From such a vigorous experiment and analysis procedure, several postulate findings were established. Secondly, the results indicate that there is a general trend of increased accuracy and other evaluation measures once outliers as well as the data set size are appropriately adjusted. This further emphasises the need for data pre-processing in the improvement of the effectiveness and generalisation capabilities of such classification models.

Secondly, based on the analysis of multiple classifiers, it was observed that the performance of the models under study is depends on the different flow timeout. Compared to Simple, Clear, and J48 classifiers, the performance of other classifiers like Random Forest, Gradient Boosting, XGBoost was very stable across all flow timeout values in terms of accuracy and AUC score, although at the initial stages the accuracy of some classifiers was high and then reduced and increased at later stages. This gives emphasis on the significance of proper classifiers depending much on the network conditions and traffic.

In addition, the present research highlighted the importance of feature reduction in reducing classification costs as well as enhancing the performance of the classifiers. To this end, we demonstrated that even after filtering out superfluous features and thereby removing information that, in some cases, might be crucial for classification, maximum classification performance can be approached with far less computational costs than if all features were used.

The studies for future directions can look at different classification techniques, solve the problem of scalability and incorporate anomaly detection mechanism to make the traffic classification system capable of addressing new methodologies employed by the security threats in the future. Further, replication of a benchmark dataset similar to the one employed in this study would enable comparison with other studies was well as expand the under researched Tor-NonTor traffic classification.

REFERENCES

[1] I. Karunanayake, N. Ahmed, R. Malaney, R. Islam, and S. K. Jha, "Darknet Traffic Analysis: Investigating the Impact of Modified Tor Traffic on Onion Service Traffic Classification," *IEEE Access*, vol. 11, pp. 70011–70022, 2023, doi: 10.1109/ACCESS.2023.3293526.

[2] M. Lotfollahi, M. Jafari Siavoshani, R. Shirali Hossein Zade, and M. Saberian, "Deep packet: a novel approach for encrypted traffic classification using deep learning," *Soft comput*, vol. 24, no. 3, pp. 1999–2012, Feb. 2020, doi: 10.1007/s00500-019-04030-2.

[3] J.-Y. Nie, Institute of Electrical and Electronics Engineers, and IEEE Computer Society, *Tor Traffic Analysis and Detection via Machine Learning Techniques*.

[4] IEEE Communications Society and Institute of Electrical and Electronics Engineers, *FlowPic: Encrypted Internet Traffic Classification is as Easy as Image Recognition*.

[5] T. Yu, A. Association for Computing Machinery. Special Interest Group on Security, National Science Foundation (U.S.), Association for Computing Machinery, and ACM Digital Library., *Enhancing Tor's Performance using Real-time Traffic Classification*.

[6] L. Wang, H. Mei, and V. S. Sheng, "Multilevel Identification and Classification Analysis of Tor on Mobile and PC Platforms," *IEEE Trans Industr Inform*, vol. 17, no. 2, pp. 1079–1088, Feb. 2021, doi: 10.1109/TII.2020.2988870.

[7] T. Shapira and Y. Shavitt, "FlowPic: A Generic Representation for Encrypted Traffic Classification and Applications Identification," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1218–1232, Jun. 2021, doi: 10.1109/TNSM.2021.3071441.

[8] D. Sarkar, P. Vinod, and S. Y. Yerima, "Detection of Tor Traffic using Deep Learning," in *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, IEEE Computer Society, Nov. 2020. doi: 10.1109/AICCSA50499.2020.9316533.

[9] Y. Hu, F. Zou, L. Li, and P. Yi, "Traffic Classification of User Behaviors in Tor, I2P, ZeroNet, Freenet", doi: 10.1109/TrustCom50675.2020.00064/20/$31.00.

[10] Jāmiʿat Qaṭar, Institute of Electrical and Electronics Engineers. Qatar Section., and Institute of Electrical and Electronics Engineers, *Traffic Analysis Attacks on Tor : A Survey*.

[11] S. D. Mehta and D. Upadhyay, "A Review on Classification of Tor-Nontor Traffic and Forensic Analysis of Tor Browser." [Online]. Available: https://privacy.net/what-is-tor/

[12] P. Choorod and G. Weir, "Tor Traffic Classification Based on Encrypted Payload Characteristics," in *Proceedings - 2021 IEEE 4th National Computing Colleges Conference, NCCC 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021. doi: 10.1109/NCCC49330.2021.9428874.

[13] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of tor traffic using time based features," in *ICISSP 2017 - Proceedings of the 3rd International Conference on Information Systems Security and Privacy*, SciTePress, 2017, pp. 253–262. doi: 10.5220/0006105602530262.

[14] C. Johnson, B. Khadka, E. Ruiz, J. Halladay, T. Doleck, and R. Basnet, "Application of deep learning on the characterization of tor traffic using time based features," *Journal of Internet Services and Information Security*, vol. 11, no. 1, pp. 44–63, 2021, doi: 10.22667/JISIS.2021.02.28.044.

[15] T.-H. Meen, Institute of Electrical and Electronics Engineers. Tainan Section. Sensors Council Chapter, International Institute of Knowledge Innovation and Invention, and Institute of Electrical and Electronics Engineers., *Tor Traffic Classification from Raw Packet Header using Convolutional Neural Network*.

[16] E. Hodo, X. Bellekens, E. Iorkyase, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Machine Learning Approach for Detection of nonTor Traffic", doi: 10.1145/123_4.

[17] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," in *ICISSP 2016 - Proceedings of the 2nd International Conference on Information Systems Security and Privacy*, SciTePress, 2016, pp. 407–414. doi: 10.5220/0005740704070414.

View publication stats

scientific reports

Check for updates

OPEN

Advancing sentiment classification through a population game model approach

Neha Punetha & Goonjan Jain ✉

Computational Sentiment Analysis involves the automation of human emotion comprehension by categorizing sentiments as positive, negative, or neutral. In the contemporary digital environment, the extensive volume of social media content presents significant challenges for manual analysis, thereby necessitating the development and implementation of automated analytical tools. To address the limitations of existing techniques, which heavily rely on machine learning and extensive dataset pre-training, we propose an innovative unsupervised approach for sentiment classification. This novel methodology is grounded in game theory concepts, particularly the population game model, offering a promising solution by circumventing the need for extensive training procedures. We extract two textual features from review comments, namely context score and emotion score. Leveraging lexicon databases and numeric scores, this cognitive mathematical framework is language-independent. Competitive results are demonstrated across various domains (hotels, restaurants, electronic devices, etc.), and the efficacy of the proposed work is validated in two languages (English and Hindi). The highest accuracy recorded for the English domain dataset is 89%, while electronic Hindi reviews attain an 84% accuracy rate. The proposed model exhibits domain and language independence, validated through statistical analyses confirming the significance of the findings. The framework demonstrates noteworthy rationality and coherence in its outcomes.

Keywords Sentiment classification, Context score, Emotion score, Population game model

In the expanding digital ecosystem, where text is continuously generated across diverse platforms and devices, the task of accurately interpreting the sentiments embedded within the text has become a critical area of study. Computational Sentiment Analysis, an integral subfield of Natural Language Processing (NLP), addresses this challenge by automating the complex process of analyzing human sentiment conveyed through language. This task seeks to elucidate the intricate layers of emotions, opinions, and attitudes present in textual data, which has wide-ranging applications in numerous fields[1]. The vast and growing volume of online content—from consumer reviews to social media posts—renders manual analysis impractical and often infeasible. As a result, there is an increasing demand for automated sentiment analysis tools that can handle extensive datasets efficiently. Various sophisticated algorithms have been developed to fulfill this need, playing a crucial role in deriving actionable insights from textual data. Such advancements are essential for enabling organizations to make data-driven decisions and sustain a competitive edge in today's digital landscape[2].

For the extraction of sentiment from text, three types of approaches are most commonly used in the literature: lexicon-based approaches, machine-learning approaches, and hybrid approaches. Lexicons are predefined dictionaries of words or phrases with associated sentiment scores. However, these lexicons are not exhaustive and may not include all the words or expressions in a given domain or language[3]. This can lead to inaccurate sentiment analysis results, particularly for newer or domain-specific terms. In machine learning approaches, one major drawback is their dependence on high-quality, labeled training data. Acquiring and annotating large amounts of data can be time-consuming, expensive, and subjective[2]. Additionally, machine learning models can suffer from overfitting and fail to generalize well to new or unseen data. Furthermore, adapting models to new domains or handling noisy and unstructured data can pose significant challenges. Hybrid approaches in sentiment analysis, which combine multiple techniques such as lexicons, machine learning, and rule-based systems, also have certain drawbacks[4]. One drawback is increased complexity and computational overhead. Integrating different approaches and coordinating their interactions can be challenging, requiring additional effort in feature extraction, model design, and optimization[5]. Another drawback is the potential for conflicting results or

Department of Applied Mathematics, Delhi Technological University, New Delhi, India. ✉email: goonjan_jain@hotmail.com

errors when different techniques yield contrasting sentiment predictions. Combining disparate methods may introduce additional sources of noise and ambiguity, leading to decreased accuracy and reliability. Furthermore, hybrid approaches may require domain-specific knowledge and expertise to effectively select, combine, and fine-tune the various components. Overall, while hybrid approaches have the potential to leverage the strengths of multiple techniques, careful consideration, and rigorous evaluation are necessary to ensure their effectiveness and mitigate their drawbacks[6]. To overcome the limitations associated with traditional methods, we present a novel approach for sentiment analysis at the sentence level in review datasets. The primary goal is to develop a decision-making system capable of effectively determining the sentiment of online reviews. The main contributions of this study are as follows:

1. Innovative mathematical sentiment analysis approach: This study presents a groundbreaking unsupervised sentiment classification model based on population game theory. Unlike traditional models, it does not require labelled datasets, making it highly adaptable and reducing the dependency on extensive pre-existing data. This innovation addresses a significant limitation in sentiment analysis by offering a more flexible and scalable solution.
2. Dynamic and real-time sentiment tagging: The population game model dynamically adapts to changing language patterns and sentiment expressions over time. By framing sentiment categories as players and sentiment labels as strategies, the model achieves a Nash equilibrium that ensures stable and accurate sentiment classification. This dynamic adaptation is crucial for maintaining model effectiveness in real-world applications where language and context continuously evolve.
3. Cross-language and cross-domain robustness: The model's design emphasizes language and domain independence, proven effective across various datasets, including English and Hindi reviews from multiple domains like hotels, restaurants, and electronic devices. This broad applicability highlights the model's robustness and potential for wide-scale deployment in diverse sentiment analysis tasks.
4. Enhanced accuracy through strategic interaction modelling: By capturing the strategic interactions between context and emotion within an asymmetric game framework, the model provides a more nuanced understanding of sentiment dynamics. This leads to enhanced accuracy, with the study achieving high-performance metrics (89% for English and 84% for Hindi), demonstrating the model's superiority over existing sentiment analysis techniques.

The study is structured as follows: Section "Related work" offers a concise review of previous scientific work in the realm of sentiment analysis. Section "Methodology" elaborates on the core tenets of the population game theory model and the proposed algorithmic approach for sentiment analysis. In Section "Result and evaluation", we assess the efficacy of the proposed model by juxtaposing it against alternative methods across different review datasets. This evaluation is substantiated through adherence to specified guidelines and statistical significance tests, accompanied by a thorough analysis of the algorithm's strengths and limitations. Section "Discussion." encapsulates the error rate, significance, real-life implementation, and limitations of the study and delineates potential avenues for future research. Finally, Section "Conclusion" presents the conclusion and the future work.

Related work

Textual sentiment analysis methods include Topic-based approaches, machine learning approaches, deep learning approaches, and mathematical modeling approaches. Each of these methods has its strengths and limitations, and understanding these can help identify the gap that our proposed work aims to fill as illustrated in Table 1.

Addressing research gap

Below are some of the research gaps that we identified by analyzing different approaches, mentioned in Table 1.

1. Topic modeling: Topic modeling methods, such as (W2VLDA), often lack contextual understanding, relying on word co-occurrence without grasping the meaning, which can result in incoherent topics. The need to predefine the number of topics can be challenging, and the models are computationally intensive and sensitive to hyperparameters. They also ignore word order, affect meaning, and may struggle with short texts. Additionally, interpretability and pre-processing quality can impact their effectiveness.
2. Machine learning: Machine learning methods, such as Lagrangian Support Vector Machine (LSVM) and Bidirectional Neural Networks, face major challenges while handling high-dimensional data, particularly, overfitting, especially with limited training data. Recurrent neural network (RNN)-based attention models are particularly prone to overfitting. Ontology embeddings struggle with polysemy and ambiguity in domain-specific languages. Selective Domain Adaptation (SDA) encounters difficulties in managing domain shifts while maintaining model interpretability. Keyphrase extraction methods using fuzzy entropy and k-means clustering, which rely on predefined algorithms, may miss emerging sentiment patterns. Furthermore, methods like BERT Post Training (BERT-PT) have high demands for computational resources and labeled data for fine-tuning.
3. Deep learning: Deep learning models, including BERT and Attention-Based Point Networks (AttPNet), have significant drawbacks due to their substantial computational resource requirements and the complexity involved in their deployment. Techniques such as Recurrent Memory Neural Networks (ReMemNN) experience performance degradation when dealing with noisy or sparse data. BERT with knowledge incorporation demands extensive computational power and pre-training. Syntactic and semantic analysis approaches strug-

Sr. no.	References	Category	Approach	Strengths	Limitations	Research gap filled by the proposed method
1	García-Pablos et al.[7]	Topic modelling	Word2Vector-Latent Dirichlet Allocation (W2VLDA)	Almost unsupervised, versatile across domains	Requires extensive tuning, and variable performance across domains	Consistent performance across domains with less tuning required
2	Yiran and Srivastava[8]	Deep Learning	Linear Discriminant Analysis (LDA)	Robust for topic clustering and sentiment analysis	Challenges in interpretability, limited practical applicability	Improved interpretability and practical applicability
3	Song et al.[9]	Deep Learning	Attention-based Long Short-Term Memory (LSTM)	Enhances aspect-level sentiment analysis with sentiment lexicon embedding	Deteriorates with longer sequences, impacting effectiveness	Maintains effectiveness with long sequences without attention mechanism limitations
4	Mee et al.[10]	Machine Learning	Hybrid approach (regression and sentiment analysis)	Versatile, combines multiple methods	Relies heavily on TF-IDF, and can overlook contextual nuances	Better capture of contextual nuances without relying solely on TF-IDF
5	Fiok et al.[11]	Deep Learning	Unsupervised sentiment analysis for Twitter accounts	Scalable, handles large volumes of data	Struggles with evolving language trends and slang usage	Adapts more effectively to evolving language trends and slang
6	Daniel and Meena[12]	Machine Learning	Lagrangian Support Vector Machine (LSVM)	High accuracy	Struggles with high-dimensional data require careful feature selection	Handles high-dimensional data more efficiently without extensive feature selection
7	Basiri et al.[13]	Machine Learning	Bidirectional neural network	Addresses LSTM and GRU limitations	Suffers from vanishing gradient problems in deep architectures	Avoids vanishing gradient problems through different modelling techniques
8	Perikos et al.[2]	Machine Learning	Recurrent neural network (RNN)-based attention models	Effective in sentiment analysis tasks	Prone to overfitting with limited training data	Reduces overfitting risk by not relying on large labeled datasets
9	Nguyen et al.[14]	Machine Learning	Ontology embedding	Enhances aspect capturing	Faces difficulties with polysemy and ambiguity in domain-specific languages	Training independent
10	Dai et al.[15]	Machine Learning	Selective Domain Adaptation (SDA)	Feature-level transfer of knowledge	Challenges in handling domain shifts and maintaining model interpretability	Maintains interpretability and adapts better to domain shifts
11	Vashishtha and Susan[4]	Machine Learning	Keyphrase extraction using fuzzy entropy and k-means clustering	Offers interpretability	Relies on predefined clustering algorithms, and might overlook emerging sentiment patterns	Dynamically captures emerging sentiment patterns without relying on predefined algorithms
12	Peng et al.[1]	Machine Learning	BERT Post Training (BERT-PT)	Adaptability, state-of-the-art performance	Requires substantial computational resources and labeled data for fine-tuning	Requires fewer computational resources and works well without extensive labeled data
13	Chiha et al.[16]	Machine Learning	Sentiment Analysis Based on Sentiment Clustering (SASC)	Combines deep learning and fuzzy logic	Sensitive to choice of clustering algorithms and sentiment cluster granularity	Less sensitive to clustering algorithm choices and granularity
14	Hu et al. [17]	Deep Learning	Attention-based Point Network (AttPNet)	Innovative attention mechanisms	Complexity in training and deployment, especially with limited computational resources	Simplifies training and deployment while maintaining effectiveness
15	Liu and Shen[18]	Deep Learning	Recurrent Memory Neural Network (ReMemNN)	Enhances sentiment analysis with memory components	Performance degrades with noisy or sparse data	Maintains performance even with noisy or sparse data
16	Zhao and Yu[19]	Deep Learning	BERT with knowledge incorporation	Enhanced aspect-based sentiment analysis	High computational demands and pre-training requirements	Reduces computational demands and pre-training requirements
17	Zhang et al.[20]	Deep Learning	Syntactic and semantic analysis	Comprehensive understanding	High computational demands may limit scalability	Improves scalability while maintaining comprehensive understanding
18	Atandoh et al.[21]	Deep Learning	BERT-MultiLayer Convolutional Neural Network (B-MLCNN)	Computational efficiency	Overlooks nuanced sentiment expressions within sentences or phrases	Better captures nuanced sentiment expressions within sentences or phrases
18	Zuheros et al.[22]	Mathematical model	SA-MpMcDM for TripAdvisor	Effective in specific domains	Effectiveness varies across different review platforms, requires extensive customization	More adaptable across different review platforms without extensive customization
20	Punetha and Jain[23]	Mathematical model	Bayesian game model	Suitable for sentiment classification	Relies on weights allocated to textual features criteria	Rating independent, adaptability over time
21	Various (SOTM)[3], (GRA-ST)[24], (TOGT-ST)[3], (COPGT-ST)[25])	Mathematical model	MCDM-based frameworks for sentiment classification	Different approaches for sentiment orientation and tagging	Reliance on weights and extensive customization requirements	Independent of weights assigned to the criteria

Table 1. Comparative evaluation of advanced sentiment analysis methodologies.

gle with scalability due to high computational demands. BERT-Multi Layer Convolutional Neural Network (B-MLCNN) often overlooks nuanced sentiment expressions within sentences or phrases.

4. Mathematical approaches: Existing mathematical models for sentiment analysis, such as SA-MpMcDM, face drawbacks due to the need for extensive customization across different review platforms, which limits their adaptability. Bayesian game models based approaches are dependent on the probability of the occurrence of textual features. Various MCDM-based frameworks for sentiment classification are challenged by their reliance on weights and the extensive customization required.

In summary, the proposed method fills these research gaps by offering consistent performance across domains, effective handling of high-dimensional data, simplified deployment, and adaptability to language trends and domain shifts, thus ensuring scalability and robustness in sentiment analysis. These contributions aim to enhance the reliability, applicability, and practicality of sentiment analysis methodologies in real-world scenarios.

Methodology

The utilization of the population game models for sentiment analysis is given in Section "Population game model". Section "Proposed methodology" contains the proposed methodology for sentiment analysis. We give a detailed illustrative example of the proposed model in Section "Numerical illustration".

Population game model

The population game model is a mathematical framework that integrates concepts from game theory and evolutionary biology to analyze the dynamics of strategic interactions within a population[26]. The population game model offers a compelling framework for sentiment analysis due to its alignment with the dynamics of strategic interactions inherent in sentiment classification.

Sentiment analysis inherently involves strategic interactions between different sentiment categories or strategies. By framing sentiment categories as players and sentiment labels as strategies within the population game model, we accurately capture the strategic nature of sentiment classification. This representation allows for a nuanced understanding of how different sentiment strategies compete and evolve. As a first step, the model defines the "players" as various elements contributing to sentiment, such as context and emotion, while sentiment categories like positive and negative sentiments are regarded as distinct "strategies." The performance of each strategy is then quantified through a payoff matrix, indicating how well the strategies predict sentiment in given text data. From there, the model calculates the fitness of each sentiment category based on its predictive capability in the given text data. This fitness assessment forms the basis for determining the dynamic changes in strategy proportions over time. Utilizing replicator dynamics, the model describes how the proportions of different sentiment strategies evolve. This dynamic process ensures that the sentiment analysis model adapts and optimizes its strategies over time in response to changing language patterns, sentiment expressions, and user preferences. As the proportions of sentiment categories evolve, the model reaches a steady state, representing a Nash equilibrium where the distribution of sentiment categories stabilizes. At this point, no single sentiment category can unilaterally improve its performance, indicating a convergence to a stable state where the sentiment analysis model maintains a consistent distribution of sentiment categories. Throughout this iterative process, the population game model enables continuous adaptation and optimization of sentiment analysis algorithms. By updating strategy proportions according to replicator dynamics, the model ensures adaptability to changes in sentiment trends and linguistic patterns, ultimately leading to improved performance and adaptability in real-world applications. Evaluation and validation of the model's performance at this steady state help ensure its effectiveness in accurately capturing sentiment in text data.

Overall, the population game model offers a suitable theoretical framework for sentiment classification due to its ability to capture the dynamic nature of sentiment analysis, optimize sentiment analysis algorithms, achieve stable equilibrium, and facilitate systematic evaluation and validation. By leveraging the strategic interactions inherent in sentiment classification, this model enhances the accuracy, adaptability, and performance of sentiment analysis systems in various real-world applications.

We employ the population game model to accurately reflect the diverse and varying influences of context and emotion on strategic interactions. Unlike symmetric games, where all players have identical strategies and payoffs, asymmetric games allow us to capture the inherent differences between the players' roles and payoffs in our model. This distinction is crucial because the strategies and outcomes for context and emotion are not identical and affect each other differently. By employing an asymmetric game framework, we can more precisely model and analyze the dynamics of how context and emotion interact, evolve, and reach equilibrium, leading to a more realistic and applicable understanding of the strategic behavior within the population.

A population game model is represented by a tuple $G = <N, S, A, B, p, q, F_1, F_2, E_1, E_2, \dot{p}, \dot{q}>$ where each component is defined as follows.

1. Players (N): The model consists of N players. In the specific context of this task, there are two players: context (C) and emotion (E) for sentiment analysis.
2. Strategies (S): The strategies available to players. In the proposed work, we have two strategies viz., positive and negative.
3. Payoff matrices (A and B):

   Matrix A represents the payoffs for player 1 (context). It is denoted by $A = [a_{ij}]$, where a_{ij} represents the payoff for context playing strategy i against emotion playing strategy j.

   Matrix B represents the payoffs for player 2 (emotion). It is denoted by $B = [b_{ij}]$, where b_{ij} represents the payoff for emotion playing strategy j against context playing strategy i.

4. Strategy proportion (p and q):
   p: A vector of proportions for Player 1 (context). Since we use only two strategies, thus this vector is of size two (02). The individual values are referred to as p_1 and p_2. p_1 is the proportion of context choosing the positive strategy and p_2 is the proportion of context choosing the negative strategy. The additional constraint is: $p_1 + p_2 = 1$ with $p_1, p_2 \geq 0$.
   q: A vector of proportions for Player 2 (emotion). Similarly, q_1 is the proportion of emotion choosing the positive strategy and q_2 is the proportion of emotion choosing the negative strategy. The constraint is $q_1 + q_2 = 1$ with $q_1, q_2 \geq 0$.
5. Combined strategy proportion (π): A combined vector representing the strategy proportions of both players.

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ q_1 \\ q_2 \end{bmatrix}$$

where $\pi_1 = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$ represents the strategy proportions of player 1 and $\pi_2 = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$ represents the strategy proportion of player 2.

6. Fitness functions (F_1 and F_2):
   The fitness of individual players in the population game model is calculated to understand and analyze the dynamics of the population and how different strategies contribute to the overall success of the individual players. It helps to understand which strategies are more effective in maximizing payoffs and providing individuals with a competitive advantage. For player 1 (context), the fitness ($F_1(\pi, i)$) is calculated using Eq. (1). Equation (2) shows the equation to calculate the fitness $F_2(\pi_2, j)$ of player 2 (emotion).

$$F_1(\pi_1, i) = \sum_j a_{ij} * q_j \forall i, j \in \{1,2\} \tag{1}$$

$$F_2(\pi_2, j) = \sum_i p_i * b_{ij} \forall i, j \in \{1,2\} \tag{2}$$

   The fitness of players helps us to understand which strategies are more effective in maximizing payoffs and providing players with a competitive advantage.
7. Expected fitness (E_1 and E_2):
   Expected fitness is the average expected payoff of individual players in the population, considering the proportions of different strategies. It is calculated by taking the weighted average of expected payoffs for each individual, where the weights are given by the proportions of strategies in the population. Equations (3) and (4) show the formulas to calculate the expected fitness of the two players – context and emotion, respectively.

$$E_1(\pi_1, \pi_2) = \sum_i p_i * F_1(\pi_1, i) = \sum_i p_i * (\sum_j a_{ij} * q_j) \tag{3}$$

$$E_2(\pi_1, \pi_2) = \sum_j q_j * F_2(\pi_2, j) = \sum_j q_j * (\sum_i p_i * b_{ij}) \tag{4}$$

8. Replicator dynamics ($\dot{p}_i$ and $\dot{q}_i$):
   The rate of change of strategies concerning time t is known as replicator dynamics[27]. It can be represented as a set of differential equations where the rate of change of the proportion of strategy j is determined by the difference between the fitness of strategy i and the average fitness of all strategies in the population. The proportions of strategies in the population can change over time according to replicator dynamics, which describe how the proportions evolve based on the fitness of different strategies. The rate of change of strategy proportion over time for player 1 (context) is given by Eq. (5). For player 2 (emotion), the rate of change of strategy proportion over time is given by Eq. (6).

$$\dot{p}_i = \frac{dp_i}{dt} = p_i * (F_1(\pi_1, i) - E_1(\pi_1, \pi_2)) \forall i \in \{1,2\} \tag{5}$$

$$\dot{q}_j = \frac{dq_j}{dt} = q_j * \left(F_2(\pi_2, j) - E_2(\pi_1, \pi_2)\right) \forall j \in \{1,2\} \tag{6}$$

9. Steady-state:
   The steady state is reached when the proportions of strategies no longer change over time. At this point, the model reaches a Nash equilibrium, where no player can improve their payoffs by unilaterally changing their strategy. In other words, the sentiment analysis model converges to a stable state where the distribution of sentiment categories remains constant. Mathematically, this is represented by Eqs. (7) and (8) for player 1 (context) and player 2 (emotion), respectively.

$$\frac{dp_i}{dt} = 0 \Rightarrow p_i * (F_1(\pi_1, i) - E_1(\pi_1, \pi_2)) = 0 \tag{7}$$

$$\frac{dq_j}{dt} = 0 \Rightarrow q_j * \left(F_2(\pi_2, j) - E_2(\pi_1, \pi_2)\right) = 0 \tag{8}$$

Additionally, the sum of all strategy proportions in the population must be 1 as given by Eqs. (9) and (10) for context and emotion, respectively.

$$\sum_i p_i = 1 \text{ this simplifies to } p_1 + p_2 = 1 \tag{9}$$

$$\sum_j q_j = 1 \text{ this simplifies to } q_1 + q_2 = 1 \tag{10}$$

Note on time dependence: The strategy proportions p_i and q_j are explicit functions of time t. This means that the proportions of strategies for both players are not static but can change over time. The dynamics of these proportions are governed by the replicator equations, reflecting how the fitness and relative success of strategies drive the evolution of these proportions in the population. As the game progresses, these proportions adjust until they reach a steady state where they remain constant, indicating that the population has reached a Nash equilibrium.

Proposed methodology

The proposed methodology comprises of two phases—Phase I and Phase II. Figure 1 outlines the flowchart delineating the two-phase framework. Phase I involves the extraction of textual features from the review comments, utilizing suitable algorithms. Subsequently, we generate a decision matrix, facilitating further analysis. In Phase II, a population game model is implemented, tailored to the task of sentiment classification. Further elaboration on the different phases is provided in subsequent subsections. Table 2 lists the notations employed in the study.

Phase I: features extraction

Phase I of the methodology is initiated by following a cleaning process to enhance the quality and uniformity of the dataset. This cleaning process encompasses various pre-processing steps, including tokenization,

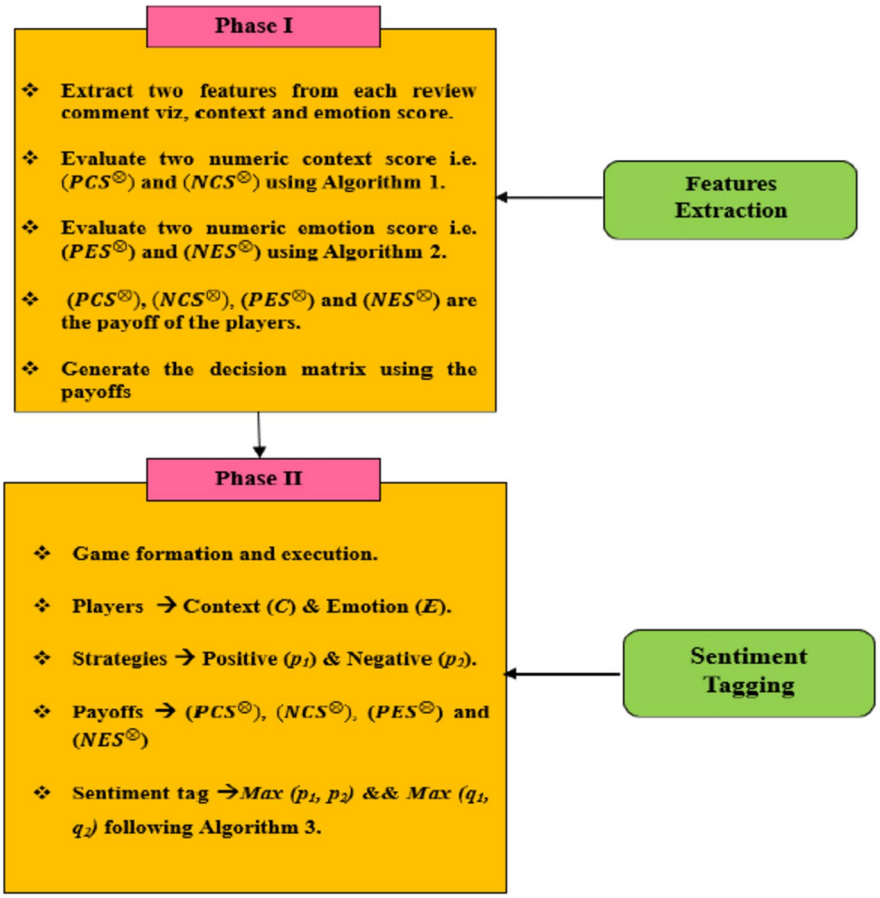


Fig. 1. Flowchart illustrating the structure of the proposed framework.

Notations	Description
$F_1(\pi_1, i)$	Fitness of player 1 (context)
$F_2(\pi_2, j)$	Fitness of player 2 (emotion)
$E_1(\pi_1, \pi_2)$	Expected Fitness of Player 1
$E_2(\pi_1, \pi_2)$	Expected Fitness of Player 2
p_1, q_1	Proportion of Positive strategy of both players
p_2, q_2	Proportion of Negative strategy of both players
$\dot{p}_i$	Replicator dynamics of player i
$\dot{q}_j$	Replicator dynamics of player j
$(PCS^{\otimes})$	Normalized Positive context score
$(NCS^{\otimes})$	Normalized Negative context score
$(PES^{\otimes})$	Normalized Positive emotion score
$(NES^{\otimes})$	Normalized Negative emotion score
(c^+, c^-, e^+, e^-)	Payoff values of players

Table 2. Comprehensive list of notations and their definitions utilized in the study.

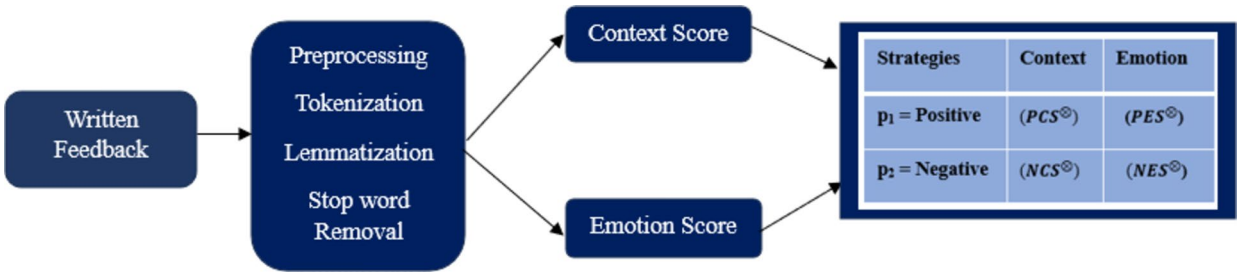


Fig. 2. Flowchart depicting phase I of the framework.

lemmatization, and stop word removal, as depicted in Fig. 2. This pre-processing ensures that the data is appropriately structured and prepared for subsequent analytical procedures. After the cleaning process, the focus shifts towards the extraction of textual features from the refined dataset. The next step entails the retrieval of two distinct features from the text, namely context scores ($(PCS^{\otimes})$, $(NCS^{\otimes})$) and emotion scores ($(PES^{\otimes})$, $(NES^{\otimes})$), as delineated in Fig. 2. The concluding step of this phase involves the generation of a decision matrix of order 2×2. This matrix is the normal form representation of the game consisting of two players (context and emotion), two strategies (positive and negative), and payoffs of the players ($(PCS^{\otimes})$, $(NCS^{\otimes})$, $(PES^{\otimes})$ and $(NES^{\otimes})$) for playing certain strategies. Detailed explanations of these payoff computations are explained in step 1 and step 2 below.

Step 1: Evaluate the normalized context scores of reviews ($(PCS^{\otimes})$, $(NCS^{\otimes})$)

Context scores are objective numerical measures allocated to assess the contextual sentiment in a given text, effectively quantifying the magnitude of positive or negative sentiment conveyed within. These scores are calculated utilizing the Python-based libraries, leveraging the capabilities of SentiWordnet (*SWN*) for accurate computation. The resulting context score values reside within the closed interval [0, 1]. $(PCS^{\otimes})$ denotes the normalized positive context score and $(NCS^{\otimes})$ denotes the normalized negative context score. Algorithm 1 is employed to calculate these scores.

Input: *A sequence of words (W) in a review R.*
Output: *Normalised Positive Context Score ($PCS^{\otimes}$) and Negative Context Score ($NCS^{\otimes}$).*
1: *Initialise PCS = 0 and NCS = 0 for each reviews.*
2: *Retrieve synsets for the word. Synsets=SentiWordNet.lookup(word).*
3: *Retrieve positive score and negative score for the synset.*

$$PS = synset.\ positiveScore$$
$$NS = synset.\ negativeScore$$

4: *Update positive context score (PCS) and negative context score (NCS).*

$$PCS \leftarrow PCS + PS \text{ and } NCS \leftarrow NCS + NS$$

5: *Normalized positive context score ($PCS^{\otimes}$) and negative context score ($NCS^{\otimes}$).*

$$(PCS^{\otimes}) = \frac{PCS}{|W|} \text{ and } (NCS^{\otimes}) = \frac{NCS}{|W|}$$

Algorithm 1. Retrieve textual Context scores

Step 2: Evaluate the normalized emotion scores of reviews (($PES^{\otimes}$), ($NES^{\otimes}$))

For this study, emotions are categorized into five distinct categories viz., happy (H), angry (A), sad (S), surprised (S_p), and fear. However, the emotion of fear is not taken into account. To quantify the emotions, the text2emotion library in Python is utilized. This library enables the computation of emotion scores, which provide numerical representations of the intensity or prevalence of the identified emotions in the text. After obtaining the emotion scores for the five emotions, these emotions are further categorized into two groups: Normalized positive emotion ($PES^{\otimes}$) and normalized negative emotion ($NES^{\otimes}$). The categorization of emotions into these two categories is performed following Algorithm 2. The resulting values of ($PES^{\otimes}$) and ($NES^{\otimes}$), representing the degree of positive and negative emotions, respectively, fall within the range of 0–1.

Input: *A sequence of words (W) in a review R.*
Output: *Normalised Positive Emotion Score ($PES^{\otimes}$) and Negative Emotion Score ($NES^{\otimes}$).*
1: *Load a set of emotion words and associated scores.*

$$EmotionWords \leftarrow LoadEmotionWords()$$

2: *Initialise emotion score variables H =0, A=0, S=0 and S_p =0.*
3: *Tokenize each review into words.*
4: *For all W in R, if $W \in EmotionWords$*

$$EmotionScores \leftarrow GetEmotionScores(W)$$

5: *Update emotion score variables.*
For each emotionScore variables

$$H \leftarrow H + emotion_score(H)$$
$$A \leftarrow A + emotion_score(A)$$
$$S \leftarrow S + emotion_score(S)$$
$$S_P \leftarrow S_P + emotion_score(S_P)$$

6: *Total emotion score E of W*

$$E \leftarrow H + A + S + S_P$$

7: *Positive emotion score (EP) and negative emotion score (EN).*

$$EP \leftarrow H + S_P \text{ and } EN \leftarrow A + S + S_P$$

8: *Normalised Positive Emotion Score ($PES^{\otimes}$), Negative Emotion Score ($NES^{\otimes}$).*

$$(PES^{\otimes}) = \frac{EP}{E} \text{ and } (NES^{\otimes}) = \frac{EN}{E}$$

Algorithm 2. Retrieve textual Emotion scores

Table 3 contains the information about game components i.e. two players (context and emotion) having two strategies (positive and negative) and their respective numeric values ($PCS^{\otimes} = c^{+}$, $PES^{\otimes} = c^{-}$, $NCS^{\otimes} = e^{+}$ and $NES^{\otimes} = e^{-}$). Integrating context and emotion features leads to a more holistic sentiment analysis, offering a comprehensive view by combining the specific details (context) with the intensity and nature of the sentiment (emotion). This integration resolves ambiguities where one feature alone might be misleading, ensuring accurate sentiment classification even in mixed or contradictory scenarios. For instance, a review mentioning "excellent battery life" (positive context) with "I am extremely happy with my purchase" (positive emotion) solidifies the positive sentiment, while a contextually positive statement with a negative emotional tone indicates a complex sentiment that requires deeper analysis. By considering both the rationale (context) and the emotional response (emotion), the proposed methodology ensures accurate sentiment classification, leveraging the strengths of both features to enhance the model's efficiency and reliability in sentiment tagging. This comprehensive approach

Players with strategies	Given values
Positive context $PCS^{\otimes}$	c^{+}
Negative context $NCS^{\otimes}$	c^{-}
Positive emotion $PES^{\otimes}$	e^{+}
Negative emotion $NES^{\otimes}$	e^{-}

Table 3. Normalized contextual and emotional scores of a review.

improves the accuracy of predictions and the robustness of the analysis, making it essential for applications in natural language processing, customer feedback analysis, and automated review systems, where understanding the full spectrum of sentiments is vital.

Phase II: sentiment tagging

To apply the population game to the sentiment tagging task, we consider a review R as a game with two players: context and emotion. Each player has two strategies viz. positive (p_1) and negative (p_2). These strategies have values between 0 and 1, ensuring that their sum is always 1 ($p_1 + p_2 = 1$). We evaluated the numeric scores from Phase I are fed as input to Phase II. We follow Algorithm 3 to perform sentiment tagging of reviews using the population game model. In the algorithm, input is the payoff. The first step is the evaluation of the fitness of both players. The fitness of player 1 (context) is denoted as $F_1(\pi_1, i)$, and for player 2 (emotion), it is denoted as $F_2(\pi_2, j)$. These fitness measures capture the effectiveness of each strategy in the game. In the next step, the expected fitness of players 1 and 2 is $E_1(\pi_1, \pi_2)$ and $E_2(\pi_1, \pi_2)$. The third step involves determining the replicator dynamics, which describes the rate of change of strategies over time. For player 1, this is represented by $\dot{p}_i = \frac{dp_i}{dt}$, and for player 2, it is represented by $\dot{q}_j = \frac{dq_j}{dt}$. As the time approaches infinity ($t \to \infty$), the strategies of both players reach a point known as Nash equilibrium, where they no longer change their strategies. At this point, the rate of change of strategies for both players tends to zero i.e., $\left(\frac{dp_i}{dt} = 0 \text{ and } \frac{dq_j}{dt} = 0\right)$. The maximum value between p_1, p_2, q_1, and q_2 indicates the sentiment tag assigned to review R. A detailed explanation of each step of Algorithm 3 is illustrated below from step 1 to step 6.

Input: *Normalised* $PCS^{\otimes}$, $PES^{\otimes}$, $NCS^{\otimes}$ and $NES^{\otimes}$
Output: *Sentiment tag to the reviews.*

1: *Evaluate the Fitness of the player playing strategy and j.*

$$F_1(\pi_1, i) = \sum_j a_{ij} * q_j \quad \text{and} \quad F_2(\pi_2, j) = \sum_i p_i * b_{ij} \quad \forall\, i, j \in \{1,2\}$$

2: *Calculate the expected fitness of both players*

$$E_1(\pi_1, \pi_2) = \sum_{i=1}^{n} p_i * F_1(\pi_1, i) \text{ and } \quad E_2(\pi_1, \pi_2) = \sum_{j=1}^{n} q_j * F_2(\pi_2, j) \ \forall\, i, j \in \{1,2\}$$

3*: Calculate replicator dynamics.*

$$\dot{p}_i = \frac{dp_i}{dt} = p_i * \left(F_1(\pi_1, i) - E_1(\pi_1, \pi_2)\right) \text{ and } \dot{q}_j = \frac{dq_j}{dt} = q_j * \left(F_2(\pi_2, j) - E_2(\pi_1, \pi_2)\right)$$

4: *At steady state* $\lim_{t\to\infty} \frac{dp_i}{dt} = 0$ *and* $\lim_{t\to\infty} \frac{dq_j}{dt} = 0$

5: *Evaluate* p_i *and* q_j *with the initial condition* $p_1(0) = 0.5$, $p_2(0) = 0.5$, $q_1(0) = 0.5$, $q_2(0) = 0.5$.

6: *Calculate* p_1 *and* p_2 *values then give the sentiment tag.*

$max(p_1, p_2)$ *and* $max(q_1, q_2)$ → *Sentiment Tag*

Algorithm 3. Sentiment tagging of reviews using population game

Step 1: Define the Payoff matrix

The payoff matrix for the game is shown in Table 4, which is the normal form representation of the game. In this study the choices made by the Context player and the Emotion player are not fully correlated; their strategies impact each other's results. Here, the payoffs of player 1 (Context) are represented by the rows, and player 2 (emotion) are represented by the columns. Here, a_{ij} represents the payoff received by player 1 (context) and player 2 (emotion). Similarly, b_{ij} represents the payoff for player 2 under the same condition. For the proposed approach, the values in the table are derived from the context and emotion scores, as shown in matrix A and matrix B.

Table 4 shows that the strategies of the players are non-correlated, i.e., each player's choice is independent of the other player's choice. This means that the decision made by the context (Player 1) does not influence the

	Emotion (player 2)		
	Positive		Negative
Context (player1)	Positive	$(a_{11}=c^{+}, b_{11}=e^{+})$	$(a_{12}=c^{+}, b_{12}=e^{-})$
	Negative	$(a_{21}=c^{-}, b_{21}=e^{+})$	$(a_{22}=c^{-}, b_{22}=e^{-})$

Table 4. Game theoretical normal form matrix depicting strategic payoffs.

decision made by the emotion (Player 2), and vice versa. In other words, we can say the strategy proportion of player 1 (p_i) is independent of (q_j). The outcome is determined by the combination of strategies chosen independently by each player.

Step 2: Calculation of fitness of the players

The fitness of a player reflects its effectiveness in achieving desired outcomes within a given context. It measures how well the chosen strategy aligns with the goals of the situation. This evaluation considers the weighted combination of scores extracted from Phase I, encompassing factors like preferences, capabilities, and contextual constraints relevant to decision-making. Through this integration with chosen strategies, player fitness is assessed. This process determines how much each strategy contributes to achieving desirable outcomes based on the combined scores. Ultimately, these fitness values offer insights into the effectiveness of different strategies in meeting specific goals or requirements. The fitness of player 1 for the positive strategy is given by Eq. (11), and for the negative strategy is given by Eq. (12). The fitness of player 2 for the positive strategy is given by Eq. (13), and for negative is given by Eq. (14). Here, p_i and q_j are the proportions of individuals using strategy i and j respectively. The strategy proportion of player 1 is p_1 and p_2 and q_1 and q_2 is the strategy proportion of player 2.

For player 1 (Context)

$$F_1(\pi_1, 1) = a_{11} * q_1 + a_{12} * q_2 = c^{+} * q_1 + c^{+} * q_2 \tag{11}$$

$$F_1(\pi_1, 2) = a_{21} * q_1 + a_{22} * q_2 = c^{-} * q_1 + c^{-} * q_2 \tag{12}$$

For player 2 (Emotion)

$$F_2(\pi_2, 1) = b_{11} * p_1 + b_{21} * p_2 = e^{+} * p_1 + e^{+} * p_2 \tag{13}$$

$$F_2(\pi_2, 2) = b_{12} * p_1 + b_{22} * p_2 = e^{-} * p_1 + e^{-} * p_2 \tag{14}$$

Step 3: Estimation of Expected Fitness

In this step, we evaluate the average expected fitness of players in the population by considering the proportions of different strategies. The expected fitness for player 1 ($E_1(\pi, \pi)$) is given by Eq. (15), and for player 2 ($E_2(\pi, \pi)$) is given by Eq. (16).

$$E_1(\pi_1, \pi_2) = p_1 * F_1(\pi_1, 1) + p_2 * F_1(\pi_1, 2) \tag{15}$$

$$E_2(\pi_1, \pi_2) = q_1 * F_2(\pi_2, 1) + q_2 * F_2(\pi_2, 2) \tag{16}$$

Step 4: Evaluate Replicator dynamics

In this step, we use replicator dynamics to figure out the proportion of change in strategies over time. This is represented by Eqs. (17) and (18) for player 1. Equations (19) and (20) represent the replicator dynamics for player 2.

For player 1 (Context)

$$\dot{p}_1 = \frac{dp_1}{dt} = p_1 * (F_1(\pi_1, 1) - E_1(\pi_1, \pi_2)) \textit{ for positive strategy} \tag{17}$$

$$\dot{p}_2 = \frac{dp_2}{dt} = p_2 * (F_1(\pi_1, 2) - E_2(\pi_1, \pi_2)) \textit{ for negative strategy} \tag{18}$$

For player 2 (Emotion)

$$\dot{q}_1 = \frac{dq_1}{dt} = q_1 * (F_2(\pi_2, 1) - E_2(\pi_1, \pi_2)) \textit{ for positive strategy} \tag{19}$$

$$\dot{q}_2 = \frac{dq_2}{dt} = q_2 * (F_2(\pi_2, 2) - E_2(\pi_1, \pi_2)) \textit{ for negative strategy} \tag{20}$$

Step 5: Steady state

After examining the rate of change of strategies over time, we analyze the population game's dynamics by observing the long-term behavior of replicator dynamics. This involves identifying a solution where strategies no longer change over time. This condition implies that as time progresses, a point comes when the rate at which

strategies change slows down and eventually stops, i.e. $\frac{dp_i}{dt} = 0$ for player 1 and $\frac{dq_j}{dt} = 0$ for player 2. This point is the time when players cannot enhance their performances and are hence restricted to the same strategy. This stable state is called Nash equilibrium (NE). Equations (21)–(24) shows the conditions for a steady state for both the players.

For player 1 (Context)

$$\frac{dp_1}{dt} = 0 \Rightarrow p_1 * (F_1(\pi_1, 1) - E_1(\pi_1, \pi_2)) = 0 \tag{21}$$

$$\frac{dp_2}{dt} = 0 \Rightarrow p_2 * (F_1(\pi_1, 2) - E_1(\pi_1, \pi_2)) = 0 \tag{22}$$

For player 2 (Emotion)

$$\frac{dq_1}{dt} = 0 \Rightarrow q_1 * (F_2(\pi_2, 1) - E_2(\pi_1, \pi_2)) = 0 \tag{23}$$

$$\frac{dq_2}{dt} = 0 \Rightarrow q_2 * (F_2(\pi_2, 2) - E_2(\pi_1, \pi_2)) = 0 \tag{24}$$

Step 6: Determine Sentiment Tag

In the population game model, the sentiment tag of reviews is determined by analyzing the proportion and stability of Nash equilibria. It is important to note that while Nash equilibria represent stable states, such that no player can improve their payoff by unilaterally changing their strategy, the dynamics of reaching such equilibria in replicator systems are highly dependent on the initial conditions. This dependency arises due to the nature of the differential equations governing the system, which can lead to convergence to different equilibria based on the starting point. The system's trajectory and the eventual equilibrium state are influenced by the initial values.

During our experimentation with various initial conditions, we observed that the results fluctuated and showed bias. To ensure no initial bias towards any strategy, we fix the initial conditions at $p_1(0) = 0.5$, $p_2(0) = 0.5$, $q_1(0) = 0.5$, $q_2(0) = 0.5$. This consistent starting point mitigates the influence of initial condition biases on the resulting sentiment tags. By fixing these neutral initial conditions, we achieve a more stable and repeatable process for determining the sentiment tag based on the dynamics of the population game model. By analyzing the most frequently occurring sentiment tags across player interactions and maintaining consistent initial conditions, we can reliably classify the sentiment of the text. We solve Eqs. (21)–(24) by using an analytic differential equation with the given initial conditions and evaluating the values of p_1, p_2, q_1 and q_2. The ultimate tag to the text is given by Eq. (25).

$$\text{Max } (p_1, p_2) \text{ and Max } (q_1, q_2) \tag{25}$$

In sentiment classification, context, and emotion tags are assigned based on the analysis of the text. Each sentiment tag's strategy proportion is denoted by p_1, p_2, q_1, and q_2 for both players. To determine the most appropriate tag, we consider the maximum probabilities from both context and emotion analyses. Specifically, by taking the maximum of p_1 and p_2, we select the tag with the highest confidence based on the context. Similarly, by taking the maximum of q_1 and q_2, we select the tag with the highest confidence based on the emotion. When both the maximum confidence values from context (max (p_1, p_2)) and emotion (max (q_1, q_2)) converge to the same tag, it indicates that both analyses agree on the sentiment. This convergence is crucial as it provides higher reliability and confidence in the classification. The agreement of two independent assessments—context and emotion—reinforces the conclusion, making the sentiment classification more robust and dependable. In some cases, when context and emotion analyses disagree, it might be due to several factors such as sarcasm, mixed emotions, or complex contexts that are difficult to interpret accurately. Some of the cases are addressed in Section "Recommendations for the future work" as research limitations.

Numerical illustration

In this subsection, we took three illustrative examples of the above process being implemented on three reviews one with a different star rating.

Illustrative example 1 Let us consider an example review with a 4-star rating.

Rating: ★★★★☆: *"The Headphones are fantastic! The sound quality is crystal clear, and they are incredibly comfortable to wear for long periods. I'm thoroughly impressed with the battery life and the sleek design."*

Initially, the context scores and emotion scores of the written review comment are computed as explained in *Phase I* using Algorithms 1 and 2, respectively. Table 5 contains the normalized context and emotion scores of the review.

Step 1: Create Payoff Matrices

Table 6 is the normal form representation of the game, consisting of two players with two strategies along with their payoffs. With the help of Table 6, we generate the two decision matrices—*A* and *B* as shown in Eqs.

Players with strategies	Scores
Positive context $PCS^{\otimes}$	$c^{+} = 0.75$
Negative context $NCS^{\otimes}$	$c^{-} = 0$
Positive emotion $PES^{\otimes}$	$e^{+} = 0.9$
Negative emotion $NES^{\otimes}$	$e^{-} = 0.1$

Table 5. Context and emotion scores for an illustrative review.

		Emotion (player 2)	
		Positive	Negative
Context (player1)	Positive	(0.75, 0.9)	(0.75, 0.1)
	Negative	(0, 0.9)	(0, 0.1)

Table 6. Normal form representation of strategic game interactions.

(26) and (27). Now we use these payoff matrices as input in Phase II for playing the population game model and sentiment tagging.

$$A = \begin{bmatrix} c^{+} & c^{+} \\ c^{-} & c^{-} \end{bmatrix} = \begin{bmatrix} 0.75 & 0.75 \\ 0 & 0 \end{bmatrix} \tag{26}$$

$$B = \begin{bmatrix} e^{+} & e^{-} \\ e^{+} & e^{-} \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{bmatrix} \tag{27}$$

Step 2: Calculate the Fitness of players. The fitness of player 1 ($F_1(\pi, i)$) and player 2 ($F_2(\pi, j)$) is expressed mathematically by Eqs. (28)–(31).

For player 1 (Context)

$$F_1(\pi,\ 1) = a_{11} * q_1 + a_{12} * q_2 = 0.75 * q_1 + 0.75 * q_2 \tag{28}$$

$$F_1(\pi,\ 2) = a_{21} * q_1 + a_{22} * q_2 = 0 * q_1 + 0 * q_2 \tag{29}$$

For Player 2 (Emotion)

$$F_2(\pi, 1) = b_{11} * p_1 + b_{21} * p_2 = 0.9 * p_1 + 0.9 * p_2 = 0.9 \tag{30}$$

$$F_2(\pi, 2) = b_{12} * p_1 + b_{22} * p_2 = 0.1 * p_1 + 0.1 * p_2 = 0.1 \tag{31}$$

Step 3: Calculate Expected Fitness. The expected fitness of the players in the game, is given by Eqs. (32) and (33).

For player 1 (Context)

$$\begin{aligned} E_1(\pi, \pi) &= p_1 * F_1(\pi, 1) + p_2 * F_1(\pi, 2) \\ &= p_1 * \left(0.75 * q_1 + 0.75 * q_2\right) \\ &= p_1 * \left(0.75 * q_1 + 0.75 * q_2\right) \end{aligned} \tag{32}$$

For Player 2 (emotion)

$$\begin{aligned} E_2(\pi, \pi) &= q_1 * F_2(\pi, 1) + q_2 * F_2(\pi, 2) \\ &= q_1 * (0.9) + q_2 * (0.1) \\ &= 0.9q_1 + 0.1q_2 \end{aligned} \tag{33}$$

Step 4: Replicator Dynamics Equation: We evaluate the replicator dynamics for player 1 and player 2 using Eqs. (34)–(37).

For player 1 (Context),

$$\dot{p_1} = \frac{dp_1}{dt} = p_1 * (F_1(\pi, 1) - E_1(\pi, \pi)) = p_1 * \left(0.75 - 0.75p_1\right) = 0.75p_1 * (1 - p_1) \tag{34}$$

$$\dot{p_2} = \frac{dp_2}{dt} = p_2 * (F_1(\pi, 2) - E_1(\pi, \pi)) = p_2 * \left(0 - 0.75p_1\right) = -0.75p_1p_2 \tag{35}$$

For player 2 (Emotion),

$$\dot{q_1} = \frac{dq_1}{dt} = q_1 * (F_2(\pi, 1) - E_2(\pi, \pi)) = q_1 * \left(0.9 - (0.9q_1 + 0.1q_2)\right) \tag{36}$$

$$\dot{q_2} = \frac{dq_2}{dt} = q_2 * (F_2(\pi, 2) - E_2(\pi, \pi)) = q_2 * \left(0.1 - (0.9q_1 + 0.1q_2)\right) \tag{37}$$

Using the constraints, $p_1 + p_2 = 1$ and $q_1 + q_2 = 1$ we get the Eqs. (38)–(41).
For player 1 (Context),

$$\frac{dp_1}{dt} = 0.75p_1 * \left(1 - p_1\right) \tag{38}$$

$$\frac{dp_2}{dt} = -0.75p_1p_2 \tag{39}$$

For Player 2 (emotion),

$$\frac{dq_1}{dt} = 0.8q_1 * (1 - q_1) \tag{40}$$

$$\frac{dq_2}{dt} = -0.8q_1q_2 \tag{41}$$

Using the analytic differential equation method with initial condition $p_1(0) = 0.5$, $p_2(0) = 0.5$, $q_1(0) = 0.5$, $q_2(0) = 0.5$, we get the values of $p_1(t)$, $p_2(t)$, $q_1(t)$, $q_2(t)$ as given below.

$$p_1(t) = \frac{e^{0.75t}}{1 + e^{0.75t}} \quad p_2(t) = \frac{1}{1 + e^{0.75t}} \quad q_1(t) = \frac{e^{0.8t}}{1 + e^{0.8t}} \quad q_2(t) = \frac{1}{1 + e^{0.8t}}$$

Step 5: Determine Steady-State Value: For a further long time ($t \to \infty$) and at a steady state, the rate of change of strategies becomes zero with time. Given that both p_1 and q_1 initially start at 0.5, we observe the dynamic at ($t \to \infty$) for p_1 and q_1 increases over time towards 1, i.e., $p_1 = 1$, $p_2 = 0$ and $q_1 = 1$, $q_2 = 0$.

$$\max\left(p_1, p_2\right) = \max\left(1, 0\right) = 1 \text{ positive tag}$$

$$\max\left(q_1, q_2\right) = \max\left(1, 0\right) = 1 \text{ positive tag}$$

This ensures that the equilibrium solutions are consistent with the initial conditions and the dynamics defined by the replicator equations, providing a stable solution for the probabilities p_1, p_2, q_1, and q_2. The sentiment tag is determined by the maximum proportion of the strategies. Since both p_1 and q_1 represent positive strategies with higher values, the sentiment tag of the review is deduced as positive.

The plot of the replicator dynamics of the illustrative example is generated to see the sentiment dynamics. Figure 3 represents the plot generated from the replicator dynamics model. It illustrates the evolution of strategy proportions for both context and emotion over time. The x-axis represents the time span of the observation, while the y-axis shows the probability values of the strategies, ranging from 0 to 1.

The blue line represents that p_1 (Context Positive), starting at an initial probability of 0.5, rapidly increased to 1. This indicates the increasing proportion of the population adopting the positive context strategy, stabilizing at 1, which suggests the dominance of the positive context strategy. Conversely, the red line for p_2 (Context Negative) starts at 0.5 and quickly decreases to 0, indicating the extinction of the negative context strategy. Similarly, the green line represents q_1 (Emotion Positive), also starting at 0.5 and rapidly increasing to 1, showing the dominance of the positive emotion strategy. The yellow line for q_2 (Emotion Negative) starts at 0.5 and decreases to 0, indicating the extinction of the negative emotion strategy.

The rapid change observed, with positive strategies p_1 and q_1 quickly increasing and negative strategies p_2 and q_2 decreasing, demonstrates the system's tendency to Favor positive interactions. The system reaches a steady state where p_1 and q_1 are near 1, and p_2 and q_2 are near 0, indicating the overwhelming adoption of positive strategies. The dominance of p_1 and q_1 over p_2 and q_2 leads to the sentiment classification being "positive." This dynamic behavior aligns with the provided payoff matrices, where positive interactions are more beneficial for both context and emotion players. The rapid convergence to positive strategies suggests that the model strongly favors positive sentiment in the given setup.

Illustrative example 2 Let us consider an example review with a 2-star rating.

Rating: ★★☆☆☆: *"I recently purchased this product hoping it would meet my needs, but unfortunately, it didn't live up to my expectations. On the positive side, the quality is decent, and the delivery was on time, which I appreciate. However, I was quite disappointed with the overall performance. The product did not perform as advertised, and I found several issues that impacted its usability. The experience left me feeling frustrated and dissatisfied."*

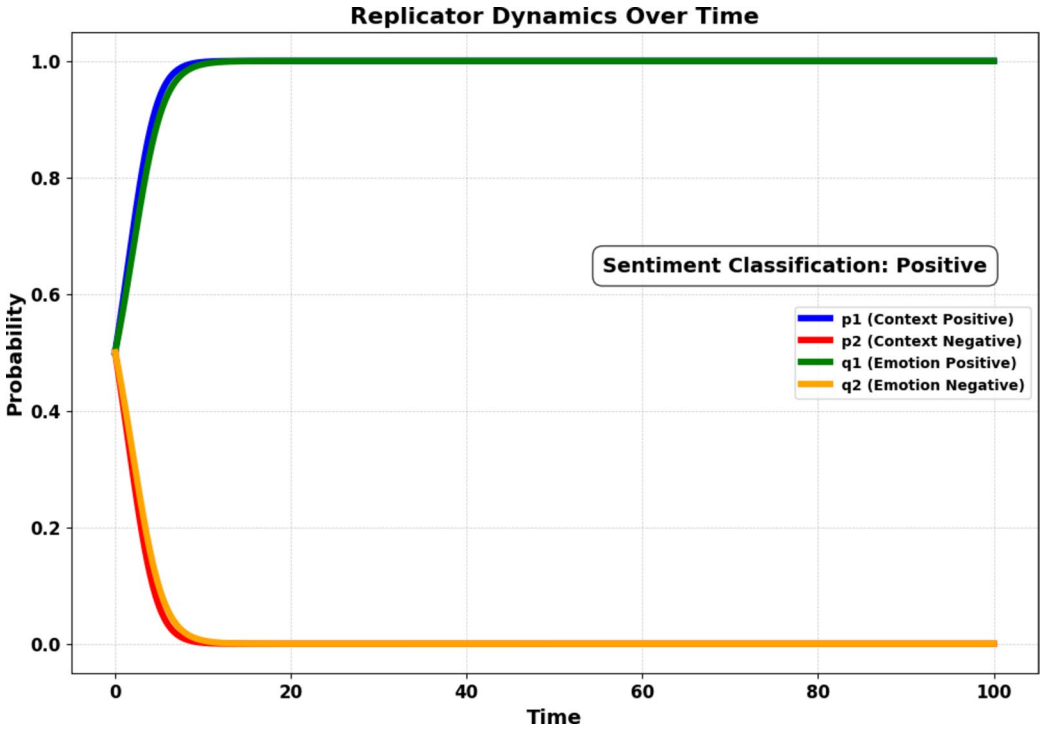


Fig. 3. Dynamic visualization of replicator dynamics.

Players with strategies	Scores
Positive context $PCS^{\otimes}$	$c^{+}=0.8$
Negative context $NCS^{\otimes}$	$c^{-}=0.2$
Positive emotion $PES^{\otimes}$	$e^{+}=0$
Negative emotion $NES^{\otimes}$	$e^{-}=0.3$

Table 7. Context and emotion scores for an illustrative review.

Initially, the context and emotion scores for the review comment are computed using Algorithms 1 and 2, as described in Phase I. These scores are detailed in Table 7. In Phase II, we then apply the proposed methodology to these extracted scores to determine the sentiment tag. The evolution of the two strategies, based on different features, over time, is illustrated in Fig. 4. Ultimately, we observe the convergence of these two strategies, which will determine the final sentiment tag of the text.

In the analysis conducted using Algorithm 3, the replicator dynamics plot (Fig. 4) demonstrates that as time progresses, the scores for context-negative and emotion-negative strategies converge towards 1, while the scores for context-positive and emotion-positive strategies approach 0. This trend indicates that the review sentiment classification is negative. This result shows that, despite any conflicting scores in the initial stages, the system stabilizes over time, consistently converging to an accurate negative classification for the review. Thus, the dynamics effectively resolve the sentiment classification to reflect the correct sentiment as negative, regardless of the initial discrepancies.

Illustrative example 3 Let us consider an example review with a 3-star rating.

Rating: ★★★☆☆: *"Royal Inn is a decent place to stay, but it has its ups and downs. The lobby is really nice and the staff are super friendly, which is a big plus. However, there are some issues that need attention. The cleanliness of the rooms could definitely be better, and the noise from outside was pretty annoying. If they can fix these problems, Royal Inn could be a much more enjoyable place to stay."*

In accordance with the methodology described in Phase I, numeric scores are extracted, as presented in Table 8. Subsequently, in Phase II, these scores are utilized to determine the sentiment tag, with the results illustrated in Fig. 5.

Table 8 illustrates that, initially, the positive scores for both context and emotion features are dominated by the negative scores. After implementing the proposed methodology, we observe the following dynamics: At $t=0$, the game begins with initial strategies, and over time, the positive strategy for both context and emotion

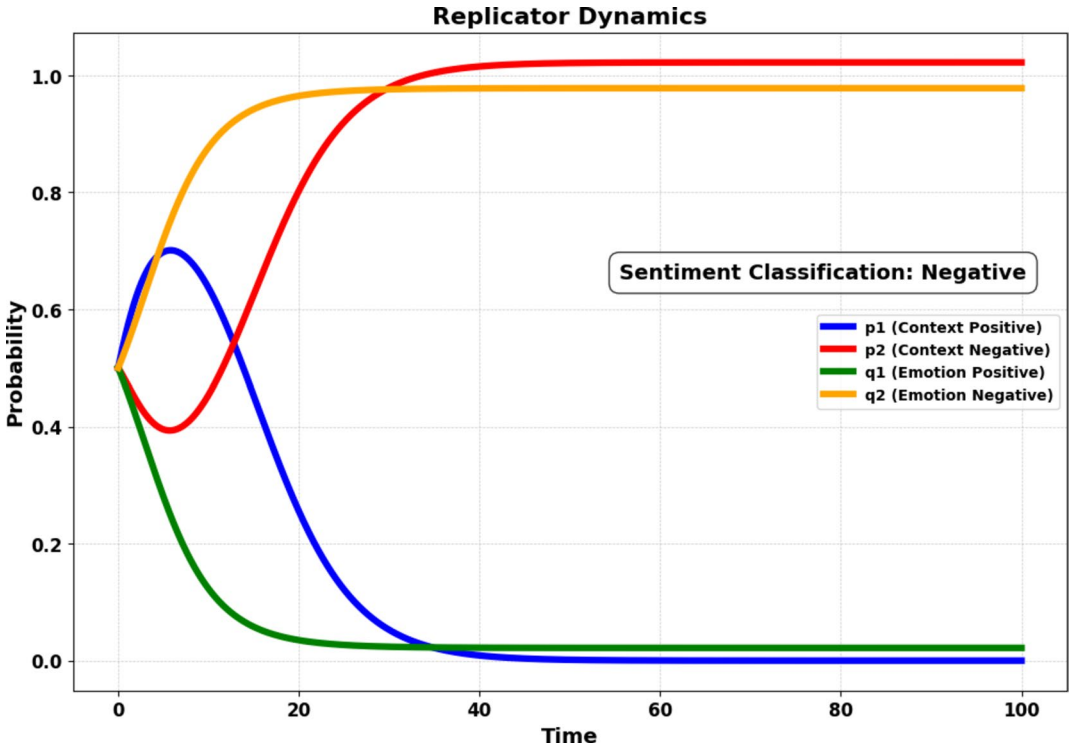


Fig. 4. Dynamic visualization of replicator dynamics.

features converges to 1 as illustrated in Fig. 5. Initially, the negative strategy for context increases but subsequently declines and converges to 0. Similarly, the negative strategy for emotion also converges to 0. This indicates that, ultimately, the positive strategy prevails over the negative strategy.

The population game model effectively captures these dynamics, demonstrating its robustness across different scenarios. The proposed methodology has proven effective in resolving conflicting scores and accurately tagging sentiment, as evidenced by the consistency and accuracy of the results across three different scenarios.

Human and animal ethics

No humans or animals were harmed in any way.

Result and evaluation

This section presents an overview of the datasets utilized for evaluating the performance of our population game model in both English and Hindi languages. We assessed our model's performance using several metrics, including accuracy, F1 score, and precision score. These metrics were compared against those achieved by state-of-the-art techniques to establish a benchmark for the effectiveness of our model. To ensure the validity of our comparisons, we adhered to the data annotation guidelines detailed in Section "Annotation guidelines for model output verification". These guidelines ensured consistency in annotation across different methodologies. We evaluated the overall performance of our model by calculating macro and micro versions of accuracy, precision, F1 score, and recall. These metrics, presented in Sections "Comparison of the proposed model with mathematical optimization techniques on the English dataset" to "Comparison of macro and micro evaluation", offer a comprehensive assessment of our model's performance, reflecting both aggregated and individual performance metrics. Furthermore, we conducted a statistical analysis, as outlined in Section "Statistical validation of the proposed model", to assess the significance of our proposed model. This analysis provided insights into the extent to which our model outperformed existing techniques, thereby highlighting its relative effectiveness.

Datasets

The proposed algorithm was applied to six distinct datasets, covering both English and Hindi reviews. We used three datasets in English from diverse domains and three datasets in the Hindi language from varied domains. Each dataset contains written reviews paired with their respective ratings. Detailed data statistics are illustrated in Table 9.

Annotation guidelines for model output verification

To evaluate the performance of the algorithm, the outcomes were compared with the sentiments annotated in the dataset. Manual annotation by domain experts was employed to classify the reviews based on sentiment. This manual annotation process involved the formation of an expert team with a strong understanding of sentiment analysis and the specific domain. Sentiment labels, positive and negative, were determined for the reviews.

Players with strategies	Scores
Positive context $PCS^{\otimes}$	$c^{+}=0.4$
Negative context $NCS^{\otimes}$	$c^{-}=0.6$
Positive emotion $PES^{\otimes}$	$e^{+}=0.3$
Negative emotion $NES^{\otimes}$	$e^{-}=0.5$

Table 8. Context and emotion scores for an illustrative review.

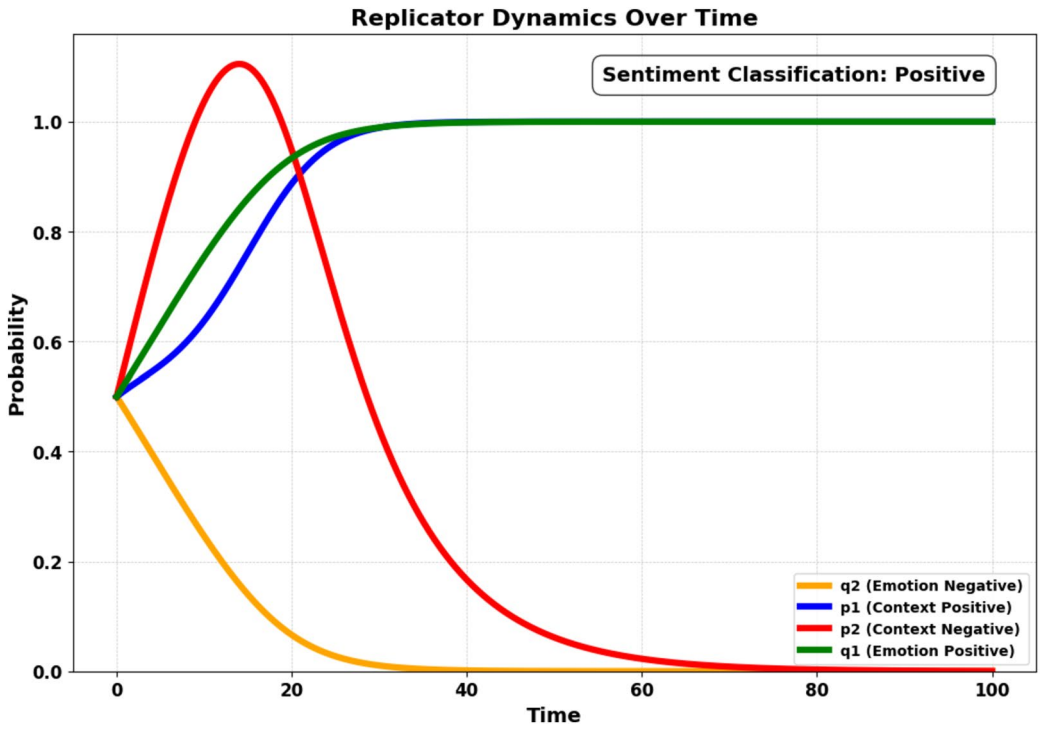


Fig. 5. Dynamic visualization of replicator dynamics.

Annotation guidelines were created to compare the results of the population game algorithm with the rating scores specified in Eq. (42). Each written review was associated with a normalized rating score between 0 and 1, and based on this range, the reviews were labelled with positive or negative sentiment. Conflicts were resolved through expert discussions and collaboration within the team. Through this scientific approach, the manual annotation process by experts ensured reliable and accurate sentiment labelling for the reviews.

$$S = \begin{cases} If\ rating\ score\ \in [0, 0.5) \rightarrow Negative\ Sentiment\ Tag \\ If\ rating\ score\ \in (0.5, 1] \rightarrow Positive\ Sentiment\ Tag \end{cases} \tag{42}$$

The purpose of this manual annotation was to evaluate and assess the accuracy of the population game model by comparing its results with the sentiments annotated in the dataset. This allowed us to measure the proposed model's performance, understand its strengths and weaknesses, and determine how well it aligned with the human-annotated sentiments. The comparison between the population game model's outputs and the annotated dataset served as a means of validating and verifying the model's accuracy by evaluating various metrics for performance measurement of the proposed population game model.

Why evaluating the neutral tag in reviews considered less critical?

While evaluating neutral sentiment can provide a more comprehensive understanding of sentiment distribution within textual data, in some contexts, it might be less critical or impactful for decision-making or analysis purposes compared to strongly polarized sentiments. Therefore, the importance of evaluating the neutral tag in reviews may vary based on the specific objectives, applications, and priorities of the sentiment analysis task at hand. Some of the main reasons we only consider the positive and negative tags are listed below:

Data set	Language	No. of reviews
TripAdvisor reviews	English	2643
Electronic reviews	English	2603
IMBD movies review	English	2478
Hotel reviews	Hindi	2977
Electronic reviews	Hindi	1789
Movie reviews	Hindi	1101

Table 9. Summary statistics of datasets employed in experimental evaluation.

Parameters	Sample 1	Sample 2
Sample size (n_1)	1000	500
Sample proportion (p_1)	0.91	0.896
Favorable cases (X_1)	910	448

Table 10. Two-Proportion Z-test statistics across multiple datasets.

1. Focus on polarized sentiments: Many sentiment analysis tasks prioritize distinguishing between positive and negative sentiments, as they often carry more significant implications for decision-making processes. Neutral sentiments, while valuable, might not influence decision-making or opinions to the same extent.
2. Application specificity: In certain applications, particularly those where decision-making relies heavily on polarized sentiments (e.g., customer feedback analysis, product reviews), assessing only positive or negative sentiments might hold more relevance than the neutral category.
3. Complexity and interpretation: Identifying neutral sentiment can be inherently more complex than identifying clear positive or negative sentiments. The ambiguity or subjectivity in what constitutes neutral sentiment makes its evaluation more challenging.
4. Resource allocation: Resources, including time and computational power, might be better allocated towards distinguishing and analyzing polarized sentiments, which often have a more direct impact on decision-making.

Evaluation of the population game model's performance

Evaluating performance metrics for frameworks is crucial as it provides a systematic and quantitative approach to assess the performance, reliability, and effectiveness of the framework. These metrics offer objective measures to gauge how well the framework achieves its intended goals, whether it's classification accuracy, anomaly detection, or any other task. By evaluating these metrics, researchers and practitioners can pinpoint the strengths, weaknesses, and areas for improvement within the framework. Additionally, it enables comparison with alternative frameworks or approaches, facilitating the selection of the most suitable tool for a given problem domain. The use of various metrics such as accuracy, F1-score, recall, precision, and MCC score is fundamental in assessing models as they provide diverse insights into performance. Accuracy reflects overall correctness by measuring the proportion of correctly classified instances. Precision underscores the accuracy of positive predictions, crucial when minimizing false positives is paramount. Recall focuses on capturing all relevant instances, especially vital in scenarios where missing positives incur significant costs. F1-score, harmonizing precision and recall, delivers a balanced assessment, particularly valuable in dealing with imbalanced datasets. Together, these metrics offer a comprehensive understanding of a model's efficiency, facilitating model selection, parameter tuning, and optimization, ultimately elevating the reliability and effectiveness of machine learning systems. These metrics are evaluated using Eqs. (43), (44) which employs TP, TN, FP, and FN as essential indicators for assessing classification models. True positives (TP) refer to cases that are correctly identified as positive, whereas true negatives (TN) are instances that are accurately classified as negative. False positives (FP) arise when the model erroneously classifies an instance as positive when it should be negative, while false negatives (FN) occur when the model mistakenly classifies an instance as negative when it should be positive. TP and TN denote accurate forecasts, but FP and FN signify prediction inaccuracies. These measures are crucial for evaluating the model's accuracy, precision, recall, and other performance indicators, offering vital insights into its capacity to effectively categorize cases. In the appendix subsection, we assessed the performance of the population game model across several datasets illustrated in Tables 9, 10, and 11.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{43}$$

$$Precision = \frac{TP}{TP + FP} \tag{44}$$

Players with strategies	Scores
Positive context $PCS^{\otimes}$	$c^{+}=0.5$
Negative context $NCS^{\otimes}$	$c^{-}=0.5$
Positive emotion $PES^{\otimes}$	$e^{+}=0.5$
Negative emotion $NES^{\otimes}$	$e^{-}=0.5$

Table 11. Context and Emotion Scores for an Illustrative Review.

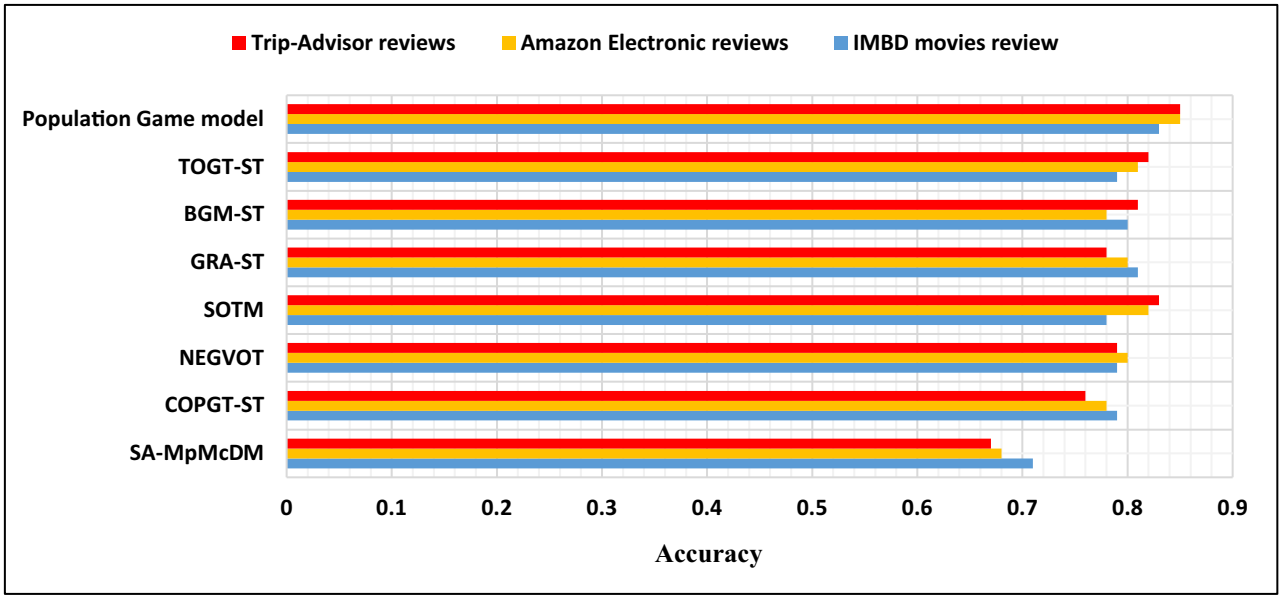


Fig. 6. Comparison of the proposed model with optimization techniques-based sentiment taggers in terms of accuracy.

Comparison of the proposed model with mathematical optimization techniques on the English dataset

In this section, we conduct a comparative analysis between the proposed model and several established sentiment analysis approaches, including SA-MpMcDM[22], COPGT-ST[28], NEGVOT[29], SOTM[5], GRA-ST[24], BGM-ST[23], and TOGT-ST[3]. Among these, six models are rooted in Multiple Criteria Decision Making (MCDM) techniques, while BGM-ST is based on the Bayesian game model. The performance of these models is depicted through the evaluation metrics of accuracy (refer Fig. 6), precision (refer Fig. 7), and F1-score (refer Fig. 8). The numeric scores of Eqs. (26)–(28) metrics are in Table 1 of the appendix section. SA-MpMcDM, an MCDM-based model, achieved accuracies of 0.71 on the IMDB movie reviews dataset, and 0.68 and 0.67 on the Electronic and Trip-Advisor datasets, respectively. However, the dependency of SA-MpMcDM on the selection and weighting criteria may lead to varying performance across different datasets, making it less robust in diverse settings. COPGT-ST, integrating the COPRAS MCDM technique with a non-cooperative game model, attained accuracies of approximately 0.79 on IMDB movies, 0.78 on electronic reviews, and 0.76 on TripAdvisor reviews. While commendable, the performance of COPGT-ST may suffer from scalability issues and the need for fine-tuning parameters to achieve optimal results. NEGVOT, utilizing the VIKOR MCDM technique, scored around 0.79 on IMDB movies, 0.8 on electronic reviews, and 0.79 on TripAdvisor reviews datasets. Despite its competitive accuracy, NEGVOT's reliance on specific optimization strategies may limit its generalizability and adaptability to different domains.SOTM, employing the Simple Additive Weighting (SAW) technique with a non-cooperative game model, achieved the highest accuracy of 0.85 on the electronic dataset. However, SOTM's performance may be sensitive to the choice of weighting scheme and may struggle with capturing complex relationships in the data. GRA-ST, employing the Grey Relational Analysis (GRA) MCDM technique, achieved the highest accuracy of 0.81 on the IMDB movie dataset. While effective, GRA-ST may face challenges in handling high-dimensional data and may not scale well to larger datasets. Similarly, BGM-ST, a Bayesian game model-based sentiment tagger incorporating context and rating score, reached the highest accuracy of around 0.81 on the TripAdvisor dataset. However, BGM-ST's reliance on additional contextual information such as ratings alongside text may limit its applicability in scenarios where such information is unavailable or irrelevant. TOGT-ST, an integrated MCDM and game theory-based model assigning sentiment tags to text, achieved an accuracy of 0.82 on the TripAdvisor review dataset. Nonetheless, TOGT-ST's performance may be affected by the complexity of the decision-making process and the need for domain-specific adjustments. In contrast, the proposed model consistently outperforms

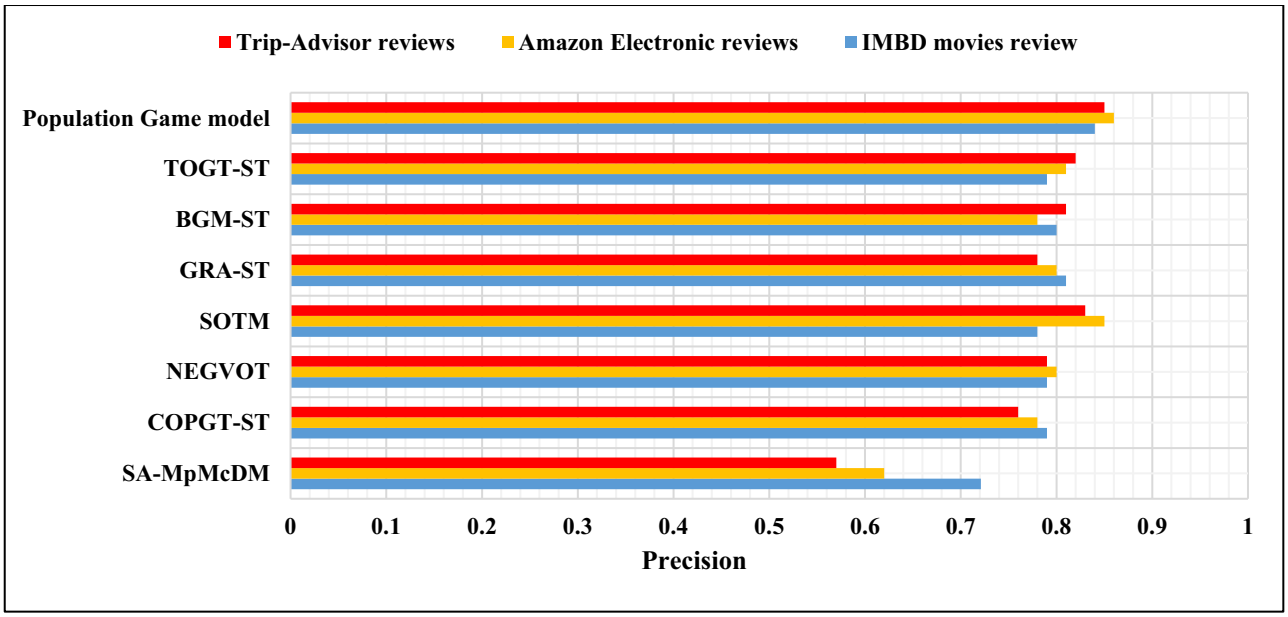


Fig. 7. Comparison of the proposed model with optimization techniques-based sentiment taggers in terms of precision.

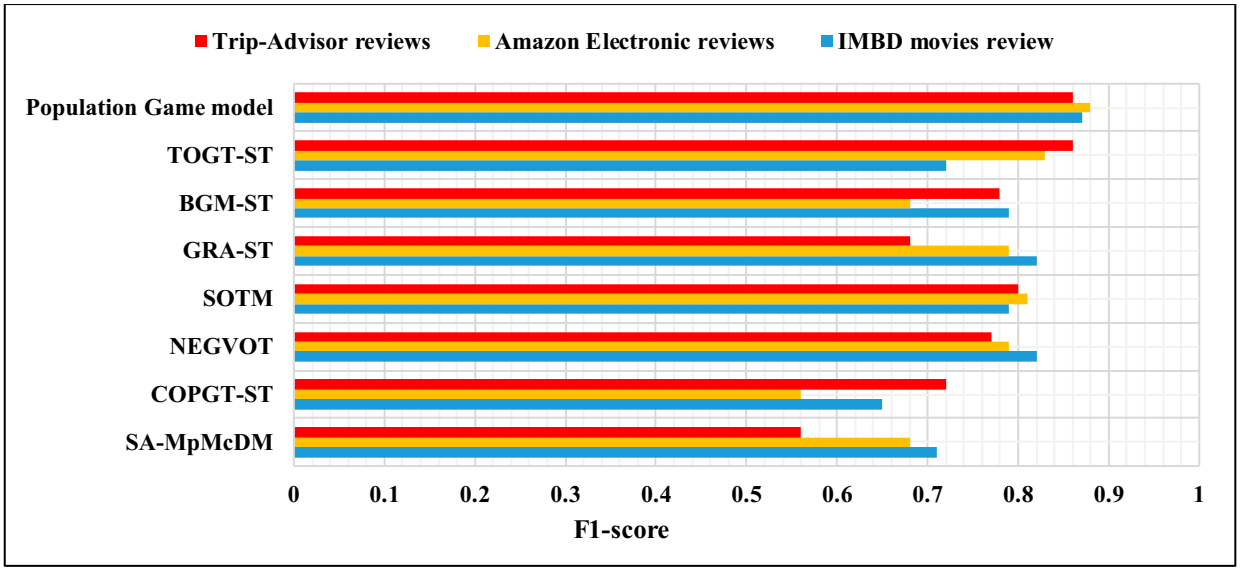


Fig. 8. Comparison of the proposed model with optimization techniques-based sentiment taggers in terms of F1-measure.

existing approaches across all evaluations. The comparatively lower accuracies of MCDM-based models stem from their dependency on the number and weighting of criteria used. Additionally, BGM-ST's dependency on the presence of ratings alongside text restricts its domain of implementation. In contrast, the proposed model overcomes these limitations. It does not rely on weights and maintains consistent sentiment tags over time, making it a robust framework for sentiment analysis. Its decentralized decision-making process harnesses collective intelligence, mitigating individual biases and offering superior performance in diverse data settings. With its scalable and adaptable nature, the proposed model emerges as a reliable solution for sentiment analysis tasks, achieving the highest accuracy of 0.85 on the trip-advisor reviews dataset out of three in the English language.

Comparison of the proposed model with existing approaches in English Text

In this subsection, we conduct a comparison between the proposed model and several existing approaches in terms of accuracy. These approaches include W2VLDA, LSVM classifier, SDA model, B-MLCNN, LDA, Apriori model, and DOC-ABSADeepL model. We applied these models to three datasets from different domains and evaluated their accuracy for comparative analysis, as illustrated in Fig. 9. Numeric scores of Eqs. (28)–(30) metrics are tabulated in Table 2 of the appendix section.

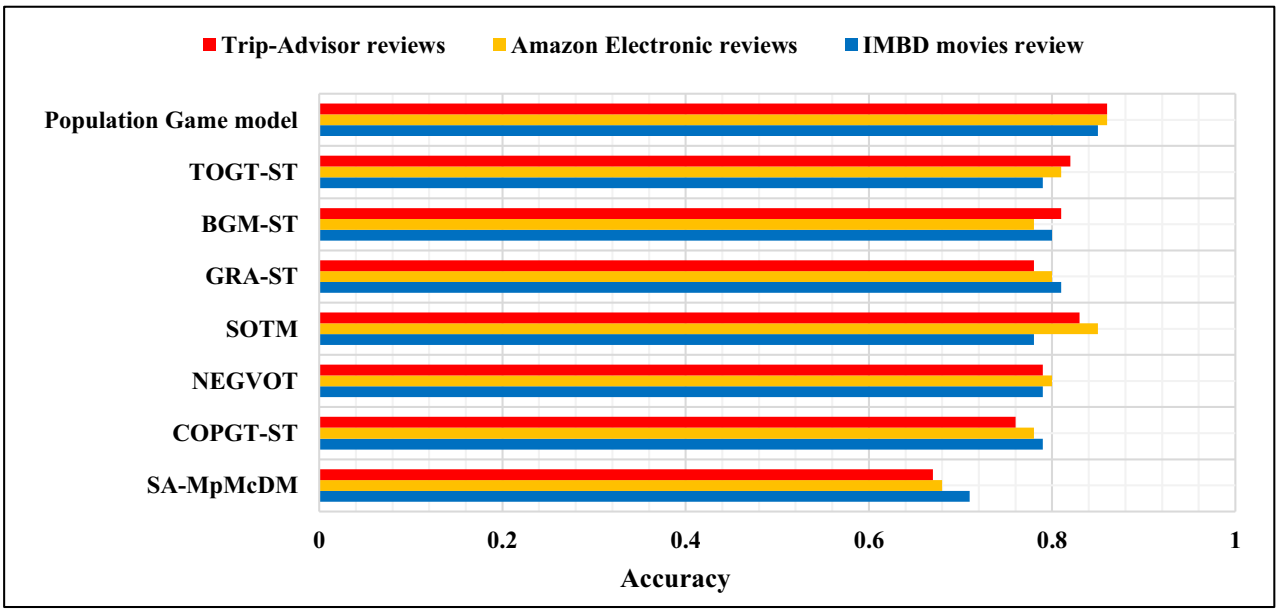


Fig. 9. Comparison of the performance of the proposed model with existing approaches in terms of accuracy.

Atandoh et al.[21] introduced the BERT-MultiLayer Convolutional Neural Network (B-MLCNN) as an integrated deep learning paradigm for sentiment analysis. Despite achieving a commendable accuracy of approximately 0.81 for IMDB movie reviews, B-MLCNN's reliance on treating entire textual reviews as a single document limits its ability to capture nuanced sentiments within longer texts. Furthermore, while B-MLCNN outperforms in certain domains, its performance diminishes in others, highlighting its lack of domain adaptability. Additionally, the proposed model exhibited better performance as compared to Buon Appetito[30]. Apriori on IMBD and Trip-Advisor its accuracy is 0.79, and on the electronic dataset, it is recorded as 0.8. The Latent Dirichlet Allocation (LDA)[31] based model achieved respectable accuracies, including 0.78 on IMDB movie reviews and 0.85 and 0.83 on electronic and Trip-Advisor reviews, respectively. However, LDA's performance may suffer from its inherent reliance on topic modeling, which may not fully capture sentiment nuances. In contrast, the proposed model offers a more holistic approach to sentiment analysis, overcoming the limitations of topic-based models by directly analyzing sentiment expressions, thereby ensuring more accurate sentiment classification across various datasets. Daniel and Meena's[12] LSVM classifier approach achieved accuracies of around 0.71 and 0.68 on IMDB and Amazon electronic review datasets, respectively. However, LSVM's performance may be limited by its dependency on specific feature representations, which may not fully capture the complexity of natural language sentiment. In contrast, the proposed model does not rely on predefined feature representations, allowing it to adapt more effectively to different datasets and capture subtle sentiment nuances more accurately. Similarly, W2VLDA by Garcia-Pablos et al.[7] achieved accuracies of 0.8 for the electronic dataset, 0.79 for IMDB, and 0.81 for the TripAdvisor dataset. While W2VLDA combined word embeddings with LDA for sentiment analysis, it may still face challenges in accurately capturing sentiment nuances, especially in complex language expressions. In contrast, the proposed model leverages population dynamics to capture the collective sentiment of a group, thereby providing a more robust and accurate sentiment analysis across various linguistic contexts.

Selective Domain Adaptation (SDA)[15] focused on selectively transferring knowledge from source to target domains at the feature level and achieved its highest accuracy of 0.81 for the TripAdvisor dataset. However, SDA's performance may be hindered by its reliance on specific domain features, which may not generalize well to other domains. In contrast, the proposed model does not rely on domain-specific features, allowing it to perform consistently well across diverse datasets without the need for domain-specific adaptations. DOC-ABSADeepL[22] methodology incorporated expert evaluations based on natural language reviews and numerical ratings. However, the proposed model demonstrates superior precision, recall, and f1-measure when compared to DOC-ABSADeepL. DOC-ABSADeepL's reliance on expert evaluations may limit its scalability and adaptability to different datasets and linguistic contexts. In contrast, the proposed model's data-driven approach ensures robust and accurate sentiment analysis across diverse datasets and linguistic expressions achieving an accuracy of 0.86.

Comparison of the proposed model with mathematical optimization techniques on Hindi text

In this section, we compare the performance of the proposed model on the Hindi dataset against several mathematical optimization models, including TOGT-ST[3], BGM-ST[23], GRA-ST[24], SOTM[5], NEGVOT[29], and COPDT-ST[28]. Numeric scores of Eqs. (28)–(30) metrics are tabulated in Table 3 of the appendix section. To evaluate the performance of the proposed model on Hindi reviews, we tweaked the proposed approach a bit. To fetch the context scores of the Hindi review comments, instead of SWN, we employed Hindi SentiWordnet (HSWN). HSWN is a collection of Hindi words and their associated positive and negative sentiment values. Once, we had the context scores of Hindi reviews, the rest of the approach remained the same.

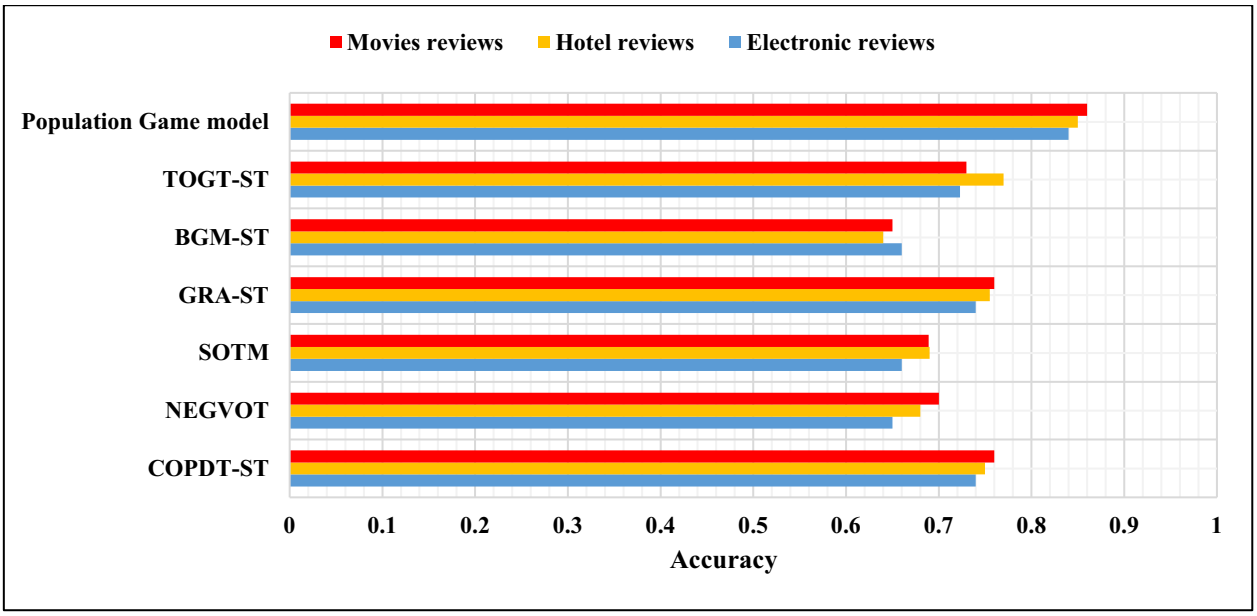


Fig. 10. Comparison of accuracy of the population game model with optimization techniques across various domains.

COPDT-ST, formulated by integrating the COPRAS MCDM technique with a non-cooperative game model, achieved an accuracy of approximately 0.74 for electronic reviews and around 0.76 for movie reviews. While achieving commendable accuracy, COPDT-ST may suffer from computational complexity and scalability issues due to its integration of multiple decision-making techniques. NEGVOT demonstrates the highest accuracy of approximately 0.7 for movie reviews. However, it may struggle with generalizability and robustness across different domains due to its reliance on specific optimization strategies tailored for movie review datasets. SOTM, an amalgamation of the SAW technique and other non-cooperative game models, achieved accuracies of 0.66 for electronics, 0.69 for hotels, and 0.689 for movies. Despite its versatility, SOTM may lack adaptability to dynamic data environments and could be sensitive to parameter settings. GRA-ST, utilizing the GRA, MCDM technique, achieves its peak accuracy at 0.76. While GRA-ST offers a robust methodology for decision-making, it may face challenges in handling complex decision landscapes and may be less effective in scenarios with high-dimensional data. Similarly, BGM-ST, grounded on the Bayesian game model, achieves an accuracy of around 0.66 on electronic review datasets. Bayesian game models, while theoretically sound, may require strong prior knowledge and assumptions about the underlying data distribution, limiting their applicability in real-world scenarios. TOGT-ST, an integration of the TOPSIS MCDM technique and a non-cooperative game model achieves the highest accuracy in hotel domain reviews at around 0.77. However, its performance in other domains may vary, indicating potential domain-specific biases or limitations in generalizability. Evaluation metric measures for all these models consistently reveal the superior performance of the proposed model when compared to other models, as shown in Figs. 10, 11, and 12. The proposed model not only demonstrates competitive accuracy but also offers advantages in terms of scalability, adaptability, and robustness across diverse datasets and archived accuracy of 0.88 on movie reviews dataset. By leveraging collective intelligence and decentralized decision-making, it mitigates individual biases and outperforms traditional optimization models in real-world applications.

Comparison of macro and micro evaluation

In this section, we conduct a comparative analysis between the macro and micro-level scores of the Bayesian game model and the proposed model. The examination reveals the superior performance of the proposed model in terms of accuracy, precision, and recall metrics. With a macro F1-score of 0.88 for the proposed model and micro precision is 0.89 across the four datasets, the former demonstrates enhanced reliability. Figure 13 presents an overview of the performance across six datasets, reinforcing the superiority of the proposed model.

Statistical validation of the proposed model

Two distinct samples were extracted from datasets containing hotel reviews and movie reviews, respectively. In the first sample, a total of 1000 reviews (n_1) were gathered and subsequently analyzed. Of these reviews, 910 were accurately classified, yielding a proportion of 0.91 (p_1). Likewise, the sample size (n_2) in the second group was 252, and out of a total of 500 reviews, 410 were accurately categorized, resulting in a sample proportion of 0.896 (p_2). A Z-test was conducted to analyze the proportions of two populations (p_1 and p_2). The data statistics are presented in Table 10.

To perform the z-test, we define a null hypothesis (H_0) and an alternate hypothesis (H_a) as given below.

$\boldsymbol{H_o}$ $p_1 = p_2$ i.e., the accuracy of *sample 1* is equal to the accuracy of *sample 2*

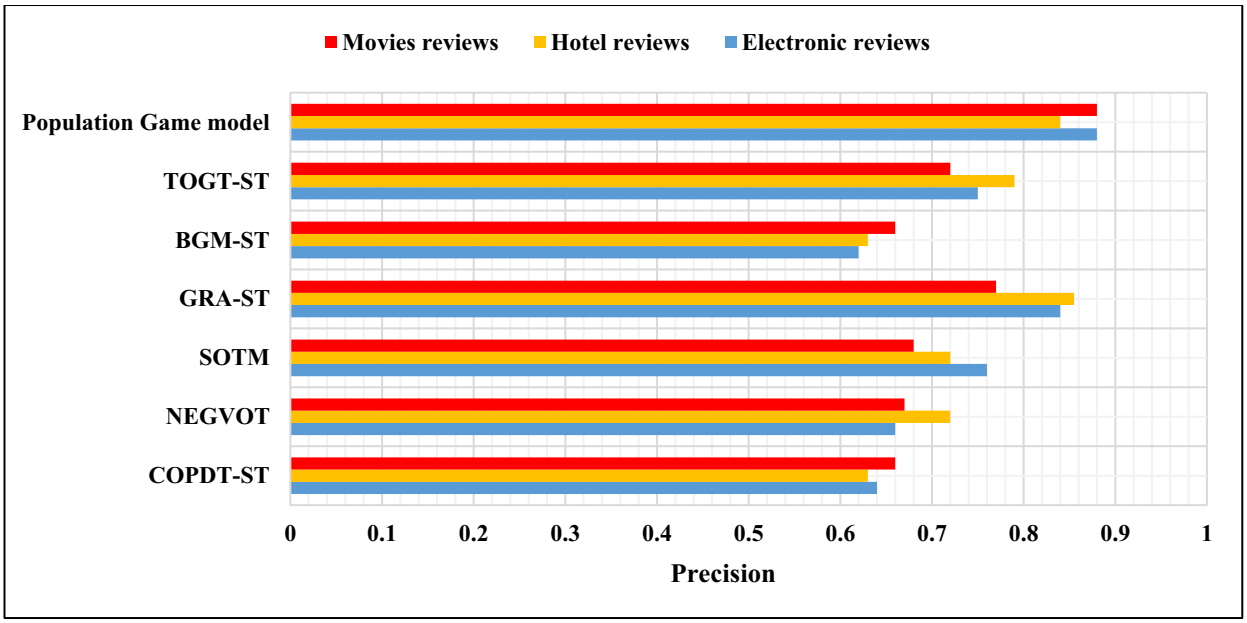


Fig. 11. Comparison of Precision the population game model with optimization techniques across various domains.

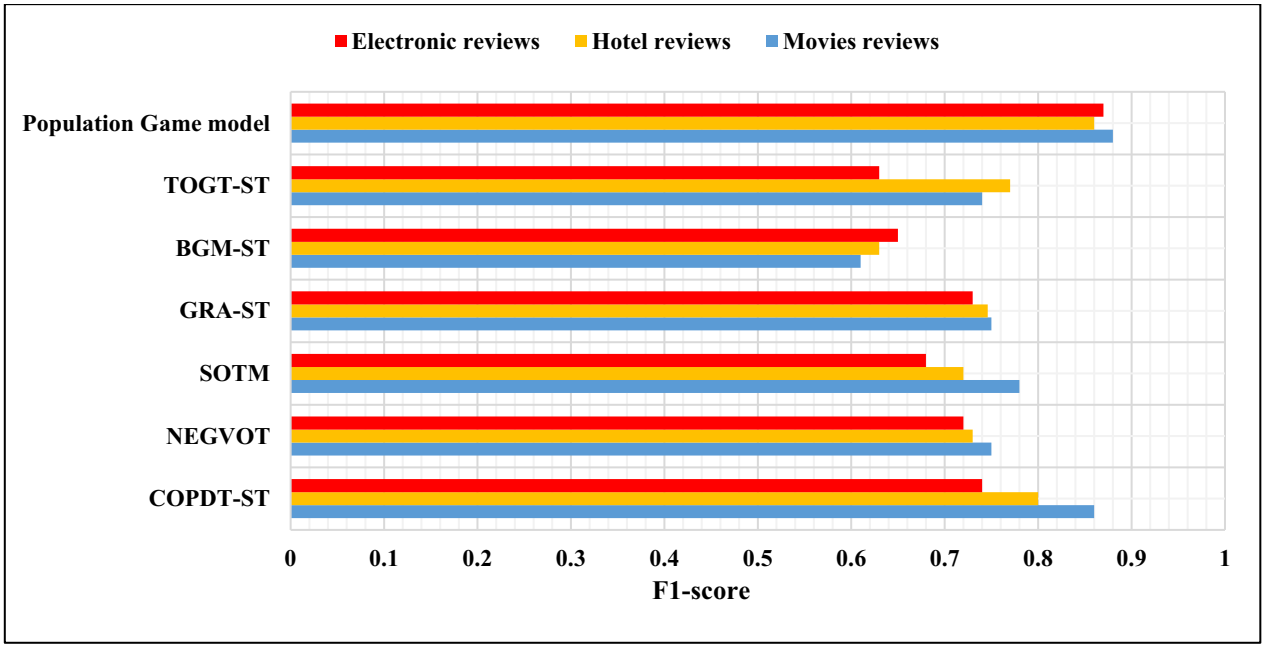


Fig. 12. Comparison of F1-score the population game model with optimization techniques across various domains.

H_a $p_1 \neq p_2$ i.e., the accuracy of *sample 1* is not equal to the accuracy of *sample 2*

Two population proportions were analyzed using a two-tailed test and a z-test as shown in Fig. 14. The z-statistic is computed utilizing Eq. (45).

$$z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.91 - 0.896}{\sqrt{0.9053} * (1 - 0.9053)\left(\frac{1}{1000} + \frac{1}{500}\right)} = 0.873 \quad (45)$$

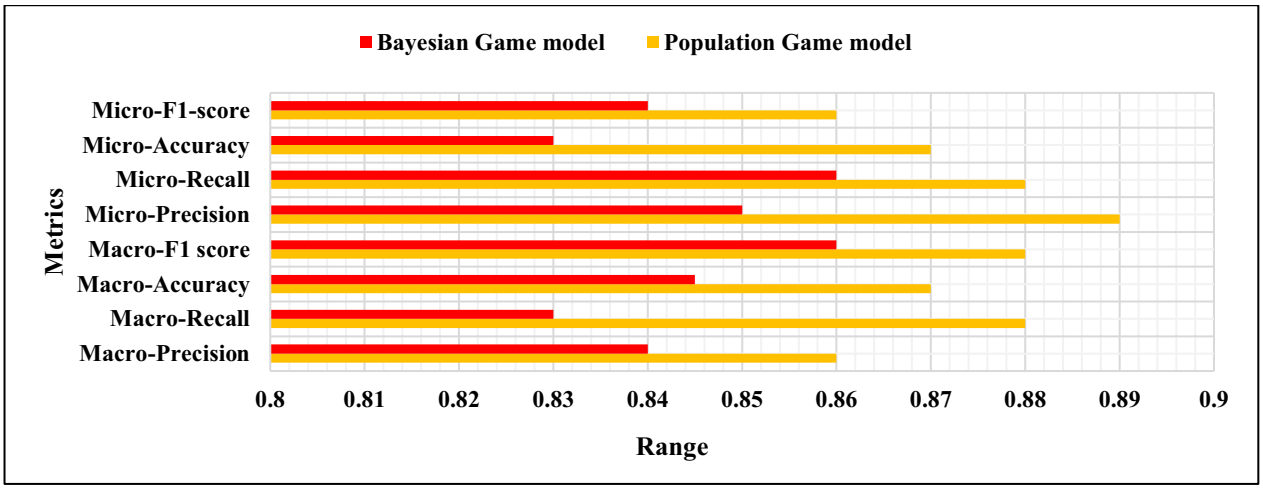


Fig. 13. Performance of population game model over Bayesian game model.

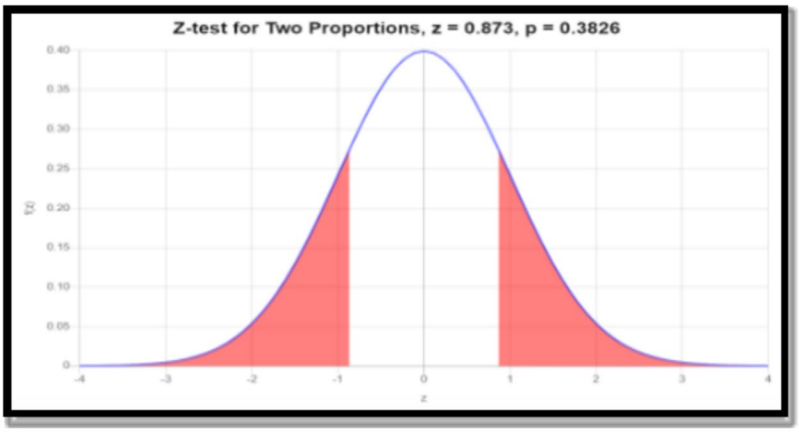


Fig. 14. An illustration of the hypothesis's critical zone.

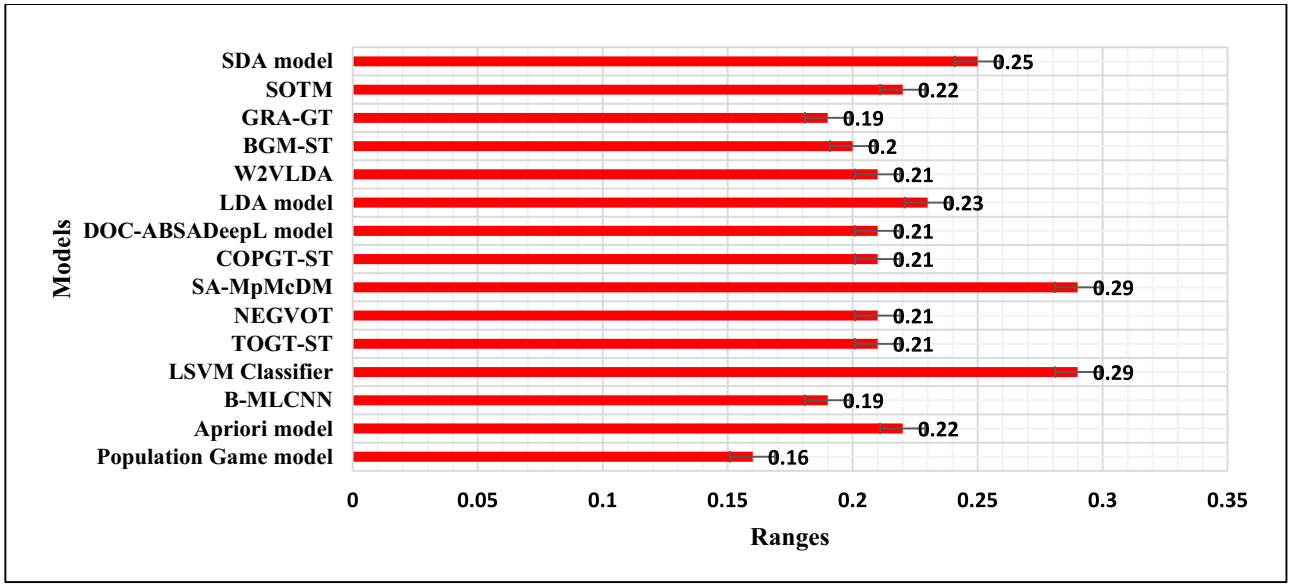


Fig. 15. Error rate comparison across different models.

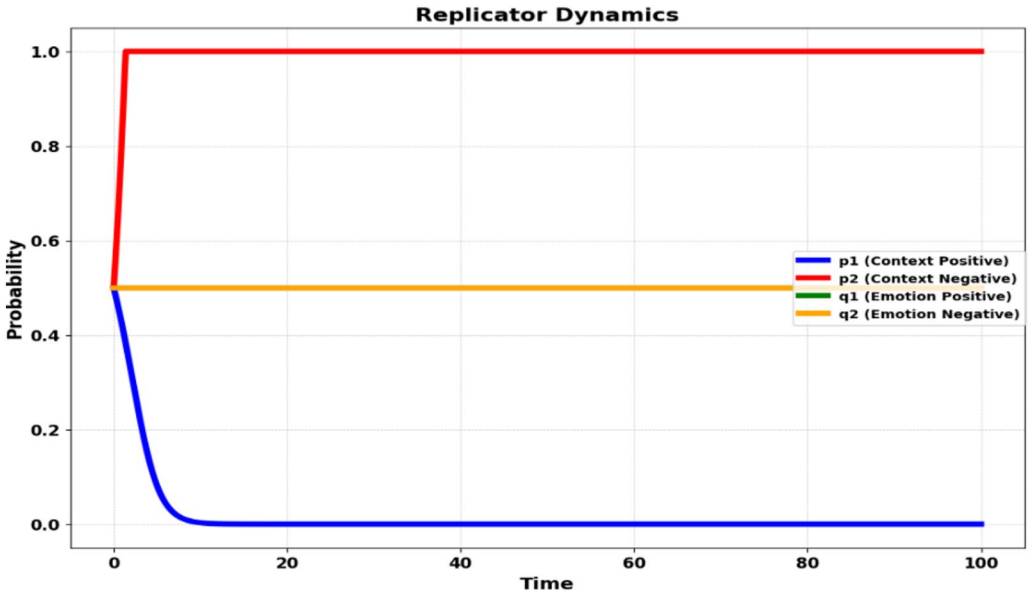


Fig. 16. Replicator dynamic plot of the review.

The critical region of the assumed hypothesis is depicted in graphical form in Fig. 14. Based on the statistical analysis results, the null hypothesis (H_o) was not rejected, indicating a lack of sufficient evidence to support the claim that the population proportions p_1 and p_2 are different at the 0.05 significance level. Hence, it can be concluded that there is insufficient statistical support to assert a significant disparity between the two proportions. This suggests that the percentage of correctly classified reviews remains consistent across datasets with different sample sizes. As a result, it can be inferred that the proposed model consistently produces reliable outcomes.

Discussion

In this section, we discuss the proposed model's error rate, significance, practical applications, limitations, and areas for improvement.

Error analysis

The error rate in model evaluation is a fundamental metric used to gauge the performance of machine learning models. It represents the proportion of incorrect predictions made by a model on a given dataset. This metric holds significant importance as it quantifies the model's ability to generalize to unseen data and its overall predictive accuracy. Figure 15 illustrates a comparison of error rates between the proposed model and existing models. Notably, the LSVM classifier and SA-MpMcDM model exhibit the highest error rates, approximately 0.29, while the proposed model demonstrates the lowest, at 0.16. The remaining models fall within the error rate spectrum of 0.29–0.16. This analysis underscores the effectiveness of the proposed model in minimizing prediction errors compared to its counterparts.

Significance of the proposed model

This study proposes a methodology for conducting sentiment analysis by utilizing a mathematical optimization framework. The proposed model's unsupervised nature is a significant advantage. No extensive dataset annotation is necessary. The proposed model utilizes data that are unlabelled and unclassified. The sole prerequisite is the availability of SWN or analogous lexicon resources that can be utilized for the computation of context scores of reviews. The model is efficient in both space and time. The study paves the way for the implementation of mathematical optimizations in diverse NLP applications.

Real-time insights: utilizing dynamic models for effective sentiment analysis

The proposed framework applies to real-time textual data as it relies on two key inputs: context and emotion scores. These scores can be generated using Algorithms 1 and 2, making the framework adaptable to various textual data domains. The four scores extracted in Phase I serve as inputs for Phase II, allowing implementation across different domains. Potential real-time domains for implementing this framework include:

(i) Market Intelligence and Brand Management:
The model analyzes textual data such as social media posts, comments, reviews, and feedback to understand consumer sentiments towards brands, products, or services. It identifies strategic interactions among consumers, modeling how sentiments evolve and spread within the population. By considering the dynamics of sentiment propagation as strategic choices made by individuals, the model provides more

accurate predictions of market trends and brand perceptions. It also identifies key influencers whose sentiments significantly impact overall trends and perceptions, enabling targeted marketing strategies.

(ii) Social Media Monitoring and Trend Analysis:

For social media data, the model captures the strategic interactions among users through their textual engagements. It analyzes how sentiments are expressed, shared, and influenced within online communities, identifying influential individuals or groups driving trends. By modeling sentiment spread as strategic interactions, the model can anticipate emerging trends, crises, or opportunities in real time. It provides insights into sentiment dynamics and helps organizations respond effectively to changing public opinion on social media platforms.

(iii) Political Analysis and Public Opinion Research:

In the political domain, the model examines textual data such as social media discussions, news articles, and polling data to understand voter sentiments and behavior. It models political sentiments as strategic choices made by individuals within the population, considering factors like candidate preferences and policy perceptions. By analyzing these interactions, the model predicts electoral outcomes more accurately and helps policymakers gauge public sentiment toward policies and initiatives. It enables evidence-based decision-making and responsive governance by understanding the strategic dynamics of sentiment within the population.

(iv) Customer Feedback Analysis and Service Improvement:

For customer feedback data, the model analyzes textual inputs from various channels like surveys, reviews, and social media comments. It identifies patterns, trends, and outliers in customer sentiments, considering factors like product experiences and service interactions. By modeling customer sentiments as strategic choices, the model provides actionable insights for service improvement strategies. It helps businesses understand the underlying dynamics of customer sentiment within the population, enabling targeted interventions and personalized approaches to address customer needs effectively.

Overall, the population game model excels in sentiment analysis by capturing the strategic interactions among individuals based on textual data, providing deeper insights into market trends, public opinion, and customer preferences across diverse domains.

Research limitations

In this subsection, we examine the error detection in one case where the proposed methodology fails.

Case 1: Ambiguous Sentiment Tags: A significant limitation of the proposed model is when $p_1 = p_2 = q_1 = q_2 = 0.5$. This situation indicates an equal probability for both positive and negative sentiments, leading to ambiguity illustrated in example 2.

Rating: ★★★☆☆: *"The presentation was fine. It wasn't particularly exciting, but it was informative."*

In Phase I, we initially extract the numeric scores as demonstrated in Table 11. Subsequently, these numeric scores are input into Phase II, where sentiment tags are assigned to the text. The final output is illustrated in Fig. 16. Output illustrated by the replicator dynamics plot shown in Fig. 16, the results indicate that the positive context score converges to 0, while the negative context score converges to 1. Meanwhile, both the positive and negative emotion scores stabilize at an initial condition value of 0.5. This pattern reveals significant conflicting results among the four parameters. The dynamics exhibit a clear shift where context strategies reach a definitive state, but the emotion strategies remain unresolved at their initial levels, highlighting inconsistencies in sentiment classification. In this case, both the context and emotion analyses result from different strategies. This indicates an ambiguous sentiment where neither positive nor negative sentiment is dominant. This is also a unique scenario where there is no clear dominant sentiment, and both are equally probable. In this case, we get the absurd and incorrect answers. This might require additional rules or more sophisticated handling to determine the final tag or to recognize it as an inherently ambiguous sentiment.

Recommendations for the future work

The future of this research holds promising advancements driven by technological innovations and evolving needs in various industries. Some potential directions for future work in sentiment analysis include:

(1) Fine-grained sentiment analysis: Current sentiment analysis often categorizes text into positive, negative, or neutral sentiments. Future work aims to enhance this by detecting and analyzing more nuanced emotions, opinions, and intentions within the text, such as sarcasm, irony, humor, or subtle sentiments.

(2) Multimodal sentiment analysis: We may aim to integrate multiple data modalities like text, images, videos, and audio for sentiment analysis. This involves understanding sentiment expressed not just through text but also through visual and auditory cues, enhancing the accuracy and depth of sentiment interpretation.

(3) Contextual understanding: Improving algorithms to better understand the context in which sentiment is expressed. Considering the broader context of conversation, cultural nuances, user-specific contexts, and historical data can help in more accurate sentiment analysis.

(4) Aspect-based sentiment analysis: The target will be to focus on identifying sentiments towards specific aspects or entities within the text. For instance, in a product review, understanding sentiments towards different features of the product rather than general sentiment about the whole product.

Overall, the future of sentiment analysis is likely to focus on more precise, context-aware, and adaptable models that can understand human sentiments more accurately across various mediums and contexts while addressing ethical considerations and user-specific needs.

Conclusion

In this study, we employed a population game model specifically designed for sentence-level sentiment analysis. We conducted a rigorous evaluation across diverse datasets in both English and Hindi, covering various domains such as movie reviews, hotel reviews, electronic device reviews, and IMDB movie ratings. Our findings demonstrate the model's language and domain independence. Utilizing a range of evaluation metrics, we achieved a maximum macro accuracy of 89%. We thoroughly examined error cases, and limiting scenarios, and discussed real-life implementation considerations, alongside offering recommendations for future program enhancements. Furthermore, our framework incorporates statistical cross-validation techniques, including Z-tests, to assess its performance robustly. The results reveal the promising performance of our model and affirm its competency in sentiment classification tasks. In our forthcoming research, we aim to delve deeper into sentiment analysis nuances, such as aspect-based sentiment analysis, and address challenges such as negation handling, irony detection, and multimodal framework generation. Our future endeavors will focus on utilizing advanced mathematical and statistical frameworks by integrating additional textual features and employing fuzzy-based approaches to effectively address various challenges of sentiment classification within the text.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Code availability

The code generated during the current study is available from the corresponding author upon reasonable request.

Received: 7 January 2024; Accepted: 21 August 2024
Published online: 04 September 2024

References

1. Peng, Y., Xiao, T. & Yuan, H. Cooperative gating network based on a single BERT encoder for aspect term sentiment analysis. *Appl. Intell.* **52**, 5867–5879 (2022).
2. Perikos, I., Kardakis, S. & Hatzilygeroudis, I. Sentiment analysis using novel and interpretable architectures of Hidden Markov Models. *Knowl. Based Syst.* **229**, 107332 (2021).
3. Punetha, N. & Jain, G. Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. *Appl. Intell.* **53**, 20152–20173 (2023).
4. Vashishtha, S. & Susan, S. Highlighting keyphrases using senti-scoring and fuzzy entropy for unsupervised sentiment analysis. *Expert Syst. Appl.* **169**, 114323 (2021).
5. Punetha, N. & Jain, G. Aspect and orientation-based sentiment analysis of customer feedback using mathematical optimization models. *Knowl. Inf. Syst.* **65**, 2731–2760 (2023).
6. Jha, V., Savitha, R., Shenoy, P. D., Venugopal, K. R. & Sangaiah, A. K. A novel sentiment aware dictionary for multi-domain sentiment classification. *Comput. Electr. Eng.* **69**, 585–597 (2018).
7. García-Pablos, A., Cuadros, M. & Rigau, G. W2VLDA: Almost unsupervised system for aspect based sentiment analysis. *Expert Syst. Appl.* **91**, 127–137 (2017).
8. Yiran, Y. & Srivastava, S. Aspect-based Sentiment Analysis on mobile phone reviews with LDA. In *ACM International Conference Proceeding Series* 101–105. https://doi.org/10.1145/3340997.3341012 (2019).
9. Song, M., Park, H., Shin, & Shik, K. Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean. *Inf. Process. Manag.* **56**, 637–653 (2019).
10. Mee, A., Homapour, E., Chiclana, F. & Engel, O. Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit. *Knowl. Based Syst.* **228**, 107238 (2021).
11. Fiok, K., Karwowski, W., Gutierrez, E. & Wilamowski, M. Analysis of sentiment in tweets addressed to a single domain-specific Twitter account: Comparison of model performance and explainability of predictions. *Expert Syst. Appl.* **186**, 115771 (2021).
12. Daniel, D. A. J. & Meena, M. J. A novel sentiment analysis for amazon data with TSA based feature selection. *Scalable Comput. Pract. Exp.* **22**, 53–66 (2021).
13. Basiri, M. E., Nemati, S., Abdar, M., Cambria, E. & Acharya, U. R. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Gener. Comput. Syst.* **115**, 279–294 (2021).
14. Nguyen, D. N., Phan, T. T. & Do, P. Embedding knowledge on ontology into the corpus by topic to improve the performance of deep learning methods in sentiment analysis. *Sci. Rep.* **11**(1), 23541 (2021).
15. Dai, Y., Liu, J., Zhang, J., Fu, H. & Xu, Z. Unsupervised sentiment analysis by transferring multi-source knowledge. *Cogn. Comput.* **13**, 1185–1197 (2021).
16. Chiha, R., Ayed, M. B. & Pereira, C. D. A complete framework for aspect-level and sentence-level sentiment analysis. *Appl. Intell.* https://doi.org/10.1007/S10489-022-03279-9/TABLES/11 (2022).
17. Hu, J., Zhang, X., Yang, Y., Liu, Y. & Chen, X. New doctors ranking system based on VIKOR method. *Int. Trans. Oper. Res.* **27**, 1236–1261 (2020).
18. Liu, N. & Shen, B. ReMemNN: A novel memory neural network for powerful interaction in aspect-based sentiment analysis. *Neurocomputing* **395**, 66–77 (2020).
19. Jin, Z., Zhao, X. & Liu, Y. Heterogeneous graph network embedding for sentiment analysis on social media. *Cogn. Comput.* **13**, 81–95 (2021).
20. Zhu, F., Zhong, P. A., Xu, B., Wu, Y. N. & Zhang, Y. A multi-criteria decision-making model dealing with correlation among criteria for reservoir flood control operation. *J. Hydroinf.* **18**, 531–543 (2016).
21. Atandoh, P., Zhang, F., Adu-Gyamfi, D., Atandoh, P. H. & Nuhoho, R. E. Integrated deep learning paradigm for document-based sentiment analysis. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 101578 (2023).
22. Zuheros, C., Martínez-Cámara, E., Herrera-Viedma, E. & Herrera, F. Sentiment Analysis based Multi-Person Multi-criteria Decision Making methodology using natural language processing and deep learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews. *Inf. Fusion* **68**, 22–36 (2021).

23. Punetha, N. & Jain, G. Bayesian game model based unsupervised sentiment analysis of product reviews. *Expert Syst. Appl.* **214**, 119128 (2023).
24. Punetha, N. & Jain, G. Sentiment analysis of stock prices and news headlines using the MCDM framework. In *AIST 2022—4th International Conference on Artificial Intelligence and Speech Technology* 1–4. https://doi.org/10.1109/AIST55798.2022.10065221 (2022).
25. Punetha, N. & Jain, G. Integrated Shannon entropy and COPRAS optimal model-based recommendation framework. *Evol. Intell.* https://doi.org/10.1007/s12065-023-00886-4 (2023).
26. Vincent, T. L. & Brown, J. S. *Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics* (Cambridge University Press, 2005).
27. Grant, A. P. *Game Theory in Action: An Introduction to Classical and Evolutionary Models* 131–133 (Princeton University Press, 2017).
28. Punetha, N. & Jain, G. Unsupervised sentiment analysis of Hindi reviews using MCDM and game model optimization techniques. *Sādhanā* **48**, 1–13 (2023).
29. Punetha, N. & Jain, G. Optimizing sentiment analysis: A cognitive approach with negation handling via mathematical modelling. *Cogn. Comput.* https://doi.org/10.1007/s12559-023-10227-3 (2023).
30. Trevisiol, M., Chiarandini, L. & Baeza-Yates, R. Buon appetito: Recommending personalized menus. In *HT 2014—Proceedings of the 25th ACM Conference on Hypertext and Social Media* 327–329. https://doi.org/10.1145/2631775.2631784 (2014).
31. Huang, F. *et al.* Multi-granular document-level sentiment topic analysis for online reviews. *Appl. Intell.* **52**, 7723–7733 (2022).

Acknowledgements
Vaibhav sharma.

Author contributions
All authors contributed equally.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70766-z.

Correspondence and requests for materials should be addressed to G.J.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2024

2024 Asia Pacific Conference on Innovation in Technology (APCIT)
Mysuru, India. Jul 26-27, 2024

An improved Network Intrusion Detection System using a Novel Hybrid Genetic Coati-Pelican Optimization Algorithms with an Ensemble Fuzzy c-means and CNN-RF

Indu Singh[1], Eric S. Moore[2], Ebrima Jallow [3], and Alhagie S. Jobe[4]

[1,2,3,4]Department of Computer Science and Engineering, Delhi Technological University

Email: [1]indusingh@dtu.ac.in, [2]ericmoore116@gmail.com, [3]pischekjallow@gmail.com
[4]alhagiesalieu@gmail.com

2024 Asia Pacific Conference on Innovation in Technology (APCIT) | 979-8-3503-6153-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/APCIT62007.2024.10673641

***Abstract*—Network intrusion remains a critical security concern in the modern cyber world, necessitating innovative approaches to enhance network confidentiality. In this paper, we introduce a novel Hybrid Genetic Coati-Pelican Optimization Algorithm (hGCPOA) designed to improve feature selection by expanding population size, increasing generation count, and enhancing individual exploration characteristics. Leveraging fitness functions, Holt's approach, and Gaussian mutation, our methodology incorporates a customized fitness function to refine the search for optimal solutions. This step ensures that the hGCPOA converges towards the most significant features while maintaining genetic diversity, thereby avoiding premature convergence. The optimally selected features undergo analysis using an advanced anomaly detection framework comprising an ensemble model of Fuzzy c-means clustering, Densely Convolutional Neural Networks (DCNN), and Random Forest (RF) algorithms. Fuzzy c-means clustering partitions the data into categories indicative of normal and anomalous traffic patterns, while DCNN executes multi-class parameter analysis, leveraging its robust feature representation capabilities. RF enhances this ensemble by offering strong classification performance and increasing predictive accuracy, effectively differentiating between benign and malicious activities. Evaluation of our model on NSL-KDD and CIC-IDS2017 datasets demonstrates exceptional performance across accuracy, precision, recall, and F1 score metrics, achieving a classification accuracy of 99.99%. Our approach underscores the significance of hybrid optimization in developing resilient and adaptive Network Intrusion Detection Systems (NIDS) to combat evolving cyber threats effectively.**

***Index Terms*—Network Intrusion Detection System, Densely Convolutional Neural Networks, Metaheuristics Optimization, Genetic Algorithm, Machine Learning**

I. Introduction

The national institute of standard and technology (NIST) defines intrusion detection as the process of keeping a close watch on events that take place within networks or computer system and examining them by looking for indications and intrusions. Software and hardware can be used for monitoring purposes in order to protect the system from malicious activity and detect violations of integrity policies [1]. To safeguard a computer or network, the intrusion detection procedure is carried out by the Intrusion detection system. It offers a scenario that is more prominent for defending against unauthenticated access control services on a network.

Cybersecurity Ventures [2] in 2023, stated that cybercrimes should have cost the world $8 trillion USD. cybercrimes would have been the world's third largest economy if it were measured as a country and would cost $9.5 trillion annually in 2024. furthermore it is expected that the global cybercrime damage costs to increase by 15 percent per year over the next few years, reaching $10.5 trillion USD annually by 2025, growing from $3 trillion USD in 2015. The global average expenditure on data breach by IBM [3], in 2023 was $4.45 million, a 15% increase over three years, indicating the growing financial burden on organizations.

In protecting networks, numerous researchers have made an effort aiming to detect the infiltration using ML, DL, and a few automated techniques-based on the traits of these intruders. However, the persistent challenges of managing the significant volume of data traversing networks continue to affect both security and performance [4]. Thus, there is a need to identify and stop intrusions while securing network infrastructure. The expansion of feature space during intrusion detection compounds issues related to performance and loss.

This study presents a novel approach for Network Intrusion Detection System utilizing a Hybrid Genetic Coati-Pelican Optimization Algorithm (hGCPOA), which hybridizes swarm intelligence algorithms with genetic algorithm for feature selection. The methodology introduces a novel ensemble model that incorporates fuzzy c-means clustering with Densely Convolutional Neural Networks (DCNN) and Random Forest (RF) to optimize parameters for the and prediction of attacks [5]. We used the datasets for evaluating our network intrusion detection system model, and the prevalence of various intrusion detection techniques which are not addressed in the literature on NIDS. The major contributions of the paper are as follows:

1) *A novel Hybrid Genetic Coati-Pelican Optimization*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

Algorithm (hGCPOA) is proposed to enhance population size, generation count, and individual exploration characteristics for feature selection

2) *Holt's method, alpha-beta-gamma parameters, Gaussian mutation and a fitness function-based approach is applied to enhance the value of each chromosome. This approach addresses challenges in anomaly detection and improves the anomaly score, designated as 'holt-score,' to enhance dimensional performance*
3) *A Fuzzy c-means clustering is used for grouping the data into attacks and benign, the Densely Convolutional Neural Networks (DCNN) for efficiently modifying and analysing the multi-class parameters tendency of the network and Random Forest (RF) for an effective classification, along with metrics that enhances patterns to predict the intrusion accuracy with in the network efficiently.*

We give an overview of previous studies conducted on network intrusion detection systems in section 2. We further described our proposed technique with suitable mathematical formulation in section 3. We discuss and described our experiments with results in section 4 and Finally, we conclude our research with remarks and future discussion in section 5.

II. Literature Review

Network intrusion detection is a critical component of modern cybersecurity which has been a subject of extensive research aimed at safeguarding information systems from malicious activities. In this section, we presented and discussed various stage-of-the art researches.

S. Bhuvaneshwari1 et al. (2022) [6] proposed that to further improve the accuracy and prediction of intruders in the network for IDS, an information gain dragonfly optimizer-based feature selection algorithm was proposed. Likewise [7] discrete Pre-Processing Method to improves Deep learning convolutional neural networks. This was because an optimization-based feature selection algorithm has been shown to be the best for finding the relevant features.

I. Hassan et al. (2022) [8] introduces an intrusion detection model leveraging their binary manta ray foraging (BMRF) optimization technique with the adaptive S-shaped transfer functions and a Random Forest (RF) Classifier. The BMRF efficiently selects relevant features enhancing the intrusion detection dataset. Their proposed model provides a significant contribution by combining BMRF and RF, demonstrating statistical significance. Despite slight drawbacks in execution time compared to specific classifiers, the model showcases substantial improvements in intrusion detection, highlighting its efficacy for network security applications.

H. Turabieh (2020) [9] proposed a hybrid Convolution Neural Network intelligent IDS system, and Binary Particle swarm Optimization (BPSO). BPSO was employed as feature selection to determine there most valuable features by reducing the high dimensionality of collected data. Following W. Wang et al. (2021), [10] For optimal performance of the convolutional neural networks, they merged the grey wolf and particle swarm optimisation techniques (HPSOGWOCNN). Their proposed HPSOGWO-CNN model was built in two phases to achieve effective and precise rudder system anomaly detection.

Kilichev et al. (2023) [11] explores hyperparameter optimization in 1D-CNN by utilizing (GA) and (PSO) techniques. Both techniques significantly enhance model performance across diverse datasets, demonstrating improvements. Their research contributes a novel application of GA and PSO, provides a comparative analysis, and emphasizes the crucial role of hyperparameter optimization in enhancing effectiveness. Their findings offer practical implications to network security practitioners, by addressing the landscape of cyber threats and highlighting the importance for an informed algorithm selection.

Osama et al. (2023) [12] introduces IG-COA, a hybrid gene selection method designed to address the challenges posed by high-dimensional microarray gene expression data in cancer classification. The model combines the Information Gain (IG) approach with the Coati Optimization Algorithm (COA), a swarm intelligence algorithm. COA is employed for gene selection following the initial IG filtering, contributing to enhanced efficiency.

Qazi et al. (2022) [13] proposed a robust detection technique by leveraging a (1D-CNN) architecture for an effective detection and classification in network intrusions. Focused on the CICIDS2017 dataset, the study comprehensively analyses flow-based statistics and presents a novel multi-class classification technique capable of identifying threads within the network [14].

III. Proposed Methodology

We propose an improved network intrusion detection system that addresses the dynamic and constantly shifting world of cyber threats using a multi-faceted hybrid optimization and classification framework. The methodology leverages a Hybrid Genetic Coati-Pelican Optimization Algorithm (hGCPOA), which enhances feature selection by combining genetic algorithms with the foraging behavior inspired by coatis and pelicans. This novel approach optimizes the identification of critical data points indicative of network intrusions by navigating the search space more effectively. The hybrid optimization algorithm enhances population diversity and improves exploration capabilities, enabling the selection of the most relevant features for detecting anomalies in network traffic.

After feature selection, our methodology incorporates a customized fitness function that integrates Holt's linear trend method and Gaussian mutation to refine the search for optimal solutions. This critical step ensures that the hGCPOA converges towards the most significant features while maintaining genetic diversity, thereby avoiding premature convergence.

The optimally selected features undergo analysis using an advanced anomaly detection framework comprising an ensemble model of Fuzzy c-means clustering, Densely Convolutional Neural Networks (DCNN) and Random Forest

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

(RF) algorithms. Fuzzy c-means clustering partitions the data into categories indicative of normal and anomalous traffic patterns. Concurrently, DCNN executes multi-class parameter analysis, leveraging its robust feature representation capabilities. Random Forest enhances this ensemble by offering strong classification performance and increasing predictive accuracy, effectively differentiating between benign and malicious activities.

We evaluated our model on the NSL-KDD and CIC-IDS2017 datasets and found exceptional performance in terms of accuracy, precision, recall and F1 score metrics. Comparing our technique to previous methods, our approach has a number of advantages, including enhanced features, more efficient computation due to lower complexity encapsulations, and superior classification performance within the signature Network local nature.

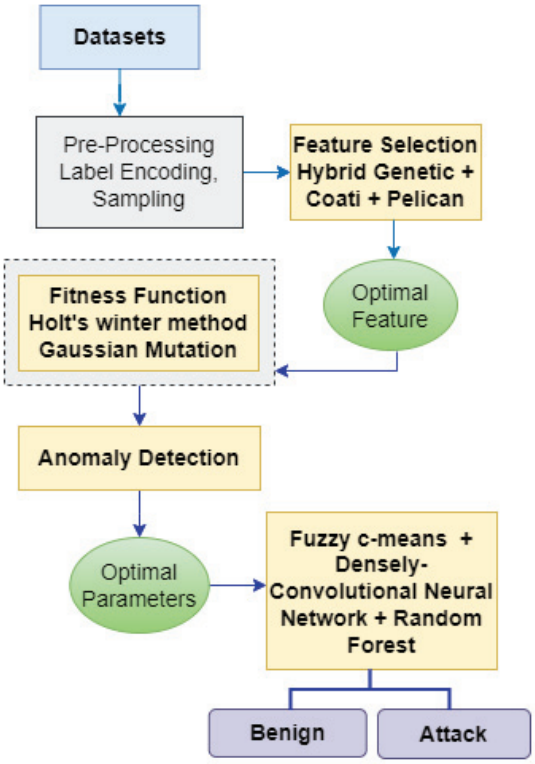


Fig. 1: Model Architecture

A. Hybrid Genetic Coati-Pelican Optimization Algorithm

In network intrusion detection using genetic algorithm, an evolution algorithm proposed by [15], we applied the sigma scaling aids in modifying the fitness values to guarantee that individuals with extreme or outlier performance are handled appropriately during the evolutionary process, improving convergence and the algorithm's efficacy in detecting network intrusions. Those mathematical functions can be traced in Eq. (10), (12), and (13). Furthermore, [16] proposed the COA, where it is shown that the Iguana holds the best position among the coati population. Therefore, positioning the coatis to rise from the tree was mathematically simulated in equation (1),(2),(3).

$$X_i^{P_1} : X_{i,j}^{P_1} = X_{i,j} + \text{rand} \cdot (Iguana_i - x_{i,j}), \quad i = 1, 2...\frac{N}{2} \; and \; j = 1, 2...m \quad (1)$$

$$Iguana^G : Iguana_j^G = Ib_j + \text{rand} \cdot (ub_j - Ib_j) j = 1..m \quad (2)$$

$$x_{i,j}^{p1} = \begin{cases} x_{i,j} + \text{rand} \cdot \left(\text{lguana}_j - I \cdot X_{i,j}\right) & F_{\text{lguana}}{}^G < F_i \\ x_{i.j} \text{ rand } \cdot \left(x_{i,j} - \text{lguana}_j^G\right) & \text{else} \end{cases}, \quad i = \left\lfloor \frac{N}{2} \right\rfloor + \left\lfloor \frac{N}{2} \right\rfloor + 2, \ldots..N \text{ and } j = 1, 2..m \quad (3)$$

The new position calculated for each coati was acceptable for our updating processes by improving their value of all functions. It would have remain in its previous position and expressing the simulated updating condition $i = 1, 2, N$ equation (4).

$$X_i = \begin{cases} X_i^{p_1}, F_i^{p_1} < F_i \\ X_i, \text{ else} \end{cases} \quad (4)$$

There new position X_i^{p1} was calculated for ith coati, $x_{i,j}^{p1}$ and jth dimension, $F_i^{p_1}$ represents a real number generated by random from the interval $[0, 1]$ and iguana indicates the prey's position in the search space, which was originally the best member's position. $lguana_j$ is its jth dimensions, as represented, the integer, which was randomly selected from a set of $1, 2$. There $lguana^G$ position on the ground, was randomly generated, $lguana_j^G$ and jth dimension.

In 2022, the Pelican Optimization Algorithm was proposed [17], which is a swarm intelligence based algorithm where the values for the population members for the variables in the optimising problem have been determined by their location inside the search space. All population members were first initialised at random based on the problem's upper and lower bounds as stated in equation (5).

$$x_{i,j} = l_j + \text{rand} \cdot (v_j - l_j), \; I = 1, 2, 3..n, j = 1, 2, ..m \quad (5)$$

1) Exploration Phase: There strategy moving towards their prey's location was modeled mathematically in equation (5). Here, $x_{i,j}^{p1}$ denotes the updated status of the ith pelican in the jth dimension based on phase 1, I is a randomly generated number, taking values of either one or two, p_j represents the prey's location in the jth dimension and Fp denotes the corresponding objective function value. The parameter I is a number that can be randomly equal to 5.

$$X_{i,j}^{p1} = X_{i.j} \, rand \cdot (p_j - Ii, j) \, F_p < F_i \quad i.j \, rand \cdot (X_{i,j} - p_j) \, F_p < F_i \quad (6)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

2) Exploitation Phase: The major disadvantages of coati Inefficient Search Strategies' exploration phase is the issues of local optima trapping as not all the coati are involve in the exploration of the iguana the leading to suboptimal or even unacceptable solutions and finding the global optimum in complex optimization landscapes. We initiated mechanisms to promote exploration, such as introducing randomness in the search process or adjusting exploration-exploitation trade-offs dynamically. Here $X_{i,j}^{p^2}$ denotes the updated status of the ith pelican in the jth dimension based on phase 2, R is a constant, set at 0.2, $R(1-t/T)$ denotes the neighbourhood radius of $x_{i.j}$ while, t represents the iteration count,and T indicates all maximum number of iterations.

$$X_{i,j}^{p^2} = xi.j + r \cdot (1 - \frac{t}{T}) \cdot (2rand - 1) \cdot X_{i.j} \quad (7)$$



Fig. 2: Steps and process for adapting the model

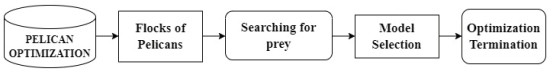


Fig. 3: *Shows the process for searching the prey and model selection*

B. Fitness Function

To enhance the accuracy of optimization, we calculate the relative proportion $P(i)$ for each individual i in the population. This is achieved by updating the values of $pbest_i - x_i(t)$ and $gbest_i - x_i(t)$ across all positions based on the population mean and standard deviation, employing sigma scaling. This meticulous process significantly improves convergence, ensuring that the hGCPOA moves towards the most significant features while maintaining genetic diversity, thus mitigating the risk of premature convergence.

$$S(x) = \frac{f(x) - \mu}{\sigma + \epsilon} \quad (8)$$

1) Holt's Winter Method: For validating the error or fitness parameters of our model's performance , we denoted there parameters method $\theta = (\alpha, \beta, \gamma, L_0,\ T_0, S_0)$. During each iteration, the fitness of each candidate was evaluated by adjusting their position of particles, and search space base on their velocities $pbest_i - x_i(t)$ and $gbest_i - x_i(t)$.

$$MSE = \frac{1}{T}\sum (y_t - \hat{y}_t)^2 \quad (9)$$

2) Gaussian Mutation: We performed the GA operations, to update the position $x_{new} = x_{old} + u_{new}$, and crossover rate, such as inertia weight, the cognitive and coefficients C_1 and C_2 in ensuring a proper handling of boundary constraints and other specific requirements.

Algorithm 1: Hybrid Genetic Coati-Pelican Optimization Algorithm (hGCPOA)

Input: N: Number of individuals in the population
T: Number of iterations
M: Dimension of the problem
N: Number of candidate solutions
Set: $i = t : 1$
Update position of the coatis
Exploration phase 1:
for $I = 1$ *to* $n/2$ **do**
  calculating all coati positions by using (4)
  for $i = i + 1$ **do**
    update i using equation (6)
  end
  for $i = 1 + N/2$ **do**
    Generate position of the coatis at random with equation (5).
    calculate new position of i with equation(6) and update i using equation (7)
  end
end
for $T = 1 : T$ **do**
  generating all the prey's random position
end
for *I = 1:N* **do**
  Moving towards the prey
  for $j = 1 : m$ **do**
    calculate new status of the j with dimension
  end
  update the I with population member.
  phase 2 (Exploitation phase)
  calculate the bound of variables using equation (8)
  if $i < N\ in\ equation(7)$ set i = 1
  for $i = 1$ **do**
    calculate all positions for i by using equation (8) and (9)
    updating all positions of i by using equation (10)
  end
  if $i + 1\ is\ true$
  for $I < N$ **do**
    saving all best candidates solutions.
    if $t = t + 1\ and\ i = 1$ is true
    $t < T$ is false
  end
  for $J = 1 : m$ **do**
    calculate updated status of the J with dimension
  end
  update I with population member.
  Updating best candidate solutions
end
Output: Global optimal Solution

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

$$\chi_{\text{new}} = \chi_{\text{old}} + \text{rand}(0,2) \times \sigma \tag{10}$$

$$v_i = \omega \times v_i + c_1 \times \text{rand}(0,1) \times (p_{\text{best}} - x_i) + c_2 \times \text{rand}(0,1) \times (g_{\text{best}} - \chi_i)\, \pi r^2 \tag{11}$$

$$AS(\chi_i) = \alpha \times HS(\chi_i) + \beta \times FS(\chi_i) \tag{12}$$

C. Ensemble Anomaly Detection

Classification metrics were employed to evaluate the performance of our model, providing insights into its ability to accurately classify data into distinct categories. Our methodology introduces an ensemble model that combines Fuzzy C-Means (FCM) clustering [18], Densely Convolutional Neural Networks (DCNN), and Random Forest (RF) [19] classifiers to achieve efficient and accurate prediction. FCM assigns fuzzy membership values to data points, enabling soft clustering and capturing inherent uncertainty in the dataset. The objective function of FCM minimizes the distance between data points x_i and cluster centroids v_j while considering membership degrees μ_{ij}^m, facilitating the classification of data points into clusters. Mathematically, the objective function of FCM can be defined as:

$$J_{FCM} = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m \left\| x_i - v_j \right\|^2 \tag{13}$$

The ensemble model utilizes Densely Convolutional Neural Networks (DCNN) for multi-class parameter analysis, leveraging its robust feature representation capabilities to analyze complex network traffic data effectively. The output of the DCNN layer h_i is computed as:

$$h_i = \mathbf{f}\left(\sum_{j=1}^{m} W_{ij} * x_{i+j-1} + b_i \right) \tag{14}$$

Finally, Random Forest (RF) algorithms are integrated into the ensemble model to enhance classification performance and increase predictive accuracy. RF leverages the collective decision-making of multiple decision trees to effectively differentiate between benign and malicious network activities. The output of the RF classifier is denoted as:

$$\text{Y} = \text{RF}(\text{X}) \tag{15}$$

Our ensemble approach combines the strengths of FCM, DCNN, and RF to achieve improved accuracy and robustness in detecting network intrusions, making it adaptable to diverse datasets and capable of addressing evolving cyber threats.

IV. Experimentation and Results

A. Dataset

Our research leverages widely used open-source network intrusion datasets, specifically the NSL-KDD [20] and CIC-IDS2017 [21] datasets, for the training and testing of our proposed novel Network Intrusion Detection System (NIDS) model. Addressing certain limitations associated with the original dataset comprising 41 attributes and encompassing $148,516$ instances for attacks.

Noteworthy advantages of using both datasets in our paper include its non-redundant instance presence, mitigating issues of unfair classification, and its suitability for effectively contrasting various IDS methods. We also perform a critical review in order to find and remove attributes that didn't add much to the prediction capability.

B. Performance evaluation

The classifications and hGCPOA model demonstrated effectiveness in detecting intrusions, with slight variations in their performances. Notably, our model achieved perfect matching in identifying intrusion instances. Table I presents a comparison of Precision, Recall and F1-score across different datasets, illustrating the number of transactions and various evaluation metrics.

TABLE I: Number of instances in datasets

Attack Types	Training	Test
Normal	56910	13174
R2L	700	4076
U2R	41	31
DoS	69365	20847
Probing	32984	1872
Total	160000	40000

For each sample, we established the disparity between the actual and predicted class probabilities employing weighted binary cross-entropy loss. Computing the reduction in cross-entropy function:

$$\delta = -\frac{1}{n} \sum_{i=1}^{N} \left[\omega_1 \cdot h_i \log\left(\hat{p}_i\right) + \omega_0 \cdot (1 - h_i) \log\left(1 - \hat{p}_i\right) \right] \tag{16}$$

Where N represents total number of samples in the dataset.The variable h_i denotes the true binary label assigned to each sample, which can take values of either 0 or 1, corresponding to the absence or presence of the target event, respectively. The predicted probability for the positive class is denoted as $\hat{p}_i$. During the training process, we applied a weight of $w_1 = 10$ to instances belonging to the attack class to emphasize their importance, while the weight $w_0 = 1$ was assigned to instances representing the normal class.
To assess the efficacy of our trained model, we employed standard metrics which was utilized in our classification tasks. These metrics has the following precision, recall, and F1 score, these were derived from our confusion matrix, along with the area under the precision-recall curve.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Hyper-parameter values

Hyper-parameter	Values
Conv 1D filters	64
Kernal Size	2
Batch size	32
Epochs	10
Strides	1
Validation split	0.2
Dense	1
Dense 1 activation function	Softmax

In our table II, the 1D convolutional layer has 64 filters, including the kernel size of 2, a batch size of 32 to provide computational efficiency since we are leveraging parallelism in modern hardware such as GPUs, a epochs, strides of 1 implies that the filter was applied to each adjacent set of elements in the input signal without leaving any validation split of 0.2 where 80% was used for training 20% was set aside for validation, a dense of 1 produces a single numeric value, and dense activation function is softmax.

C. Result Generating

For predicting our evaluated performances, We utilised the following techniques to evaluate the effectiveness of our hybrid-optimized hGCPOA based network intrusion-detection model. With the low percentage of false positives and negatives, these four measures provide an accurate depiction of how effectively our model detects network intrusions. We evaluated the following metrics for training accuracy, validation accuracy in Fig 4., and training loss, validation loss in Fig 5. for our study:

$$Precision(P) = \frac{TP}{TP + FP} \tag{17}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{18}$$

$$F1\ Score(FS) = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{19}$$

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

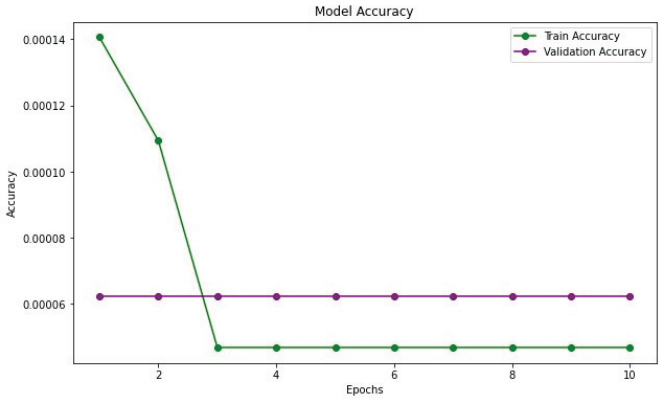


Fig. 4: Validation of the Model accuracy on training

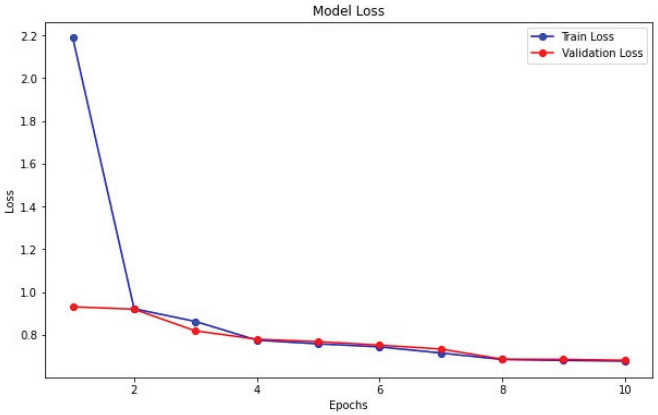


Fig. 5: Validating of the model loss on training

The validation loss and validation accuracy graphs in this paper provide valuable insights into the performance of our anomaly detection system. The validation loss graph illustrates the model's learning progress over epochs, with a decreasing trend indicating improved performance. Meanwhile, the validation accuracy graph tracks the proportion of correctly classified instances, offering a measure of the model's predictive accuracy. By analyzing these graphs, we can assess the effectiveness and generalization ability of our model, guiding further refinement and optimization efforts.

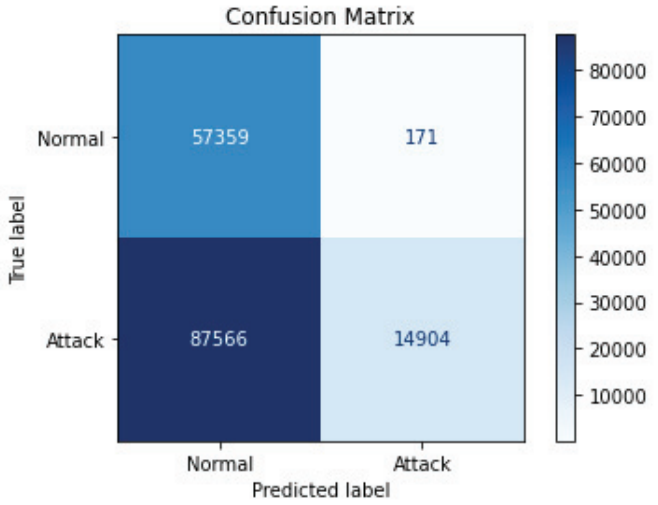


Fig. 6: Binary confusion Matrix on CIC-IDS2017 dataset

The performance of binary and multiclass models is summarised via matrices, from which we derive metrics for precision and recall across two datasets: Attack and Normalisation (datasets 0 and 1). The equilibrium between specificity and sensitivity is highlighted by our assessment as being essential to successful intrusion detection. It's also important to note that our model has a marginal edge in benign behaviour

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

TABLE III: Comparing of our method with other past approaches

Authors	Technique Used	Performance Measure			
		Accuracy	Recall	Precision	F1 Score
Xin Yang et al. (2018) [19]	ML + DL	99.41	98.76	94.90	86.00
Yuuelei et al. (2019) [14]	ResNet-based IDS	99.52	98.85	98.85	98.85
Turabieh (2020) [9]	Binary Particle Swarm Optimization	94.30	88.00	87.00	84.00
W. Wang et al. (2021) [10]	HPSOGWO-CNN PSO + Gray Wolf + CNN	99.84	99.49	99.74	99.61
I. H. Hassan et al. (2022) [8]	BMRF+RF-CIC-IDS2007 BMRF+RF-NSL-KDD	99.30 98.80	94.30 96.20	99.60 96.80	96.90 96.50
Kilichev et al. (2023) [11]	GA-1DCNN PSO-1DCNN	99.31 99.28	99.00 99.00	99.00 99.00	98.00 98.00
Proposed Model (hGCPOA)	**hGCPOA-CIC-IDS2007** **hGCPOA-NSL-KDD**	**99.97** **99.99**	**98.9** **98.9**	**90.99** **90.99**	**95.23** **95.23**

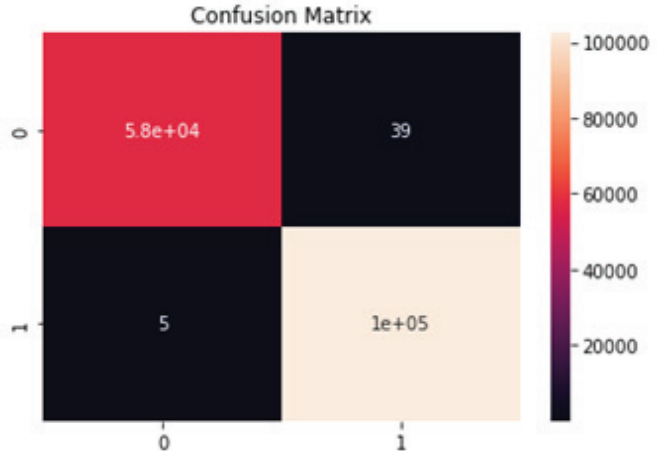


Fig. 7: Binary confusion Matrix on NSL-KDD dataset

identification, indicating its practical applicability. Decisions about the application of the model are guided by these insights.

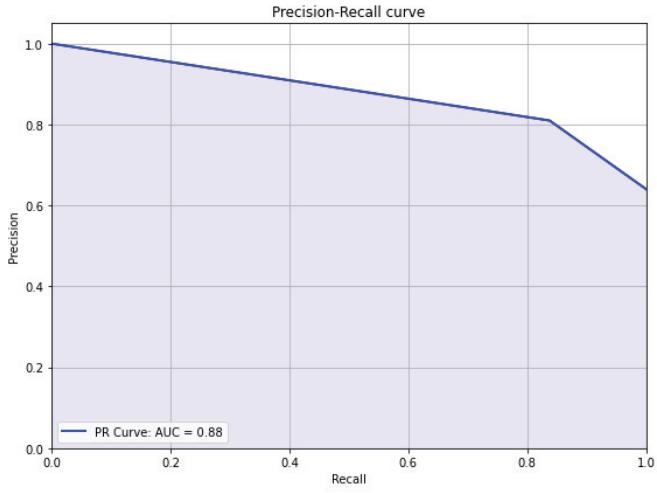


Fig. 8: Precision-Recall Curve

The Precision Recall Curve trade-off of our model prioritizes maximizing the retrieval of positive instances (recall) at the expense of including more false positives (low precision). This trade-off is acceptable on the basics of our model to retrieve a large portion of relevant instances and capturing as many potential Anomalous Traffic Patterns, Denial of Service, Port Scanning and others. Our emphasis is on detecting any deviation from normal behaviour, ensuring potential security incidents.

D. Comparison analysis with cutting-edge methods

Our table III shows that our method performed significantly better than previous methods, which was attributed to the huge number of association rules and high-quality clusters created by our improved hGCPOA algorithm for clustering that reflect the profiles of users.

V. Conclusion and Future Work

In this paper, a novel approach for network intrusion detection system is proposed utilizing a Hybrid Genetic Coati-Pelican Optimization Algorithm (hGCPOA) for feature selection, fitness functions, Holt's approach and Gaussian mutation for parameter evaluation and ensemble modeling of Fuzzy c-means clustering, Densely Convolutional Neural Networks (DCNN) and Random Forest (RF) for optimal classification parameters is presented. Experimental results on NSL-KDD and CIC-IDS2017 datasets demonstrate significant improvements in detection accuracy, precision, recall, and F1 Score, with a classification accuracy of 99.99%. This approach represents a critical advancement in combating cyber threats and enhancing network security.

Future research endeavors could explore dynamic fitness functions to optimize intrusion detection performance and investigate novel formulations in hybrid optimization algorithms to enhance efficiency and accuracy. Additionally, integrating real-time monitoring and response mechanisms into NIDS could bolster network security by enabling proactive threat detection and mitigation, thus continuing to innovate and advance the field of network intrusion detection.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] P. V. Ansam Khraisat, Iqbal Gondal and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," in *Internet Commerce Security Laboratory, Federation University Australia*, pp. 1–22, Springer, 2019.

[2] C. Ventures in *2023 SECURITY AWARENESS TRAINING REPORT*, pp. 1–31, 2023-2024.

[3] IBM in *2023 Cost of a Data Breach Report*, pp. 1–78, 2023-2024.

[4] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," in *Internet Commerce Security Laboratory, Federation University Australia*, pp. 1–28, MDPI, 2019.

[5] Z. G. CHAO LIU1 and J. WANG, "A hybrid intrusion detection system based on scalable k-means+ random forest and deep learning," vol. 9, pp. 75729–75740, IEEE, 2021.

[6] S. H. P. T. K. S. Bhuvaneshwari, K. Venkatachalam and P. Prabu, "Improved dragonfly optimizer for intrusion detection using deep clustering cnn-pso classifier," *Computers,Materials Continua*, vol. 70, no. 3, pp. 5949–5965, 2022.

[7] S. K. D. S. JIHOON YOO1, BYEONGJUN MIN and D. SHIN, "Study on network intrusion detection method using discrete pre-processing method and convolution neural network," in *IEEE Access*, pp. 142348–142361, IEEE, 2021.

[8] I. H. H. et al, "An improved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection," *Intelligent Systems with Applications*, vol. 16, pp. 2667–3053, 2022.

[9] H. Turabieh, "A hybrid convolution neural network with binary particle swarm optimization for intrusion detection," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, no. 12, pp. 128–134, 2020.

[10] C. G. WEILI WANG, RUIFENG YANG and H. QIN, "Cnn-based hybrid optimization for anomaly detection of rudder system," in *IEEE Access*, vol. 9, pp. 121845–121858, IEEE, 2021.

[11] D. Kilichev and W. Kim, "Hyperparameter optimization for 1d-cnn-based network intrusion detection using ga and pso," *mathematics*, vol. 11, no. 17, pp. 3724–3755, 2023.

[12] A. A. A. Sarah Osama and H. Shaban, "A hybrid of information gain and coati optimization algorithm for gene selection in microarray gene expression data classification," *Kafrelsheikh Journal of Information Sciences ISSN (Online)*, 2023.

[13] A. A. Emad Ul Haq Qazi and T. Zia, "A one-dimensional convolutional neural network (1d-cnn) based deep learning system for network intrusion detection," *Applied Sciences*, vol. 10, p. 356, 2022.

[14] Y. Xiao and X. Xiao, "An intrusion detection system based on a simplified residual network," *Information*, vol. 12, no. 16, p. 7986, 2019.

[15] Z. Michalewicz and M. Schoenauer, "Evolutionary algorithms for constrained parameter optimization problems," in *Evolutionary Computation*, 1996.

[16] E. T. P. T. Mohammad Dehghani, Zeinab Montazeri, "Coati optimization algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 259, p. 110011, 2022.

[17] P. Trojovský and M. Dehghani, "Pelican optimization algorithm: A novel nature-inspired algorithm for engineering applications," *Sensors*, vol. 22, no. 3, p. 855, 2022.

[18] C. W. . L. Y. . R. W. Can Jin 1, Zhiwei Ye 1, "A network intrusion detection method based on hybrid rice optimization algorithm improved fuzzy c-means," in *The 4th IEEE International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems*, vol. 978, pp. 5386–7587, IEEE, 2018.

[19] Z. L. M. I. YANG XIN, LINGSHUANG KONG *et al.*, "Machine learning and deep learning methods for cybersecurity," in *IEEE Accesss*, vol. 6, pp. 35365–35381, IEEE, 2018.

[20] D. D. Protić, "Review of kdd cup '99, nsl-kdd and kyoto 2006+ datasets," *VOJNOTEHNIČKI GLASNIK / MILITARY TECHNICAL COURIER,*, vol. 66, no. 3, pp. 580–596, 2018.

[21] A. H. L. Iman Sharafaldin and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *4th International Conference on Information Systems Security and Privacy (ICISSP 2018),*, vol. 9, pp. 108–116, 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:09:58 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

ANN Controlled Single-Phase Microinverter for Off-Grid Solar Application

Ajay Kumar Sahu, *Student Member, IEEE*, Ramnarayan Patel, *Senior Member, IEEE*, and Lalit Kumar Sahu, *Senior Member, IEEE*

***Abstract*—The voltage produced by solar panels fluctuates due to change in irradiance, temperature, and shading. This makes it challenging to obtain a stable voltage suitable for powering devices that use regular household electricity. This article proposes a solar microinverter with Artificial Neural Network (ANN) controller that can handle input voltages ranging from 24V to 46V and consistently produces an output voltage of $230V_{rms}$. A modified Z-source converter with an active switch helps achieve the requisite gain. The ANN controller enables dynamic adjustment of the duty ratio in response to fluctuations in input voltage or load, ensuring optimal control and performance. The proposed microinverter provides a common grounding feature between the solar panel and load; this suppresses the Photovoltaic (PV) leakage current. The article provides a detailed circuit operation explanation, and real-time validation confirms the performance of microinverter under practical conditions. This solution addresses the challenge of obtaining stable voltages for household device power supply from solar panels.**

***Index Terms*—Common-ground, Microinverter, off-grid inverter, Single-phase single-stage, Solar Photovoltaic (PV).**

I. Introduction

Control techniques are essential for the operation of microinverters, microinverters convert the direct current (DC) output of photovoltaic (PV) panels into alternating current (AC) for grid connection. Microinverters have several advantages over conventional string or central inverters, such as higher efficiency, flexibility, reliability, and modularity. However, they also pose several challenges, such as complex control objectives, high switching frequency, low power rating, and grid synchronization [1]. Several control techniques have been proposed in literatures to address these challenges and improve the performance of microinverters. Off-grid solar applications are essential for remote areas and regions with limited or no access to the utility grid. A novel operational mode termed "island mode" is introduced in [2] where the proposed droop scheme undergoes a reconfiguration in the microinverter control mechanism, transforming it into a voltage source. The voltage controller is reconfigured to regulate the inverter output voltage, and the dc-dc converter controls its input voltage, while its output voltage is regulated by an additional control loop. The droop scheme allows the inverter to work in grid-isolated mode [2]. Use of dual droop control of capacitor voltage and output frequency to balance the active power, and Q/V droop control to regulate the reactive power and voltage magnitude is proposed in [3]. It also uses voltage-current double closed loop to control the capacitor voltage at the inverter's AC side.

Transformer-less microinverters have gained considerable attention due to their compact design and reduced cost. These microinverters eliminate the need for bulky transformers, offering greater flexibility in PV system integration. A work demonstrates an innovative approach to mitigate the limitations posed by electrolytic capacitors, which are prone to degradation over time [4]. Various transformer-less PV micro-inverter topologies have emerged, offering efficiency, cost-effectiveness, and reliability enhancements. These designs include single-stage boost microinverters for shading mitigation, common-ground single-phase inverters for cost reduction, capacitor-free three-level flying-capacitor microinverters for extended system lifespan, active power decoupling for improved energy harvesting, and grid-connected micro-inverters with power decoupling for efficient grid integration [5], [6], [7], [8]. A variable frequency-oriented PWM control strategy is applied for a single-stage DC-AC microinverter, optimizing switching network losses and improving efficiency and THD [9] and a coupled inductor-based boost microinverter is developed for renewable energy applications, highlighting its high conversion efficiency and reduced current ripple [10].

A control scheme for microinverters that can operate in different modes and parallel with other microinverters without communication or control interconnections is introduced in [11], it can operate in grid-connected mode to feed power to the grid with a programmable power factor, line-interactive mode to share load power without feeding power to the grid and standalone mode to supply uninterruptible power in case of grid failure [11]. Decoupled active and reactive power distribution control without additional communication components in microinverters under island mode is introduced [12], an ANN based controller for a modified packed U-cell five-level inverter (MPUC-5) that can operate under variable solar insolation conditions and produce a constant output voltage with minimum harmonic distortion is introduced in [13]. A novel VVC strategy based on an ANN control for distributed energy resources (DERs) inverters at the grid edge, the proposed approach uses an ANN controller to regulate the inverter output voltage and current, and achieves optimal control based on approximate dynamic programming [14]. Direct Digital Synthesis (DDS) introduced in [15] provides flexibility, high resolution, fast response, and low cost in the control of the PV microinverter, as compared to conventional methods, dsPIC microcontroller is used to implement the control technique. A sensor-less control approach is based on the analysis of active power decoupling (APD) capacitor energy and voltage and uses the power and grid information from the main inverter sensors to predict the required duty ratio for the APD circuit [16].

Improved BCM peak current control method can provide accurate and isolated current sensing and dynamic dead time optimization [17]. Single-stage, single-phase micro-inverters simplify the conversion process by directly transforming DC power from the solar panel into AC power. Various single-phase topologies are presented including hybrid multiport, buck-boost inverters, DAB based single stage microinverter,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

CCM Zeta microinverter, interleaved boost based microinverter. Substantial voltage gain is offered by these microinverters [18], [19], [20], [21], [22]. However, they share a common limitation—lack of common-grounding capability, which can result in a significant amount of PV leakage current. Additionally, flyback, boost-flyback, and CCM zeta microinverters require transformers, contributing to system bulkiness. Common-ground micro-inverters have a single reference point for DC and AC circuits. The common-ground configurations aim to reduce leakage current, simplify grounding and enhance performance [5], [6], [23], [24]. Single-stage high-frequency-link (HFL) microinverter with a split-phase structure for both grid-tied and islanded operation is introduced in [25] this microinverter is designed for 600W class PV modules and offers natural voltage balancing and soft-switching for all switches.

Diverse control strategies for microinverters in photovoltaic systems are explored such as optimization of peak current mode control, enabling reconfigurable operation for on-grid and off-grid modes, implementation of phase synchronization for interleaved flyback microinverters, optimal control for high efficiency across a broad load range and the utilization of exact linearization for hybrid multiport microinverter [2], [17], [18], [26], [27]. These existing control methods are effective to some extent and have limitations such as narrow operating range optimizations, limited adaptability, complexity in implementation, dependency on grid presence etc., but the proposed microinverter in this study employs ANN-based controller, enabling improved efficiency, robustness, and adaptability to varying conditions. The existing literatures on off-grid solar applications have made notable strides in microinverter technologies. However, a critical evaluation reveals inherent limitations and challenges in these existing designs, prompting the need for a new topology with digital controller. The proposed topology aims to bridge these gaps, advancing the state-of-the-art in microinverter technology for off-grid scenarios.

The proposed microinverter has following distinct features,

- High voltage gain.
- Efficient power conversion.
- Common grounding between the PV panel and the load.
- Minimization of the common-mode voltage issues.
- Compatible with reactive loads.
- ANN control provides versatility in application and adapts effectively to the varying load conditions frequently encountered in photovoltaic systems.

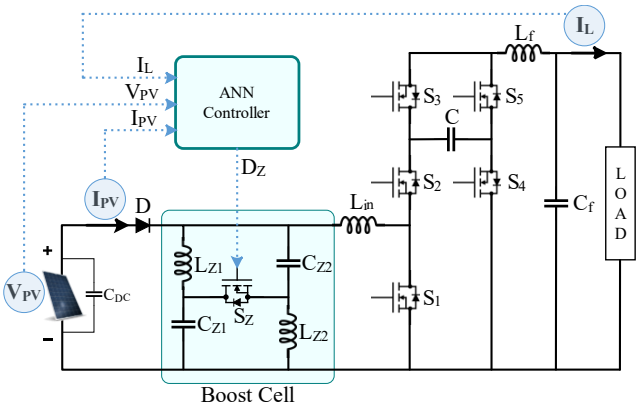


Fig. 1. System under investigation

These features collectively position the proposed microinverter as a versatile solution for photovoltaic applications.

The rest of the article is organized as—section II describes the system configuration and operation. ANN control of the proposed system is described in section III. Simulation results are presented in section IV, experimental results are given in section V and finally a brief conclusion is presented in section VI.

II. System Configuration And Operation

In the presented micro-inverter topology, as depicted in Fig. 1, a Z-source type boosting cell is integrated with an accompanying switch S_Z. This inclusion empowers a considerably extended boost operation range. A capacitor C is used to facilitate the energy transfer from input to output. The core of the micro-inverter comprises MOSFET switches, denoted as S_1 to S_5, collectively constituting the inverter section. This arrangement is complemented by an output filter, an inductor L_f and a capacitor C_f. Together, these elements define the micro-inverter configuration. The proposed microinverter achieves high voltage gain, common grounding, and off-grid capability.

A. Operation of the boost cell:

The boost cell operates in two modes, depending on the state of boost-cell switch S_Z. During the ON state, inductors L_{Z1} and L_{Z2} undergo charging, while capacitors C_{Z1} and C_{Z2} discharge. Conversely, in the OFF state, L_{Z1} and L_{Z2} discharge, and C_{Z1} and C_{Z2} charge. Fig. 2(a) and Fig. 2(b) depict the ON and OFF states of the switch. In ON State, when S_z is ON, the inductors charge, causing capacitors to discharge, and a reverse-biased diode interrupts the solar panel circuit connection. The governing equations are given by (1). In OFF State, with S_z OFF, the solar panel connects to the circuit, leading to capacitor charging and inductor discharge, described by (2).

$$\left.\begin{matrix} V_{LZ1} = V_{CZ2} \;\; and \;\; V_{LZ2} = V_{CZ1} \\ V_{LZ2} + V_{CZ2} - V_Z = 0 \end{matrix}\right\} \quad (1)$$

$$\left.\begin{matrix} V_{PV} - V_{LZ1} - V_{CZ1} = 0 \\ V_{LZ2} + V_{CZ2} - V_Z = 0 \end{matrix}\right\} \quad (2)$$

TABLE-I
MODES OF OPERATION OF BOOST-CELL

Parameters \ Mode	S_Z is ON	S_Z is OFF
L_{Z1}	Charging	Discharging
C_{Z1}	Discharging	Charging
L_{Z2}	Charging	Discharging
C_{Z2}	Discharging	Charging
L_{in}	Charging	Discharging

Fig. 3 shows the operational waveforms of the boost cell, illustrating the changes in voltage and current for inductors L_{Z1} and L_{Z2}, capacitors C_{Z1} and C_{Z2}, and input inductor L_{in} during the ON and OFF states of switch S_Z. For the inductors L_{Z1} and L_{Z2}, the waveforms i_{LZ1} and i_{LZ2}, along with their respective changes in current, $\Delta i_{LZ1}$and Δi_{LZ2}, signify the charging and discharging cycles. These cycles correspond directly to the states of the semiconductor switches, ensuring the appropriate energy transfer for voltage boosting. Similarly, the voltage waveforms across the capacitors, v_{CZ1} and v_{CZ2}, along with the change in voltages Δv_{CZ1} and Δv_{CZ2}, showcase the capacitor charging and discharging cycles,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

which are important in maintaining the converter output voltage stability during varying load conditions. The input inductor current waveform, $i_{L_{in}}$, and its change, $\Delta i_{L_{in}}$, illustrates the energy storage and release that facilitates the smooth transition between the high and low states of the output voltage.

The boost cell design demonstrates intrinsic symmetry, consequently, it is assumed that $V_{LZ1} \cong V_{LZ2} = V_{LZ}$ and $V_{CZ1} \cong V_{CZ2} = V_{CZ}$. Applying the volt-second balance equation to inductors L_{Z1} and L_{Z2} enables the derivation of the gain expression (V_Z/V_{PV}), expressed as:

$$\frac{V_Z}{V_{PV}} = \frac{2(1-D_Z)}{(1-2D_Z)} \tag{3}$$

Where D_Z: Duty ratio of boost cell.

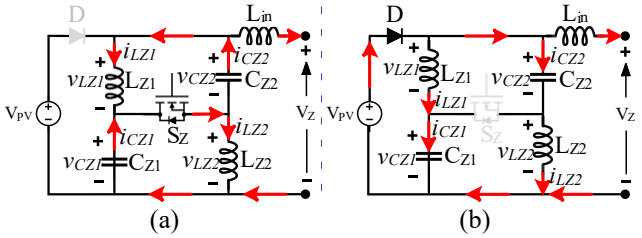


Fig. 2. Operation of the boost cell when (a) S_Z is ON (b) S_Z is OFF.

From (3), the operational constraint of boost cell is when the duty ratio surpasses 0.5. Values exceeding 0.5 lead to an undesirable negative gain within the proposed topology. The graphical representation of the observed voltage gain variation is presented in Fig. 4.

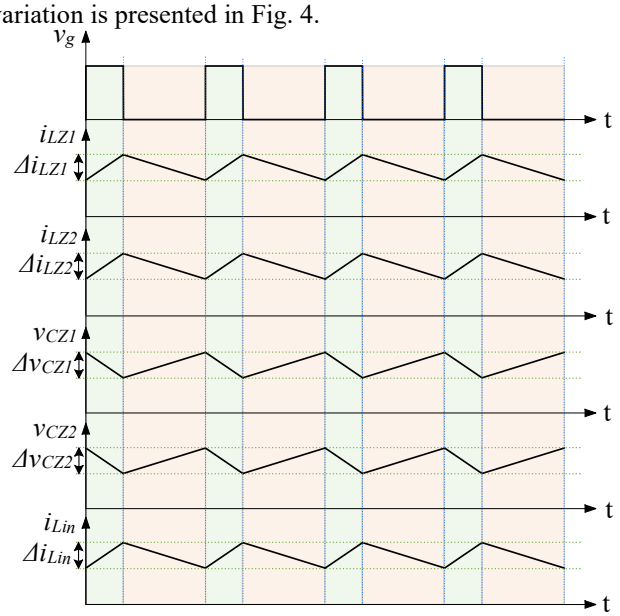


Fig. 3. Boost-cell operation waveforms

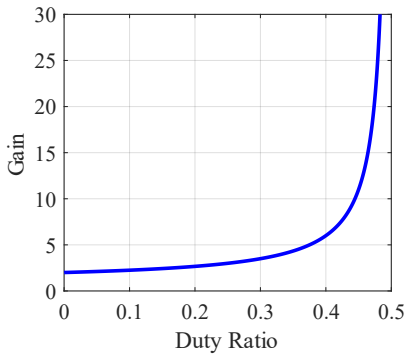


Fig. 4. Gain plot of the boost cell

B. Operation of the inverter:

The microinverter operates in four modes, distinguished by specific ON-OFF switch patterns as outlined in Table-I. These modes, namely Mode I+, Mode II+, Mode I–, and Mode II–, correspond to the microinverter's performance during the positive and negative half cycles of the output voltage waveform as illustrated in Fig. 5.

TABLE-II
MODES OF OPERATION

Component \ Mode	Mode I+	Mode II+	Mode I–	Mode II–
S_1	1	0	1	0
S_2	0	1	1	1
S_3	1	0	0	0
S_4	1	1	0	1
S_5	0	1	1	1
C	Discharging	Charging	Discharging	Charging

1) Mode I+: Fig. 5(b) illustrates the operation of microinverter in Mode I+, with switch states given in Table-I. In this mode, the input inductor L_{in} charges, subsequently powering the load with the energy stored in capacitor C. The corresponding voltage equation for this operational mode is given in (4).

2) Mode II+: In this operational mode, capacitor C charges and inductor L_{in}, discharges its energy to power the load. As given in Table-I and illustrated in Fig. 5(c), switches 1 and 3 are OFF, while switches 2, 4, and 5 are ON. Switch 4, coupled with the internal diode of switch 5, forms a freewheeling path for the load and the filter circuit. The voltage equation for this mode can be expressed mathematically as (5).

$$\left.\begin{aligned} V_Z - V_{L_i} &= 0 \\ V_C - V_{L_f} - V_0 &= 0 \end{aligned}\right\} \tag{4}$$

$$\left.\begin{aligned} V_Z - V_{L_i} - V_C &= 0 \\ V_0 + V_{L_f} &= 0 \end{aligned}\right\} \tag{5}$$

In the positive half cycle of the output voltage, Mode I+ extends for DT, and Mode II+ for $(1-D)T$, where D is the duty ratio. Applying the volt-second balance equation to the input inductor L_{in} yields the gain expression (V_0/V_Z):

$$G+= \frac{D}{1-D} \tag{6}$$

3) Mode I–: In this mode of operation, negative output voltage cycle phase, switches 1, 2, and 5 are ON. Inductor L_{in} charges with the supply voltage, and capacitor C discharges towards the load. The circuit equations for this mode are given in (7). Fig. 5(d) shows this mode of operation.

4) Mode II–: This mode of operation is shown in Fig. 5(e). In this mode, capacitor C charges and inductor L_{in} discharges. Switch 5 and the internal diode of switch 4 provide a freewheeling path for the load and the filter. Equations governing this mode are presented in (8)

$$\left.\begin{aligned} V_Z - V_{L_i} &= 0 \\ V_C + V_{L_f} + V_0 &= 0 \end{aligned}\right\} \tag{7}$$

$$\left.\begin{aligned} V_Z - V_{L_i} - V_C &= 0 \\ V_{L_f} + V_0 &= 0 \end{aligned}\right\} \tag{8}$$

During the negative half cycle of the output voltage, Mode I– extends for a duration of DT, and Mode II- for $(1-D)T$. Applying the volt-second balance equation to filter inductor L_f yields the gain expression (V_0/V_Z):

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

$$G-= -\frac{D}{1-D} \tag{9}$$

The expression for the gain (V_0/V_Z) from (6) and (9) is:

$$\frac{V_0}{V_Z} = \pm\frac{D}{1-D} \tag{10}$$

Combining equations (3) and (10) gives the expression for the overall gain (V_0/V_{PV}) of the microinverter.:

$$G = \frac{V_0}{V_{PV}} = \pm\frac{2(1-D_Z)}{(1-2D_Z)} \times \frac{D}{1-D} \tag{11}$$

The transfer function for the system from (11) is given by,

$$G(s) = \frac{\frac{2(1-D_Z)}{(1-2D_Z)} \times \frac{D}{1-D}}{(1+\tau_b s)(1+\tau_i s)} \tag{12}$$

Where, τ_b and τ_i are the time constants of the boost-cell and inverter respectively.

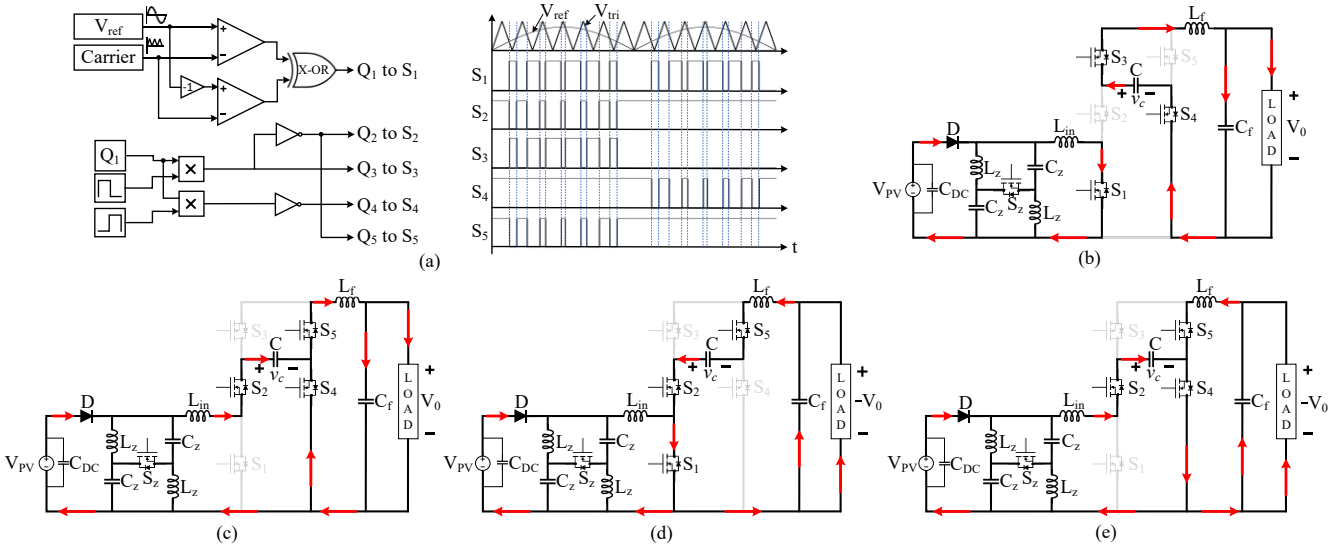


Fig. 5. (a) Switching strategy for switches S_1 to S_5, (b, c) Operation of the circuit when $V_0 > 0$, (d, e) Operation of the circuit when $V_0 < 0$

C. *Bode plot of the system:*

Fig. 6 shows the bode plot of the system. The Bode plot indicates that the system is stable with a significant phase margin of approximately 135 degrees, suggesting robustness to parameter changes. The bandwidth of the system is around 10 rad/s, indicating the frequency range over which it effectively operates without significant attenuation. The low-frequency gain is about 20dB, showing amplification of low-frequency signals, while the high-frequency roll-off at -20 dB/decade suggests a first-order low-pass filter behaviour.

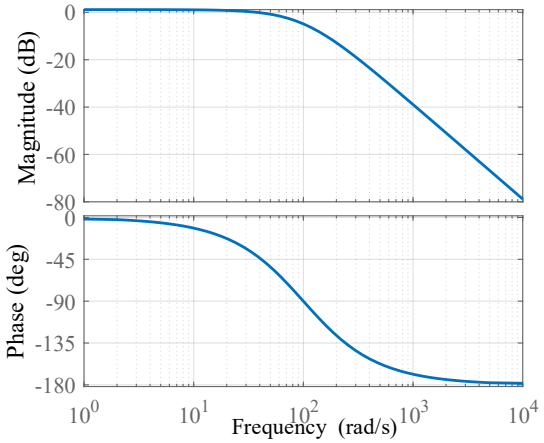


Fig. 6. Bode Plot of the system

The equations governing the design parameters of the components of proposed microinverter system is given in Table-III, which include inductors (L_{in} and L_Z) and capacitors (C_Z and C_C). The switches used in the microinverter circuit is selected based on the voltage and current stress calculated based on the equation given in Table-III.

For the inductors of the boost cell (L_Z), the design considers voltage variation across capacitors during the off state of the switch, boost cell duty cycle (D_Z), and the ripple current (Δi_{LZ}). This ensures capacitors charge and discharge properly, providing the necessary voltage boost for varying loads.

The capacitor (C_Z) is sized to manage the current from the inductors during the boost phase, against permissible voltage fluctuations (Δv_{C_Z}), maintaining steady voltage elevation. capacitor C is sized according to the output current (i_0) and the allowed ripple voltage (Δv_C), throughout the operational period, to ensure the output voltage remains consistent, yielding a stable AC output.

TABLE-III
EQUATIONS OF COMPONENT PARAMETERS OF MICROINVERTER

Parameters	Equations
Switch Voltage stress	$V_{SZ} = \frac{V_{PV}}{1-2D_Z}$ $V_{S1-S5} = \frac{2V_{PV}(1-D_Z)}{(1-D)(1-2D_Z)}$
Switch current Stress	$i_{SZ} = i_{LZ}$ $I_{S1-S3} = I_Z + I_0 \quad \& \quad I_{S4,S5} = I_0$
Inductors	$L_{in} = \frac{V_Z DT}{\Delta I_{L_i}}$ $L_Z = \frac{v_{CZ} D_Z T_Z}{\Delta i_{LZ}}$
Capacitors	$C_Z = -\frac{i_L D_Z T_Z}{\Delta v_{CZ}}$ $C = \frac{i_0 DT}{\Delta v_C}$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

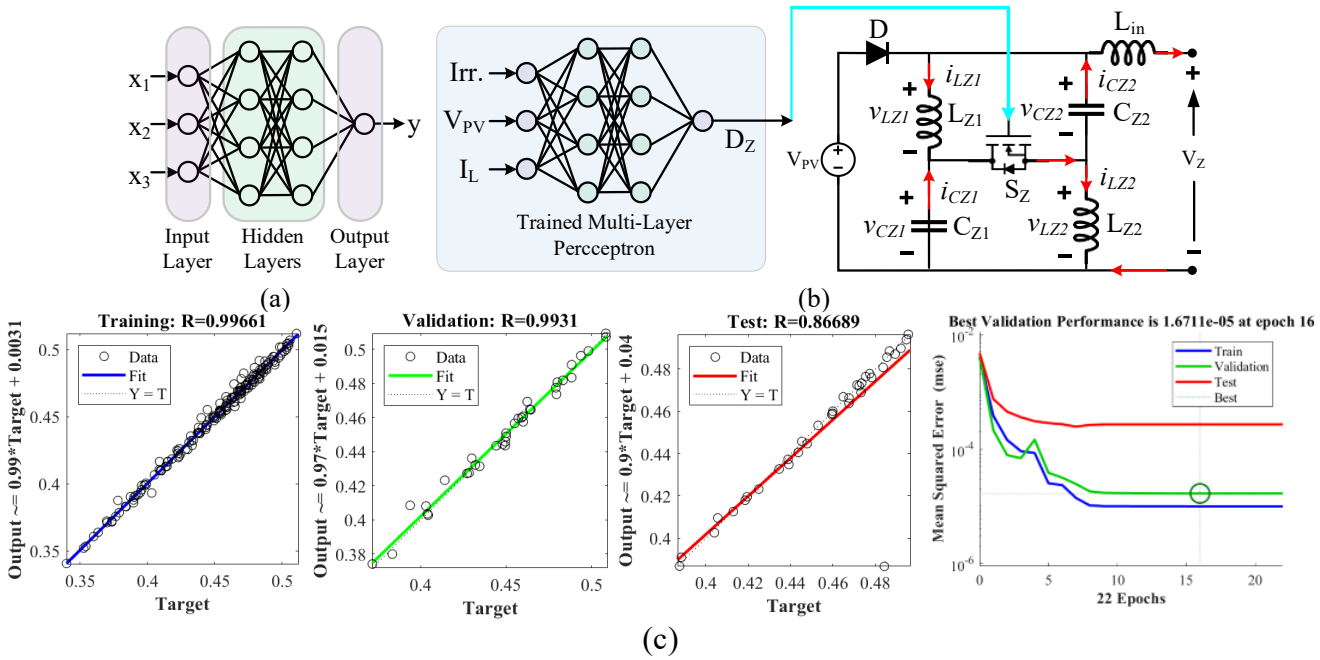


Fig. 7. (a) Basic MLP architecture of ANN, (b) Implementation of ANN as a controller for microinverter, (c) Result of Training of ANN

III. Control Scheme

The control of the proposed microinverter is achieved by adjusting the duty cycle of switch S_Z in response to the PV voltage, available irradiance, and the connected load. This control is carried out through the ANN algorithm. The algorithm has been trained to modify the duty cycle D_Z of switch S_Z based on variations observed in irradiance, PV voltage, and load current. Thus, the input parameters for this algorithm consist of irradiance, PV voltage, and load current, while the output parameter is the duty cycle D_Z.

Supervised neural network used in the microinverter is trained to produce the desired duty cycle to the estimated PV voltage, irradiance. Implementation of this supervised neural network is carried out using around 600 input-output training data that were obtained from simulation of the complete microinverter topology.

Objective Functions:

- Constant output voltage with change in input voltage
- Constant output voltage with change in irradiation.
- Constant output voltage with change in load.

A. *Structure of ANN*

The basic structure of the multi-layer perceptron ANN is given Fig. 7(a). x_1, x_2, x_3 are inputs while y is output. The architecture used here is a Multi-Layer Perceptron (MLP) Network. It can be divided into three layers, input layer, hidden layer and output layer. The input layer consists of three inputs—irradiance of the sun, PV voltage and load current. The input data go through a processing unit called the nucleus or processor. This unit adds up the weighted inputs and then passes the result through a binary threshold function, which is best suited for binary control of boost-cell switching. There are other activation functions discussed in [28]. The weights are defined during the training process to justify the output of the network. The error is sent back from the output to the input neuron, and the weights are adjusted through number of iterations until the input–output relationship is reached. This method is also known as Lavenberg-Marquardt (LM) method. In presented work, the inputs are irradiance, PV voltage and load current; the network must be capable of generating the optimal duty ratio, as shown in Fig. 7(c).

B. *Training of ANN*

Supervised neural network used in the microinverter is trained to produce the desired duty cycle to the estimated PV voltage, irradiance. Implementation of this supervised neural network is carried out using around 600 input-output training data that were obtained from simulation of the complete microinverter topology. An LM back propagation technique is used in this work. The network has 3 inputs and 1 output. Hidden layer of 10 neurons is employed. ANN is trained based on the training parameters shown in Table-IV. It is to be noted that no training occurs during real-time, and ANN is trained offline. The sampling time of 10 μs is used for both training and control stage.

TABLE-IV
ANN Training Parameters

Parameters	Value
Number of Input	3
Number of Output	1
Hidden Layers	2
Hidden Neurons	10
Iterations	1000
Method	Back Propagation
μ	10^{-6}
σ	0.1

C. *Implementation of ANN*

The trained ANN is implemented as shown in Fig. 7. The parameters of the ANN is given in Table-IV and it is trained in offline mode. A random number of iterations are required before the weights are finalized. The training dataset is divided into three groups—training data, test data and validation dataset. The training dataset is used to train the model, the validation dataset is used during training to fine-tune the parameters of the model and prevent overfitting, and the test dataset is employed after training to evaluate the performance of the model on unseen data and assess its

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

capabilities. The training process concludes when the specified error criterion is attained on the validation dataset. The ANN is then used as shown in Fig. 7(b).

Fig. 7(c) shows the regression plot of ANN in MATLAB environment which includes training, validation, and testing. Among all the dataset, 70% is used for training, 15% is used for validation and remaining 15% is used for testing.

IV. Loss Analysis of the Proposed Microinverter

The losses due to switches are calculated using the PLECS model of the proposed topology. Fig. 12 shows the power loss under different output power levels for DC input voltage of 36V and AC output voltage of 230V$_{RMS}$.

A. Switching Loss

Switches exhibit a finite switching time, which results in switching losses arising from the dynamic voltages and currents they need to manage while transitioning between the on and off states. Gate-drive losses are considered part of switching losses because they are necessary for turning the switches on and off. The switching losses in MOSFETs depend on current and switching frequency as indicated by,

$$P_{SW} = \frac{1}{2} \times (V_{SZ} \times I_{SZ} + V_{S1} \times I_{S1} + V_{S2} \times I_{S2} + V_{S3} \times I_{S3} + V_{S4} \times I_{S4} + V_{S5} \times I_{S5}) \times (t_r + t_f) \times f_{sw} \quad (13)$$

Where,

V_S = voltage across switch (V)
I_S = current through the switch (A)
t_r = rise time (sec)
t_f = fall time (sec)
f_{SW} = switching frequency

B. Conduction Loss

Conduction loss can be calculated using the equation,

$$P_{CON} = (I_{SZ}^2 + I_{S1}^2 + I_{S2}^2 + I_{S3}^2 + I_{S4}^2 + I_{S5}^2) \times R_{DSon} \quad (14)$$

As the MOSFET conducts, energy is dissipated in the form of heat due to the resistance within the device, resulting in conduction losses. The on-state resistance of the switch, R_{DSon}, causes conduction power loss. It affects the efficiency of the microinverter because the lost energy reduces the amount of power that can be delivered to the load.

C. Inductor Loss

Loss due to inductors is associated with its coil resistance. This loss is given by the equation,

$$P_{ind} = I_{LZ}^2 \times (R_{LZ1} + R_{LZ2}) + I_{in}^2 \times R_{Lin} + I_{filter}^2 \times R_{Lf} \quad (15)$$

Inductor losses are directly related to the current flowing through each inductor and the resistance of their windings.

D. Capacitor Loss

The power loss in a capacitor is given by the equation,

$$P_{cap} = (I_{CZ1} + I_{CZ2} + I_C + I_{Cf})^2 \times ESR \quad (16)$$

The total loss can be calculated using (30),

$$P_{Loss,Total} = P_{SW} + P_{CON} + P_{ind} + P_{cap} \quad (17)$$

Furthermore, the efficiency of the proposed microinverter can be calculated using (18),

$$\%\,\eta = \frac{P_{out}}{P_{out} + P_{Loss,Total}} \times 100 \quad (18)$$

Fig. 8 shows the percentage distribution of switching and dynamic losses, it indicates that switching loss contributes to majority of the losses.

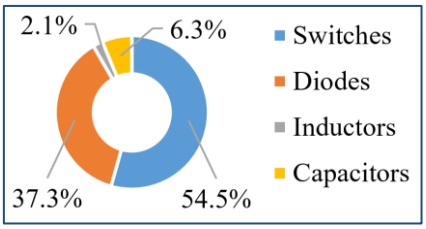


Fig. 8. Power loss distribution of all the components

V. Comparison with Existing Topologies

A detailed comparison of the proposed microinverter with some existing inverters is presented in Table-2. A family of buck-boost inverters [16] with virtual ground is presented, there is no common ground due to which there is a possibility of leakage current flow. Proposed topology has common grounding which helps in mitigating leakage current. Buck-boost inverter [20] and transformer-less single-phase inverter [25] with common-ground can reduce the PV leakage current. However, buck-boost inverter requires higher number of component count (5 switches, 5 to 8 diodes, 4 to 5 inductors and 3 to 4 capacitors) and transformer-less single-phase inverter suffers with low efficiency and low voltage gain. The proposed topology has comparatively lower number of components counts and offers higher gain and efficiency.

A flyback-type microinverter [2], boost–flyback/flyback microinverter [26], Z-source-derived microinverter [27] and CCM zeta microinverter [18] offer high gain but lack common-grounding capability which results in considerable amount of PV leakage current. Also, flyback, boost flyback and CCM zeta microinverters require transformer which make the system bulky however, it operates in off-grid mode of operation. The transformer-less inverter works in both grid-tied and off-grid mode of operation but as mentioned earlier it has low voltage gain and efficiency. The proposed microinverter for off-grid application with common ground offers negligible ground leakage current, high gain and competitive power conversion efficiency.

Table-V
Comparison with Existing Topologies

Ref. / Parameters	Buck-Boost Inverters [19]	Flyback-type Microinverter [29]	Buck–Boost Inverters [23]	Hybrid Boost-Flyback/Flyback Microinverter [30]	Z-Source-Derived Microinverter [31]	CCM Zeta Microinverter [21]	Single-Phase Transformerless Inverter [5]	Proposed Microinverter
No. of Switches	6-8	4	5	7	5	5	3	6
No. of Diodes	0	4	5-8	2	2	0	1	1

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

No. of Inductors	1	1	4-5	3	4	2	3	4
No. of Capacitors	2	2	3-4	5	3	3	4	4
Common-grounding	No	No	Yes	No	No	No	Yes	Yes
Off-grid Capability	No	No	No	No	No	Yes	No	Yes
Control	Simple	Moderate	Moderate	Simple	Complex	Complex	Moderate	Moderate
Leakage current	Small	Small	Negligible	Moderate	High	Moderate	Negligible	Negligible
Peak Efficiency	97.50% at V_{in}=160V & V_0=115V_{RMS}	95.8% at V_{in}=35V & V_0=230V_{RMS}	96.74% at V_{in}=77-200V & V_0=110V_{RMS}	96.2% at V_{in}=22-48V & V_0=220V_{RMS}	96.5% V_{in}=35V & V_0=220V_{RMS}	93% V_{in}=30-48V & V_0=230V_{RMS}	93% V_{in}=38V & V_0=110V_{RMS}	97% V_{in}=24-47V & V_0=230V_{RMS}
Gain Equation	$\frac{D_{1max}}{1-D_{1max}}$	$\frac{n}{1-D}$	$\frac{D}{1-D}$	$\frac{1+ND}{1-D}$	$\frac{1+n}{1-D}$	$\frac{\frac{nD}{1-D}}{n=\frac{V_{0peak}(1-D_{\max})}{V_{inmin}D_{\max})}}$	$\frac{1-2D}{D(1-D)}$	$\pm\frac{2(1-D_Z)}{(1-2D_Z)}\times\frac{D}{1-D}$
Power (W)	250	250	400	240	250	220	280	500
Voltage stress on switches	$V_{S(1,2)}=\frac{V_{in}(1+D)}{1-D}$ $V_{S(a-d)}=V_0$	$V_S=V_{in}+\frac{V_0}{n}$ $V_D=\frac{nV_{PV}+V_0}{2}$	$V_S=V_{in}+V$	$V_S=V_{inmax}+\frac{V_{inmax}+\Delta V}{N}$	$V_S=\frac{V_{in}}{n+1}$	$V_S=nV_{inmax}+V_{0peak}$	$V_{S1}=\frac{DV_0}{1-2D}$ $V_{S2}=\frac{V_0}{1-2D}$ $V_{S3}=\frac{(1-D)V}{1-2D}$	$V_{SZ}=\frac{V_{PV}}{1-2D_Z}$ $V_{S1-S5}=\frac{2V_{PV}(1-D_Z)}{(1-D)(1-2D_Z)}$
Peak Voltage Stress	260V – 360V	-	280V – 380V	110V – 120V	120V – 130V	50V (-40V negative peak)	200V – 270V	165V for S_Z 280V – 300V for S_1–S_5
Current stress on switches	$I_{S(1,2)}=I_0(1+G)$ $I_{S(a-d)}=I_0$	-	$I_{S1,4}=I_{in}+I_0$ $I_{S2}=I_{in}$ $I_{S3,5}=I_0$	$2I_0\left(\frac{1}{N}\frac{V_0-V_{inmin}}{V_{inmin}}\right)$	I_{in}	$I_S=-\frac{I_{ppeak}}{n}+\frac{V_0}{L_0}(t-t_{c1})+\frac{V_0}{n^2L_m}(t-t_{c1})$	$I_{S1max}=-\frac{I_0}{1-D}$ $I_{S2max}=I_{S3max}=\frac{I_0}{D}$	$i_{SZ}=i_{LZ}$ $I_{S1-S3}=I_Z+I_0$ & $I_{S4,S5}=I_0$
Peak Current stress	-	-	-	-	-	30A – 35A (with -8A negative peak)	-	14A for S_Z -10A to +10A for S_1–S_5

VI. Results and Discussion

Experiments are conducted to confirm the proper functionality of the proposed microinverter topology. The microinverter's performance is analysed under diverse operating conditions, including variations in input voltage and load. Fig. 9 shows the experimental results of the proposed microinverter under different operating conditions. Fig. 9 shows the PV voltage, load voltage, load current and voltage across boost-cell switch S_Z. The PV voltage is around $36V$, the load voltage is $230V_{RMS}$, the load current is about $2.25A$, and the voltage across S_Z is about 165V. The waveforms are in steady state and show the proper functionality of the microinverter. The parameters of the proposed topology are presented in Table 6.

TABLE-VI
Real-Time Simulation Parameters

Parameters	Value
Input voltage (V_{PV})	24V - 48V
Output Voltage (V_0)	230 V_{RMS}
Output Power (P)	500 W
Fundamental Frequency (f)	50 Hz
Switching Frequency (f_s)	10 kHz
Boost-cell Inductors (L_{Z1}, L_{Z2})	1 mH
Boost-cell Capacitors (C_{Z1}, C_{Z2})	100 µF
Input Inductor (L_{in})	5 mH
Capacitor (C)	500 µF
Filter Inductor (L_f)	10 mH
Filter Capacitor (C_f)	1000 µF

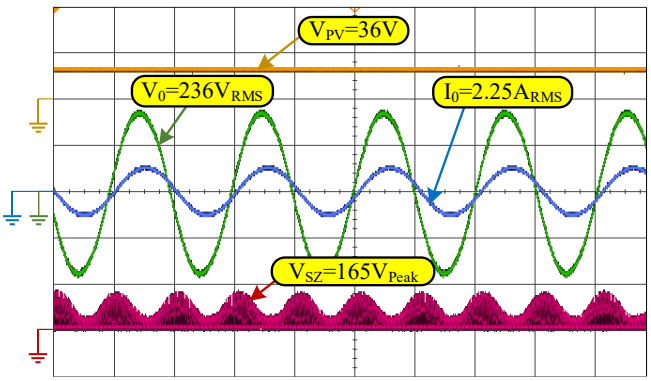


Fig. 9. Input voltage, output voltage, output current and switch 'S_Z' voltage for $RL = 100\Omega + j100mH$

Fig. 10 shows the output voltage and current waveforms of the proposed microinverter under different operating conditions. Fig. 10(a) Output voltage and current waveforms under varying input voltage: This subfigure shows how the output voltage (V_0) and current (I_0) remain stable when the input voltage (V_{PV}) changes from $36V$ to $44V$ and then to $28V$. The output voltage has a peak value of about 325V and output current has peak value of about 3.18A. The input PV voltage varies in steps from 36V to 44V to 28V and 36V.

Fig. 10(b) shows the output voltage and current waveforms under varying load conditions. In this case, output voltage stays at 230V and output current varies with change in load. The load is changed from 500W to 100W to 250W; the load current decreases with decrease in load and increases with increase in load. The load used in this work is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

combination of resistor (R) and inductor (L). The subfigure also shows zoomed-in views of the output voltage and current waveforms for each case, highlighting the changes in output voltage and current.

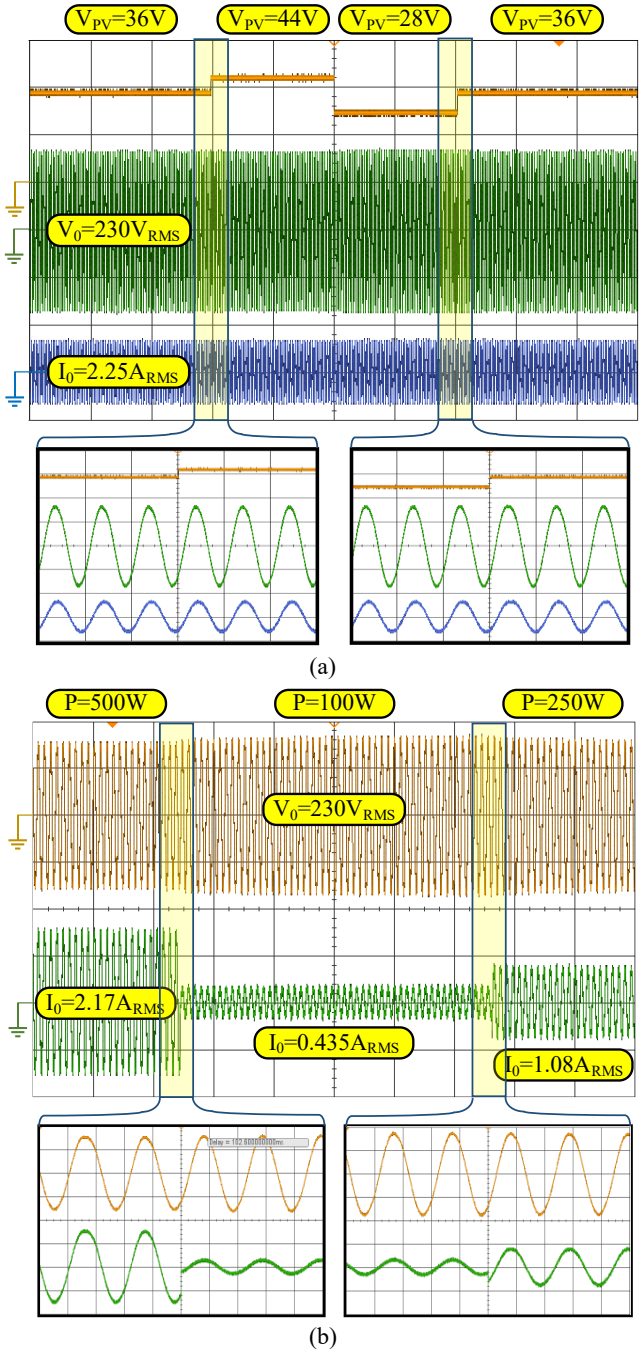


Fig. 10. (a) Output voltage and current waveforms under varying input voltage (b) Output voltage and current waveforms under varying load.

Table 7 shows the components used in the experimental setup, highlighting specific part numbers/values.

Fig. 11 illustrates the voltage stress distribution across the switches in the proposed microinverter. The peak voltage stress across switch S_Z is approximately 165V, whereas the peak voltage stress across all other switches fluctuates within the range of 280V to 300V. The analysis of these waveforms and stress distributions underscores the stable and effective performance of the proposed microinverter topology under diverse operating conditions. Fig. 12 shows the power loss under different output power levels for DC input voltage of 36V and AC output voltage of 230V_{RMS}. Fig. 12 also shows the efficiency curve of the proposed microinverter under different operating voltages.

TABLE-VII
COMPONENTS USED IN THE EXPERIMENTAL SETUP

S. No.	Components	Part No./Value
1.	Power Switches	IRFP460
2.	Diode	MUR3060
3.	Magnetic cores of Inductors	PQ6070
4.	ESR of Capacitors	20 $m\Omega$
5.	Copper wire resistivity	1.7 $\mu\Omega - cm$

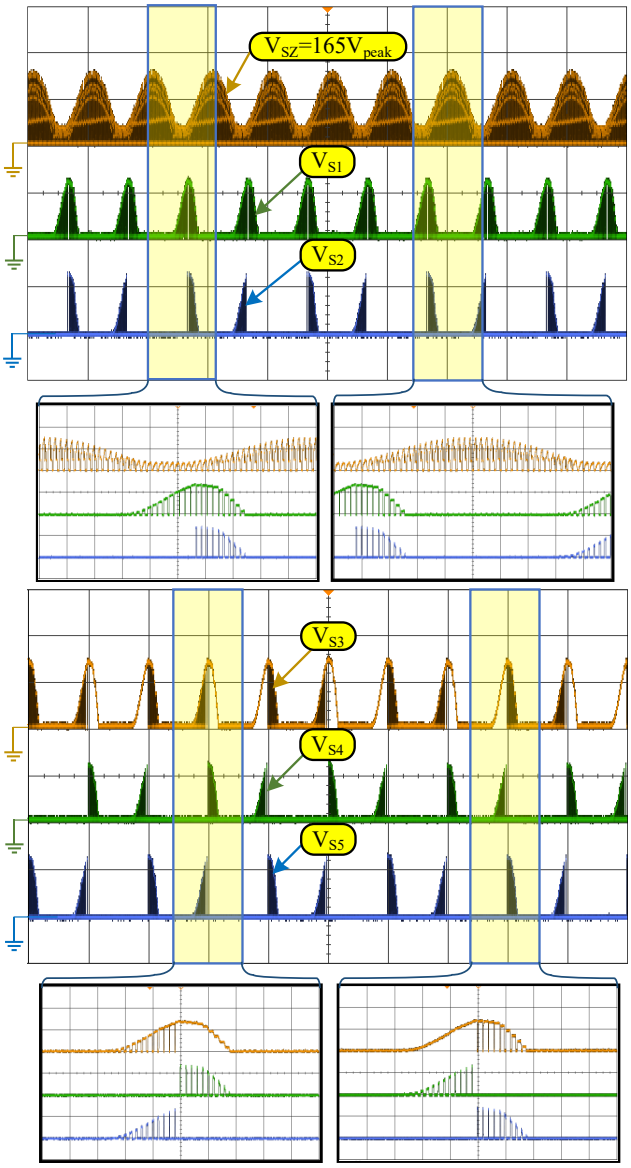


Fig. 11. Voltage waveforms across the semiconductor switches

A. Measurement of leakage current and THD

Leakage current in PV systems, especially in transformer-less inverters, is a critical parameter due to its potential to cause safety issues and reduce efficiency. It is measured by observing the current flowing through the ground path. Fig. 13 shows a very low amplitude current, implying that the inverter design effectively minimizes leakage current.

Fig. 14 (a) and Fig. 14(b) illustrate the Total Harmonic Distortion (THD) measurements for the output voltage and current, respectively. The output voltage has a THD of 2.6%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

at an RMS voltage of 229.9V and frequency of 49.5Hz. This indicates that the implemented microinverter ensures high power quality and efficient operation of the connected loads. Furthermore, the output current has a THD of 2.25% at an RMS current of 1.548A and frequency of 49.52Hz. This low current THD indicates minimizing potential heating issue in the microinverter and connected devices.

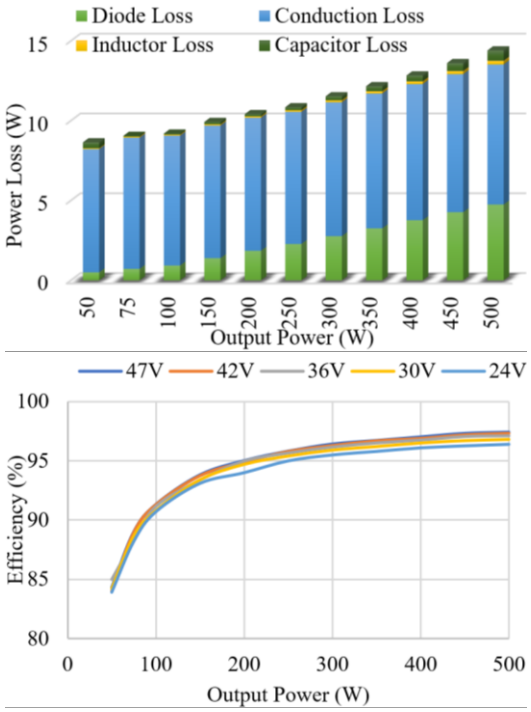


Fig. 12. Power Loss and Efficiency

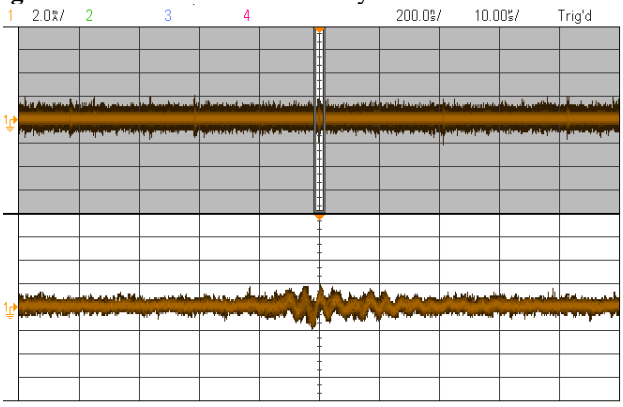
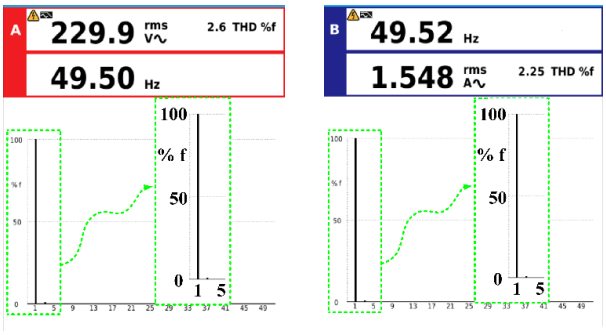


Fig. 13. Measurement of leakage current



(a) (b)

Fig. 14. Measurement of (a) Voltage THD, (b) Current THD

VII. Conclusion

This article presents a significant advancement in the field of solar microinverters. The proposed off-grid microinverter can handle input voltages between 24V and 46V, consistently producing a 230V output voltage for up to 500W power. This is possible through a modified Z-source converter with an active switch that dynamically adjusts its duty ratio to adapt to voltage and load fluctuations. A significant feature is the microinverter's control system, driven by ANN controller. This ANN controller ensures optimal performance and adaptability under varying conditions. Additionally, the microinverter offers a common grounding feature between the solar panel and the load, effectively reducing PV leakage current. The article provides comprehensive insights into the operation and control of microinverter. It also compares the proposed microinverter with existing solutions in the field.

Acknowledgement

The authors would like to express their sincere gratitude to the reviewers for their constructive comments and suggestions, which significantly contributed to enhancing the quality of this paper. Their insightful feedback and meticulous review have been invaluable in refining the manuscript.

References

[1] K. Alluhaybi, I. Batarseh, and H. Hu, "Comprehensive Review and Comparison of Single-Phase Grid-Tied Photovoltaic Microinverters," *IEEE J Emerg Sel Top Power Electron*, vol. 8, no. 2, pp. 1310–1329, Jun. 2020, doi: 10.1109/JESTPE.2019.2900413.

[2] C. Trujillo Rodriguez, D. Velasco De La Fuente, G. Garcera, E. Figueres, and J. A. Guacaneme Moreno, "Reconfigurable control scheme for a PV microinverter working in both grid-connected and Island modes," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 4, pp. 1582–1595, 2013, doi: 10.1109/TIE.2011.2177615.

[3] Z. Wang *et al.*, "Active Power Control of Voltage-Controlled Photovoltaic Inverter in Supporting Islanded Microgrid Without Other Energy Sources," *IEEE J Emerg Sel Top Power Electron*, vol. 10, no. 1, pp. 424–435, Feb. 2022, doi: 10.1109/JESTPE.2021.3069700.

[4] W. Qi, M. Wang, and S. Li, "Transformerless Three-Level Flying-Capacitor Step-Up PV Micro-Inverter without Electrolytic Capacitors," *IEEE J Emerg Sel Top Circuits Syst*, vol. 11, no. 1, pp. 49–58, Mar. 2021, doi: 10.1109/JETCAS.2020.3039284.

[5] A. Sarikhani, M. M. Takantape, and M. Hamzeh, "A Transformerless Common-Ground Three-Switch Single-Phase Inverter for Photovoltaic Systems," *IEEE Trans Power Electron*, vol. 35, no. 9, pp. 8902–8909, Sep. 2020, doi: 10.1109/TPEL.2020.2971430.

[6] H. Tian, M. Chen, G. Liang, and X. Xiao, "A Single-Phase Transformerless Common-Ground Type PV Inverter With Active Power Decoupling," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 4, pp. 3762–3772, Apr. 2023, doi: 10.1109/TIE.2022.3181361.

[7] A. R. Paul, A. Bhattacharya, and K. Chatterjee, "A Novel Single Phase Grid Connected Transformer-Less Solar Micro-Inverter Topology With Power Decoupling Capability," *IEEE Trans Ind Appl*, vol. 59, no. 1, pp. 949–958, Jan. 2023, doi: 10.1109/TIA.2022.3207445.

[8] F. C. Melo, L. S. Garcia, L. C. De Freitas, E. A. A. Coelho, V. J. Farias, and L. C. G. De Freitas, "Proposal of a Photovoltaic AC-Module with a Single-Stage Transformerless Grid-Connected Boost Microinverter," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 3, pp. 2289–2301, Mar. 2018, doi: 10.1109/TIE.2017.2750611.

[9] S. Dey, A. Mallik, and T. Warren, "A Variable Switching Frequency-Based Triple-Phase-Shifted Loss Optimized Modulation Strategy for DAB-Based DC–AC Microinverter," *IEEE J Emerg Sel Top Power Electron*, vol. 12, no. 1, pp. 1110–1128, Feb. 2024, doi: 10.1109/JESTPE.2023.3328271.

[10] F. M. Alhuwaishel and N. A. Ahmed, "Coupled Inductor Based Boost Microinverter With Dual Mode Time Sharing Operation for Renewable Energy Applications," *IEEE Access*, vol. 12, pp. 73889–73905, 2024, doi: 10.1109/ACCESS.2024.3405808.

[11] X. Li, J. Liu, F. Ji, X. Cao, Y. Wang, and J. Liu, "A Single-Stage High-Frequency-Link Split-Phase Microinverter for Both Grid-Tied and Islanded Operation," *IEEE Trans Power Electron*, vol. 39, no. 8, 2024, doi: 10.23919/ICPE2023.

[12] H. C. Chiang, F. J. Lin, and J. K. Chang, "Novel Control Method for Multimodule PV Microinverter with Multiple Functions," *IEEE Trans*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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 Journal of Emerging and Selected Topics in Power Electronics. 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/JESTPE.2024.3468391

Power Electron, vol. 33, no. 7, pp. 5869–5879, Jul. 2018, doi: 10.1109/TPEL.2017.2742546.

[13] M. Qiu, M. Wei, S. Yang, X. Liu, and D. Cao, "Fully Decoupled Active and Reactive Power Distribution Control for Single Phase Cascaded Connected Microinverter Under Island Mode," *IEEE Trans Ind Appl*, vol. 60, no. 4, pp. 6393–6408, 2024, doi: 10.1109/TIA.2024.3396420.

[14] M. Ali *et al.*, "Robust ANN-Based Control of Modified PUC-5 Inverter for Solar PV Applications," in *IEEE Transactions on Industry Applications*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 3863–3876. doi: 10.1109/TIA.2021.3076032.

[15] S. Li, Y. Sun, M. Ramezani, and Y. Xiao, "Artificial Neural Networks for Volt/VAR Control of DER Inverters at the Grid Edge," *IEEE Trans Smart Grid*, vol. 10, no. 5, pp. 5564–5573, Sep. 2018, doi: 10.1109/TSG.2018.2887080.

[16] S. Öztürk and I. Çadirci, "A Generalized and Flexible Control Scheme for Photovoltaic Grid-Tie Microinverters," in *IEEE Transactions on Industry Applications*, Institute of Electrical and Electronics Engineers Inc., Jan. 2018, pp. 505–516. doi: 10.1109/TIA.2017.2753175.

[17] Y. Shen, D. Zakzewski, A. Hasnain, R. Resalayyan, and A. Khaligh, "Sensorless Control for DC-Parallel Active Power Decoupling in PV Microinverters," *IEEE Trans Power Electron*, vol. 38, no. 11, pp. 14628–14637, Nov. 2023, doi: 10.1109/TPEL.2023.3306281.

[18] S. M. Tayebi and I. Batarseh, "Analysis and Optimization of Variable-Frequency Soft-Switching Peak Current Mode Control Techniques for Microinverters," *IEEE Trans Power Electron*, vol. 33, no. 2, pp. 1644–1653, Feb. 2018, doi: 10.1109/TPEL.2017.2676097.

[19] R. Aliaga, J. Munoz, M. Rivera, P. Wheeler, J. Rohten, and A. Trentin, "A Single Phase Hybrid Multiport Microinverter for Photovoltaic Energy Controlled by Exact Linearization," *IEEE Access*, vol. 11, pp. 50075–50089, 2023, doi: 10.1109/ACCESS.2023.3278099.

[20] F. Akbar, A. Elkhateb, H. F. Ahmed, A. A. Khan, H. Cha, and J. W. Park, "Single-Phase Virtual-Ground Transformerless Buck-Boost Inverters," *IEEE Trans Power Electron*, Sep. 2023, doi: 10.1109/TPEL.2023.3283927.

[21] O. Kwon, K. S. Kim, and B. H. Kwon, "Highly Efficient Single-Stage DAB Microinverter Using a Novel Modulation Strategy to Minimize Reactive Power," *IEEE J Emerg Sel Top Power Electron*, vol. 10, no. 1, pp. 544–552, Feb. 2022, doi: 10.1109/JESTPE.2021.3090097.

[22] R. K. Surapaneni and A. K. Rathore, "A Single-Stage CCM Zeta Microinverter for Solar Photovoltaic AC Module," *IEEE J Emerg Sel Top Power Electron*, vol. 3, no. 4, pp. 892–900, Dec. 2015, doi: 10.1109/JESTPE.2015.2438012.

[23] A. K. Bhattacharjee and I. Batarseh, "An Interleaved Boost and Dual Active Bridge-Based Single-Stage Three-Port DC-DC-AC Converter with Sine PWM Modulation," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 6, pp. 4790–4800, Jun. 2021, doi: 10.1109/TIE.2020.2992956.

[24] H. F. Ahmed, M. El Moursi, B. Zahawi, and K. Hassan Al Hosani, "Single-Phase Photovoltaic Inverters with Common-Ground and Wide Buck-Boost Voltage Operation," *IEEE Trans Industr Inform*, 2021, doi: 10.1109/TII.2021.3066511.

[25] M. Chen and P. C. Loh, "A Dual-Boost H-Bridge Inverter with Common Ground for Photovoltaic Interfacing," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 10, pp. 9515–9526, Oct. 2021, doi: 10.1109/TIE.2020.3026275.

[26] Z. Zhang, X. F. He, and Y. F. Liu, "An optimal control method for photovoltaic grid-tied-interleaved flyback microinverters to achieve high efficiency in wide load range," *IEEE Trans Power Electron*, vol. 28, no. 11, pp. 5074–5087, 2013, doi: 10.1109/TPEL.2013.2245919.

[27] Z. Zhang, M. Chen, W. Chen, C. Jiang, and Z. Qian, "Analysis and implementation of phase synchronization control strategies for BCM interleaved flyback microinverters," *IEEE Trans Power Electron*, vol. 29, no. 11, pp. 5921–5932, 2014, doi: 10.1109/TPEL.2014.2300483.

[28] "Neural Networks - A Comprehensive Foundation - Simon Haykin".

[29] R. Hasan, W. Hassan, and W. Xiao, "PV Microinverter Solution for High Voltage Gain and Soft Switching," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 3, no. 2, pp. 352–361, Feb. 2021, doi: 10.1109/jestie.2021.3061910.

[30] F. Zhang, Y. Xie, Y. Hu, G. Chen, and X. Wang, "A hybrid boost-flyback/flyback microinverter for photovoltaic applications," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 1, pp. 308–318, Jan. 2020, doi: 10.1109/TIE.2019.2897543.

[31] R. K. Surapaneni and P. Das, "A Z-Source-Derived Coupled-Inductor-Based High Voltage Gain Microinverter," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 6, pp. 5114–5124, Jun. 2018, doi: 10.1109/TIE.2017.2745477.



Ajay Kumar Sahu (Student Member, IEEE) received the B.E. degree in Electrical Engineering from Chhattisgarh Swami Vivekanand Technical University, Bhilai, India, in 2016 and the M. Tech. degree in Power system from Delhi Technological University, Delhi, India, in 2019. He is currently working towards a PhD degree in the area of power electronics in the Department of Electrical Engineering at National Institute of Technology, Raipur, India. His research interests include power electronic inverters for renewable energy sources, control techniques, and electric drives.



Ramnarayan Patel (Senior Member, IEEE) received the B.E. degree in electrical engineering from the Shri Govin-dram Seksaria Institute of Technology and Science, Indore, India, in 1997, the M.Tech. degree in energy studies, and the Ph.D. degree in electrical engineering from the Indian Institute of Technology Delhi, New Delhi, India, in 1999 and 2003, respectively. He is currently an Associate Professor with the Department of Electrical Engineering, National Institute of Technology, Raipur, India. His research interests include design and development of efficient power electronics interface, solar rural electrification, and rural economy.



Lalit Kumar Sahu (Senior Member, IEEE) received the M.Tech. degree in electrical drives and the Ph.D. degree in Electrical Engineering from the Maulana Azad National Institute of Technology, Bhopal, India in 2010 and 2016, respectively. He is currently an Assistant Professor at the Department of Electrical Engineering, National Institute of Technology, Raipur, India. Dr. Sahu's research interests include power electronic converters, electrical drives, electric and hybrid electric vehicular systems, and renewable and hybrid energy systems.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 30,2024 at 04:31:15 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.

arXiv:2409.00639v1 [cs.CV] 1 Sep 2024

Artificial Intelligence in Gastrointestinal Bleeding Analysis for Video Capsule Endoscopy: Insights, Innovations, and Prospects (2008-2023)

Tanisha Singh*, Shreshtha Jha*, Nidhi Bhatt†, Palak Handa‡,
Nidhi Goel*, S. Indu§

*Department of Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India
†Department of Computer Science Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India
‡Research Center for Medical Image Analysis and Artificial Intelligence, Department of Medicine, Danube Private University, Krems, Austria
§Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

***Abstract*—The escalating global mortality and morbidity rates associated with gastrointestinal (GI) bleeding, compounded by the complexities and limitations of traditional endoscopic methods, underscore the urgent need for a critical review of current methodologies used for addressing this condition. With an estimated 300,000 annual deaths worldwide, the demand for innovative diagnostic and therapeutic strategies is paramount. The introduction of Video Capsule Endoscopy (VCE) has marked a significant advancement, offering a comprehensive, non-invasive visualization of the digestive tract that is pivotal for detecting bleeding sources unattainable by traditional methods. Despite its benefits, the efficacy of VCE is hindered by diagnostic challenges, including time-consuming analysis and susceptibility to human error. This backdrop sets the stage for exploring Machine Learning (ML) applications in automating GI bleeding detection within capsule endoscopy, aiming to enhance diagnostic accuracy, reduce manual labor, and improve patient outcomes. Through an exhaustive analysis of 113 papers published between 2008 and 2023, this review assesses the current state of ML methodologies in bleeding detection, highlighting their effectiveness, challenges, and prospective directions. It contributes an in-depth examination of AI techniques in VCE frame analysis, offering insights into open-source datasets, mathematical performance metrics, and technique categorization. The paper sets a foundation for future research to overcome existing challenges, advancing gastrointestinal diagnostics through interdisciplinary collaboration and innovation in ML applications.**

***Index Terms*—Gastrointestinal Bleeding, Video Capsule Endoscopy, Machine Learning, Deep Learning**

I. Introduction

Gastrointestinal (GI) bleeding, a critical medical condition, involves the hemorrhage within the digestive tract encompassing the esophagus, stomach, small intestine, large intestine (colon), rectum, and anus. This pathological state presents a spectrum of risks, from immediate life-threatening emergencies to long-term detrimental consequences [1], [2]. The severity of GI bleeding ranges from asymptomatic microhemorrhages detectable only by laboratory tests to significant visible bleeding, manifesting in stool or vomit, potentially culminating in fatal outcomes. The emergence of GI bleeding stems from various underlying conditions, including peptic ulcers, diverticulosis, inflammatory bowel disease (IBD), and colorectal cancer, each contributing to the complexity and danger of this condition. The World Health Organization (WHO) underscores the gravity of this issue, attributing approximately 300,000 annual deaths worldwide to GI bleeding [3]. These daunting statistics underscore the need for continued research and innovation in diagnostic and therapeutic approaches.

GI bleeding is broadly classified into upper and lower GI bleeding, each with distinct characteristics and implications [4]. Upper GI bleeding, more prevalent than its lower counterpart primarily originates from the upper digestive tract [5]. The incidence of upper GI bleeding ranges from 50 to 150 cases per 100,000 population annually [6], disproportionately affecting older individuals [7]. Mortality rates for upper GI bleeding lie between 6% to 10%, reflecting its significant health impact [8]–[10]. Conversely, lower GI bleeding, with an annual incidence of 20-27 episodes per 100,000 persons, has a slightly lower mortality rate of 4%-10% [11], [12]. Acute GI bleeding, a significant cause of hospital admissions in the United States, is estimated at 300,000 patients annually [13], [14]. The clinical presentation of GI bleeding can be overt, such as hematemesis, melena, hematochezia, or occult, often detected through positive fecal occult blood tests or iron deficiency anemia [15]. The complexity extends to obscure GI bleeding, where the bleeding source remains

unidentified after standard endoscopic evaluations [16], [17].

Diagnostic modalities for GI bleeding have evolved, with endoscopy being the primary method for upper GI bleeding [18]. It involves direct visualization of the GI tract to identify bleeding sources. However, the advent of Video Capsule Endoscopy (VCE) marks a significant advancement. This method has shown a higher diagnostic yield in patients with chronic occult and obscure GI bleeding compared to push enteroscopy and barium studies [19]. VCE, a non-invasive and painless procedure, involves swallowing a small, wireless capsule containing a camera that captures frames as it travels through the digestive tract. It provides a comprehensive visualization of the small intestine, which is not easily accessible by traditional endoscopy methods. It offers the advantage of minimal or no hospitalization and specialized care. It does not require sedation, and the capsule is naturally excreted from the body. It plays a crucial role in identifying the source and location of bleeding. Thus, it can help in detecting lesions, ulcers, tumors, abnormal blood vessels, and other abnormalities [20]. In developed countries, VCE is being utilized as the first line of treatment for the management of GI conditions. However, the process of developing a differential diagnosis based on the 6-8 hours of video captured is time-consuming. The evaluation of anomalies in a frame-by-frame manner requires expertise and is susceptible to human error. Moreover, due to the movement of the camera along the GI tract, there are challenges such as poor lighting, motion artifacts, blurring, and the possibility of missed abnormalities at the camera's rear side [21]. Hence, automation in VCE can streamline the process by providing standardized and objective analysis. It saves the time of healthcare professionals, enabling faster diagnosis and treatment. Furthermore, it enhances detection accuracy and minimizes the rate of false positives or false negatives [22].

These challenges have catalyzed interest in employing Machine Learning (ML) for the automatic classification, detection, and segmentation of GI bleeding in VCE. ML advancements promise improved accuracy, faster diagnostic processes, and reduced human error. This review paper examines the current ML methodologies for automatic bleeding detection in VCE, assessing their effectiveness, challenges, and future directions. It analyzes 113 papers published between 2008 and 2023, with Figure 1 depicting the year-wise distribution of these studies, highlighting the growing interest and research efforts in this field.

A. Motivation

The literature indicates a rising incidence of GI bleeding and the inadequacies of conventional diagnostic techniques, highlighting the urgent demand for novel solutions that employ advanced methods to increase diagnostic precision, decrease manual effort, and enhance patient outcomes. The following key factors influence the drive to pursue this comprehensive review:

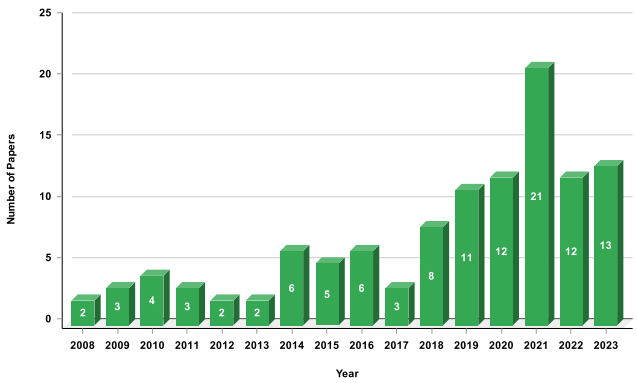


Fig. 1. Year-wise distribution of papers included in the review.

- **Advancements in Deep Learning (DL) and ML**: Recent breakthroughs in DL have demonstrated remarkable success in medical image analysis. The exploration of the application of these advanced methodologies in this domain is thus imperative.
- **Evaluation Parameters**: Diverse evaluation parameters play a critical role in validating the effectiveness of GI bleeding detection methodologies. By exploring these parameters, we aim to identify the most relevant metrics that ensure the reliability and robustness of diagnostic techniques.
- **Exploration of Open-Source Datasets**: The availability of open-source data for GI bleeding offers a foundation for developing and testing AI models. We intend to review these datasets for the development and validation of more accurate, universally applicable ML models.
- **Technique Categorization and Progress Discussion**: A systematic categorization of the techniques used in the detection, classification, and segmentation of GI bleeding is essential. By examining the progress and advancements in each area, we aim to identify trends, gaps, and opportunities for future research in improving GI bleeding diagnosis through ML and image processing technologies.

B. Contribution

This paper provides an exhaustive analysis of AI techniques used to detect GI bleeding in VCE frames. Our contributions are summarized as follows:

1) **GI Bleeding Overview**: We present an overview of GI bleeding, including its causes, symptoms, and the role of VCE in its diagnosis.
2) **Critical Literature Analysis**: Evaluates current research on ML for GI bleeding detection, identifying gaps and future opportunities.

3) **Open Source Datasets Analysis**: Provides the first extensive assessment of open source datasets for abnormal bleeding detection in VCE frames, laying the groundwork for future model development.
4) **Mathematical Performance Metrics**: Details mathematical formulas for diagnostic accuracy and precision metrics in GI bleeding detection.
5) **Technique Examination**: An in-depth examination of current techniques employed in GI bleeding detection, including a distinction between classification, segmentation, detection, and combined approaches.
6) **Future Research Directions**: Suggests potential research avenues to leverage ML advancements for better GI bleeding detection. Identification of gaps in current research, proposing possible future research directions.

To the best of our knowledge, this is the first review to provide such an in-depth analysis explicitly focused on the application of ML for GI bleeding detection in VCE. Through this review, we aim to offer insights that could guide future researchers, provide foundational knowledge, and stimulate advancements in GI healthcare research. Fig. 2 outlines our review structure: Section I introduces the motivation and contribution of the study. Section II details the research methodology, including research questions, existing reviews, and the review approach. Section III discusses the open source datasets for abnormal bleeding detection in VCE frames. Performance evaluation measures are covered in Section IV, divided into statistical, classification, and segmentation metrics. Section V reviews the techniques used in abnormal bleeding detection, including classification, segmentation, and combined approaches. The paper concludes with Section VI, presenting concluding remarks and future research directions.

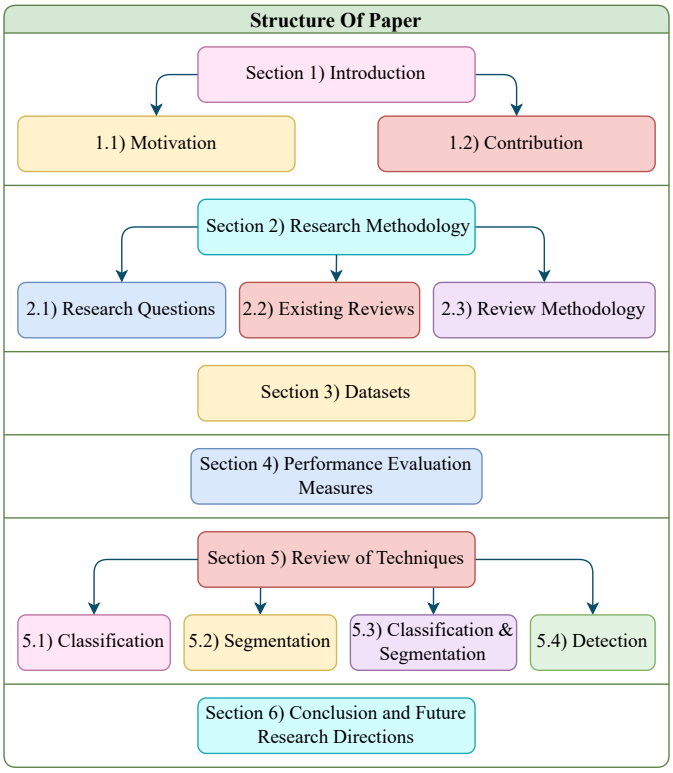


Fig. 2. Flowchart depicting the structure of the review.

II. Research Methodology

This section outlines the methodology for analyzing traditional and ML techniques in detecting GI bleeding in VCE frames, covering research questions, existing reviews, and the selection process for literature review.

A. Research Questions

This review aims to be a definitive guide for researchers exploring the intersection of VCE and ML in bleeding detection. It presents a series of research questions aimed at unpacking the nuances of this specialized domain. These questions not only illuminate key aspects of VCE bleeding detection but also highlight open challenges and future directions for research. Table I outlines these research questions, providing a structured path for emerging researchers to understand the foundational and advanced aspects of this field. Meanwhile, Table II offers a comparative analysis of existing literature, showing how each survey, including ours, addresses these questions. This comparison

TABLE I
Research questions related to GI bleeding detection in VCE frames.

S. No.	Research Questions
Q1	What is video capsule endoscopy, and how is it used in bleeding detection?
Q2	What are the primary types of bleeding detectable via VCE?
Q3	What are the various ML techniques applied in VCE bleeding detection?
Q4	What datasets are most commonly used in ML models for VCE bleeding detection?
Q5	What are the key performance evaluation metrics for ML models in this field?
Q6	What are the challenges in detecting bleeding using ML in VCE?
Q7	How do transparency and interpretability in DL models impact VCE bleeding detection?
Q8	How does ML in VCE bleeding detection compare to traditional image processing and non-AI methods?
Q9	What are the future research directions and unexplored areas in ML for VCE bleeding detection?

highlights the unique and comprehensive nature of our review and its contribution to the field's knowledge base.

B. Existing Reviews

This section details several relevant review articles aimed at summarizing techniques for detecting bleeding in VCE frames. The section also attempts to explain the limitations of these review articles by bench-marking them with our presented review. Extensive reviews have addressed the need for ML algorithms in automatic bleeding classification, detection, and segmentation in VCE in a timely manner. Increasing scientific advancements in image processing methods, ML, transfer learning (TL), and DL in bleeding classification, detection, and segmentation in VCE frames have been observed in [23]–[27].

Brzeski et al. (2013) [23] conducted a comprehensive analysis of existing methods for detecting bleeding in VCE, emphasizing image processing techniques utilized in algorithms. They categorized these techniques into various groups: Color-based, Texture-based, Contour-based, Segmentation, Decision, and Other. The authors concluded that algorithms may need broader techniques to address visual features in endoscopic bleeding. While Brzeski et al.'s analysis provides valuable insights into image processing methodologies, their review lacks coverage of DL techniques specifically applied to bleeding detection in VCE. Additionally, there is no discussion on datasets commonly used in DL models for bleeding detection or the comparative analysis between DL and traditional image processing methods in this context. Du et al. (2019) [25] provided an extensive review of the application of DL methods in the analysis of GI frames. The paper outlined various DL architectures, including Convolutional Neural Networks (CNN), supervised architectures like LeNet, AlexNet, Visual Geometry Group Network (VGGNet), GoogleNet, Residual Network (ResNet), unsupervised architectures such as Generative Adversarial Networks (GAN), and others. The review covered existing works on classification, detection, and segmentation related to GI diseases, including polyps, hemorrhages, GI cancer, and GI disease analysis. The findings highlighted polyps as the most extensively studied GI disease, with GI cancer being the second most explored. The authors found that CNN-based supervised learning networks currently dominate the landscape of deep models in GI frame processing. The authors suggested areas for further exploration, including developing a 3D-CNN-based DL diagnostic system and exploring other DL methods, such as Recurrent Neural Networks (RNNs) and Graph Neural Networks (GNNs). Although Du et al.'s review offers a detailed overview of DL methods in GI frame analysis, it lacks a specific focus on bleeding detection in VCE.

Rahim et al. (2020) [24] conducted a review focusing on computer-aided detection (CAD) methods utilizing VCE frames. The primary objective was to classify these frames as either diseased/abnormal or disease-free/normal, specifically targeting bleeding, tumors, polyps, and ulcers. The survey encompasses a comprehensive analysis of various studies, providing detailed insights into their methodologies, findings, and conclusions. The authors noted that no existing research has addressed the joint classification of all three anomalies. Thus, they proposed an innovative approach for the concurrent classification of tumors, polyps, and ulcers in VCE videos. This proposed method employs a cascaded approach of neural networks, and the paper includes pertinent information on the experimental setup, the dataset used, and the obtained results. While Rahim et al.'s review provides insights into CAD methods in VCE, it lacks coverage of DL techniques and their application, specifically in bleeding detection. Additionally, there is no discussion on the comparative analysis between DL and traditional image processing methods in bleeding detection or future research directions in this field. Soffer et al. (2020) [26] presented the findings of a systematic review of the current literature for DL implementation in VCE. The work provides a succinct overview of the characteristics of articles included in a quantitative meta-analysis, specifically focusing on clinical applications, dataset sizes, and network performance, particularly on ulcers and bleeding. By examining existing research, CNN is concluded to show remarkable performance in abnormality detection, with accuracy consistently exceeding 90% in most studies. Quantitative analysis showed high pooled sensitivities and specificities for the detection of ulcers, bleeding, or the detection of the source of bleeding. Although the systematic review by Soffer et al. offers valuable insight into DL implementation in VCE, it lacks a detailed discussion of bleeding detection techniques, challenges, and future research directions specific to this application. Additionally, there is no coverage of datasets commonly used in DL models for bleeding detection or comparative analysis with traditional image processing methods.

Musha et al. (2023) [27] conducted a systematic review of CAD bleeding algorithms in VCE. The study aimed to identify a taxonomy for these algorithms, discuss prevalent methods, and pinpoint the most effective ones for practical use. The researchers categorized the existing literature into three tasks, i.e., classification, segmentation, and combined classification and segmentation. The combined task proved to be the most effective, with the highest accuracy and specificity. In the review, various color spaces were divided into four groups, i.e., RGB (red, green, blue), HSV (hue, saturation, value), combined (multiple color spaces), and other (YIQ (Luminance (Y), In-phase (I), Quadrature (Q)), YCbCr (Luma (Y), Blue Chroma (Cb), Red Chroma (Cr)), etc.). As per the review above, the selected color space did not offer any performance advantages for the classification or segmentation tasks except RGB, which showed slightly higher recall and lower variance. Combining global and local feature extraction yielded better accuracy. The research findings reveal that for ML, K-Nearest Neighbour (KNN), Support Vector Ma-

TABLE II
COMPARATIVE ANALYSIS OF REVIEW PAPERS ON BLEEDING DETECTION IN VCE FRAMES IN TERMS OF RESEARCH QUESTIONS.

Study	Year	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Brzeski et al.	2013	✓	✗	✓	✗	✗	✓	✗	✗	✓
Du et al.	2019	✓	✗	✓	✗	✗	✓	✗	✓	✓
Rahim et al.	2020	✓	✗	✓	✓	✗	✗	✗	✗	✓
Soffer et al.	2020	✓	✗	✓	✓	✗	✓	✗	✗	✓
Musha et al.	2023	✓	✗	✓	✓	✗	✓	✗	✗	✓

chine (SVM), Multi-Layer Perceptron (MLP), and neural network algorithms were commonly used, while CNNs dominated DL. Although KNN and CNN outperformed other algorithms, KNN faced limitations like overfitting and reliance on hand-crafted features.

In the above review, the chosen color space did not provide any performance benefits for the classification or segmentation tasks. The study highlighted the growing focus on DL algorithms, particularly CNNs, for bleeding detection and emphasized the need to address dataset limitations by using synthetic VCE frames. Challenges include minimizing computational requirements and ensuring visually interpretable outcomes for clinician trust. While Musha et al.'s systematic review provides valuable insights into bleeding detection algorithms in VCE, it lacks coverage of DL techniques and their application specifically in VCE. Additionally, there is no discussion on datasets commonly used in DL models for bleeding detection or comparative analysis with traditional image processing methods. By addressing these gaps, our review offers a comprehensive understanding of bleeding detection in VCE, covering all research questions listed in Table I and providing valuable insights into DL techniques, datasets, challenges, and future research directions in this field.

C. Review Approach

This review on GI bleeding detection follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [28]. PRISMA is selected for its robust framework, which helps outline the processes for content retrieval and assessment to enhance the credibility and thoroughness of the review. To ensure comprehensive coverage, we utilized four academic databases: Google Scholar, PubMed, Researchgate, and IEEE Xplore. Additionally, preprint platforms such as ArXiv, TechRxiv, MedRxiv, and BioRxiv were explored to include the most recent studies. The search terms included "Capsule Endoscopy" AND "Bleeding Detection Techniques", "GI bleeding" AND "Video Capsule Endoscopy", "Deep Learning" AND "WCE Bleeding Detection", "Machine Learning" AND "GI Bleeding Detection", and "Artificial Intelligence" AND "GI Bleeding in Video Capsule Endoscopy". Fig. 3 illustrates the detailed search string used.

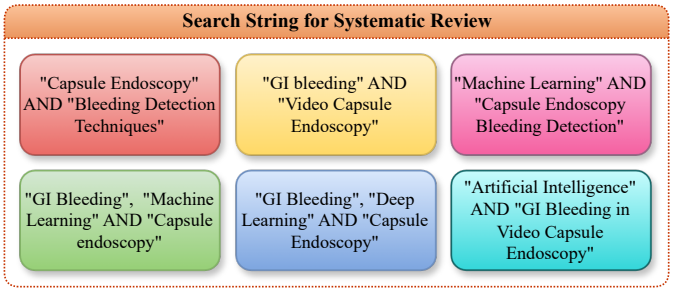


Fig. 3. Search string for searching the research articles.

A rigorous, multi-phase screening process was implemented to evaluate both traditional and ML techniques for VCE bleeding detection. Initially, 500 manuscripts were identified through both electronic databases (N=400) and manual searches (N=100). After removing duplicates, 300 manuscripts remained. After reviewing titles, abstracts, and introductions, these were further narrowed down to 150 manuscripts. Ultimately, 113 manuscripts were selected based on predefined selection criteria. Selected papers ranged from 2008 to 2023, covering various methodologies, datasets, and findings. The selection process ensured that only papers with the most impacted and innovative approaches were included.

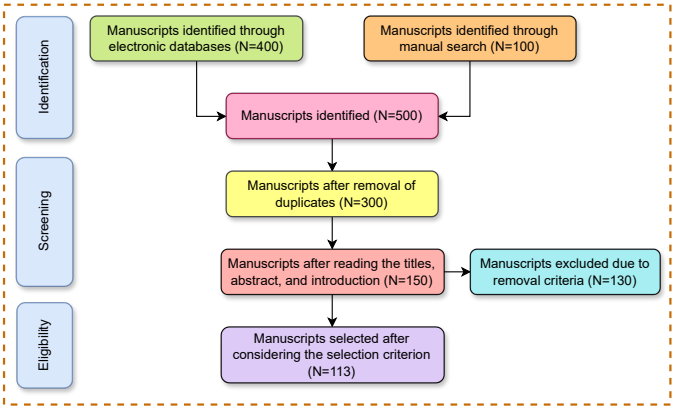


Fig. 4. PRISMA: Outlining the review strategy for bleeding detection in VCE frames.

Fig. 4 illustrates the PRISMA flow diagram, depicting the different phases of literature selection and the number of articles screened, assessed for eligibility, and included in the review.

TABLE III
SELECTION AND REMOVAL CRITERIA FOR THE SYSTEMATIC REVIEW OF VCE GI BLEEDING DETECTION.

S. No.	Parameter	Selection Criteria	Removal Criteria
1	Time Frame	Studies spanning from 2008 to 2023	Studies published before 2008
2	Language	Papers written in English	Papers in languages other than English
3	Study Design	Quantitative studies with empirical data	Qualitative studies or theoretical papers
4	Database Access	Papers accessible through databases like Google Scholar, PubMed, ResearchGate, IEEE Xplore	Papers not accessible through specified databases
5	Relevance	Studies focused on VCE GI bleeding detection	Studies not focused on VCE GI bleeding detection
6	Publication Type	Papers published in journals	Papers published in conferences

The selection and removal criteria for research articles are mentioned in Table III. This methodology ensured a thorough and unbiased collection of literature, offering a clear understanding of advancements in GI bleeding detection techniques.

III. Open Source Datasets for Abnormal Bleeding Detection in VCE frames

In the field of VCE, datasets play a significant role. The development and validation of DL models require a diverse set of datasets with varied characteristics. Researchers in the field of VCE bleeding detection employ various datasets, each with unique features [29]. While some of these datasets are publicly accessible, others are restricted due to privacy and proprietary concerns. Fig. 5 depicts the composition of papers studied in this review that have used publicly available datasets and private datasets. There are also some papers whose information regarding the datasets used could not be found; they are labeled as excluded. Open Source datasets contribute to collaborative advancement in medical research while maintaining transparency [30]. This section is an overview of publicly available datasets that facilitate the recent developments in bleeding detection using VCE.

Bleeding as an abnormality was first introduced in the KID Project [31]. It was developed to advance the medical research support systems (MDSS) for VCE. Clinicians annotated VCE frames and videos. Six centers contributed 2500 annotated VCE frames and 47 videos, which included vascular lesions that covered angioectasias and/or bleeding. KID dataset produced semantic and graphic image annotations supported by an open-access annotation tool. Kvasir-Capsule [32], the largest publicly available PillCAM dataset, contains 4,741,504 frames extracted from 117 videos, out of which 47,238 frames are medically verified, and are labeled with bounding box. It contains 14 different classes. Under the luminal findings section, there are two classes: blood-fresh, for the liquid red appearance of the bleeding in the upper GI tract or small bowel, and blood-hematin, for the cases of minimal bleeding where small black stripes may be observed. But, in both these datasets, bleeding is observed as a coexisting abnormality.

Red Lesion Endoscopy (RLE) Dataset [33] contains 2 subsets of anonymized VCE frames with annotated red

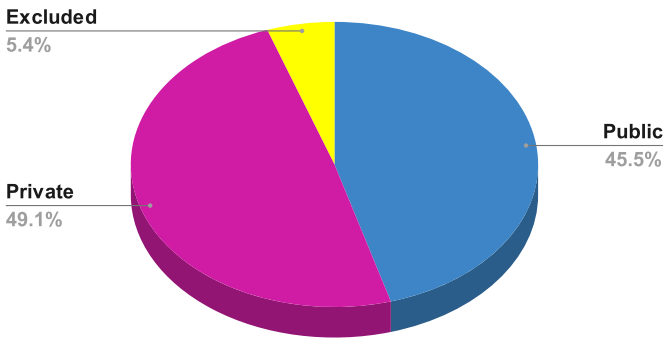


Fig. 5. Distribution of public and private datasets used in the papers.

lesions. The first folder has 3,295 non-sequential frames with different red lesions, such as angioectasias, angiodysplasias, bleeding, and others, which total 1,311 frames. The second folder has 600 sequential frames, out of which 73% of frames have red lesions, each one labeled manually as Blood/Non-Blood. All frames in both folders are 320×320 pixels wide. The Bleeding dataset by Farah Deeba [34] was released in 2016 and contains 50 bleeding frames and 50 non-bleeding frames, along with ground truth frames and annotations. The limitation of this dataset was that its size does not reflect the practical cases of generating huge numbers of frames in VCE. Endoscopy Recommendation System (ERS) Dataset [35] is a multi-label image dataset that contains frames from flexible endoscopy, colonoscopy, and Capsule Endoscopy (CE). The frames are labeled according to Minimal Standard Terminology 3.0 (MST 3.0). This dataset comprises 104 medical labels and 19 ML application labels, divided into 5 categories: healthy, blood, quality, gastro, and colono. WCE-BleedGen Dataset [36] primarily focuses on bleeding and non-bleeding frames. It consists of multiple bleeding frames from internet resources with different types of GI bleeding. It also contains medically validated binary masks and bounding boxes in You Only Look Once (YOLO)-txt, XML, and txt formats.

Auto-WCEBleedGen dataset [36] contains independently collected bleeding and non-bleeding frames of more than 30

patients suffering from acute, obscure, and occult GI bleeding referred to the Department of Gastroenterology and HNU, All India Institute of Medical Sciences (AIIMS), New Delhi, India. It is divided into two subsets—one containing 49 frames from 7 patients with marked and unmarked annotations, and the second has 514 frames from 23 patients with known class labels but no marked annotations. The Gastrointestinal Bleeding WCE images Dataset [37] comprises VCE frames focused on GI bleeding, sourced from Turkey Hospital Muzaffargarh (THM) in Pakistan. Collected between January 2022 and May 2022, it contains 226 GI frames. Among these, 113 frames represent instances of GI bleeding or lesions, while the remaining 113 frames capture VCE visuals from normal patients. Auto-WCEBleedGen and GI Bleeding WCE frames datasets have not been used in any manuscripts, but they focus on GI bleeding as a primary abnormality. Table IV provides a detailed description of these public datasets.

IV. Performance Evaluation Measures

In the realm of GI bleeding detection through traditional and ML techniques, a variety of evaluation metrics are employed to assess the performance and reliability of the models. These metrics are crucial for understanding the efficacy of the models in accurately identifying bleeding instances. This section elaborates on the key metrics used in the literature.

A. Statistical Metrics

1) Mean: The mean, commonly known as the average, is a measure of central tendency in a dataset and is calculated by summing all numerical values and dividing this sum by the count of values [38]. The mean's range can vary widely depending on the dataset. In the context of medical imaging, such as VCE, the mean is used to determine the average pixel intensity across a set of frames, aiding in tasks like normalization or contrast adjustment in the pre-processing stages of model development. It is represented by Eq. 1.

$$\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

2) Variance: Variance is a statistical measurement of the spread between numbers in a dataset, where a high variance indicates a large spread of data points around the mean, and a low variance indicates a tight clustering of data points [38]. In VCE, variance can indicate the diversity in the dataset, such as variations in lighting, texture, and color, which can significantly impact the model's learning process. The variance is calculated using Eq. 2.

$$\text{Variance} = \frac{\sum_{i=1}^{n} (X_i - \text{Mean})^2}{n} \tag{2}$$

3) Kurtosis: Kurtosis is a measure of the "tailedness" of the probability distribution of a dataset [38]. High kurtosis means that the data have heavy tails or outliers; low kurtosis means that the data are light-tailed or lack outliers. It is given by Eq. 3.

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \times \sum_{i=1}^{n} \left(\frac{X_i - \text{Mean}}{SD} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{3}$$

4) Standard Deviation: Standard deviation quantifies the amount of variation or dispersion in a set of data points, with a high standard deviation indicating that the data points are spread over a larger range of values [39]. Its range can be from zero to positive infinity. In the realm of VCE, the standard deviation helps in understanding the variability in image features, which may affect the consistency and reliability of bleeding detection models. The formula for the standard deviation is given by Eq. 4.

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \text{Mean})^2}{n}} \tag{4}$$

B. Classification Metrics

1) Accuracy : Accuracy is a fundamental metric representing the overall correctness of the model. It is the ratio of correctly predicted observations to the total observations [40] and is given by Eq. 5.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

2) Precision : Precision, or positive predictive value, indicates the proportion of positive identifications that were actually correct [40]. It is crucial in medical diagnosis to minimize the burden of unnecessary treatments. The formula for precision is given by Eq. 6.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{6}$$

3) Sensitivity(True Positive Rate, Recall): Sensitivity, or recall, is a measure of the proportion of actual positives correctly identified [41]. In the context of VCE bleeding detection, it represents the model's ability to detect actual bleeding cases. It is defined as Eq. 7.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{7}$$

4) Specificity (True Negative Rate): Specificity, also known as the true negative rate, measures the proportion of actual negatives that are correctly identified. It is critical in medical diagnostics to ensure that healthy cases are not falsely diagnosed as diseased [42]. The mathematical formula for specificity is given by Eq. 8.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{8}$$

TABLE IV
OVERVIEW OF OPEN SOURCE VCE DATASETS FOCUSED ON GI BLEEDING ABNORMALITY UP TO 2023.

Name	Size	No. of bleeding frames	Dimension	CSV*	Augmented frames	Additional Features
KID [31]	335 MB	5	320×320	No	No	Class labels, annotations
Kvasir-Capsule [32]	89 GB	458	336×336	No	No	Class labels, Bounding boxes
RLE Set 1 and Set 2 [33]	506 MB	448	320×320 and 512×512	Yes	No	Annotations
Farah Deeba [34]	1.04 MB	50	568×430	No	No	Annotations
ERS [35]	–	357	–	Yes	Yes	Class Labels, Annotations, masks
WCE-BleedGen [36]	168 MB	1309	224×224	Yes	Yes (version 2)	Class labels, annotations, and bounding box
Auto-WCE BleedGen Dataset [36]	89.1 MB	339	224×224	No	Yes	Class labels, annotations, and bounding box
Gastrointestinal bleeding WCE frames [37]	596 MB	113	variable	No	–	No

5) Area Under the Curve (AUC) - ROC Curve: The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a performance measurement for classification problems at various threshold settings. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity), and the AUC ranges from 0 to 1 [42], with a higher value indicating better model performance. In VCE, the AUC provides a comprehensive measure of the model's ability to distinguish between bleeding and non-bleeding cases across different thresholds.

6) False Positive Ratio (FPR): The False Positive Ratio is the proportion of negative cases that are incorrectly classified as positive [43]. It is a measure of the rate at which false alarms occur. In VCE, a lower FPR is desirable as it indicates fewer healthy patients being incorrectly diagnosed with bleeding, which is essential for patient safety and trust in the diagnostic process.

$$\text{FPR} = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}} \tag{9}$$

7) False Negative Ratio (FNR) : The False Negative Ratio is the proportion of positive cases that are wrongly classified as negative. It is critical in medical diagnostics as it represents missed diagnoses. In VCE bleeding detection [43], a high FNR would mean failing to identify actual bleeding cases, which could have serious implications for patient health.

$$\text{FNR} = \frac{\text{False Negatives}}{\text{True Positives} + \text{False Negatives}} \tag{10}$$

8) F1 Score: The F-Measure, also known as the F1 Score, is a harmonic mean of precision and recall [44]. It is particularly useful in scenarios where an equal balance between precision and recall is desired. The F1 Score is defined as Eq. 11.

$$\text{FPR} = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}} \tag{11}$$

9) Matthews Correlation Coefficient (MCC): The Matthews Correlation Coefficient is a measure of the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even when the classes are of very different sizes. The MCC ranges from -1 to +1 [44], where +1 represents a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation. The MCC is defined as Eq. 12.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{12}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

10) Negative Predictive Value (NPV) : NPV is the ratio of true negatives to total predicted negatives. It measures the model's ability to correctly identify negatives [45]. It ranges from 0 to 1, with higher values indicating better performance. The formula for NPV is given by Eq. 13.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{13}$$

11) Positive Predictive Value (PPV): The Positive Predicted Value, also known as precision, is the proportion of positive test results that are true positives. It ranges from 0

to 1 [46], with higher values indicating better performance. The formula for PPV is given by Eq. 14.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

12) Cohen's Kappa Score: Cohen's Kappa Score is a statistic used to measure inter-rater agreement for qualitative items [47]. It considers the possibility of the agreement occurring by chance. The Kappa Score ranges from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating no agreement better than chance. The formula for Cohen's Kappa is given by Eq. 15.

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{15}$$

Where P_o is the observed agreement, and P_e is the expected agreement by chance. In VCE, Cohen's Kappa can be used to evaluate the consistency of diagnosis between different clinicians or automated systems.

13) Fréchet Inception Distance (FID): FID is used to measure the quality of the generated frames [48]. It calculates the distance between feature vectors calculated for real and generated frames where

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{16}$$

Where μ_r, μ_g are the means and Σ_r, Σ_g are the covariance matrices of real and generated frames, respectively.

C. Segmentation and Object Detection Metrics

1) Jaccard Index: The Jaccard Index, also known as the Intersection over Union (IoU), is a widely used metric in various fields, including medical image analysis, to compare the similarity and diversity of sample sets [49]. In the context of VCE bleeding detection, it measures how well the predicted bleeding areas overlap with the actual bleeding areas. The formula for the Jaccard Index is given by Eq. 17.

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|} \tag{17}$$

In this formula, $|A \cap B|$ represents the intersection of the two sets A and B, which in the context of medical imaging would be the area where both the b predicted bleeding region and the actual bleeding region overlap. $|A \cup B|$ is the union of the two sets A and B, which is the total area covered by both the predicted and actual bleeding regions combined. The Jaccard Index ranges from 0 to 1, where 0 indicates no overlap between the predicted and actual regions (no similarity), and 1 indicates perfect overlap (complete similarity).

2) Dice Score (DS): The Dice Score, also known as the Dice Similarity Coefficient, is a measure of overlap between two samples. This statistic [50] is widely used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. The Dice Score is calculated by Eq. 18.

$$\text{Dice Score (DS)} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{18}$$

where A and B are the predicted and actual segmentation regions, respectively. The Dice Score ranges from 0 (no overlap) to 1 (perfect overlap).

3) Mean Intersection over Union (mIOU): mIOU is an extension of the IoU metric, used primarily in multi-class segmentation tasks [51]. It calculates the IoU for each class and then averages them. This provides a more comprehensive measure for models predicting multiple classes. The formula for mIOU is given by Eq. 19.

$$\text{mIOU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}_i \tag{19}$$

where N is the number of classes, and IoU_i is the IoU for the ith class.

4) Average Precision: Average Precision (AP) is a widely used metric in object detection and segmentation tasks, particularly in datasets where instances are annotated with bounding boxes or segmentation masks [52]. AP summarizes the precision-recall curve as the weighted mean of precision's achieved at each threshold, emphasizing the performance at different recall levels. The range of AP is from 0 to 1, with higher values indicating better model performance.

5) Mean Overlap Coefficient : The Mean Overlap Coefficient is a metric used to assess the similarity and overlap between two datasets or frames [53]. The range of the Mean Overlap Coefficient is between 0 and 1, where 0 indicates no overlap and 1 signifies complete overlap or perfect agreement between the segmented results and the ground truth.

$$\text{Mean Overlap Coefficient} = \frac{1}{N}\sum_{i=1}^{N} \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \tag{20}$$

In this formula, N is the number of observations or samples, A_i and B_i represent the two sets (or segmented regions) for the i-th observation, and $|\cdot|$ denotes the cardinality of the set.

V. Overview of Techniques in Papers Conducting Bleeding Analysis

In this section, a comprehensive review of the techniques employed in recent studies focusing on bleeding detection in VCE frames is undertaken. Bleeding detection in VCE frames is considered a critical task in GI disease diagnosis, with applications ranging from early detection to treatment monitoring. To effectively organize the review, the reviewed articles are categorized into four distinct subsections based on their primary focus: Classification, Segmentation, Classification and Segmentation, and Detection. In the Classification subsection, papers focus solely on

classifying bleeding frames in VCE frames. Segmentation papers concentrate on delineating bleeding regions within the frames. Those falling under Classification and Segmentation combine both tasks, while papers categorized under Detection primarily focus on identifying bleeding regions. The objective is to provide a detailed analysis of the novelty, methodologies, and results achieved by these papers, offering insights into the current state-of-the-art techniques in the field. This information is complemented by chord charts illustrating the relationships between authors and techniques within each subsection, as well as a comprehensive table detailing the techniques employed, datasets utilized, and the corresponding results achieved by each paper, along with the year of publication. Additionally, a taxonomy is included in the Fig. 6, referencing each paper under the technique the authors have implemented in that paper. Through this review, a deeper understanding of the advancements and trends in bleeding detection methodologies within the VCE imaging domain is aimed to be facilitated.

Bleeding classification in VCE frame analysis has evolved from early range ratio color condition methods to using color texture feature extraction techniques in RGB and HSI color spaces. Over time, CNNs gained prominence for feature extraction, while classifiers such as SVM were employed for classification alongside other GI anomalies like polyps, tumors, and ulcers. A recent work has integrated blockchain security with CNN for bleeding classification in VCE frames, marking a significant advancement in medical image analysis. The classification

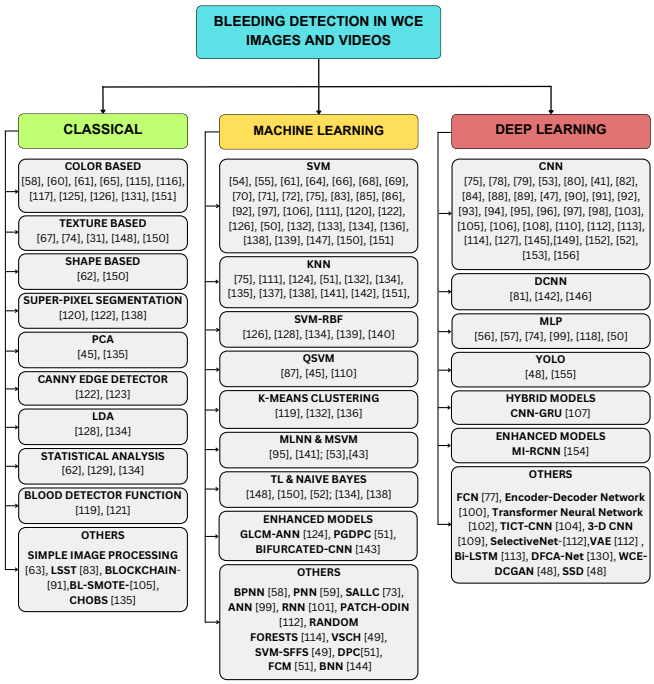


Fig. 6. Classifying Research Methods for Bleeding Detection in VCE frames.

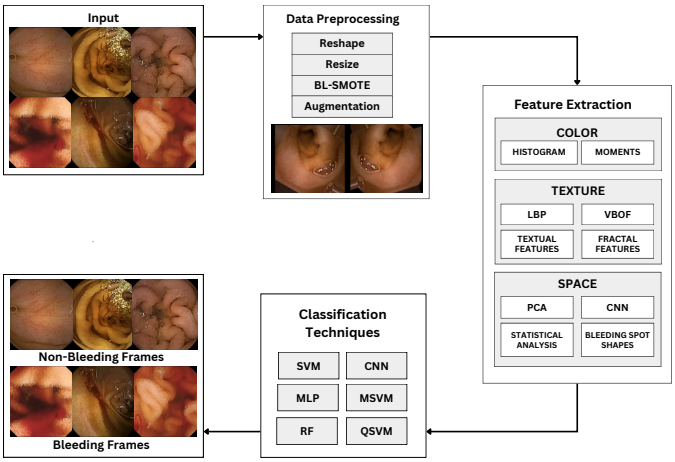


Fig. 7. Pipeline for Classification of Bleeding Frames in VCE frames.

pipeline, as shown in the Fig. 7, involves pre-processing steps such as reshaping, resizing, BorderLine Synthetic Minority Over-sampling Technique (BL-SMOTE), and data augmentation to handle imbalanced datasets. Feature extraction methods encompass color histogram, color moments, texture features (Local Binary Pattern (LBP), Visual Bag of Features (VBOF)), and space feature extraction (Principal Component Analysis (PCA), bleeding spot shapes). Metaheuristic algorithms like Binary Dragonfly Algorithm (BDA), Moth-Flame Optimization (MFO), Particle Swarm Optimization (PSO) and Crow Search Algorithm (CSA) were also employed to optimize feature extraction. Extracted features are then inputted into classifiers (SVM, MLP, Random Forest (RF) etc.) for bleeding classification. Some works employed end-to-end DL architectures like CNNs, handling the entire classification pipeline within the CNN framework.

For the purpose of segmentation of bleeding regions in VCE frames, research has developed from color and texture-based methods to advanced DL algorithms for bleeding region annotation. Some highlights in this domain include the implementation of super-pixel segmentation techniques, dual networks for accurately localizing bleeding regions, and the use of GPU-based parallel programming for reduced computational power. In the domain of detection of bleeding frames, i.e., classification, followed by segmentation of bleeding regions in VCE frames, research began by increasing algorithm proficiency using techniques like saliency maps and classifier fusion algorithms. In recent years, there has been advancement with the implementation of block-based statistics, pixel-based feature extraction, and the introduction of DL algorithms such as VCENet. After reviewing all these techniques, we have presented a generalized segmentation pipeline in Fig. 8 that includes pre-processing of the input image, followed by feature extraction based on color, texture, or spatial

characteristics. Then, pixel-level classification of bleeding regions is applied, after which DL methods such as CNN or MLP are usually employed for accurate segmentation. This is sometimes followed by post-processing with morphological operations to enhance the segmentation results.

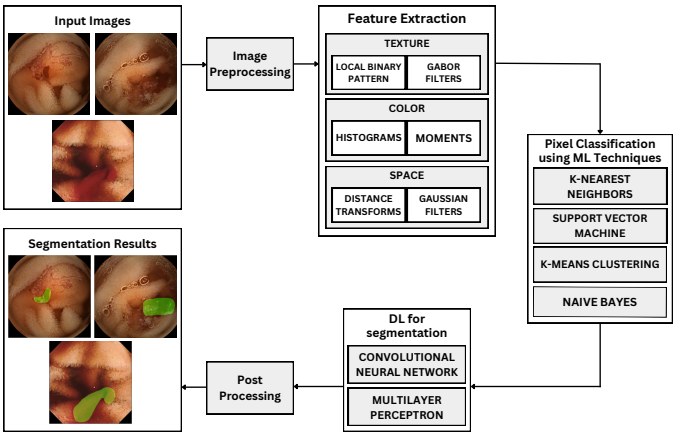


Fig. 8. Pipeline for Segmentation of Bleeding Frames in VCE images.

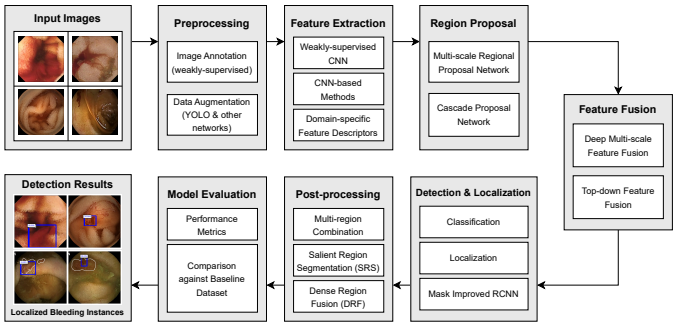


Fig. 9. Pipeline for localization of Bleeding instances in VCE frames.

In the last decade, the application of detection of bleeding regions in VCE frames has witnessed significant advancements with the implementation of CNN and various other DL algorithms in this domain. The latest studies have focused on introducing DL models and improving currently released algorithms to increase the level of visual features extracted from VCE frames for detecting pathologies, including bleeding. One of the articles reviewed in this paper applied the You Only Look Once (YOLO) algorithm for detecting Multiple Bleeding Spots (MBS), while another improved upon a previously applied CNN architecture for detecting multiple pathologies to localize bleeding regions. Typically the detection pipeline includes a sequential process for identifying bleeding anomaly in frames, encompassing preprocessing, feature extraction, region proposal, feature fusion, detection, localization, post-processing, and model evaluation, ultimately outputting classified and localized instances of bleeding. Fig. 9 shows a generalized pipeline for detection method. The key differences lie in the various architectures applied.

A. Classification

Mackiewicz, Fisher, and Jamieson [54] proposed a bleeding detection method in VCE frames using an adaptive color histogram model and SVM. They addressed the challenge of blood color alteration due to GI fluids. It had an accuracy of 97%. Liu and Yuan [55] employed SVM for obscure bleeding detection, using a dataset of 800 VCE frames, manually classified by specialists. This work highlighted the importance of color feature selection in enhancing the performance of SVM classifiers. Li and Meng [56] explored the detection of bleeding and ulcers in VCE frames using a three-layer MLP, focusing on chromaticity moments. This approach used Tchebichef and Krawtchouk discrete moments for feature extraction. Validated with VCE video data from 10 patients, the method achieved a sensitivity of 88.62% and a specificity of 87.81%.

In a subsequent work, Li and Meng [57] introduced the concept of chrominance moment using Tchebichef polynomials and the hue/saturation/intensity (HSI) color space to detect bleeding regions in VCE frames. The method achieved about 90% for accuracy, specificity, and sensitivity. Pan et al. [58] explored the use of backpropagation (BP) neural network for bleeding detection in VCE. They extracted color texture features in both RGB and HSI color spaces, achieving a sensitivity of 93% and a specificity of 96%. Further expanding the use of neural networks, Pan, Yan, Qiu, and Cui [59] employed a Probabilistic Neural Network (PNN) to distinguish between bleeding regions from non-bleeding regions in VCE frames, achieving a sensitivity of 93.1% and a specificity of 85.8%.

Al-Rahayfeh and Abuzneid [60] proposed a method to classify bleeding in VCE frames. Each frame was divided into pixels, and the range ratio colour condition was applied to each pixel. If any pixels meet this condition, the frame is classified as bleeding; otherwise, it is classified as non-bleeding. Tested on 100 VCE frames, it was found that the overall accuracy improved to 98%. Li [61] introduced a novel total variation model (TV-Retinex) to address the Retinex problem, along with employing a SVM for classification of bleeding frames in VCE images, achieving a sensitivity of 96.60% and specificity of 99.50% using the RGB+HSI feature set. Lee and Yoon [62] propose a real-time technique for bleeding detection in VCE frames. The algorithm relies on statistical analysis and bleeding spot shapes. Tested on 30 CE cases the proposed method achieved a sensitivity of 99% and a specificity of 97%.

Lee and Yoon [63] in their work utilized image processing for VCE to address noise and sensitivity issues, employing variance-based filtering and feature-based weighting for blood detection. Improvement was achieved by additional preprocessing techniques in the algorithm for detecting bleeding areas. Abouelenien et al. [64] introduced a cluster-based sampling and ensemble method for

bleeding detection in VCE videos. This method involved feature extraction, data rebalancing, and ensemble classification, leading to improved sensitivity and specificity. Yeh et al. [65] in their paper introduced a new technique for identifying bleeding and ulcers in VCE frames, using color features. The color coherence vector (CCV) features were found to improve the efficiency of VCE image analysis, achieving accuracy (92.86%) and sensitivity (93.64%).

In his paper, Brzeski [66] introduced high-level visual features for bleeding detection in VCE frames. Evaluated on 135 VCE examinations, utilizing an SVM classifier on manually assigned feature values, the efficiency was found to be 95% under perfect feature implementation. Nawarathna et al. [67] proposed a novel multitexture analysis method to effectively distinguish frames with mucosal abnormalities from those without by employing a "texton histogram" of image blocks as features. Experimental results demonstrated a 92% recall and 91.8% specificity on VCE frames. In their work, Hassan and Haque [68] employed SVM for a bleeding detection algorithm that uses textual features extracted from the magnitude spectrum of VCE frames. Their approach achieved an accuracy of 99.19%, sensitivity of 99.41%, and specificity of 98.95%.

Dmitry et al. [69] developed an automated bleeding recognition software, following a consistent three-step process: image reading, feature identification, and result submission to an SVM algorithm. It identifies pathologies and presents them in a review-friendly format, achieving an accuracy of 94%. The dataset used consisted of 110 frames, with 50 being healthy and 60 showing abnormalities. Priya et al. [70] employed SVM along with superpixel segmentation to automatically detect obscure bleeding in VCE frames. By eliminating edge-pixel influence and introducing red ratio features in the RGB color space, the proposed approach showed high improvements in sensitivity and accuracy. Usman et al. [71] introduced a pixel-based approach for detecting bleeding regions in VCE videos using SVM. The findings demonstrate that the method exhibits high sensitivity, specificity, accuracy, and precision in identifying various bleedings in the small colon.

Hu et al. [72] used SVM for bleeding and tumor detection in VCE frames. They introduced a novel geometric image feature called local contrast-enhanced higher-order local auto-correlation (LCE-HLAC) and a pre-processing technique in the HSV color space. The system achieved 98.54% accuracy on a dataset of VCE frames from 28 patient cases. Yuan et al. [73] proposed a computer-aided system for VCE image classification into bleeding, polyp, ulcer, and normal classes. They used color scale-invariant feature transform and K-means clustering for feature extraction and visual word generation. The saliency and adaptive locality-constrained linear coding (SALLC) algorithm was applied for feature encoding. The system achieved a bleeding classification accuracy of 96.60%. Charfi et al. [74] introduced a novel texture extraction scheme for discriminating pathological inflammation, polyp, and bleeding regions in VCE frames. Their approach combines wavelet-based LBP for feature extraction and recognition. The technique, applied with an MLP classifier on VCE frames in RGB color space, achieved a high recognition rate.

Bchir, Ismail, and AlZahrani [75] introduced a technique for automating the detection of MBS in VCE frames, utilizing unsupervised learning for clustering similar frames and supervised learning to distinguish positive (MBS-containing) from negative frames. Their method achieved a 90.92% accuracy rate. Ding et al. [76] employed a CNN-based approach to evaluate small bowel CE (SB-CE) frames, validated on a large multicenter dataset. The model, designed for diagnosing abnormal SB lesions, improved sensitivity from 76.89% to 99.90%, increased lesion detection from 54.57% to 70.91%, and reduced reading time from 96.6 minutes to 5.9 minutes for detecting SB abnormalities using VCE. Diamantis et al. [77] introduced Look-Behind Fully Convolutional Neural Network (LB-FCN), for detecting GI abnormalities like polyps, ulcers, and bleeding. LB-FCN utilizes parallel convolutional layers with different filter sizes for multi-scale feature extraction. While bleeding-only results were not isolated, LB-FCN achieved 99.72% and 93.50% AUC on flexible and VCE datasets, respectively.

Sathiya et al. [78] proposed an automated abnormality detection technique for VCE frames using CNN. They extracted features from image patches to enhance generality and explore optimal color space components for feature extraction and classifier design. Their approach achieves an AUC of approximately 0.8 on a dataset of 137 VCE frames. Aoki et al. [79] utilized a deep CNN trained on 27,847 VCE frames (6,503 blood content, 21,344 normal). Evaluation on an independent test set of 10,208 small-bowel frames (208 blood content, 10,000 normal) yielded an ROC-AUC of 0.9998 and a sensitivity, specificity, and accuracy of 96.63%, 99.96%, and 99.89%, respectively. The CNN-based system outperformed SBI in individual image analysis for detecting blood content in VCE frames. Khan et al. [53] applied TL with a ResNet101 pre-trained CNN model for classification of GI abnormalities, including ulcer, polyp, and bleeding. The extracted deep features were input into a multi-class SVM (MSVM) with a cubic kernel function, achieving a classification accuracy of 99.13%. Though bleeding-only results were not isolated, the datasets included bleeding frames.

Khan et al. [80] proposed StomachNet, an automated method for classifying GI infections, including ulcer, bleeding, polyp, and healthy classes. They utilized ResNet101 with deep TL, enhanced contrast techniques, and meta-heuristic algorithms for optimal feature extraction. Fused feature vectors were input into an Extreme Learning Machine (ELM) classifier, achieving 99.46% accuracy on a combined database. While bleeding-only results were not isolated, the datasets included bleeding frames. Pannu et

al. [41]introduced a CNN-augmented supervised learning ensemble architecture designed to diagnose bleeding symptoms in VCE frames. The proposed method achieved an accuracy of 0.95 in the public RLE data set and 0.93 in the real video data set. Shahril et al. [81] evaluated a deep CNN (DCNN) for bleeding detection in VCE frames. Their pre-processing technique improved accuracy. DCNN outperformed SVM and Fuzzy logic. Specificity, sensitivity, and average accuracy were 0.8703, 0.8271, and 0.8907, respectively.

Jain et al. [82] proposed a novel approach for automatically detecting abnormalities in VCE frames using fractal features. They extract fractal dimensions (FD) using the differential box counting method and employ a RF-based ensemble classifier. The results showed AUC scores of 85% and 99% on the KID dataset and a private dataset, respectively. Kundu et al. [83] proposed a two-stage technique for classifying GI diseases using salient points of interest (POI) extracted via Least Square Saliency Transformation (LSST). A Probability Density Function (PDF) model was fitted on the POI for supervised classification, yielding high performance with a hierarchical cascaded scheme of binary SVMs. Majid et al. [84] proposed a fully automated system for recognizing gastric infections from VCE frames, including ulcer, polyp, esophagitis, and bleeding. The system involved database creation, feature extraction (discrete cosine transform, discrete wavelet transform, strong color feature, and VGG16-based CNN features), feature fusion, robust feature selection via genetic algorithm (GA), and disease recognition using an Ensemble Classifier. The system achieved 96.5% accuracy on a database comprising four datasets, although specific bleeding-only classification results were not isolated.

Ponnusamy et al. [85] proposed a method for classifying GI tract frames from VCE into various classes (Z-Line, Bleeding, Pylorus, Cecum, Esophagitis, Polyps, Ulcerative Colitis) using visual words. They utilized a VBOF approach with center symmetric LBP (CS-LBP), auto-color correlogram (ACC), and scale-invariant feature transform (SIFT) to capture interest points, texture, and color information. Features were clustered into visual words using K-means and classified with SVM. Their method achieved 94.80% accuracy for bleeding detection using the Kvasir dataset. Patel et al. [86] proposed a sparse coding approach for bleeding detection in VCE frames. They utilized SIFT keypoint detector to extract patches around keypoints, reducing patch numbers significantly. SIFT and uniform LBP feature descriptors were computed around keypoints and sparse coded to obtain new feature vectors, then max-pooled for classification with SVM. The method achieved 98.18% accuracy on a dataset of 912 frames. Mamun et al. [87] developed a computer-aided system for bleeding detection in VCE frames using color thresholding and Quantum SVM (QSVM) classification. The system achieved 95.8% accuracy, 95% sensitivity, 97% specificity, and an 85% F1 score.

Mamun et al. [45] proposed a computer-automated system for bleeding detection in VCE frames using fuzzy logic technique and QSVM classifier. It incorporated fuzzy logic edge detection and statistical feature vector extraction with PCA in HSV color space. The proposed approach has been tested on 2393 VCE frames, achieving a sensitivity of 98%, accuracy of 98.2%, specificity of 98%, NPV of 99%, precision of 93%, and F1 Score of 95.4%. In their work, Rustam et al. [88] introduce BIR (Bleedy Image Recognizer), a model combining MobileNet and a custom CNN, which was evaluated on a dataset of 1650 VCE frames, yielding accuracy, precision, recall, F1 score, and Cohen's kappa values of 0.993, 1.000, 0.994, 0.997, and 0.995, respectively. Additionally, when tested on a Google-collected VCE image dataset, BIR had an accuracy of 0.978. In their paper, Saraiva et al. [89] applied CNN to detect blood and hematic residues in VCE frames. The proposed model achieved an accuracy and precision of 98. 5% and 98. 7%, respectively, to detect these elements in the lumen of the small intestine. The sensitivity and specificity were 98.6% and 98.9%, respectively.

Yogapriya et al. [47] employed traditional image processing algorithms, data augmentation, and adjusted pre-trained CNNs (VGG16, ResNet-18, GoogLeNet) to classify GI diseases in VCE frames. While specific results for bleeding-only classification were not isolated, the datasets employed in the work did contain bleeding frames. VGG16 achieved the highest performance with 96.33% accuracy, 96.37% recall, 96.5% precision, 96.5% F1-measure, MCC (0.95), and Cohen's kappa score (0.96). In their paper, Latha et al. [90] employed a CNN-based supervised learning ensemble to detect bleeding in VCE frames. The frames underwent preprocessing, followed by feature extraction using CNN architecture. Hyperparameter tuning involved varying parameters such as optimizers, epochs, and batch sizes. Using the Kvasir Capsule dataset, they achieved 95.7% accuracy with 97.1% sensitivity and 94.6% specificity. Khan et al. [91] proposed a system for detecting stomach abnormalities such as ulcers and bleeding from VCE frames. They integrated blockchain-based security into a CNN model, processed frames with pre-trained deep models, fused features using a Mode value-based approach, optimized with a GA and entropy function, and employed a Softmax classifier for classification. Experimental results on a private dataset showed 96.8% accuracy, including bleeding frames.

Nayyar et al. [92] utilized hybrid contrast enhancement and fine-tuned a pre-trained AlexNet model with TL. Deep features were fused using a vector length-based approach and optimized with a genetic algorithm. Evaluation on 24,000 VCE frames achieved 99% average accuracy using a cubic SVM (CSVM) classifier. The validation included datasets from various sources such as Kvasir and CVC. Modi et al. [93] proposed a system achieving 97.82% accuracy on the Kvasir-capsule dataset for training and predicting 13 classes of digestive tract anomalies. They

preprocessed the dataset for uniform image sizes and used the ImageDataGenerator library for label matching. Their CNN Sequential model effectively detected abnormalities, with the RMSprop optimizer improving accuracy over Adadelta, despite dataset imbalance. Kim et al. [94] employed the Inception-ResNet-V2 model for binary classification to selectively identify inflamed mucosa, atypical vascularity, or bleeding in VCE frames, achieving 98% diagnostic accuracy and an AUC of 0.99 during internal testing. However, external testing showed a 10% accuracy decrease and a 0.07 AUC drop. Although specific results for bleeding-only classification were not isolated, the datasets employed in the work contained bleeding frames. This was the first AI model developed and tested using the Mirocam capsule image-set.

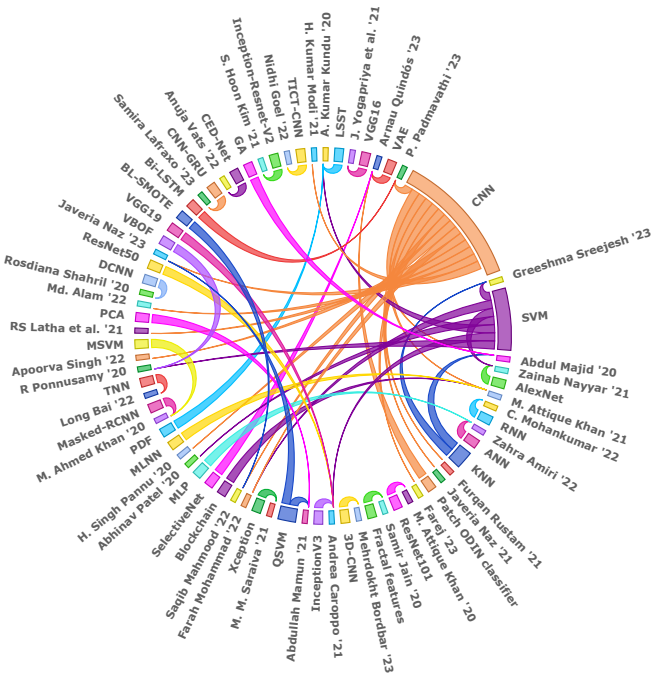


Fig. 10. Authors and Methods: Chord Chart for Bleeding Frame Classification in CE Images.

Khan et al. [95] proposed a method for multiclass classification of stomach diseases using VCE frames. It involved preprocessing, deep TL-based feature extraction with pretrained Inception V3, and feature optimization using PSO and CSA. Classification was done with a Multi-Layered Perceptron Neural Network (MLNN). Evaluation on two datasets (CUI WahStomach Diseases and Combined dataset) showed an average accuracy of 99.5%. While specific results for bleeding-only classification were not isolated, the datasets employed in the work did contain bleeding frames. Naz et al. [96] introduced a hybrid method combining LBP and Segmentation-based Fractal Texture Analysis (SFTA) texture features with pretrained CNN features (VGG16 and InceptionV3). The enhanced image dataset undergoes power-law transformation to improve contrast. The system achieved classification accuracies of 99.25%, 99.75%, and 100% for the KVASIR, Nerthus, and Ulcer-Bleeding datasets, respectively. Caroppo et al. [97] utilized three pre-trained deep CNNs (VGG19, InceptionV3, ResNet50) for feature extraction and fusion was done using the minimum Redundancy Maximum Relevance (mRMR) method. Bleeding/non-bleeding VCE frame classification was done with SVM. Evaluation on KID and MICCAI 2017 datasets yielded average accuracies of 97.65% and 95.70% for bleeding region detection.

In their work, Alam et al. [98] proposed RAt-CapsNet, a CNN architecture designed to classify abnormalities in VCE video data. By leveraging regional information and attention mechanisms, the proposed model achieves a bleeding accuracy of 100% on the Kvasir public dataset. Amiri et al. [99] used an MLP classifier to analyze VCE frames for anomalies like bleeding and angiodysplasia lesions. Their method identifies regions of interest and extracts features using a combination of color histogram analysis and statistical features. It achieved precision and recall rates of approximately 96.5% and 95.9%, respectively. While specific results for bleeding-only classification were not isolated, the datasets used in the work did contain bleeding frames. Vats et al. [100] employed a multichannel encoder-decoder network to learn pathology patterns in VCE frames from discriminative cues like the presence of bleeding indicated by red. Despite using fewer than 2500 labels for training, the approach accurately classified various pathologies in two different VCE datasets, namely Kvasir-capsule and Computer-Assisted Diagnosis for CAPsule Endoscopy database (CAD-CAP), achieving specificity (>97%) for all classes.

Mohankumar et al. [101] employed RNN DL to identify GI abnormalities in VCE frames, achieving 93.27% accuracy. The work included ulcers and bleeding detection, utilizing MLPs for local categorization and experimenting with color spaces for feature extraction. While specific bleeding-only classification results were not isolated, the datasets used contained bleeding frames. Bai et al. [102] employed a transformer neural network (TNN) with spatial pooling to train on two publicly available CE disease datasets for bleeding detection. They achieved 79.15% accuracy on the Kvasir-Capsule multi-classification task and 98.63% bleeding accuracy on the RLE binary classification task. Singh et al. [103] utilized a CNN model to classify bleeding frames in a VCE video dataset, extracted using VLC software. The approach included pre-processing for standardization and denoising. The method achieved a test accuracy of 91.92%, sensitivity of 68.42%, specificity of 97.48%, precision of 86.67%, and F1 score of 0.7647 on private dataset comprising 2,621 frames (505 bleeding, 2,116 normal). Additionally, they proposed an explainable tool for medical experts to detect gastroenterological bleeding more efficiently and reliably.

Goel et al. [104] employed Transforming Input Color

TABLE V
Papers Implementing Bleeding Frame Classification.

Reference	**Year**	**Technique**	**Dataset**	**Accuracy**	**Other Evaluation Metrics**
Mackiewicz et al. [54]	2008	SVM	Frames from 10 VCE videos	97%	-
Liu et al. [55]	2008	SVM	800 VCE frames	-	Sensitivity - 99%, Specificity - 99%
Li et al. [56]	2009	Three-Layer MLP	100 VCE frames	-	Sensitivity - 88.62%, Specificity - 87.81%
Li et al. [57]	2009	Three-Layer MLP	-	90%	Sensitivity - 90%, Specificity - 90%
Pan et al. [58]	2009	RGB/HSI + BP	Train - 20 VCE frames, Test - 150 VCE video	-	Sensitivity - 93%, Specificity - 96%
Pan et al. [59]	2010	PNN	Image level test - 150 VCE video	-	Sensitivity - 93.1%, Specificity - 85.8% on image level
Al-Rahayfeh et al. [60]	2010	Range ratio color	Test - 100 VCE frames	98%	-
Li [61]	2010	TV-Retinex, SVM	1150 frames	-	Sensitivity - 96.6%, Specificity - 99.5%
Lee et al. [62]	2011	Statistical analysis, bleeding spot shapes	30 VCE cases/3000 frames	-	Sensitivity - 99%, Specificity - 97%
Lee et al. [63]	2012	Image processing	30 VCE frames	-	-
Abouelenien et al. [64]	2013	SVM	Frames from 8 annotated VCE videos	-	-
Yeh et al. [65]	2014	Using color features	Capsuleend-oscopy.org, RAPID Atlas	92.86%	Sensitivity - 93.64%
Brzeski [66]	2014	SVM	125 VCE frames	95%	-
Nawarathna et al. [67]	2014	Texton histogram	-	-	Recall - 92%, Specificity - 91.8%
Hassan et al. [68]	2015	SVM	1200 VCE frames	99.19%	Sensitivity - 99.41%, Specificity - 98.95%
Dmitry et al. [69]	2015	SVM	110 VCE frames	94%	-
Priya et al. [70]	2015	SVM	-	-	-
Usman et al. [71]	2016	SVM	8500 VCE frames	-	-
Hu et al. [72]	2016	SVM	28 cases	98.54%	-
Yuan et al. [73]	2016	SALLC algorithm	1650 VCE frames	96.60%	-
Charfi et al. [74]	2018	Wavelet-based LBP, MLP	-	-	-
Bchir et al. [75]	2019	SVM, KNN	1275 frames	90.9%	-
Ding et al. [76]	2019	CNN	113,426,569 frames from 6970 patients	-	Sensitivity - 99.88%

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Diamantis et al. [77]	2019	LB-FCN	MICCAI 2015, KID	-	99.72% AUC
Sathiya et al. [78]	2019	CNN	137 VCE frames	-	AUC - 0.8
Aoki et al. [79]	2019	CNN	Train - 27,847 VCE frames, Test: 10,208 VCE frames	99.89%	ROC-AUC - 0.9998, Sensitivity - 96.63%, Specificity - 99.96%
Khan et al. [53]	2020	CNN, MSVM	Frames from 30 VCE videos	99.13%	-
Khan et al. [80]	2020	CNN	-	99.46%	-
Pannu et al. [41]	2020	CNN	RLE Dataset, Private Dataset	95% on public dataset, 93% on real video dataset	-
Shahril et al. [81]	2020	DCNN	-	89.07%	Specificity - 87.03%, Sensitivity - 82.71%
Jain et al. [82]	2020	Fractal features	KID, Private dataset	-	AUC scores of 85%, 99% on KID, Private dataset
Kundu et al. [83]	2020	LSST, PDF, SVM	2588 VCE frames	-	-
Majid et al. [84]	2020	CNN, GA, KNN	Kvasir-Capsule, CVC-ClinicDB, Private, and ETIS Larib-PolypDB	96.5%	-
Ponnusamy et al. [85]	2020	VBOF, SVM	Kvasir-Capsule	94.80%	-
Patel et al. [86]	2020	SVM	912 VCE frames	98.18%	-
Mamun et al. [87]	2021	QSVM	2393 annotated VCE frames	95.8%	Sensitivity - 95%, Specificity - 97%, Precision - 80%, NPV - 99%, F1 score - 85%
Mamun et al. [45]	2021	QSVM, PCA	2393 annotated VCE frames	98.2%	Sensitivity - 98%, Specificity - 98%, NPV - 99%, Precision - 93%, F1 Score - 95.4%
Rustam et al. [88]	2021	CNN	Frames from 33 patients' VCE videos	99.3%	Precision - 100%, Recall - 99.4%, F1 score - 99.7%, Cohen's kappa - 99.5%
Saraiva et al. [89]	2021	CNN	22,095 VCE frames	98.5%	Precision - 98.7%, Sensitivity - 98.6%, Specificity - 98.9%
Yogapriya et al. [47]	2021	CNN	Kvasir-Capsule	96.33%	Recall - 96.37%, Precision - 96.5%, F1-measure - 96.5%, MCC - 95%, Cohen's kappa score - 96%

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Latha et al. [90]	2021	CNN	Kvasir-Capsule	95.7%	Sensitivity - 97.1%, Specificity - 94.6%
Khan et al. [91]	2021	CNN, Blockchain	Private dataset	96.8%	-
Nayyar et al. [92]	2021	CNN, SVM	24,000 VCE frames from Kvasir-Capsule, CVC, and Private dataset	99%	-
Modi et al. [93]	2021	CNN	Kvasir-Capsule	97.83%	-
Kim et al. [94]	2021	CNN	Mirocam capsule image set	98%	AUC - 0.99
Khan et al. [95]	2021	CNN, MLNN	CUI WahStomach Diseases, Combined dataset	99.5%	-
Naz et al. [96]	2021	CNN	Kvasir-Capsule, Nerthus, and Ulcer-Bleeding dataset	99.3%, 99.9%, 100%	-
Caroppo et al. [97]	2021	CNN, SVM	KID (2352 frames) and MICCAI 2017 (3895 frames)	97.65%, 95.70%	-
Alam et al. [98]	2022	CNN	Kvasir-Capsule	100%	-
Amiri et al. [99]	2022	MLP, ANN	GIANA 2017 dataset, 600 frames	-	Precision - 96.5%, Recall - 95.9%
Vats et al. [100]	2022	Cascaded Encoder-Decoder network	Kvasir-Capsule, CAD-CAP	-	Specificity - 97%
Mohankumar et al. [101]	2022	RNN	Private Dataset	93.27%	-
Bai et al. [102]	2022	TNN	Kvasir-Capsule, RLE	79.15% in Multi and 98.63% in binary classification	-
Singh et al. [103]	2022	CNN	2,621 VCE frames	91.92%	Sensitivity - 68.42%, Specificity - 97.48%, Precision - 86.67%, F1 score - 76.47%
Goel et al. [104]	2022	TICT-CNN	AIIMS Delhi dataset, KID	-	-.
Mahmood et al. [105]	2022	CNN, BL-SMOTE	Kvasir-Capsule	98.9%	AUC - 99.8%, F1-score - 98.9%, Precision - 98.9%, Recall - 98.8% , Loss - 0.0474

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Mohammad et al. [106]	2022	CNN, SVM	-	99.8%	-
Lafraxo et al. [107]	2023	CNN-GRU	MICCAI 2017	-	Precision - 99. 39%
Farej et al. [108]	2023	CNN	Kvasir-Capsule	99.44%	Precision - 99.47%, Recall - 99.412%, F1 score - 99.44%
Bordbar et al. [109]	2023	3D-CNN	Frames from 29 VCE videos	99.20%	Sensitivity - 98.92%, Specificity - 99.50%
Naz et al. [110]	2023	CNN, QSVM	Hybrid dataset, Kvasir-V1 dataset	100%, 99.24%	-
Sreejesh [111]	2023	SVM, KNN	-	95.75%	Sensitivity - 92%, Specificity - 96.5%, AUC - 0.9771
Quindós et al. [112]	2023	Patch ODIN classifier, SelectiveNet, VAE	Kvasir-Capsule	-	AUROC > 0.6
Padmavathi et al. [113]	2023	CNN, Bi-LSTM	-	99.6%	-
Kaur et al. [114]	2023	CNN, RF	WCE Curated Colon Disease Dataset	98.75%	F-Score - 98%

Space in Tandem with CNN (TICT-CNN), a CNN-based framework, for binary VCE frame classification. It conducted on-the-fly data augmentation and color space conversion before training. They analyzed color space performance using objective parameters and feature maps, using AIIMS Delhi and KID datasets. Results favored HSV color space, reducing parameters by 2 to 6 times and achieving a diagnosis time of 0.02 s/frame. Mahmood et al. [105] proposed GIDD-Net, a custom CNN for efficient GI tract disorder detection from VCE frames. They balanced datasets using BL-SMOTE and visualized class activation with Grad-CAM. Evaluation on Kvasir-Capsule yielded 98.9% accuracy, 99.8% AUC, 98.9% F1-score, 98.9% precision, 98.8% recall, and 0.0474 loss. While bleeding-only results weren't isolated, the datasets included bleeding frames. Mohammad and Al-Razgan [106] proposed a method for stomach disease recognition from VCE frames, classifying five classes: polyp, ulcerative colitis, esophagitis, bleeding, and healthy. The approach involved data augmentation, deep TL with Inception v3 and DenseNet-201, feature fusion, optimization with a modified dragonfly optimization method, and final classification using various ML algorithms. CSVM achieved the highest accuracy of 99.8%. While bleeding-only results weren't isolated, the datasets included bleeding frames.

Lafraxo, Ansari, and Koutti [107] presented a computer-assisted technique to identify bleeding from VCE frames, involving preprocessing, optimized CNN feature extraction, and classification using Gated Recurrent Unit (GRU). The method achieved 99.39% precision with the MICCAI 2017 dataset. Farej et al. [108] developed and assessed a CNN model for VCE frame analysis, distinguishing small intestine bleeding from normal cases. Utilizing the Kvasir-Capsule database with 892 frames, the model achieved macro-average accuracy, precision, recall, and F1 scores of 0.99441, 0.99474, 0.99412, and 0.99440 respectively. Bordbar et al. [109] introduced an automatic multiclass classification system based on a 3D-CNN for VCE diagnosis. The proposed 3D model outperformed pre-trained networks using 29 VCE videos (14,691 frames), achieving higher sensitivity (98.92 vs. 98.05), specificity (99.50 vs. 86.94), and accuracy (99.20 vs. 92.60). While specific results for bleeding-only classification were not isolated, the datasets contained bleeding frames.

Naz et al. [110] compare optimization algorithms for classifying GI abnormalities using ensemble Xception Network (XcepNet23) and ResNet18 features. Their work includes dataset augmentation, image preprocessing, feature engineering, and optimization with metaheuristic algorithms (BDA, MFO, PSO). The QSVM achieves 100% accuracy on the Hybrid dataset (five classes) and 99.24% on Kvasir-V1 (eight classes). Sreejesh [111] proposed a novel color feature extraction method to distinguish bleeding

frames from normal ones and localize bleeding regions. K-means clustering characterized color histograms, and SVM and KNN performed classification. Experimental results using the YCbCr color space, 80 clusters, and SVM achieved 95.75% accuracy and 0.9771 AUC. Quindos et al. [112] introduced a patch-based self-supervised approach for analyzing VCE frames. Using a triplet network, they improved out-of-distribution (OOD) detection. Tested on the Kvasir-capsule dataset, the approach achieved AUROC 0.6 for detecting anomalies like lymphangiectasia, foreign bodies, and blood.

Padmavathi et al. [113] introduced a method for VCE bleeding frame detection and classification using a deep neural network. Their approach included a MobileNetV2-based BIR for initial processing, followed by further analysis with a CNN. Additionally, a Bidirectional Long Short-Term Memory (Bi-LSTM) with attention mechanism was employed to boost performance, achieving an accuracy of 0.996 with data augmentation. Kaur and Kumar [114] applied TL and an RF classifier with MobileNetV2 for feature extraction. They addressed challenges like small datasets and overfitting by optimizing hyperparameters through Bayesian optimization. Achieving a disease prediction accuracy of 98.75% and reducing video viewing time, their method obtained an average F-Score of 0.98 using the WCE Curated Colon Disease Dataset. A chord chart describing the techniques used by the authors of these papers has been provided in Fig. 10. It explains that CNN is the most commonly used DL algorithm for classifying bleeding frames in VCE images, and SVM is a frequently implemented ML technique in this domain. Several modifications to SVM and novel CNN algorithms have also been introduced for more efficient and accurate results. The Tables V, V-A, V-A, V-A shows the important details of the papers reviewed in this subsection.

B. Segmentation

Pan Xu et al. [115] proposed a novel method to calculate color similarity vector coefficients using the grey intensity similarity coefficient and chroma similarity coefficient. Upon utilizing this method in the segmentation of bleeding frames with 97% sensitivity and specificity of 90%. Yong-Gyu Lee and Gilwon Yoon [116] studied the weights between the contrast and intensity of the bleeding and non-bleeding VCE frames. To give more contrast to bleeding regions, the red-rated pixels were assigned more weight, and the blood pixels were compared with the neighboring ones. While testing this method, the results showed sensitivity and specificity of 87% and 90%, respectively. Szczyzypinsk et al. [49] explored the differentiating effects of color and texture-based characteristics of VCE frames and proposed a new algorithm for feature selection, Vector Support Convex Hull (VSCH). VSCH gave specificity and sensitivity ratios of 98% each in the detection of bleeding in VCE frames.

Figueiredo et al. [117] introduced four bleeding detector functions (BDFs) based on the a-channel component of the CIE lab color space and enhancement techniques like adaptive anisotropic diffusion on VCE frames. These detectors rely on the eigenvalues of the Hessian and the Laplacian of the modified frames and achieved sensitivity and accuracy rates: 89.3% and 90.4%, for the bleeding detector 1, 92.9% and 91.1%, for detector 2, 92.9% and 92.7%, for detector 3 and 92.9% and 91.1%, for the bleeding detector 4, respectively. In their work, Sainju et al. [118] presented an algorithm to annotate bleeding regions from successive video frames, which reduced the number of features per region for pattern recognition purposes, focusing on the fact that consecutive frames of a VCE have fewer differences. This method achieved accuracy, specificity, and sensitivity of 98.1%, 99.3%, and 100%, respectively. Ramaraj et al. [119] introduced histogram variance-controlled bleeding detectors for the detection of bleeding regions. They contain three modules: determining histogram statistics of CIE-Lab components, defining BDFs based on these statistics, and applying K-means clustering for the segmentation of bleeding regions. The paper showed notable enhancement in detecting and segmenting bleeding regions from VCE frames, supported by ROC curves for binary classifier.

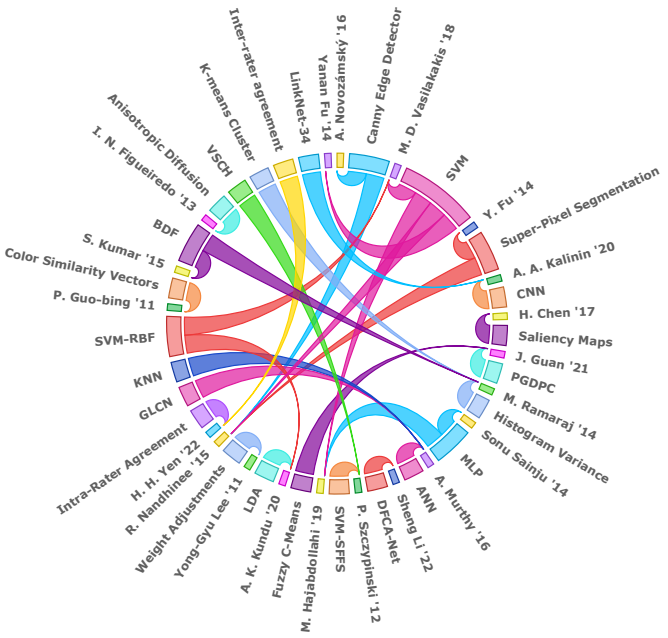


Fig. 11. Chord Chart of Authors and Techniques in Bleeding Frame Segmentation for CE Images.

Yanan Fu et al. [120] introduced a new method of bleeding detection in VCE frames by grouping pixels based on color and location using super-pixel segmentation. Features were extracted using the red ratio in RGB color space, minimizing the influence of edge pixels. These features were fed into SVM for classification, which resulted in a specificity of 99%, sensitivity of 94%, and 95% accuracy for bleeding detection. Kumar et al. [121]

introduced an algorithm that reduced the computational time of segmentation and detection of bleeding by 50 times to 72 frames per second by utilization of GPU-based parallel programming. They parallelized segmentation and blood detector function processes and presented a blood discriminator based on Hessian eigenvalues and color channel analysis. This algorithm showed an accuracy of 89.56%. Nandhinee et al. [122] proposed an algorithm to detect bleeding regions from VCE videos using the super-pixel segmentation process, emphasizing color and spatial similarity. The pixels are grouped based on color and location for classification using SVM classifier. The paper lacks quantitative figures to support the claim to have achieved high rates of accuracy, sensitivity and specificity.

Novozámský et al. [123] proposed two methods for automatic bleeding detection in VCE videos: one based on color information and the other one on characteristics of blood spot, shape and size. Evaluation on private datasets revealed the optimal results when there were five suspected pixels in one frame, yielding true positive values of 96.76% for the first method, 86.73% for the second method, and 85.84% when both methods were used to classify bleeding frames. A. Murthy et al. [124] presented a feature extraction method for detection of blood in VCE frames which gave an accuracy of 97.25%. This method included detection of bleeding by utilizing a word-based histogram, implementation of Gray Level Co-occurrence Matrix (GLCM) using ANN and KNN for classification of bleeding frames. Further, implementation of a two-stage saliency map extraction method to observe bleeding regions under different color components and identify them by the red color. Chen et al. [125] proposed a system for the localization of bleeding regions in VCE frames by computing three saliency maps: an edge saliency map based on phase congruency from Log-Gaber filter bands, another based on intensity histogram, and the last one based on the red proportions of the image. These three maps are fused together and threshold is generated using which the bleeding regions are localized.

Vasilakakis et al. [126] presented a feature extraction method by applying unsupervised color-based saliency detection, DIstaNces On Selective Aggregation of chRomatic image Components (DINOSARC). It includes saliency at point and region-level as well as color information on local and global levels, which resulted in fewer and more significant saliency points for the detection of bleeding regions in VCE frames. The AUC and accuracy achieved with DINOSARC were 81.3% and 80.9%, respectively, with local image descriptors and 81.5% and 81.8%, respectively, with global image descriptors. Hajabdollahi et al. [50] proposed simplification methods in CNN and MLP to reduce the computational complexity of neural networks used in the automatic segmentation of bleeding regions in VCE frames. Upon evaluating their individual performance, it was found that the simplified CNN needed almost half of the parameters while having an AUC of ROC exceeding 97%, and it was better than simplified MLP. However, modified MLP required even smaller computation operations to segment bleeding regions. A.A. Kalinin et al. [127] investigated the use of deep CNNs for the segmentation of angiodysplasia lesions in VCE frames. They integrated ImageNet pre-trained encoders into the U-net architecture for improved performance. It was found that models with ImageNet encoders, i.e., pre-trained VGG-11, VGG-16, and ResNet-34 networks, outperform U-Net trained from scratch, achieving an IOU of 75.35%, DICE score of 84.98%. The paper also presented the model's performance on multi-class instrument segmentation, but we have focused on bleeding only.

Kundu et al. [128] developed a CAD method utilizing linear discriminant analysis (LDA) to separate regions of interest (ROI) and learn the characteristic patterns of diseases from pixel-labeled frames. LDA models were used to extract salient ROIs during both the training and testing phases, with intensity patterns being modeled by a suitable probability distribution. The method achieved a high accuracy, up to 96% for bleeding video clips and 97.39% for bleeding versus normal frames, 96.09% for bleeding versus ulcer frames, and 97.15% for bleeding versus tumor frames using SVM with Radial Basis Function (RBF) kernel for classification, even with a small number of pixel-based labeled frames. In their paper, Guan et al. [51] introduced a fast density peak clustering method (PGDPC) to address the challenge of preservation of local spatial information of pixels during Fuzzy c-means (FCM), reducing time complexity from $O(n^2)$ method to $O(nlogn)$ compared to Density Peak Clustering (DPC). This method enhanced super-pixel technology. It utilizes a peak-graph-based allocation strategy to enhance robustness in reconstructing spatial information for complex-shaped clusters, achieving an average IOU score of 55% on bleeding frames from Kvasir-Capsule datsset.

Yen et al. [129] conducted the first quantitative analysis of bleeding ulcers from endoscopic frames, addressing differences in the interpretation of Forrest Classifications among endoscopists. They collected endoscopic frames of 276 patients with peptic ulcer bleeding and compared intra-rater and inter-rater reliability scores among endoscopists of varying expertise levels. By visually studying pixel-wise differences in Forrest Classification stages and manually delineating ulcer extent and bleeding area, they found distinct patterns for low-risk (Forrest IIC and III) in comparison to high-risk (Forrest I, IIA, or IIB) lesions, with inter-rater agreement ranging from 0.67 to 0.86. Sheng Li et al. [130] introduced pixel-based feature extraction using dual networks, complementing each other to accurately localize bleeding points in VCE frames. These branches were based on an attention mechanism to identify each pixel as bleeding or non-bleeding, and their outputs were fused by a feature fusion module for better segmentation results. This network yielded a mean IoU of 86.85%.

TABLE VI
Papers Implementing Bleeding Frame Segmentation.

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Pan et al. [115]	2011	Color similarity vectors	20 VCE bleeding frames	-	Sensitivity - 90%, Specificity - 97%
Yong-Gyu Lee et al. [116]	2011	Color-based weights adjustment	42 VCE frames	-	Sensitivity - 87%, Specificity - 90%
Szczypinsk et al. [49]	2012	VSCH, SVM-SFFS	50 VCE videos	-	Recall - 0.875, Precision - 0.732, Jaccard Index - 0.48
Figueirido et al. [117]	2013	Anisotropic diffusion	700, 1200, and 2087 VCE frames from 3 datasets	92.7%	Sensitivity - 92.9%
Sainju et al. [118]	2014	Two-layer feed-forward MLP	100 VCE frames	98.1%	Sensitivity - 100%, Specificity - 99.3%
Ramaraj et al. [119]	2014	Histogram variance, BDF, K-means clustering	-	-	-
Yanan Fu et al. [120]	2014	Superpixel segmentation, SVM	5000 VCE frames	95%	Specificity - 94%, Sensitivity - 99%
Kumar et al. [121]	2015	Parallelizing segmentation and BDF	690 VCE frames	89.56%	-
Nandhinee et al. [122]	2015	Canny Edge Detector, Superpixel segmentation, SVM	-	-	-
Novozámský et al. [123]	2016	Canny Edge Detector, Morphological Erosion	12,000 to 20,000 VCE frames from 15 patients	-	TP rate - 96.76%, FP rate - 3.87%
A. Murthy et al. [124]	2016	GLCM using ANN, KNN	-	97.25%	Sensitivity - 93.25%, Specificity - 97.25%
Chen et al. [125]	2017	Saliency maps based on intensity, red-proportions	200 VCE frames	98.97%	-
Vasilakakis et al. [126]	2018	Color descriptor, SVM, SVM-RBF	KID dataset	81.80%	AUC - 81.5%, Specificity - 90.8%, Sensitivity - 68%
Hajabdollahi et al. [50]	2019	SVM, MLP	5 VCE bleeding frames from KID dataset, 783 frames	-	DICE score - 0.869, AUC of ROC - 0.984
A. A. Kalinin et al. [127]	2020	CNN	1200 VCE frames	-	IOU - 75.35, DICE Score - 84.98

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Kundu et al. [128]	2020	LDA, SVM-RBF	50 VCE bleeding frames and 2588 VCE frames	97.39%	Precision - 87.14%, Recall - 85.41%, F1-Score - 86.27%
Guan et al. [51]	2021	FCM, DPC, PGDPC, KNN	Private dataset, Kvasir-Capsule	-	IOU - 0.55
H. H. Yen et al. [129]	2022	Intra-rater and Inter-rater reliability score	276 VCE frames	-	Intra-rater agreement - 0.92-0.97, Inter-rater agreement - 0.67-0.86
Sheng Li et al. [130]	2023	DFCA-Net	-	-	Mean IOU - 86.85%

A brief overview of the manuscripts reviewed in this subsection is provided in the Tables VI, V-B. Also, a chord chart in Fig. 11 is presented to illustrate the methodologies implemented by respective researchers in this domain. This clearly shows that SVM is the most popular ML technique for segmenting bleeding regions in VCE frames.

C. Classification and Segmentation

G.B.Pan et al. [131] presented improved Euler distance for image recognition in CIE color space using the covariance matrix of VCE frames to measure color similarity, emphasizing color difference as a key feature. They constructed a pattern classifier using this improved distance, achieving a sensitivity of 92% and specificity of 95%, proving the algorithm's ability to identify bleeding areas correctly. Yuan et al. [132] introduced a novel approach to differentiate bleeding frames from normal ones and locate bleeding areas using a color feature extraction method and K-means clustering method to derive cluster centers. They achieved 95% accuracy and 97% AUC with YCbCr color space with 80 clusters for SVM classification. For the localization of bleeding regions, they proposed a two-stage saliency map method based on a mixer of different color channels and visual contrast which yielded the best results using a weight of 0.8 for the first stage map and 0.2 for the second stage map, resulting in a localization precision of 95.24%. F. Deeba et al. [133] introduced a novel classifier fusion algorithm for the automatic detection of bleeding regions in VCE frames, combining two SVM classifiers based on RGB and HSV color spaces. Statistical features from first-order histograms of color channels characterize image regions. The classifiers were optimized by nested cross-validation for parameter tuning and feature selection. This approach achieved 95% accuracy, 94% sensitivity, and 95.3% specificity and performed better than single classifiers.

Ghosh et al. [134] proposed a pixel-based feature extraction method using a transform color domain, highlighting the effectiveness of the R/G (red to green) intensity ratio plane in detecting bleeding zones. They found that the G/R composite color domain yielded the best results, achieving an accuracy of 97.96% and a specificity of 97.75%. The following year, Ghosh et al. [135] developed a color histogram of block statistics (CHOBS) to detect bleeding in VCE videos automatically. By employing block-based local feature extraction from the RGB color space and utilizing PCA for feature reduction, CHOBS offers bleeding detection in low-feature dimensionality. This method achieved 97.85% accuracy, 99.47% sensitivity, and 99.15% specificity for bleeding frame detection, and 95.75% precision. Building on their previous study, Ghosh et al. [136] introduced a new approach for automatic bleeding detection in VCE videos. Initially, block-based statistical features are derived from the transformed G/R color plane and input into K-means clustering to obtain two clusters, from which cluster-based features (CBFs) are extracted. When combined with the differential cluster features, they contribute to enhanced accuracy in supervised classification.

In their research, Liaqat et al. [43] proposed several steps for enhancing lesion detection and classification of stomach infections, including bleeding cases using preprocessing and feature extraction from frames. Using a serial-based fusion method, they combined color, shape, and surface characteristics. Additionally, they proposed a multi-class support vector machine (MSVM) feature selection method based on PCA and correlation coefficient, which achieved high accuracy by correctly classifying up to 99% of bleeding frames presented to it. Kundu et al. [137] proposed an automatic bleeding frame and region detection method for VCE based on the intensity variation profile in the normalized RGB space. They extracted features from histograms of the variation of normalized green planes within an efficient ROI and implemented a KNN classifier, achieving high accuracy 97.86%, sensitivity 95.2%, and specificity 98.32% compared to other methods. Sivakumar et al. [138] developed an automatic obscure bleeding detection for VCE using super-pixel segmentation and Naive Bayes Classifier. They also proposed a two-stage saliency extraction method to localize bleeding

mucosa from the normal regions using supervised learning methods, KNN, and SVM. The paper claimed to have improved bleeding detection and localization, but specific results were not stated.

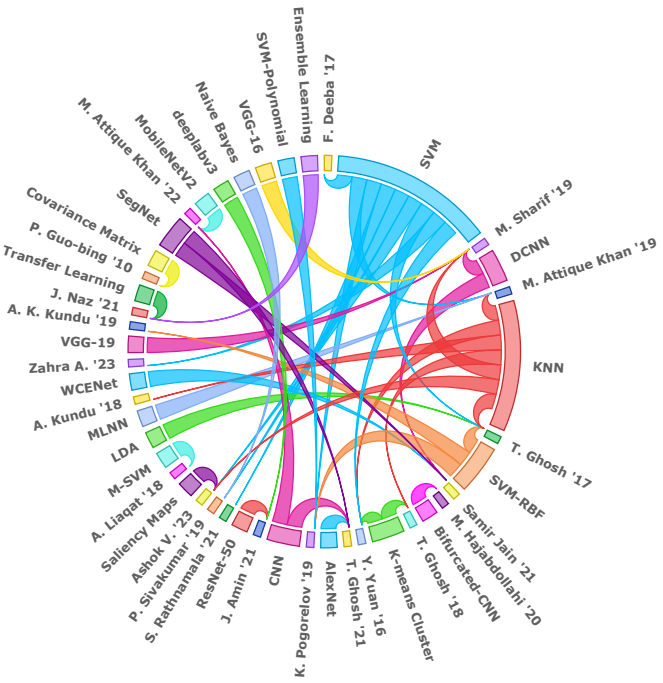


Fig. 12. Authors and Methods: Chord Chart for Bleeding Frame Classification and Segmentation in VCE Images.

Pogorelov et al. [139] introduced an automatic bleeding detection technique in VCE videos that combines texture and color features for initial detection, followed by pixel-level segmentation of bleeding regions. They achieved 97.7% accuracy, 97.6% sensitivity, 95.5% specificity, 0.978 F1-Score, and 0.898 MCC for detection of the bleeding frames. For pixel-level detection of bleeding, this algorithm achieved 97.6% accuracy, 97.6% sensitivity, 95.9% specificity, 0.976 F1-Score, 0.931 MCC, and 0.997 ROC. In their research, Kundu et al. [140] introduced a feature-based bleeding detection method that utilizes a linear separation scheme to identify pixels of interest and extract local spatial features. They fit a characteristic PDF to the resulting feature space, finding that the Rayleigh PDF model works best for bleeding detection in VCE frames. This approach provided reduced computational complexity and consistent class representation and achieved high specificity 96.59%, sensitivity 97.55%, and accuracy 96.77%. Rashid et al. [141] introduced a CAD method to improve the accuracy of the detection of GI disease. Their approach involved HSI color transformation, contour segmentation, and a saliency-based method in YIQ color space. They fused frames and then features using singular value decomposition (SVD), LBP, and GLCM for the classification of bleeding, ulcer, and healthy classes. We have focused on the method's performance in the context of the detection of bleeding, and it is found that the highest accuracy achieved is 99.6% and a sensitivity of 99.6%, FNR of 0.4, which performs better than other established methods.

Sharif et al. [142] proposed a new technique utilizing DCNN and geometric features. They extracted bleeding regions using contrast-enhanced color features, then geometric features, and fused VGG-16 and VGG-19 CNN features based on the Euclidean Fisher Vector. This method identified 100% of the bleeding class frames, thus performing better than previously introduced methods. M. Hajabdollahi et al. [143] addressed the computational challenges of CNN-based classification and segmentation by introducing a method that works simultaneously for both. They designed a structure with separate branches for classification and segmentation, each trained for distinct abnormalities but sharing a common part. Their method achieved a DICE score of 85% for segmentation and 97.8% accuracy for classification purposes of bleeding class frames. In their study, Amin et al. [144] proposed a novel method leveraging a deep semantic segmentation model for 3D segmentation on low-contrast VCE frames with irregularity in shape and size of lesions. Their approach has Labv3 as the backbone of ResNet-50, trained on ground masks to achieve successful pixel-wise classification. To enhance the classification accuracy, they used a pre-trained ResNet-50 model to extract deep features from global input frames, resulting in a global accuracy of 98% and 0.96 of the mean IoU during testing. They also investigated the uncertainty and interpretability during classification through standard thresholding and Bayesian Neural Networks (BNN) techniques, achieving 100% accuracy on privately collected datasets.

Ghosh et al. [145] introduced a CAD tool for the automatic analysis of VCE frames, focusing on the detection of bleeding in the small intestine. They used a pre-trained AlexNet to train a CNN for the detection of bleeding and non-bleeding VCE frames and incorporated SegNet for the delineation of bleeding regions in frames. This model was evaluated on two publicly available datasets, where it achieved a global accuracy of 94.42% and a weighted IoU of 90.69% for detecting bleeding zones. Jain et al. [146] introduced WCENet, a CNN-based model, for the detection and localization of GI anomalies in VCE frames. Initially, an attention-based CNN classifies the frames into four categories, and then, for anomaly localization, a Grad-CAM++ and custom SegNet fusion is employed. The WCENet classifier obtained 98% accuracy, and the segmentation model achieved a frequency-weighted IoU of 81% and an average dice score of 56% on the KID dataset for the bleeding class. S. Rathnamala et al. [147] proposed an automatic bleeding detection system for VCE frames utilizing features from Gaussian mixture model (GMM) superpixels and binary SVM classification followed by segmentation based on deltaE color differences. They achieved higher classification and segmentation accuracy of 99.88% with less computational time.

TABLE VII
Papers Implementing Bleeding Frame Classification and Segmentation.

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
G. B. Pan et al. [131]	2010	covariance matrix used on CIELab colorimetric system	960 VCE frames from 10 VCE videos	-	Sensitivity-92%, Specificity-95%
Yuan et al. [132]	2016	K-means clustering, SVM, KNN	2400 VCE frames	95.75%	Sensitivity - 92%, Specificity - 96.5%
F. Deeba et al. [133]	2017	SVM	videos from PillcamSB1 and PillcamSB2	97.22%	Sensitivity - 98.13%, Specificity - 96.52%
T. Ghosh et al. [134]	2017	Higher and Lower order statistical analysis, SVM-linear, SVM-RBF, KNN, LDA, Naive Bayes	20 VCE videos, 2350 VCE frames	98.47%	Sensitivity 98.66%, Specificity 99.52%
T. Ghosh et al. [135]	2018	CHOBS, KNN, PCA	2350 VCE frames from 32 VCE videos	99.15%	Sensitivity-97.85%, Specificity-99.47%
T. Ghosh et al. [136]	2018	CBF extraction, SVM, K-means clustering	2350 VCE frames from 32 VCE videos	98.04%	Precision-97.05%, FPR-1.11, FNR-22.38, Jaccard Index-75.83%, DICE coefficient-86.26%
Liaqat et al. [43]	2018	MSVM	private dataset	98.3%	Sensitivity-98%, Specificity-98%, Precision-99%, FPR-0.054, AUC-0.99 ,FNR-1.7, G-Measure-98.49%
Kundu et al. [137]	2018	Interplane Intensity Variation Profile, KNN	2300 VCE frames from 30 videos	97.86%	Sensitivity-95.2%, Specificity-98.32%
Sivakumar et al. [138]	2018	Superpixel segmentation, NaiveBayes Classifier, SVM, KNN	-	-	-
Pogorelov et al. [139]	2019	SVM-linear, SVM-polynomial, SVM-RBF	private database	97.7%	Sensitivity-97.7%, Specificity-95.5%
Kundu et al. [140]	2019	SVM-RBF	2350 VCE frames from 2350 VCE videos	96.77%	Sensitivity- 97.55% ,Specificity- 96.59%
Rashid et al. [141]	2019	SVD, LBP, GLCM, MLNN, SVM, KNN	9000 VCE frames	87.09%	-

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Sharif et al. [142]	2019	KNN, DCNN	private dataset with 5500 VCE frames	99.42%	Precision rate-99.51%
Hajabdollahi et al. [143]	2020	Bifurcated CNN	KID and private bleeding dataset	96%	Sensitivity-92.4%, Specificity-97.8%
Amin et al. [144]	2021	BNN, ResNet, deeplabv3	30 VCE videos, each containing 500 frames, CVC–Clinic DB, The Nerthus dataset, Kvasir-Capsule for Segmentation and Classification	98%	-
Ghosh et al. [145]	2021	CNN, AlexNet, SegNet	KID dataset	94.42%	-
Jain et al. [146]	2021	VCENet, SegNet, DCNN	KID dataset	98%	ROC-99%, freq. weighted IOU-81%, avg DICE score-56%
S. Rathnamala et al. [147]	2021	SVM-binary, GMM	Dataset1 contains 3294 VCE frames, and Dataset2 contains 100 VCE frames	99.61%	Specificity- 99.87%, Sensitivity- 98.49%, Precision- 0.99, DICE score 0.99, Jaccard Index- 0.98
Naz et al. [148]	2021	TL, LBP, Ensemble Learning Classifier	Private and Kvasir dataset	99.8%	Jackard Index-96.58%, FNR-6.61
Khan et al. [149]	2022	CNN	Kvasir 1, Kvasir 2, CUI Wah	99.61%	-
Zahara Amiri et al. [150]	2023	TL, LBP, HOG, ResNet, SVM	Kvasir-Capsule, GIANA 2017, RLE, KID	99.6%	FNR-0.008, FPR-0.003%, Precision-0.99, Recall-0.992, F-Measure-0.991, MCC-0.988%
Ashok et al. [151]	2023	Saliency maps based on optical contrast and color channel mixers, SVM, KNN	-	97.75%	-

Naz et al. [148] proposed two procedures for bleeding region segmentation and GI abnormality classification using DL. They applied a hybrid approach for bleeding segmentation, a threshold to RGB channels, and pixel-wise comparison to generate masks. For the classification of bleeding, texture features using LBP and deep features were extracted by applying TL and the best ones were input into the Ensemble Classifier. This method achieved the DICE score of 93.39% and 96.58% in Jack-index for segmentation and accuracy of 99.7% on the private dataset and 86.6% on the Kvasir-Capsule dataset for classification. Khan et al. [149] developed a framework using deep saliency maps and Bayesian optimal DL. They improved image quality through contrast enhancement and implemented pre-trained, fine-tuned MobileNet-V2, initializing hyperparameter using Bayesian Optimization. A hybrid whale optimization algorithm was introduced to select optimal features, followed by classification using an extreme learning classifier. The framework achieved 99.8% accuracy in identifying the bleeding class, showcasing its effectiveness. Zahara Amiri et al. [150] proposed a CAD method for extraction of more distinct image areas in the background as ROI based on color and texture characteristics. They utilized a ResNet-50 pre-trained model to extract deep features and proposed a new method for segmentation for VCE frames using Expectation Maximization (EM) initialization to improve accuracy. A subset of the features generated by this method was selected using the approach of minimum redundancy and maximum relevance and was then classified using SVM, achieving an accuracy of 99.8%, precision of 99.2%, recall of 99.6% and an MCC of 99.3%. Ashok et al. [151] introduced a novel color extraction method to discriminate between bleeding and non-bleeding frames. Their approach involved two steps: analyzing color information in bleeding frames and deriving cluster centers from pixel-based representation, which returned pixels as words. Saliency maps for localization of bleeding region were generated by applying distinct color channel mixers and optical contrast with appropriate thresholding and fusion techniques. This framework achieved 95.75% accuracy for classification purposes.

Fig. 12 is the chord chart depicting the authors in this sub-section and the corresponding techniques implemented by them. A similar observation is noted here that SVM is the most used technique used here, followed by KNN. It should also be noted that more deep-learning techniques have been implemented here than in the previous sub-section. Tables VII and V-C provides details of the papers reviewed in this section, such as the techniques implemented. datasets used and the results achieved.

D. Detection

Iakovidis et al. [152] introduced a weakly-supervised methodology employing a CNN for the automatic detection and localization of GI anomalies, which included bleeding among other conditions. Utilizing image-level

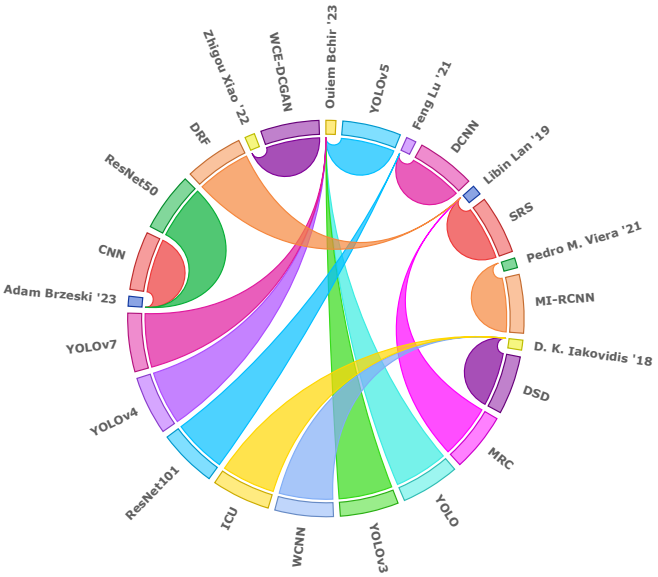


Fig. 13. Chord Chart of Authors and Techniques in Bleeding Frame Detection for VCE images.

annotations, the model classifies frames as normal or abnormal detects salient features via a Deep Saliency Detection algorithm and localizes anomalies using an Iterative Cluster Unification algorithm. While specific results for bleeding-only detection were not isolated, the datasets employed in the study did contain bleeding frames. The approach, tested on datasets with various GI anomalies, demonstrated an AUC of over 80%, with the highest AUC for anomaly detection at 96% in gastroscopy frames and 88% for anomaly localization in VCE frames. In addressing the challenge of identifying abnormal patterns in VCE frames, Libin Lan et al. [52] propose a DL approach utilizing a CascadeProposal network. This network combines a region proposal rejection module for high-recall region proposals and a detection module for abnormal pattern recognition, improved by a multiregional combination (MRC) method, salient region segmentation (SRS), and dense region fusion (DRF) for enhanced localization. Additionally, a negative category and TL strategies are integrated to refine model performance. The methodology achieved a mean average precision (mAP) of 72.3% specifically for detecting bleeding abnormalities within a dataset of over 7,000 annotated frames.

Feng Lu et al. [153] developed an edge analytic framework for detecting and localizing bleeding lesions in VCE images using a deep multi-scale feature fusion network. The approach integrates a multi-scale regional proposal network with a top-down feature fusion, enhancing the resolution and precision of detection. Validation against a clinical trial dataset showed a recall rate and F1 score of 31.69% and 22.12%, respectively, with a prediction accuracy of 98.9% and a 4.88% Average Precision (AP). Vieira et al. [154] proposed a multi-pathology detection

system for VCE frames, utilizing Mask Improved RCNN (MI-RCNN) with mask subnet and training strategy based on second momentum. The paper focused on the detection and localization of various pathologies, such as bleeding, angioectasias, polyps, and inflammatory lesions in VCE frames. Tested on the KID public database, the system demonstrated performance metrics with mAP, mAP50, mAP75, and F1-Score of 40.35, 59.42, 43.01, and 60.07, respectively. Xiao et al. [48] developed a WCE-Deep Convolutional GAN (DCGAN) network to tackle the challenge of small dataset sizes and uneven distribution in VCE image datasets, focusing on pathologies such as Ulcer, Polyp, Blood—Fresh, and Erosion. This network, aimed at enhancing data diversity and volume, was tested across object detection frameworks, including Single-Shot Detector (SSD), YOLOv5, and YOLOv4. Specifically for Blood—Fresh pathology detection, the augmented dataset yielded an accuracy of 99.2% with YOLOv5 and 99.98% with SSD, indicating the network's efficacy in generating diverse, generalizable data for improved object detection performance in VCE frames.

Bchir et al. [155] designed a pattern recognition system using YOLO DL models to detect MBS. The study employed YOLOv3, YOLOv4, YOLOv5, and YOLOv7 to recognize MBS, with YOLOv7 achieving the highest performance metrics, specifically a mAP of 0.86 and an IoU of 0.8. The effectiveness of YOLOv7 was further confirmed by data augmentation, which led to a mAP of 0.883 for bleeding spot detection. The study by Adam Brzeski et al. [156] explored bleeding detection in endoscopic videos through binary classification using a CNN enhanced with high-level visual features. The method introduces domain-specific feature descriptors that output activation maps, augmenting the input data to the CNNs. For bleeding detection, the feature-extended models showed promising performance in comparison to the baseline models. Specifically, the Resnet50 architecture augmented with these features achieved an AUC of 0.963.

In the chord chart provided in Fig. 13, it can be understood that the YOLO algorithm has been implemented by various authors in this domain. It should be noted that novel CNN algorithms have also been introduced for the purpose of the detection of bleeding regions in VCE frames. A review of the papers in this section has been provided in the Table VIII.

VI. Conclusion and Future Research Directions

Our analysis of 113 papers from 2008 to 2023 demonstrates the evolving landscape of research in this domain. However, despite significant advancements, several challenges and opportunities for future research emerge:

1) **Model Generalization**: Future studies should focus on developing generalized models that can robustly detect GI bleeding across diverse datasets, addressing variations in patient demographics and pathological presentations.
2) **Clinical Implementation**: Although ML models have shown promise in research settings, their integration into clinical practice remains limited. Efforts should be directed towards validating these models in clinical trials and understanding the barriers to their practical deployment.
3) **Data Diversity and Availability**: There is a need for larger, more diverse datasets that include a wide range of bleeding instances and patient backgrounds to train more robust and universally applicable models.
4) **Interpretable ML Models**: The black-box nature of many ML models hampers their acceptance in clinical practice. Future research should strive to improve the transparency and interpretability of these models, allowing healthcare professionals to understand and trust the predictions made by ML algorithms.
5) **Cross-disciplinary Collaboration**: Encouraging collaborations between computer scientists, biomedical engineers, and clinicians can lead to the development of more effective and clinically relevant ML solutions.

This review has provided a comprehensive examination of ML techniques for GI bleeding detection in VCE frames, highlighting significant advancements, evaluating the effectiveness of various methodologies, and discussing their challenges and future prospects. The insights drawn from this comprehensive review aim to not only guide future researchers but also contribute foundational knowledge toward the development of more efficient diagnostic methods. By addressing the identified challenges and exploring suggested future directions, the research community can enhance diagnostic methods, ultimately advancing GI healthcare.

Acknowledgment and Declarations

- Funding: Not applicable (NA)
- Competing Interest: The authors declare no personal, academic and financial conflicts of interests associated with this study.
- Availability of meta-data: NA
- Ethical Approval: NA
- Consent to participate: NA
- Consent to publish: NA

The authors declare no personal, academic, or financial conflicts of interest associated with this work. No funding was received for this work.

References

[1] J. Lewis, A. Brown, A. Localio, and J. Schwartz, "Initial evaluation of rectal bleeding in young persons: a cost-effectiveness analysis," *Ann Intern Med*, vol. 136, no. 2, pp. 99–110, 2002.

[2] E. W. Lee and J. M. Laberge, "Differential diagnosis of gastrointestinal bleeding," *Techniques in Vascular and Interventional Radiology*, vol. 7, no. 3, pp. 112–122, 2004.

TABLE VIII
Overview of Papers with Bleeding Region Detection.

Reference	Year	Technique	Dataset	Accuracy	Other Evaluation Metrics
Iakovidis et al. [152]	2018	CNN	Kvasir 1, Kvasir 2, CUI Wah	-	AUC over 80%
Lan et al. [52]	2019	CNN, TL	7000 frames	-	Mean average precision (mAP) - 72.3%
Feng Lu et al. [153]	2021	CNN	20,000 VCE frames	98.9%	AP - 4.8%, Recall - 31.69%, 22.12%
Vieira et al. [154]	2021	Mask Improved-RCNN	KID	-	mAP - 40.35, mAP50 - 59.42, mAP75 - 43.01, F1-Score - 60.07
Xiao et al. [48]	2022	DCGAN, SSD, YOLO	Kvasir-Capsule	99.2% with YOLOv5 and 99.98% with SSD	-
Bchir et al. [155]	2023	YOLO	Kvasir-Capsule	-	mAP - 0.883, IoU - 0.8
Brzeski et al. [156]	2023	CNN	ERS	-	AUC of 0.963

[3] C. M. Wilcox, B. L. Cryer, H. J. Henk, V. Zarotsky, and G. Zlateva, "Mortality associated with gastrointestinal bleeding events: Comparing short-term clinical outcomes of patients hospitalized for upper GI bleeding and acute myocardial infarction in a US managed care setting," *Clinical and Experimental Gastroenterology*, vol. 2, pp. 21–30, 2009.

[4] A. DiGregorio and H. Alvey, "Gastrointestinal bleeding," Treasure Island, FL, 2023, https://www.ncbi.nlm.nih.gov/books/NBK537291/.

[5] B. A. Wuerth and D. C. Rockey, "Changing epidemiology of upper gastrointestinal hemorrhage in the last decade: A nationwide analysis," *Digestive Diseases and Sciences*, vol. 63, no. 5, pp. 1286–1293, 2018.

[6] S. W. Salyer, "Chapter 4 - Acute Gastrointestinal Emergencies," in *Essential Emergency Medicine*, 2007, pp. 114–153.

[7] D. C. Rockey, "Gastrointestinal bleeding," *Gastroenterology Clinics of North America*, vol. 34, no. 4, pp. 581–588, 2005.

[8] E. M. Vreeburg, P. Snel, J. W. de Bruijne, J. F. Bartelsman, E. A. Rauws, and G. N. Tytgat, "Acute upper gastrointestinal bleeding in the amsterdam area: incidence, diagnosis, and clinical outcome," *The American Journal of Gastroenterology*, vol. 92, no. 2, pp. 236–243, 1997, https://pubmed.ncbi.nlm.nih.gov/9040198/.

[9] R. T. Yavorski, R. K. Wong, C. Maydonovitch, L. S. Battin, A. Furnia, and D. E. Amundson, "Analysis of 3,294 cases of upper gastrointestinal bleeding in military medical facilities," *The American Journal of Gastroenterology*, vol. 90, no. 4, pp. 568–573, 1995, https://pubmed.ncbi.nlm.nih.gov/7717312/.

[10] C. M. Wilcox and W. S. Clark, "Causes and outcome of upper and lower gastrointestinal bleeding: the grady hospital experience," *Southern Medical Journal*, vol. 92, no. 1, pp. 44–50, 1999, https://pubmed.ncbi.nlm.nih.gov/9932826/.

[11] H. Hussain, S. Lapin, and M. Cappell, "Clinical scoring systems for determining the prognosis of gastrointestinal bleeding," *Gastroenterology Clinics of North America*, vol. 29, no. 2, pp. 445–464, 2000.

[12] G. Zuccaro Jr, "Management of the adult patient with acute lower gastrointestinal bleeding. american college of gastroenterology. practice parameters committee," *The American Journal of Gastroenterology*, vol. 93, no. 8, pp. 1202–1208, 1998.

[13] T. Rubin, M. Murdoch, and D. Nelson, "Acute gi bleeding in the setting of supratherapeutic international normalized ratio in patients taking warfarin," *Gastrointestinal endoscopy*, vol. 58, pp. 369–73, 09 2003.

[14] L. L. Manning-Dimmitt, S. G. Dimmitt, and G. R. Wilson, "Diagnosis of gastrointestinal bleeding in adults," *Am Fam Physician*, vol. 71, pp. 1339–46, 04 2005, https://pubmed.ncbi.nlm.nih.gov/15832537/.

[15] K. R. Palmer, "Non-variceal upper gastrointestinal haemorrhage: guidelines," *Gut*, vol. 51, no. suppl 4, pp. iv1–iv6, 2002.

[16] K. Bull-Henry and F. H. Al-Kawas, "Evaluation of occult gastrointestinal bleeding," *American family physician*, vol. 87, no. 6, pp. 430–436, 2013, https://pubmed.ncbi.nlm.nih.gov/23547576/.

[17] G. S. Raju, L. Gerson, A. Das, B. Lewis, and A. G. Association, "American gastroenterological association (AGA) institute medical position statement on obscure gastrointestinal bleeding," *Gastroenterology*, vol. 133, no. 5, pp. 1694–1696, 2007.

[18] R. P. Adang, J. F. Vismans, J. L. Talmon, A. Hasman, A. W. Ambergen, and R. W. Stockbrügger, "Appropriateness of indications for diagnostic upper gastrointestinal endoscopy: association with relevant endoscopic disease," *Gastrointestinal endoscopy*, vol. 42, no. 5, pp. 390–397, 1995.

[19] S. L. Triester, J. A. Leighton, G. I. Leontiadis, D. E. Fleischer, A. K. Hara, R. I. Heigh, A. D. Shiff, and V. K. Sharma, "A meta-analysis of the yield of capsule endoscopy compared to other diagnostic modalities in patients with obscure gastrointestinal bleeding," *The American journal of gastroenterology*, vol. 100, no. 11, pp. 2407–2418, 2005.

[20] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, p. 417, 2000.

[21] K. Muhammad, S. Khan, N. Kumar, J. Del Ser, and S. Mirjalili, "Vision-based personalized wireless capsule endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges," *Future Generation Computer Systems*, vol. 113, pp. 266–280, Dec. 2020.

[22] S. Kim and Y. J. Lim, "Artificial intelligence in capsule endoscopy: A practical guide to its past and future challenges," *Diagnostics*, vol. 11, p. 1722, 09 2021.

[23] A. Brzeski, A. Blokus, and J. Cychnerski, "An overview of image analysis techniques in endoscopic bleeding detection," *International Journal of Innovative Research in Computer and*

Communication Engineering, vol. 1, no. 6, pp. 1350–1357, 2013, Link.

[24] T. Rahim, M. A. Usman, and S. Y. Shin, "A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging," *Computerized Medical Imaging and Graphics*, vol. 85, p. 101767, 2020.

[25] W. Du, N. Rao, D. Liu, H. Jiang, C. Luo, Z. Li, T. Gan, and B. Zeng, "Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images," *IEEE Access*, vol. 7, pp. 142 053–142 069, 2019.

[26] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 831–839, 2020.

[27] A. Musha, R. Hasnat, A. A. Mamun, E. P. Ping, and T. Ghosh, "Computer-aided bleeding detection algorithms for capsule endoscopy: A systematic review," *Sensors*, vol. 23, no. 16, p. 7170, 2023.

[28] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *BMJ (Clinical research ed.)*, vol. 372, p. n71, 2021.

[29] T. Singh, P. Handa, and N. Goel, "Automatic gi bleeding detection: A comparative analysis of pre-trained deep learning architectures," in *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, vol. 10, 2023, pp. 260–265.

[30] P. Handa, N. Goel, and S. Indu, "Datasets of wireless capsule endoscopy for ai-enabled techniques," in *Computer Vision and Image Processing*, 2022, pp. 439–446.

[31] A. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johansson, W. Marlicz *et al.*, "Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes," *Endoscopy international open*, vol. 5, no. 06, pp. E477–E483, 2017.

[32] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021.

[33] P. Coelho, A. Pereira, A. Leite, M. Salgado, and A. Cunha, "A deep learning approach for red lesions detection in video capsule endoscopies," in *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, 2018, pp. 553–561.

[34] F. Deeba, F. M. Bui, and K. A. Wahid, "Automated growcut for segmentation of endoscopic images," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 4650–4657.

[35] J. Cychnerski, T. Dziubich, and A. Brzeski, "Ers: a novel comprehensive endoscopy image dataset for machine learning, compliant with the mst 3.0 specification," *arXiv preprint arXiv:2201.08746*, 2022, https://api.semanticscholar.org/CorpusID:246210149.

[36] P. Handa, D. Nautiyal, D. Chhabra *et al.*, "Auto-wcebleedgen version v1 and v2: Challenge, datasets and evaluation," TechRxiv Preprint, Mar. 2024, available at TechRxiv. [Online]. Available: https://doi.org/10.36227/techrxiv.25646332.v1

[37] A. Khan and H. Malik, "Gastrointestinal bleeding wce images dataset," Mendeley Data, V1, 2023.

[38] P. M. Vieira, C. P. Silva, D. Costa, I. F. Vaz, C. Rolanda, and C. S. Lima, "Automatic segmentation and detection of small bowel angioectasias in wce images," *Annals of biomedical engineering*, vol. 47, pp. 1446–1462, 2019.

[39] V. R. Pandit and R. Bhiwani, "Image fusion in remote sensing applications: A review," *International journal of computer applications*, vol. 120, no. 10, pp. 22–32, 2015.

[40] S. Sunitha and S. Sujatha, "An improved bleeding detection method for wireless capsule endoscopy (wce) images based on alexnet," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, 2021, pp. 11–15.

[41] H. S. Pannu, S. Ahuja, N. Dang, S. Soni, and A. K. Malhi, "Deep learning based image classification for intestinal hemorrhage," *Multimedia Tools and Applications*, vol. 79, pp. 21 941–21 966, 2020.

[42] S. Agarwal, "Data mining: Data mining concepts and techniques," in *2013 international conference on machine intelligence and research advancement*, 2013, pp. 203–207.

[43] A. Liaqat, M. A. Khan, J. H. Shah, M. Sharif, M. Yasmin, and S. L. Fernandes, "Automated ulcer and bleeding classification from wce images using multiple features fusion and selection," *Journal of Mechanics in Medicine and Biology*, vol. 18, no. 04, p. 1850038, 2018.

[44] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan 2020.

[45] A. Al Mamun, P. P. Em, T. Ghosh, M. Hossain, M. Hasan, and M. Sadeque, "Bleeding recognition technique in wireless capsule endoscopy images using fuzzy logic and principal component analysis," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2688–2695, 2021.

[46] H. Malik, A. Naeem, A. Sadeghi-Niaraki, R. A. Naqvi, and S.-W. Lee, "Multi-classification deep learning models for detection of ulcerative colitis, polyps, and dyed-lifted polyps using wireless capsule endoscopy images," *Complex & Intelligent Systems*, vol. 10, no. 2, pp. 2477–2497, Apr 2024.

[47] J. Yogapriya, V. Chandran, S. M G, P. Anitha, P. Jenopaul, and C. Dhas, "Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–12, 2021.

[48] Z. Xiao, J. Lu, X. Wang, N. Li, Y. Wang, and N. Zhao, "Wce-dcgan: A data augmentation method based on wireless capsule endoscopy images for gastrointestinal disease detection," *IET Image Processing*, vol. 17, no. 4, pp. 1170–1180, 2023.

[49] P. Szczypiński, A. Klepaczko, M. Pazurek, and P. Daniel, "Texture and color based image segmentation and pathology detection in capsule endoscopy videos," *Computer methods and programs in biomedicine*, vol. 113, no. 1, 2012.

[50] M. Hajabdollahi, R. Esfandiarpoor, P. Khadivi, S. Soroushmehr, N. Karimi, K. Najarian, and S. Samavi, "Segmentation of bleeding regions in wireless capsule endoscopy for detection of informative frames," *Biomedical Signal Processing and Control*, vol. 53, p. 101565, 2019.

[51] J. Guan, S. Li, X. He, and J. Chen, "Peak-graph-based fast density peak clustering for image segmentation," *IEEE Signal Processing Letters*, vol. 28, pp. 897–901, 2021.

[52] L. Lan, C. Ye, C. Wang, and S. Zhou, "Deep convolutional neural networks for wce abnormality detection: Cnn architecture, region proposal and transfer learning," *IEEE Access*, vol. 7, pp. 30 017–30 032, 2019.

[53] M. A. Khan, M. A. Khan, F. Ahmed, M. Mittal, L. M. Goyal, D. J. Hemanth, and S. C. Satapathy, "Gastrointestinal diseases segmentation and classification based on duo-deep architectures," *Pattern Recognition Letters*, vol. 131, pp. 193–204, 2020.

[54] M. W. Mackiewicz, M. Fisher, and C. Jamieson, "Bleeding detection in wireless capsule endoscopy using adaptive colour histogram model and support vector classification," in *Proceedings Volume 6914, Medical Imaging 2008: Image Processing*, vol. 6914, 2008, pp. 69 140R.1–69 140 R.12.

[55] J. Liu and X. Yuan, "Obscure bleeding detection in endoscopy images using support vector machines," *Optimization and Engineering*, vol. 10, pp. 289–299, 06 2009.

[56] B. Li and M. Q.-H. Meng, "Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments," *Computers in biology and medicine*, vol. 39, no. 2, pp. 141–147, 2009.

[57] ——, "Computer-aided detection of bleeding regions for capsule endoscopy images," *IEEE Transactions on Bio-medical Engineering*, vol. 56, no. 4, pp. 1032–1039, 2009.
[58] G. Pan, G. Yan, X. Song, and X. Qiu, "Bp neural network classification for bleeding detection in wireless capsule endoscopy," *Journal of medical engineering & technology*, vol. 33, no. 7, pp. 575–581, 2009.
[59] G. Pan, G. Yan, X. Qiu, and J. Cui, "Bleeding detection in wireless capsule endoscopy based on probabilistic neural network," *Journal of medical systems*, vol. 35, pp. 1477–84, 12 2011.
[60] A. Al-rahayfeh and S. Abuzneid, "Detection of bleeding in wireless capsule endoscopy images using range ratio color," *The International journal of Multimedia & Its Applications*, vol. 2, 05 2010.
[61] M. Li, "Detection of bleeding patterns in wce video using tv-retinex," *Journal of Biomedical Science and Engineering*, vol. 3, no. 12, p. 1143, 2010.
[62] Y.-G. Lee and G. Yoon, "Real-time image analysis of capsule endoscopy for bleeding discrimination in embedded system platform," *International Journal of Biomedical and Biological Engineering*, vol. 5, no. 11, pp. 583–587, 2011.
[63] ——, "Improvement of blood detection accuracy using image processing techniques suitable for capsule endoscopy," *International Journal of Biomedical and Biological Engineering*, vol. 6, no. 5, pp. 168–171, 2012.
[64] M. Abouelenien, X. Yuan, B. Giritharan, J. Liu, and S. Tang, "Cluster-based sampling and ensemble for bleeding detection in capsule endoscopy videos," *American Journal of Science and Engineering*, vol. 2, no. 1, pp. 24–32, 2013, https://api.semanticscholar.org/CorpusID:13841609.
[65] J.-Y. Yeh, T.-H. Wu, and W.-J. Tsai, "Bleeding and ulcer detection using wireless capsule endoscopy images," *Journal of Software Engineering and Applications*, vol. 07, pp. 422–432, 01 2014.
[66] A. Brzeski, "Visual features for endoscopic bleeding detection," *British Journal of Applied Science & Technology*, vol. 4, no. 27, p. 3902, 2014.
[67] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P. Groen, and S.-j. Tang, "Abnormal image detection in endoscopy videos using a filter bank and local binary patterns," *Neurocomputing*, vol. 144, p. 70–91, 11 2014.
[68] A. R. Hassan and M. Haque, "Computer-aided gastrointestinal hemorrhage detection in wireless capsule endoscopy videos," *Computer Methods and Programs in Biomedicine*, vol. 122, pp. 341–353, 09 2015.
[69] M. Dmitry, K. Vladimir, S. Andrey, K. Timur, and K. Alla, "Development of the software for recognition of pathologies in the images obtained by the wireless endoscopic capsule and the atlas of abnormalities," *International Journal of Soft Computing*, vol. 10, no. 4, pp. 279–286, 2015, Link.
[70] K. Priya, K. Archana, and S. Neduncheliyan, "Bleeding detection through wireless capsule endoscopy (wce)," *International Journal of Advanced Computer Technology (IJACT)*, vol. 4, no. 1, pp. 5–13, 2015, Link.
[71] M. A. Usman, G. Satrya, M. R. Usman, and S. Shin, "Detection of small colon bleeding in wireless capsule endoscopy videos," *Computerized Medical Imaging and Graphics*, vol. 54, pp. 16–26, 12 2016.
[72] E. Hu, H. Sakanashi, H. Nosato, E. Takahashi, Y. Suzuki, K. Takeuchi, H. Aoki, and M. Murakawa, "Bleeding and tumor detection for capsule endoscopy images using improved geometric feature," *Journal of Medical and Biological Engineering*, vol. 36, pp. 344–356, 2016.
[73] Y. Yuan, B. Li, and M. Q.-H. Meng, "Wce abnormality detection based on saliency and adaptive locality-constrained linear coding," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 1, pp. 149–159, 2016.
[74] S. Charfi and M. E. Ansari, "Computer-aided diagnosis system for colon abnormalities detection in wireless capsule endoscopy images," *Multimedia Tools and Applications*, vol. 77, pp. 4047–4064, 2018.
[75] O. Bchir, M. M. Ben Ismail, and N. AlZahrani, "Multiple bleeding detection in wireless capsule endoscopy," *Signal, Image and Video Processing*, vol. 13, no. 1, pp. 121–126, 2 2019.
[76] Z. Ding, H. Shi, H. Zhang, L. Meng, M. Fan, C. Han, K. Zhang, F. Ming, X. Xie, H. Liu, W. Liu, R. Lin, and X. Hou, "Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model," *Gastroenterology*, vol. 157, pp. 1044–1054.e5, 06 2019.
[77] D. E. Diamantis, D. K. Iakovidis, and A. Koulaouzidis, "Look-behind fully convolutional neural network for computer-aided endoscopy," *Biomedical signal processing and control*, vol. 49, pp. 192–201, 2019.
[78] R. Sathiya and R. Magal, "A convolutional neural network based disease diagnosis in wireless capsule endoscopy images," *International Journal of Engineering and Advanced Technology*, vol. 9, pp. 327–331, 10 2019.
[79] T. Aoki, A. Yamada, Y. Kato, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara *et al.*, "Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network," *Journal of gastroenterology and hepatology*, vol. 35, no. 7, pp. 1196–1200, 2020.
[80] M. A. Khan, M. S. Sarfraz, M. Alhaisoni, A. A. Albesher, S. Wang, and I. Ashraf, "Stomachnet: Optimal deep learning features fusion for stomach abnormalities classification," *IEEE Access*, vol. 8, pp. 197 969–197 981, 2020.
[81] R. Shahril, A. Saito, A. Shimizu, and S. Baharun, "Bleeding classification of enhanced wireless capsule endoscopy images using deep convolutional neural network," *Journal of Information Science & Engineering*, vol. 36, no. 1, 2020, https://api.semanticscholar.org/CorpusID:84177838.
[82] S. Jain, A. Seal, A. Ojha, O. Krejcar, J. Bureš, I. Tachecí, and A. Yazidi, "Detection of abnormality in wireless capsule endoscopy images using fractal features," *Computers in biology and medicine*, vol. 127, p. 104094, 2020.
[83] A. K. Kundu, S. A. Fattah, and K. A. Wahid, "Least square saliency transformation of capsule endoscopy images for pdf model based multiple gastrointestinal disease classification," *IEEE Access*, vol. 8, pp. 58 509–58 521, 2020.
[84] A. Majid, M. A. Khan, M. Yasmin, A. Rehman, A. Yousafzai, and U. Tariq, "Classification of stomach infections: A paradigm of convolutional neural network along with classical features fusion and selection," *Microscopy research and technique*, vol. 83, no. 5, pp. 562–576, 2020.
[85] R. Ponnusamy, S. Sathiamoorthy, and R. Visalakshi, "An efficient method to classify gi tract images from wce using visual words," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 10, no. 6, 2020.
[86] A. Patel, K. Rani, S. Kumar, I. N. Figueiredo, and P. N. Figueiredo, "Automated bleeding detection in wireless capsule endoscopy images based on sparse coding," *Multimedia Tools and Applications*, vol. 80, pp. 30 353–30 366, 2021.
[87] A. Mamun, M. Hossain, P. P. Em, A. Tahabilder, R. Sultana, and M. Islam, "Small intestine bleeding detection using color threshold and morphological operation in wce images," *International Journal of Electrical and Computer Engineering*, vol. 11, pp. 3040–3048, 2021.
[88] F. Rustam, M. Siddique, H. Sidd, D. S. Ullah, A. Mehmood, I. Ashraf, and G. S. Choi, "Wireless capsule endoscopy bleeding images classification using cnn based model," *IEEE Access*, vol. 09, pp. 33 675–33 688, 2021.
[89] S. M. Mascarenhas, T. Ribeiro, J. Afonso, J. P. Ferreira, H. Cardoso, P. Andrade, M. P. Parente, R. N. Jorge, and G. Macedo, "Artificial intelligence and capsule endoscopy: Automatic detection of small bowel blood content using a convolutional neural network," *GE-Portuguese Journal of Gastroenterology*, vol. 29, no. 5, pp. 331–338, 2022.
[90] R. Latha, G. Sreekanth, G. Murugeasan, S. Aruna, B. Inbaraj, S. Kanivel, and S. Karthikeyan, "Deep learning based automatic detection of intestinal hemorrhage using wireless capsule endoscopy images," *Natural Volatiles & Essential Oils (NVEO)*, pp. 92–103, 2021, https://www.nveo.org/index.php/journal/article/view/329.

[91] M. A. Khan, I. M. Nasir, M. Sharif, M. Alhaisoni, S. Kadry, S. A. C. Bukhari, and Y. Nam, "A blockchain based framework for stomach abnormalities recognition," *Comput. Mater. Contin*, vol. 67, pp. 141–158, 2021.

[92] Z. Nayyar, M. A. Khan, M. Alhussein, M. Nazir, K. Aurangzeb, Y. Nam, S. Kadry, and S. I. Haider, "Gastric tract disease recognition using optimized deep learning features," *Comput. Mater. Contin*, vol. 68, pp. 2041–2056, 2021.

[93] H. K. Modi, S. Misra, and S. Gaur, "Digestive tract abnormalities classification using wireless capsule endoscopy data," *International Journal of Innovative Science and Research Technology*, vol. 6, no. 7, pp. 2456–2165, 2021, https://ijisrt.com/digestive-tract-abnormalities-classification-using-wireless-capsule-endoscopy-data.

[94] S. H. Kim, Y. Hwang, D. J. Oh, J. H. Nam, K. B. Kim, J. Park, H. J. Song, and Y. J. Lim, "Efficacy of a comprehensive binary classification model using a deep convolutional neural network for wireless capsule endoscopy," *Scientific Reports*, vol. 11, no. 1, p. 17479, 2021.

[95] M. A. Khan, A. Majid, N. Hussain, M. Alhaisoni, Y.-D. Zhang, S. Kadry, and Y. Nam, "Multiclass stomach diseases classification using deep learning features optimization," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 3381 – 3399, 2021.

[96] J. Naz, M. Sharif, M. Raza, J. H. Shah, M. Yasmin, S. Kadry, and S. Vimal, "Recognizing gastrointestinal malignancies on wce and cce images by an ensemble of deep and handcrafted features with entropy and pca based features optimization," *Neural Processing Letters*, vol. 55, no. 1, pp. 115–140, 2021.

[97] A. Caroppo, A. Leone, and P. Siciliano, "Deep transfer learning approaches for bleeding detection in endoscopy images," *Computerized Medical Imaging and Graphics*, vol. 88, p. 101852, 2021.

[98] M. Alam, R. B. Rashid, S. A. Fattah, and M. Saquib, "Ratcapsnet: A deep learning network utilizing attention and regional information for abnormality detection in wireless capsule endoscopy," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, 08 2022.

[99] Z. Amiri, H. Hassanpour, and A. Beghdadi, "Feature extraction for abnormality detection in capsule endoscopy images," *Biomedical Signal Processing and Control*, vol. 71, p. 103219, 01 2022.

[100] A. Vats, K. Raja, M. Pedersen, and A. Mohammed, "Multichannel residual cues for fine-grained classification in wireless capsule endoscopy," *IEEE Access*, vol. 10, pp. 91 414 – 91 423, 01 2022.

[101] C. Mohankumar, S. Kumar, S. Senthilkumar, V. Sornagopal, and M. Maharajan, "Rnn model-based classification of wireless capsule endoscopy bleeding images," *International Journal of Health Sciences*, vol. 6, pp. 7330–7344, 05 2022.

[102] L. Bai, L. Wang, T. Chen, Z. Yuanhao, and H. Ren, "Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset," *Electronics*, vol. 11, p. 2747, 08 2022.

[103] A. Singh, H. S. Pannu, and A. Malhi, "Explainable information retrieval using deep learning for medical images," *Computer Science and Information Systems*, vol. 19, no. 1, pp. 277–307, 2022.

[104] N. Goel, S. Kaur, D. Gunjan, and S. Mahapatra, "Investigating the significance of color space for abnormality detection in wireless capsule endoscopy images," *Biomedical Signal Processing and Control*, vol. 75, p. 103624, 2022.

[105] S. Mahmood, M. M. S. Fareed, G. Ahmed, F. Dawood, S. Zikria, A. Mostafa, S. F. Jilani, M. Asad, and M. Aslam, "A robust deep model for classification of peptic ulcer and other digestive tract disorders using endoscopic images," *Biomedicines*, vol. 10, no. 9, p. 2195, 2022.

[106] F. Mohammad and M. Al-Razgan, "Deep feature fusion and optimization-based approach for stomach disease classification," *Sensors*, vol. 22, no. 7, p. 2801, 2022.

[107] S. Lafraxo, M. El Ansari, and L. Koutti, "Computer-aided system for bleeding detection in wce images based on cnn-gru network," *Multimedia Tools and Applications*, vol. 83, pp. 21 081–21 106, 2023.

[108] A. Farhan Sheet, N. Mazin Sheet *et al.*, "On the accuracy of wireless capsule endoscopy images for small intestine bleeding detection using convolutional neural network," *International Journal of Computing and Digital Systems*, vol. 14, no. 1, pp. 1–8, 2023.

[109] M. Bordbar, M. S. Helfroush, H. Danyali, and F. Ejtehadi, "Wireless capsule endoscopy multiclass classification using three-dimensional deep convolutional neural network model," *BioMedical Engineering OnLine*, vol. 22, no. 1, p. 124, 2023.

[110] J. Naz, M. I. Sharif, M. I. Sharif, S. Kadry, H. T. Rauf, and A. E. Ragab, "A comparative analysis of optimization algorithms for gastrointestinal abnormalities recognition and classification based on ensemble xcepnet23 and resnet18 features," *Biomedicines*, vol. 11, no. 6, p. 1723, 2023.

[111] G. Sreejesh, "Bleeding frame and region detection in wireless capsule endoscopy video," *International Journal of Human Computations & Intelligence*, vol. 2, no. 1, pp. 26–33, 2023, https://www.milestoneresearch.in/JOURNALS/index.php/IJHCI/article/view/61.

[112] A. Quindós, P. Laiz, J. Vitrià, and S. Seguí, "Self-supervised out-of-distribution detection in wireless capsule endoscopy images," *Artificial Intelligence in Medicine*, vol. 143, p. 102606, 06 2023.

[113] P. Padmavathi and J. Harikiran, "Wireless capsule endoscopy infected images detection and classification using mobilenetv2-bilstm model," *International Journal of Image and Graphics*, vol. 23, no. 05, p. 2350041, 2023.

[114] P. Kaur and R. Kumar, "Wireless capsule endoscopy video summarization using transfer learning and random forests," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023.

[115] G. Pan, F. Xu, and J. Chen, "A novel algorithm for color similarity measurement and the application for bleeding detection in wce," *International Journal of Image, Graphics and Signal Processing*, vol. 5, pp. 1–7, 2011.

[116] Y.-G. Lee and G. Yoon, "Improvement of blood detection accuracy using image processing techniques suitable for capsule endoscopy," *International Journal of Biomedical and Biological Engineering*, vol. 6, no. 5, pp. 168–171, 2011.

[117] I. N. Figueiredo, S. Kumar, C. Leal, and P. N. Figueiredo, "Computer-assisted bleeding detection in wireless capsule endoscopy images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 1, no. 4, pp. 198–210, 2013.

[118] S. Sainju, F. Bui, and K. Wahid, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing," *Journal of Medical Systems*, vol. 38, pp. 1–11, 2014.

[119] M. Ramaraj, S. Raghavan, V. Raghunath, and W. Khan, "Histogram variance controlled bleeding detectors for wireless capsule endoscopic images," *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 4, pp. 500–510, 2014.

[120] Y. Fu, W. Zhang, M. Mandal, and M. Q.-H. Meng, "Computer-aided bleeding detection in wce video," *IEEE journal of biomedical and health informatics*, vol. 18, no. 2, pp. 636–642, 2014.

[121] S. Kumar, I. N. Figueiredo, C. Graca, and G. Falcao, *A GPU Accelerated Algorithm for Blood Detection in Wireless Capsule Endoscopy Images*. Springer, 2015, pp. 55–71.

[122] R. Nandhinee and D. Yashwanthni, "Detection of bleeding using morphological and segmentation in wce video," *Elysium Journal of Engineering Research and Management*, vol. 2, no. 2, 04 2015, https://journals.indexcopernicus.com/api/file/viewByFileId/372298.

[123] A. Novozámskỳ, J. Flusser, I. Tachecí, L. Sulík, J. Bureš, and O. Krejcar, "Automatic blood detection in capsule endoscopy video," *Journal of biomedical optics*, vol. 21, no. 12, pp. 126 007–126 007, 2016.

[124] A. Murthi and D. Suganya, "Automatic bleeding frame and region detection for glcm using artificial neural network," *Journal of Advances in Chemistry*, vol. 12, no. 24, pp. 5613–5620, 2016, https://core.ac.uk/works/85607934.

[125] H. Chen, S. Wang, Y. Ding, D. Qian *et al.*, "Saliency-based

bleeding localization for wireless capsule endoscopy diagnosis," *International journal of biomedical imaging*, vol. 2017, 2017.

[126] M. Vasilakakis, D. Iakovidis, E. Spyrou, and A. Koulaouzidis, "Dinosarc: Color features based on selective aggregation of chromatic image components for wireless capsule endoscopy," *Computational and Mathematical Methods in Medicine*, vol. 2018, pp. 1–11, 2018.

[127] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets, *Medical Image Segmentation Using Deep Neural Networks with Pre-trained Encoders*. Springer, 2020, pp. 39–52.

[128] A. K. Kundu, S. A. Fattah, and K. A. Wahid, "Multiple linear discriminant models for extracting salient characteristic patterns in capsule endoscopy images for multi-disease detection," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 8, pp. 1–11, 2020.

[129] H.-H. Yen, P.-Y. Wu, T.-L. Wu, S.-P. Huang, Y.-Y. Chen, M.-F. Chen, W.-C. Lin, C.-L. Tsai, and K.-P. Lin, "Forrest classification for bleeding peptic ulcer: A new look at the old endoscopic classification," *Diagnostics*, vol. 12, no. 5, p. 1066, 2022.

[130] S. Li, P. Si, Z. Zhang, J. Zhu, X. He, and N. Zhang, "Dfca-net: Dual feature context aggregation network for bleeding areas segmentation in wireless capsule endoscopy images," *Journal of Medical and Biological Engineering*, vol. 42, no. 2, pp. 179–188, 2022.

[131] G.-b. Pan, G.-z. Yan, X.-s. Song, and X.-l. Qiu, "Bleeding detection from wireless capsule endoscopy images using improved euler distance in cielab," *Journal of Shanghai Jiaotong University (Science)*, vol. 15, no. 2, pp. 218–223, 2010.

[132] Y. Yuan, B. Li, and M. Q.-H. Meng, "Bleeding frame and region detection in the wireless capsule endoscopy video," *IEEE journal of biomedical and health informatics*, vol. 20, no. 2, pp. 624–630, 2015.

[133] F. Deeba, M. Islam, F. Bui, and K. A. Wahid, "Performance assessment of a bleeding detection algorithm for endoscopic video based on classifier fusion method and exhaustive feature selection," *Biomedical Signal Processing and Control*, vol. 40, pp. 415–424, 2018.

[134] T. Ghosh, S. A. Fattah, and K. A. Wahid, "Automatic computer aided bleeding detection scheme for wireless capsule endoscopy (wce) video based on higher and lower order statistical features in a composite color," *Journal of Medical and Biological Engineering*, vol. 38, no. 3, pp. 482–496, 2017.

[135] T. Ghosh, S. A. Fattah, and K. Wahid, "Chobs: Color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–12, 2018.

[136] T. Ghosh, S. A. Fattah, K. A. Wahid, W.-P. Zhu, and M. O. Ahmad, "Cluster based statistical feature extraction method for automatic bleeding detection in wireless capsule endoscopy video," *Computers in Biology and Medicine*, vol. 94, pp. 41–54, 2018.

[137] A. Kundu, S. A. Fattah, and M. N. Rizve, "An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized rgb color space," *Journal of Healthcare Engineering*, vol. 2018, p. 9423062, 2018.

[138] P. Sivakumar and B. M. Kumar, "A novel method to detect bleeding frame and region in wireless capsule endoscopy video," *Cluster Computing*, vol. 22, no. 5, pp. 12 219–12 225, 2019.

[139] K. Pogorelov, S. Suman, F. Azmadi Hussin, A. Saeed Malik, O. Ostroukhova, M. Riegler, P. Halvorsen, S. Hooi Ho, and K.-L. Goh, "Bleeding detection in wireless capsule endoscopy videos — color versus texture features," *Journal of Applied Clinical Medical Physics*, vol. 20, no. 8, pp. 141–154, 2019.

[140] A. Kundu and S. A. Fattah, "Probability density function based modeling of spatial feature variation in capsule endoscopy data for automatic bleeding detection," *Computers in Biology and Medicine*, vol. 115, p. 103478, 2019.

[141] M. A. Khan, M. Rashid, M. Sharif, K. Javed, and T. Akram, "Classification of gastrointestinal diseases of stomach from wce using improved saliency-based method and discriminant features selection," *Multimedia Tools and Applications*, vol. 78, no. 19, p. 27743–27770, 2019.

[142] M. Sharif, M. Khan, M. Rashid, Y. Mussarat, F. Afza, and U. Tanik, "Deep cnn and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 4, pp. 577–599, 2019.

[143] M. Hajabdollahi, R. Esfandiarpoor, E. Sabeti, N. Karimi, S. Soroushmehr, and S. Samavi, "Multiple abnormality detection for automatic medical image diagnosis using bifurcated convolutional neural network," *Biomedical Signal Processing and Control*, vol. 57, p. 101792, 2020.

[144] J. Amin, M. Sharif, E. Gul, and R. S. Nayak, "3d-semantic segmentation and classification of stomach infections using uncertainty aware deep neural networks," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 3041–3057, 2022.

[145] T. Ghosh and J. Chakareski, "Deep transfer learning for automated intestinal bleeding detection in capsule endoscopy imaging," *Journal of Digital Imaging*, vol. 34, no. 2, pp. 404–417, 2021.

[146] S. Jain, A. Seal, A. Ojha, A. Yazidi, J. Bures, I. Tacheci, and O. Krejcar, "A deep cnn model for anomaly detection and localization in wireless capsule endoscopy images," *Computers in Biology and Medicine*, vol. 137, p. 104789, 2021.

[147] S. Rathnamala and S. Jenicka, "Automated bleeding detection in wireless capsule endoscopy images based on color feature extraction from gaussian mixture model superpixels," *Medical & Biological Engineering & Computing*, vol. 59, no. 4, pp. 969–987, 2021.

[148] J. Naz, M. A. Khan, M. Alhaisoni, O.-Y. Song, U. Tariq, and S. Kadry, "Segmentation and classification of stomach abnormalities using deep learning." *Computers, Materials & Continua*, vol. 69, no. 1, pp. 607–625, 2021.

[149] M. A. Khan, N. Sahar, W. Z. Khan, M. Alhaisoni, U. Tariq, M. H. Zayyan, Y. J. Kim, and B. Chang, "Gestronet: A framework of saliency estimation and optimal deep learning features based gastrointestinal diseases detection and classification," *Diagnostics*, vol. 12, no. 11, p. 2718, 2022.

[150] Z. Amiri, H. Hassanpour, and A. Beghdadi, "Combining deep features and hand-crafted features for abnormality detection in wce images," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5837–5870, 2024.

[151] A. Vajravelu, K. T. Selvan, M. M. B. A. Jamil, A. Jude, and I. d. l. T. Diez, "Machine learning techniques to detect bleeding frame and area in wireless capsule endoscopy video," *Journal of Intelligent & Fuzzy Systems*, vol. 44, pp. 353–364, 2022.

[152] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos, "Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification," *IEEE Transactions on Medical Imaging*, vol. 37, no. 10, pp. 2196–2210, 2018.

[153] F. Lu, W. Li, S. Lin, C. Peng, Z. Wang, B. Qian, R. Ranjan, H. Jin, and A. Y. Zomaya, "Multi-scale features fusion for the detection of tiny bleeding in wireless capsule endoscopy images," *ACM Transactions on Internet of Things*, vol. 3, no. 1, pp. 1–19, 2022.

[154] P. Vieira, N. Freitas, V. Brandão Lima, D. Costa, C. Rolanda, and C. Lima, "Multi-pathology detection and lesion localization in wce videos by using the instance segmentation approach," *Artificial Intelligence in Medicine*, vol. 119, p. 102141, 08 2021.

[155] O. Bchir, G. A. Alkhudhair, L. S. Alotaibi, N. A. Almhizea, S. M. Almuhanna, and S. F. Alzeer, "Deep learning-based multiple bleeding detection in wireless capsule endoscopy," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023.

[156] A. Brzeski, T. Dziubich, and H. Krawczyk, "Visual features for improving endoscopic bleeding detection using convolutional neural networks," *Sensors*, vol. 23, no. 24, p. 9717, 2023.

BERT-based RNN for Effective Detection of Depression with Severity Levels from Text Data

Pavi Saraswat
Department of Computer Science and Engineering
Delhi Technological University,Delhi, India
pavisaraswat@gmail.com

Rohit Beniwal
Department of Computer Science and Engineering
Delhi Technological University, Delhi, India
dr.rohitbeniwal@gmail.com

***Abstract*—Due to a lack of early and accurate treatment, a large number of populations are suffering from depression. The posts available on online digital platforms play major role in providing useful signs concerning mind state of a user. The traditional depression methods involve a large amount of time, budget, and expertise, thus, making depression detection a complicated task. Therefore, there is a critical need to propose a computer-assisted system to analyze and detect depression efficiently at different severity levels. Focusing on such concerns, this study performed multi-class classification utilizing a publicly available text dataset to categorize users into non-depressed, moderately, and severely depressed. This study presented a hybrid technique that robustly combines Bidirectional Encoder Representations from Transformers (BERT) with two versions of Recurrent Neural Network (RNN) i.e. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The BERT model is implemented to extract useful word embeddings which are then fed to RNN comprising LSTM, dropout, and GRU layers. The findings revealed that the employed approach outperformed all other employed models and techniques in the past studies by attaining an accuracy of 95.0%. Thus, the presented strategy is expected to be beneficial for clinicians to perform depression detection at different stages accurately.**

Keywords*—Depression Detection, BERT, LSTM, GRU, RNN, social media text.

I. INTRODUCTION

Depression, which impacts millions of people across the globe, ranks as one of the forms of mental illness [1]. Stress, low moods, nervousness, insufficient sleep, problems with eating, regret sensations, thoughts about suicide and attempts, along with other indications are among the most frequently experienced ones caused by depression [2]. More than 350 million individuals worldwide i.e. 4.4% of the global population are reported to suffer from depressive disorders, based on data from the World Health Organization (WHO) [3]. Further, by 2030, it is projected that depression will appear as the second leading reason for disability globally [4]. Depression, being a negative disorder, affects people at several severity levels, including early, moderate, and severe. Identifying it early may prevent serious adaptation problems for the concerned person [5].

Several rating scales as well as questionnaires are employed in the traditional methods used in clinical settings for determining the severity of depression. These theoretical tools demand considerable amounts of time and primarily depend on input from patients and medical guidance from experts. A faulty and unreliable assessment may result from inadequate questionnaire filling and a lack of understanding of the subject matter. Thus, researchers are drawn to utilize automated cutting-edge techniques for facilitating accurate depression analysis [6]. Over the recent years, Machine learning (ML) and Deep Learning (DL) have made immense progress in the detection of affected mental health by providing standardized evaluation and automated analysis that could be helpful to lighten a certain amount of stress for clinicians [7]. Various social media online platforms including Twitter, Facebook, Instagram, etc. have become the prime source where many people share their emotions, thoughts, and feelings on a regular basis. Thus, a large amount of such data can be carefully analyzed to understand the behavior of a person to a large extent [8]. Focusing on text data available on social platforms, this study proposed a hybrid model incorporating transfer learning (TL) i.e. BERT and DL i.e. RNN models to detect and classify depression at three severity levels. Using a combination approach, rather than a single model, may improve the model's overall performance.

The organization of this paper is well structured as follows. Section 2 presents related studies on text-based depression detection are presented in. The proposed framework is given in Section 3 along with each stage of the process to follow. The results obtained are shown in Section 4. Section 5 includes an examination of the comparison of the employed and proposed approach with different techniques. This paper concludes at last in Section 6.

II. RELATED WORK

The scrutinization of past research works revealed the efforts made by researchers to predict depression using different modalities and learning platforms. Various types of data on social media are considered to differentiate depressive and normal subjects including sensor signals [9], facial cues [10], speech attributes [11], images [12], and many more. A lot of research work has seemed to be focused on text modality due to the more public availability of robust and authentic datasets. Research studies by [13,14] utilized text data to measure the distress state of the subject.

Further, several automated platforms such as ML and DL have been employed in past studies to detect depression accurately. Choudhury et al. [15] utilized a Support Vector Machine (SVM) to analyze depression episodes up to 12 months in advance. Similarly, employing SVM on multi-linguistic attributes, Hassan et al. [16] yielded an F-score of 0.81 in the detection of depression. A study by Islam et al. [17] achieved promising results using k-nearest Neighbor (KNN) for depression analysis. Reece et al. [18] employed Random Forest (RF) and determined the depression at an early stage based on Twitter posts.

The application of BERT and Convolutional Neural Network (CNN) are implemented in a work by Lin et al. [19] for depression detection based on Twitter data. Figueredo et al. [20] proposed a CNN model using GloVe and Fast Text embeddings. An early and late fusion approach is applied which achieved satisfactory results.

Cong et al. [21] used a combination of XgBoost and BiLSTM on the Reddit dataset to analyze depressed persons. A study by Poswiata et al. [22] presented an approach to

2024 IEEE Symposium on Wireless Technology & Applications (ISWTA) | 979-8-3503-5135-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISWTA62130.2024.10651873

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:18:37 UTC from IEEE Xplore. Restrictions apply.

diagnose various levels of depression by applying an ensemble of pre-trained transfer models.

Asad et al. [23] focused on severity levels, considering four subclasses named normal, slight intermediate, moderate, and severe or serious to conduct an in-depth investigation of depression. Depression is categorized as non-depressed at a level of 1 to 55% and vice-versa at a level of 55% or more.

Therefore, it is analyzed that combining a robust pre-trained TL with a DL model can provide a useful solution for depression detection at different stages with high accuracy.

III. Proposed Methodology

The proposed methodology comprises different stages that initiate from the dataset and end with a performance evaluation, as shown in Fig. 1.

A. Dataset

The dataset created by Durairaj et al. [24] is utilized in this study for experimentation. It consists of posts from Reddit in the English language, where all of them is associated with some categories or labels i.e. not depressed, moderate, and severe. The label "not depressed" indicates no depression sign at all whereas "moderate" and "severe" labels represent slight depressive and high depressive symptoms and behavior.

The entire dataset is split into training (n=8891), development (n=4496), and test (n=3245) sets. While using ML and DL approaches, it is good to have large training sets as compared to validation and test sets. This is because the efficacy of these techniques directly relies on the number and variation in samples during training. A brief description of the considered dataset is given in Table 1.

B. Preprocessing

Preprocessing is a group of procedures that transform unprocessed data into useful forms to carry out the next steps with high accuracy [25]. Thus, to boost the worth of the raw data, firstly, a search for all the emoji present in the data is performed and then these are transformed into text form. Similarly, URLs or Links that did not offer anything of significance to the textual content were deleted.

TABLE I. Dataset Description

Label	Example	Train Set	Dev Set
Not Depressed	Happy New Year everyone……	1971	1830
Moderate	It is the worst feeling anyway…..	6019	2306
Severe	I just want to fall asleep forever…….	901	360
Total Samples	--	8891	4496

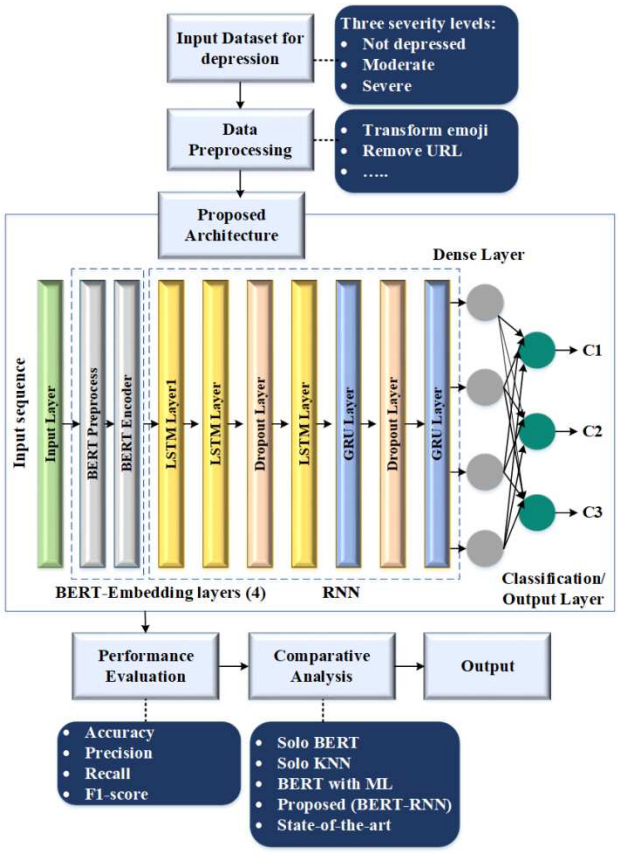


Fig. 1. Proposed Framework for Depression Classification at different Severity Levels

C. Classification

DL models are known to be enhanced by employing the TL technique. The training of a neural network is first done as a linguistic model utilizing an extensive and complete set of data in the preliminary stage of the TL designing, which is usually named as semi-supervised training. This is then followed by supervised training, in which the model is trained to employ adequately labelled training information set [26]. In this study, a novel approach combining both TL and DL techniques, namely BERT and RNN, is proposed to perform text analysis and classification tasks. A brief overview of the techniques used is provided as follows:

(i) Proposed BERT-RNN Hybrid Model

a. BERT

Many pre-trained models exist, but each of them suffers an identical issue i.e. their unidirectionality, which limits their abilities. As a consequence, a particular type of transformer network referred to as Bidirectional Encoder Representations from Transformers (BERT) came to light [27] that works by processing the text that is inputted in both directions. This ability of BERT can result in improved model performance. In this work, in this study, Bert-base architecture is employed due to its computational efficiency, compact size, and simplicity. It has 110 million parameters, 12 self-attention heads, and 12 layers of transformer encoder. WordPiece embeddings are employed to pre-train BERT. Each sentence in BERT's input encounters two different tokens, designated as [CLS] and [SEP] tokens where the [CLS] index indicates the start of each and every sequence and the [SEP] token or index serves to separate every pattern in the input. The put in

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:18:37 UTC from IEEE Xplore. Restrictions apply.

depiction of BERT is the mixture of three types of embeddings i.e. Token, Segment, and Position. Furthermore, the BERT design, shown in Fig. 2, trains the language model employing two tasks i.e. Masked Language Model (MLM) and Next Sentence Prediction (NSP). 15% of the tokens in the MLM task are enclosed using [MASK] tokens before the word sequences are passed into BERT. The system subsequently estimates the true worth of the enclosed token. Concerning the task of NSP, the BERT technique accepts input matching set of words and predicts whether or not the secondary sentence in the match corresponds to the sentence following it in the initial sentence. The concept of using a pre-trained model in comparison to traditional methods saves a lot of time for training [27,30].

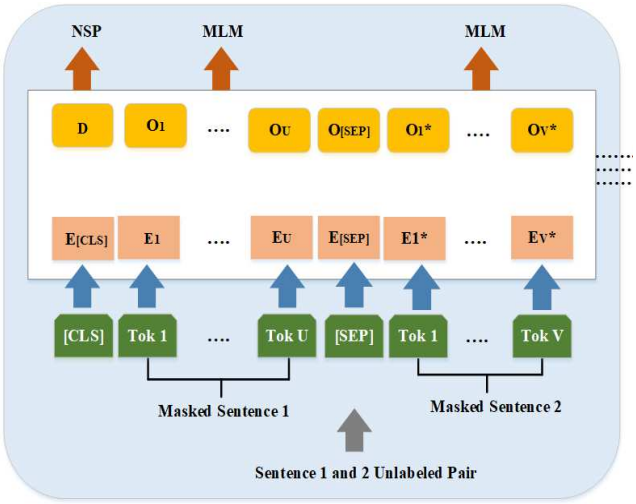


Fig.2. BERT Pre-training Architecture

b. RNN

RNN is the widely preferred DL technique for pattern recognition that can handle input data with varying lengths. These models receive output from the previous layer and feed it as input to the current step with the help of the memory concept as shown in Fig. 3. A recurrent neuron is the most important processing unit of RNN which is responsible for maintaining a hidden state. As text is in the sequential form, therefore, these methods are well suited for text analysis [28].

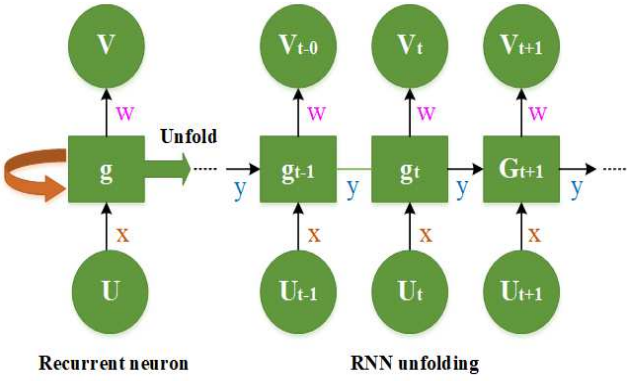


Fig. 3. RNN Structure

These models are specially constructed to tackle the issues with serial data and their inside memory feature makes them robust and very precise in estimating what's coming next. One of the disadvantages of these networks is that they suffer from the problem of vanishing gradient. Thus, in this work, two forms of RNN i.e. LSTM and GRU are utilized. LSTM has the capability to save or delete information as per its priority by employing three types of gates. The first gate i.e. Forget gate is used to drop all the useless details from memory. The next i.e. input gate is responsible for updating new details in memory and the last i.e. output gate selects relevant details and sends them to the next steps. In contrast to LSTM, GRU comprises two gates namely Reset and Update. The reset gate keeps on checking the amount of information that needs to be forgotten and the Update gate checks the amount of knowledge that requisite to be passed in the entire network [29].

c. Proposed

This work proposed to combine BERT and RNN to perform a more reliable classification of depression severity levels. Further, the advantages of LSTM such as accurateness and GRU having low complexity are fused to achieve enhanced performance. A section in Figure 1 represents the proposed BERT-RNN framework structure fusing both LSTM and GRU. The process followed is provided in Algorithm 1 depicting the step of the proposed model for depression detection. Firstly, tokenization is performed at the input layer from the input word sequence and the data which is tokenized is then sent to the following layer i.e. Embedding layer. In this layer, a pre-trained BERT embedding method is employed to generate relevant and useful word embeddings. The output is then transformed into the next or third layer which comprises of fusion of LSTM, Dropout, and GRU layers. Finally, the Softmax output layer is used as the fourth layer which provides the likeliness of the considered groups. The category having the greatest probability is considered the forecasted class.

D. Performance Evaluation

Once the system is created, it is essential to determine its efficacy utilizing some evaluation measures of performance. Depending on the results obtained, the robustness of the applied approach can be successfully analyzed. In this study, measures like accuracy, recall, precision, and F1-score are utilized to evaluate the efficiency of the proposed approach.

Algorithm 1: Proposed model for depression detection

Input: Text dataset with 3 labels: not-depressed, moderately and severely depressed

Output: Predicted label on test dataset

Step 1. Import the dataset
Step 2. Perform preprocessing of dataset by removing URL's and transforming emoji
Step 3. Divide the preprocessed dataset into training and testing in 70:30 ratio
Step 4. Perform tokenization using BERT
Step 5. Add Embedding layers() with weight
Step 6. Two LSTM() layers
Step 7. Dropout layer
Step 8. LSTM() layer
Step 9. GRU() layer
Step 10. Dropout layer
Step 11. GRU() layer
Step 12. Dense layer() with softmax activation function
Step 13. Train the model on training set
Step 14. Perform testing and evaluation of model on test set

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:18:37 UTC from IEEE Xplore. Restrictions apply.

Accuracy, being, the most preferred measure is considered to make a comparison of the presented approach with the previous studies [30].

IV. Results and Analysis

The entire implementation process is performed using Python programming on a GPU system. The dataset is split into 70:30 i.e. training and testing set. The data provided in Table 2 are taken as parameters to the BERT with the purpose of embedding extraction. The explained procedure in Fig. 1 is followed and the results yielded are provided in Table 3. From the data of results, it can be analyzed that the presented hybrid technique attained an excellent performance by achieving an accuracy, recall, precision, the presented hybrid technique attained an excellent and F1-score of 95.0%, 92.0%, 94.0%, and 93.0% respectively. To validate results, the proposed model is compared to other models considering solo BERT and RNN which yielded an accuracy rate of 93.0%. Also, BERT is evaluated by combining it with other ML techniques including SVM, decision trees (DT), KNN, and naïve Bayes (NB). The data in Table 3 showed that hybridizing BERT with RNN has given the highest accuracy i.e. 95.0% in comparison to all other techniques applied. The graphical representation of all the outcomes achieved using the proposed and other techniques is depicted in Fig. 4. As the BERT model has the benefit of providing improved embeddings, combining it with RNN can yield improved results.

TABLE II. Proposed Model Parameters

Parameter	Value
LSTM layers	3
GRU layers	2

TABLE III. Results with Different Techniques

Model	Accuracy	Recall	Precision	F1-score
BERT	93.0%	91.0%	94.0%	92.0%
RNN	93.0%	92.0%	94.0%	93.0%
BERT-KNN	90.0%	88.0%	91.0%	89.0%
BERT-SVM	92.0%	90.0%	93.0%	91.0%
BERT-DT	88.0%	86.0%	89.0%	87.0%
BERT- NB	86.0%	84.0%	87.0%	85.0%
Proposed	**95.0%**	**92.0%**	**94.0%**	**93%**

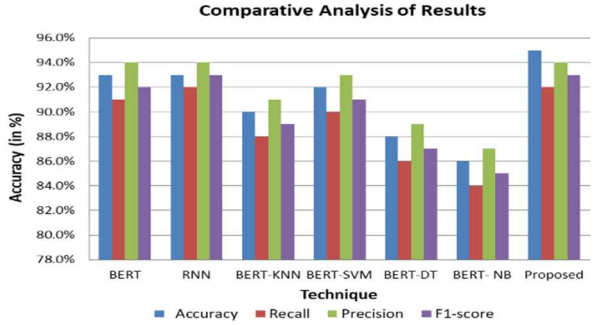


Fig.4 Graphical representations of results

Further, the proposed model accuracy is evaluated with the previous studies and the comparative data results are given in Table 4. Thus, it is evident from the table below that the proposed BERT-based DL methodology has attained the best accuracy and can be efficiently utilized by clinicians to analyze and treat at different levels of depression.

TABLE IV. Comparative Analysis with Past Studies

Ref.	Modality	Technique	Accuracy
[1]	Text	BERT, AlBert, DistilBert, RoBerta, Ensemble	61% (Best with Ensemble)
[2]	Text and audio	CNN	92% (text) 88% (audio)
[3]	Text and audio	Logistic regression, SVM	86% (Best with SVM)
[11]	Text	SVM	70%
Proposed	**Text**	**Hybrid BERT-RNN**	**94.0%**

V. Conclusion

Depression is the major cause of a lot of problems such as anxiety, suicidal attempts, etc. To deal with depression, it is necessary to analyze it at different severity levels . This paper proposed and presented a robust BERT-RNN-based hybrid approach to experiment on three labels i.e. not-depressed, moderately, and severely depressed. The text set of data is used which is taken from an online repository and is initially prepared to refine its quality. For classification, the embeddings are extracted using a BERT BERT-based model which is fed into the combination of LSTM, GRU, and a combination of LSTM, GRU, and dropout layers of RNN. The analysis of results obtained using metrics like accuracy revealed the highest values i.e. 95.0% with the proposed approach in comparison to other techniques as well as state-of-the-art. Thus, the presented strategy will be helpful for researchers and doctors to deal with the serious issue of depression accurately. Further work can be performed considering depression at early severity using a greater number of modalities and other DL techniques.

References

[1] Janatdoust, M., Ehsani-Besheli, F., & Zeinali, H. (2022, May). KADO@ LT-EDI-ACL2022: BERT-based Ensembles for Detecting Signs of Depression from Social Media Text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 265-269).

[2] Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. *Measurement: Sensors*, *25*, 100587.

[3] Thati, R. P., Dhadwal, A. S., Kumar, P., & P, S. (2023). A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms. *Multimedia Tools and Applications*, *82*(4), 4787-4820.

[4] Lau, C., Zhu, X., & Chan, W. Y. (2023). Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, *14*, 1160291.

[5] Muñoz, S., & Iglesias, C. Á. (2023). Detection of the Severity Level of Depression Signs in Text Combining a Feature-Based Framework with Distributional Representati ons. *Applied Sciences*, *13*(21), 11695.

[6] Stepanov, E. A., Lathuiliere, S., Chowdhury, S. A., Ghosh, A., Vieriu, R. L., Sebe, N., & Riccardi, G. (2018, September). Depression severity estimation from multiple modalities. In 2018 ieee 20th international

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:18:37 UTC from IEEE Xplore. Restrictions apply.

conference on e-health networking, applications and services (healthcom) (pp. 1-6). IEEE.

[7] Ilias, L., Mouzakitis, S., & Askounis, D. (2023). Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media. *IEEE Transactions on Computational Social Systems*.

[8] Kayalvizhi, S., Durairaj, T., & Chakravarthi, B. R. (2022, May). Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 331-338).

[9] Narziev, N., Goh, H., Toshnazarov, K., Lee, S. A., Chung, K. M., & Noh, Y. (2020). STDD: Short-term depression detection with passive sensing. *Sensors*, *20*(5), 1396.

[10] Wang, Q., Yang, H., & Yu, Y. (2018). Facial expression video analysis for depression detection in Chinese patients. *Journal of Visual Communication and Image Representation*, *57*, 228-233.

[11] Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

[12] Mousavian, M., Chen, J., & Greening, S. (2019, December). Depression detection using feature extraction and deep learning from sMRI images. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 1731-1736). IEEE.

[13] Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

[14] Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, *99*(2), 143.

[15] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 128-137).

[16] Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017, October). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In *2017 international conference on information and communication technology convergence (ICTC)* (pp. 138-140). IEEE.

[17] Islam, M. R., Kamal, A. R. M., Sultana, N., Islam, R., & Moni, M. A. (2018, February). Detecting depression using k-nearest neighbors (knn) classification technique. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.

[18] Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific reports*, *7*(1), 13006.

[19] Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020, June). Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 407-411).

[20] Figuerêdo, J. S. L., Maia, A. L. L., & Calumby, R. T. (2022). Early depression detection in social media based on deep learning and underlying emotions. *Online Social Networks and Media*, *31*, 100225.

[21] Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., & Tao, C. (2018, December). XA-BiLSTM: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 1624-1627). IEEE.

[22] Poświata, R., & Perełkiewicz, M. (2022, May). OPI@ LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 276-282).

[23] Al Asad, N., Pranto, M. A. M., Afreen, S., & Islam, M. M. (2019, November). Depression detection by analyzing social media posts of user. In *2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON)* (pp. 13-17). IEEE.

[24] Kayalvizhi, S., Durairaj, T., & Chakravarthi, B. R. (2022, May). Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 331-338).

[25] Bhat, P., Anuse, A., Kute, R., Bhadade, R. S., & Purnaye, P. (2022). Mental health analyzer for depression detection based on textual analysis. *Journal of Advances in Information Technology Vol*, *13*(1).

[26] Ullah, N., Khan, J. A., Khan, M. S., Khan, W., Hassan, I., Obayya, M., ... & Salama, A. S. (2022). An effective approach to detect and identify brain tumors using transfer learning. *Applied Sciences*, *12*(11), 5645.

[27] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[28] Shreyashree, S., Sunagar, P., Rajarajeswari, S., & Kanavalli, A. (2022). BERT-Based Hybrid RNN Model for Multi-class Text Classification to Study the Effect of Pre-trained Word Embeddings. *International Journal of Advanced Computer Science and Applications*, *13*(9).

[29] Bhatia, M. P. S., Kumar, A., & Beniwal, R. (2016, March). Ontology based framework for reverse engineering of conventional softwares. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 3645-3648). IEEE.

[30] Beniwal, R., & Saraswat, P. (2024). A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data. The Computer Journal, bxae018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:18:37 UTC from IEEE Xplore. Restrictions apply.

Iranian Polymer Journal
https://doi.org/10.1007/s13726-024-01386-7

ORIGINAL RESEARCH

Biodegradable, biocompatible, and self-healing, injectable hydrogel based on oxidized *Azadirachta indica* gum and carboxymethyl chitosan through dynamic imine-linkage for biomedical application

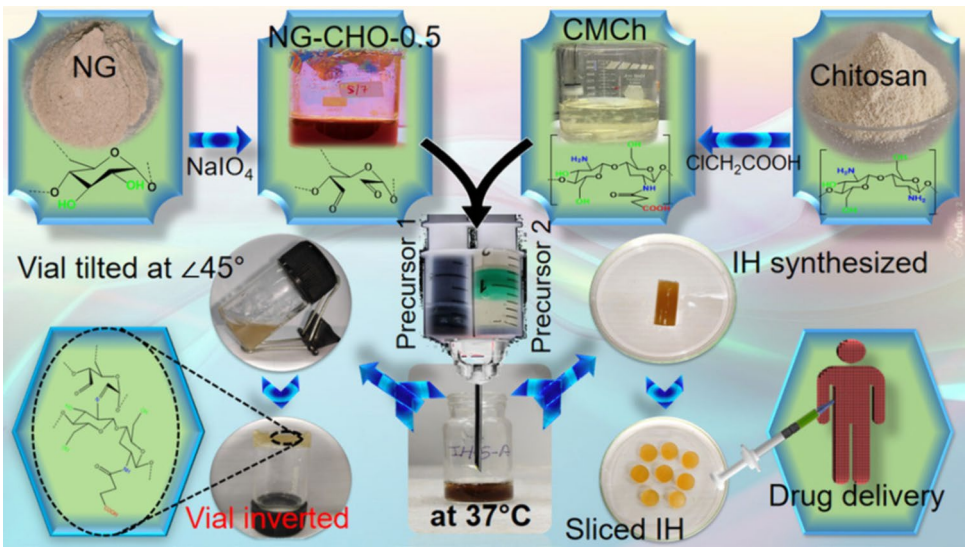
Jitendra Kumar[1] · **Roli Purwar**[1]

Received: 22 May 2024 / Accepted: 14 July 2024
© Iran Polymer and Petrochemical Institute 2024

Abstract

At physiological temperature (37 °C), neem gum (NG) containing multialdehyde groups (NG-CHO) and carboxymethyl chitosan (CMCh) were combined to formulate an injectable hydrogel through the utilization of dynamic imine linkages. The investigation comprises determining the ideal oxidizing agent concentration to maximize the amount of aldehydes in neem gum. The best-oxidized neem gum determined the minimum gelation time (3 min). The injectable hydrogel's syringeability (~ 99%), porosity (47–71%), self-healing ability, and gel content (58–76%) were evaluated. Characterization techniques, including FTIR and ^{1}H NMR, confirmed the synthesis, and FESEM showed the morphology of the injectable hydrogel (pore size ~ 80 to 120 µm). Rheometry technique was used for quantitative analysis of shear-thinning behavior and self-healing ability. The swelling ratio was assessed for dry and gel forms of injectable hydrogels in phosphate buffer saline (PBS) at pH 5.5, 7.4, and 8.5. The swelling ratio decreased as the pH increased (5.5–7.4) and then increased at pH 8.5 for all synthesized injectable hydrogels (IHs) in dry and gel forms. For the release assay, an antibiotic model drug, ciprofloxacin HCl (Cipro), was loaded in situ, and the drug release behavior of dry and gel forms of injectable hydrogel was compared within all pH ranges. The drug release kinetics were estimated using the Korsmeyer–Peppas model. Cytotoxicity evaluation using L-929 fibroblast cell lines, antibacterial assay, and hydrolytic degradability (20–65%) in phosphate buffer saline at pH 7.4 was also conducted.

Graphical Abstract



Keywords Carboxymethyl chitosan · Neem gum · Injectable hydrogel · Self-healing · Schiff base

Extended author information available on the last page of the article

Published online: 13 September 2024

Introduction

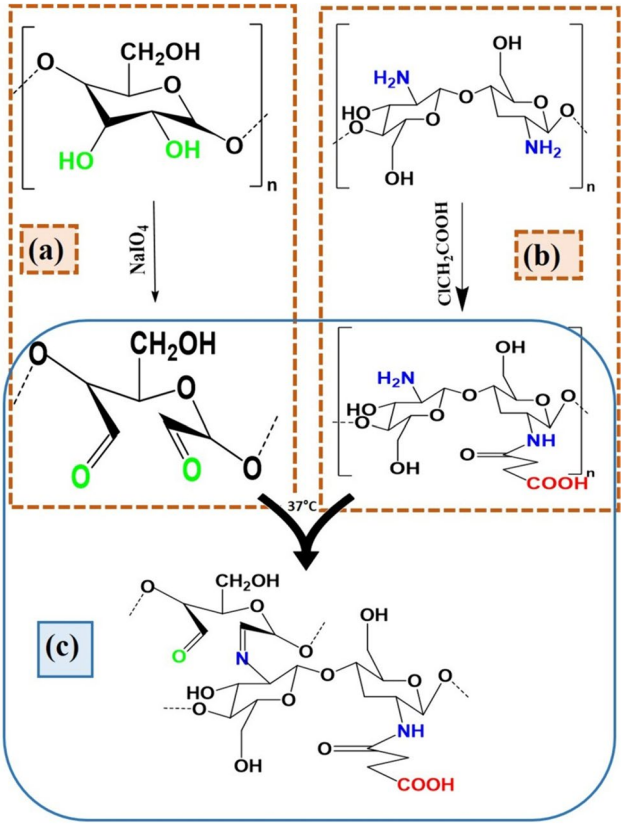
Neem gum, also known as Indian gum, is derived from the *Azadirachta indica* plant by making incisions in the tree trunk and collecting the gum that oozes out. It is a gummy substance that hardens when exposed to the air [1]. It comprises various substances: proteins, Mg, Na, Ca, and K salts, glucuronic acid, xylose, fucose, and arabinose. The gum contains arginine, aspartic acid, aminobutyric acid, alanine, glycine, and other amino acids. However, neem gum's exact chemical composition and structure can vary depending on factors like the source and processing methods. Neem gum is used topically as a poultice for wound healing and to relieve skin conditions. Due to its antibacterial and anti-inflammatory properties, neem gum is used for drug delivery applications. Among the innovative materials emerging in the biomedical context, hydrogels have garnered significant attention due to their ability to encapsulate and deliver a wide range of therapeutic agents, including small molecules, peptides, and proteins, with precise control over release kinetics. Mankotia et al. fabricated a neem gum-based hydrogel using methylene-bis-acrylamide as a crosslinker and used it to release anticancer drugs [2]. A group of researchers synthesized a hydrogel using neem gum and silk fibroin crosslinked with sodium benzoate for wound-healing applications [3]. Singh et al. fabricated neem gum-based hydrogel crosslinked with methylene-bis-acrylamide, for colonic drug delivery application [4–6]. Shreya et al. synthesized a neem gum-based hydrogel using radiation-induced crosslinking, which is further utilized for drug delivery applications [7]. Another group of researchers synthesized neem gum and *N*-vinyl pyrrolidone-based copolymeric hydrogel using methylene-bis-acrylamide as a crosslinker, again used in wound-healing applications [8, 9].

Injectable hydrogels offer a minimally invasive approach for localized drug delivery [10], providing advantages such as reduced systemic side effects and improved patient compliance [11]. Natural polymers, especially polysaccharides like chitosan [12], alginate [13], and hyaluronic acid [14], have gained prominence in hydrogel formulation due to their biocompatibility, biodegradability, and inherent ability to interact with biological systems [15]. Schiff base crosslinking, derived from the reaction between a primary amine and an aldehyde, offers several advantages for injectable hydrogel design, including its mild reaction conditions, controllable crosslinking kinetics, and the potential for tunable mechanical and drug release properties [16]. The resulting Schiff-based injectable hydrogels exhibit vigorous stability and the ability to undergo dynamic transformations in response to environmental cues, making them an attractive option for drug delivery [17].

Carboxymethyl chitosan, a water-soluble derivative of chitosan, is used for the development of injectable hydrogel for wound-healing application [18]. The degree of carboxymethylation affects its solubility in water. The carboxymethyl chitosan shows antibacterial, antioxidant, anticancer, and antitumor properties [19].

As per the literature survey, it was found that no neem gum-based injectable hydrogel has been synthesized using Schiff base chemistry. The author hypothesize that multialdehyde neem gum can establish crosslinks with carboxymethyl chitosan through dynamic Schiff base chemistry to form injectable hydrogel at 37 °C with good syringeability, injectability, and self-healing capacity without the aid of another crosslinker. It is very important to use biocompatible precursor polymers devoid of UV light, hazardous initiators, or other toxic monomers for Schiff's base injectable hydrogels. All the reported literature on neem gum utilizes toxic crosslinkers to synthesize hydrogels with no self-healing properties. The addition of any crosslinker has added one more step to complete the synthesis process, while there is no requirement for crosslinker in this research article. Thus, the present study is more economical and non-toxic. In the case of swelling studies and drug release of injectable hydrogel, researchers presented their work with a dry form of injectable hydrogel [20–23]. However, in practical application, IHs will be used in gel form. As a result, this research article includes comparison studies between the dry and gel forms of the injectable hydrogel, focusing on swelling properties, drug release behavior, and hydrolytic degradation. Thus, the novelty of this research article includes the use of neem gum biomaterial, the absence of a toxic crosslinker, and a comparative analysis of the dry form of injectable hydrogel with a freshly prepared gel form of injectable hydrogels.

Considering the information above, the current work describes the fabrication of multialdehyde groups on neem gum utilizing $NaIO_4$ as an oxidant, which converts the hydroxyl groups of neem into aldehyde groups (Scheme 1a). CMCh is synthesized by adding carboxymethyl groups to the chitosan molecule (Scheme 1b). Without any hazardous or chemical crosslinking agents, the aldehyde groups of oxidized neem gum can covalently establish crosslinks with the primary amine group of CMCh through Schiff base crosslinking at 37 °C (Scheme 1c). The rheological behavior, self-healing capability, swelling ability, and drug release assay in PBS at different pH levels of synthesized injectable hydrogels are examined. Furthermore, in vitro cytotoxicity, antibacterial assay, and hydrolytic degradation are also assessed.



Scheme I **a** Oxidation of neem gum (NG-CHO), **b** derivatization of chitosan (CMCh), and **c** synthesis of injectable hydrogel (IH)

Experimental

Materials

Chitosan was purchased from Otto Chemical Pvt. Ltd. (deacetylation degree > 85%, pH 7–9, viscosity (1% w/v, 200 cps), molecular weight 218,112 Da). In contrast, neem gum (molecular weight 191,573 g mol^{-1}) was purchased from Nutramine Life Science in New Delhi, India. From CDH, sodium periodate ($NaIO_4$) and hydroxylamine hydrochloride were produced. From Thomas Baker Chemicals Pvt. Ltd., ethylene glycol, isopropyl alcohol, and monochloroacetic acid were taken. Dulbecco's modified eagle medium (DMEM) and dimethyl sulfoxide were purchased from Thomas Baker Pvt. Ltd. Fetal bovine serum, hydrochloric acid, and sodium hydroxide were bought from Fisher Scientific Chemical. Ethanol with 99.9% purity was used. The antibiotic ciprofloxacin HCl was given as a gift from M/S Unicure India Limited. Double-distilled water (MQ) was employed for all reactions. Chemicals and solvents were used as received.

Oxidation of neem gum

Oxidation of neem gum was carried out according to the protocol described by Zang et al. with slight modifications [24]. Furthermore, titration was used to determine the aldehyde percentage of oxidized neem gum according to the protocol described by Balitaan et al. [25] (Eqs. 1 and 2). A detailed description is given in Supplementary Materials (1.1)

$$NG-(CHO)_n + H_2N-OH\cdot HCl \rightarrow NG-(CH=N-OH) + H_2O \quad (1)$$

$$HCl + NaOH \rightarrow NaCl + H_2O. \quad (2)$$

Derivatization of chitosan

Carboxymethylation of chitosan was carried out according to the procedure reported by Feng et al. [26]. For a description, see the Supplementary Materials (1.2).

Preparation of injectable hydrogel

The known quantity of oxidized NG was dissolved in MQ water (precursor 1) at 700 rpm for 24 h at 27 °C. A similar procedure was used to dissolve CMCh (5% w/v) in phosphate buffer saline at pH 7.4 at 700 rpm at 27 °C (precursor 2). After that, equal amounts of both precursors were simultaneously extruded into a vial through two separate syringes of 18 gauzes, each holding 5 mL of precursors at a physiological temperature (37 °C). Gel formation started right away. A total of six samples of injectable hydrogels were prepared. IH-m–n is the acronym for the prepared injectable hydrogel, where "m" stands for the concentration of $NaIO_4$ and "n" stands for the concentration of oxidized NG.

Self-healing behavior, syringeability, injectability evaluation, and network parameters of the injectable hydrogels are described in Supplementary Materials (1.3, 1.4, and 1.5).

Gelation time

The tube inversion test is the qualitative measurement in which precursors 1 and 2 were simultaneously extruded at different concentrations in a glass vial with a 10 mL capacity at 37 °C in a 1:1 ratio. Firstly, the vial was tilted at a 45° angle and then inverted completely at physiological temperature to check the injectable hydrogel's physical condition. Gel flow was visually examined.

Gravimetrically, IHs were unable to flow, supporting the gelation. To determine the precise gelation period of the synthesized IH, a tube inversion test was conducted five times, and the mean value of time was reported.

Gel content

The Rasool and workfellows' methodology was utilized to determine the gel content [27]. The known quantity of IHs (W_a) was immersed in 100 mL of MQ water for 48 h at 37 °C. The water-insoluble IHs were removed and dried to a constant weight (W_b) at 50 °C. The percentage of gel content (GC %) was computed using Eq. 3. The analyses were performed using the mean value of the three trials

$$GC\ \% = \frac{W_b}{W_a} \times 100 \tag{3}$$

Porosity

The liquid displacement method determined the injectable hydrogel's porosity [28]. To summarize, a known quantity of oven-dried IHs was weighed (X_b) with volume V and placed for 24 h within a 50 mL vial containing ethanol. The vial was wrapped entirely in parafilm to prevent ethanol from evaporating and losing its contents. IHs were removed from the vial after 24 h and weighed (Xa). The evaluation was performed three times, to determine the porosity of injectable hydrogel (Eq. 4)

$$P\ \% = \frac{Xa \times V}{X_b} \times 100 \tag{4}$$

Rheological properties of synthesized injectable hydrogel

A parallel plate-equipped MCR-101 rheometer (Anton Paar, Japan) was used to analyze various rheological parameters at 37 °C. A 1 mm gap was kept between the plate and the parallel plate geometry throughout each session. The IHs was incubated for an hour at 37 °C before testing. A time-dependent behavior was studied using a time sweep study in isothermal mode. An amplitude sweep test was conducted to determine the linear viscoelastic range (LVE) of the IH. Additionally, frequency sweep tests on these IHs were performed at 1% strain, covering the frequency range of 0.01 to 100 rad s^{-1}. A flow sweep experiment in the range of 0.01–100 s^{-1} was carried out to evaluate the effectiveness of shear-thinning and modifications in viscosity. A time sweep study was conducted, in which precursor 1 and precursor 2 were simultaneously extruded at the rheometer platform in a 1:1 ratio and allowed to get crosslinked on the platform itself at 37 °C. The IHs were then subjected to a range of strain amplitude (high 200 and low 1%) to determine their self-healing property. All measurements were carried out for triplicate samples.

Swelling ratio

The swelling ratio was checked in PBS at pH 5.5, pH 7.4, and pH 8.5 at 37 °C. The swelling ratio from IHs was measured in dry and freshly prepared gel form. In the first case, synthesized IHs were initially incubated for 4 h at 37 °C to enable complete crosslinking and then kept in a hot air oven at 50 °C to dry completely. Button form IHs were put in 100 mL of PBS at pH 5.5, pH 7.4, and pH 8.5 at 37 °C to check the swelling behavior. In the second case, the freshly synthesized gel form of IHs directly put in 100 mL PBS solution at pH 5.5, pH 7.4, and pH 8.5 at 37 °C. After completion of a predetermined time, IHs were removed from their respective beakers, and the surface water was wiped out using tissue paper and weighed. The analyses were performed on three samples. The swelling ratio (SR %) was determined by the amount of water absorbed by the IHs using Eq. 5

$$SR\ \% = \left[\frac{Ws - Wd}{Wd}\right] \times 100 \tag{5}$$

where, Wd is the weight of the dry injectable hydrogel and Ws is the weight of the swollen injectable hydrogel.

The network parameters of injectable hydrogels were also calculated and reported in Supplementary Materials (1.5).

Drug loading and release kinetics

An antibiotic drug, ciprofloxacin HCl, is employed as a model drug. The calibration curve for the Cipro was performed using a UV–visible spectrophotometer. To prepare precursor 3, oxidized NG solution was mixed with a 3 mL aqueous solution of Cipro (6 mg mL^{-1}). Precursors 2 and 3 were combined in equal proportions to synthesize the drug-loaded IHs. There are two distinct approaches were used to quantify drug release from IHs. In the first approach, the synthesized IHs were incubated for 4 h at 37 °C, followed by drying in a hot air oven at 50 °C. Then, button form drug-loaded IHs were placed in 100 mL of PBS with pH values of 5.5, 7.4, and 8.5 at 37 °C. In the second approach, the freshly synthesized gel form of IHs were put into 100 mL of a PBS solution with pH values of 5.5, 7.4, and 8.5 at 37 °C. For both the approaches, aliquots from each release media (pH 5.5, 7.4, and 8.5) were taken at a 2-h interval and simultaneously replenished with new release media to maintain the same total volume of the release media. Here, it was assumed that the drug molecules were dispersed evenly throughout the IHs. The analyses were performed using

three observations. The amount of Cipro released was calculated using UV–visible spectroscopy at an absorbance of 275 nm (PBS 5.5) and 271 nm (PBS 7.4 and PBS 8.5) [29].

The Korsmeyer–Peppas model was used to fit the experimental data. The power law equation serves as the foundation for the Korsmeyer–Peppas model, which is represented as Eq. 6

$$\frac{M_t}{M_\infty} = kt^n \tag{6}$$

where M_t and M_∞ represent the cumulative drug concentrations released at time t and time infinity, respectively. "n" is the release exponent, which is used to describe the release mechanism, and "k" is the release rate constant, which depends on the matrix's geometry and structure [30].

In vitro cytotoxicity evaluation

Cytotoxicity of the IHs was determined by MTT assay on cell line L-929 fibroblast cell lines (procured from NCCS Pune). The cells (5000 to 8000 cells per well) were cultured in 96 well plates for 24 h in DMEM medium supplemented with 10% FBS and 1% antibiotic solution at 37 °C with 5% CO_2. Injectable hydrogel formulations varying in concentration from 1 to 1000 µg mL $^{-1}$ were applied to the cells the next day. After incubation for 24 h, MTT solution (a final concentration of 250 µg mL $^{-1}$) was added to the cell culture and further incubated for 2 h. At the end of the experiment, the culture supernatant was removed, and the cell layer matrix was dissolved in 100 µl dimethyl sulfoxide (DMSO) and read in an Elisa plate reader (iMark, Biorad, USA) at 540 nm and 660 nm. Inhibitory concentration (IC-50) was calculated using Graph Pad Prism-6. Images were captured under an inverted microscope (Olympus ek2) using a camera (AmScope digital camera 10 MP Aptima CMOS) [31]. The measurements were conducted using three samples for each concentration, and the mean value was reported.

Antibacterial assay

The antibacterial activity was evaluated using the zone inhibition method, also known as the Kirby–Bauer method. Both *S. aureus* and *E. coli* were selected as representative models for Gram-positive and Gram-negative bacteria, respectively. The Mueller–Hinton agar plates were inoculated by spreading with 100 µL of each bacterium culture (cell density ~ 1.5×10^8 CFU/mL) and followed by placing the discs containing 10 µL of different concentrations (0–100 mg/mL). One disc (5 mm) on each plate was loaded with solvent alone, serving as the vehicle control, while a ciprofloxacin disc (10 µg) was used as the positive control. The plates were then incubated at 37 °C for 24 h. The clear zones around the discs were measured and recorded to assess the antibacterial activity. Three replicates were conducted for each compound against each bacterium.

Hydrolytic degradation in PBS at pH 7.4

The hydrolytic degradation of synthesized IHs was carried out for dry and freshly prepared gel forms of IHs. In the case of dried form, the Hozumi and workfellows' methodology was used to determine hydrolytic degradation [32]. A known quantity of dried IHs (0.2 g) was immersed in PBS at pH 7.4 and 37 °C, and after it reached equilibrium swelling, it was weighed (w_a). The samples were removed from the PBS solution, the surface was cleaned and then reweighed (w_b) at specified intervals. Weight loss was used to compute percentage degradation (HD %) using Eq. 7. In the second case, the freshly prepared gel form of IHs was cut into slices and put in PBS at pH 7.4 at 37 °C. Then, weight loss was checked using Eq. 4. The measurements were performed on three samples

$$\text{HD } \% = \frac{w_a - w_b}{w_a} \times 100 \tag{7}$$

Characterization

The Fourier transform spectroscopy test was performed in ATR mode on a Perkin Elmer 2000 instrument for qualitative analysis. The samples were analyzed in the frequency range of 500 − 4000 cm^{-1} and a resolution of 4 cm^{-1}. ^{1}H NMR was recorded on Bruker advance III, 500 MHz, using D_2O as the solvent. Chemical shift (δ) values were reported in ppm using tetramethyl silane as an internal reference. The surface morphological features were studied in a field emission scanning electron microscope (ZEISS, EVO 18). The double adhesive carbon tape was used to mount gold gold-coated samples. The photomicrographs were recorded at different magnifications. The UV–visible spectrophotometer was used for the drug release assay.

Statistical analysis

All experiments were repeated at least three times. The experimental data are presented as the mean ± standard deviation. Statistical significance was estimated using one-way analysis of variance (ANOVA), and the significance level was considered $p < 0.05$. All statistical analyses were performed using the Microsoft Excel 2010 software.

Table 1 Effect of concentration of $NaIO_4$ on gelation time of injectable hydrogel

Sample code	NG % (w/v)	NaIO % (w/v)	Aldehyde % (A)	Amine group % present in CMCh (B)	Ratio = $\frac{A}{B}$	Gelation time (min)
IH-0.1-1	1	0.1	31.5 ± 0.5	67 ± 2.1	0.47	40 ± 3
IH-0.3-1	1	0.3	40.7 ± 0.4	67 ± 2.1	0.61	32 ± 2
IH-0.5-1	1	0.5	56.8 ± 1	67 ± 2.1	0.85	18 ± 1.3
IH-1-1	1	1	56.1 ± 0.8	67 ± 2.1	0.84	20 ± 1.5

Results and discussion

Effect of concentration of $NaIO_4$ on gelation time

Table 1 shows the effect of the derivatization of NG on gelation time. With the increase in the concentration of oxidizing agent (0.1–0.5% w/v), gelation time decreases (40 ± 3 to 18 ± 1.3 min) and becomes almost constant (20 ± 1.5 min) with further increase in the concentration of $NaIO_4$ (1% w/v). This behavior was primarily because increasing the concentration of the oxidizing agent increases the conversion of hydroxyl groups to aldehyde groups. The amount of aldehyde percentage was calculated and found to be increased from 31.5% ± 0.5% to 56.8% ± 1% (Table 1). Further increase in the concentration of oxidizing agent (1% w/v), the conversion of the aldehyde group remains constant. This behavior was explained by calculating the ratio of aldehyde percentage to the amine group present on CMCh (Table 1). A group of researchers synthesized an injectable hydrogel using derivatized hyaluronic acid and chitosan, and hyaluronic acid was derivatized using $NaIO_4$. The required amount of $NaIO_4$ was added, and successful oxidation degrees of 10%, 15%, and 25% were achieved, and the observed gelation time was 8 s [33]. Another group of researchers reported similar work, which used the same procedure for deriving hyaluronic acid by adding a different amount of $NaIO_4$. Results showed that an increase in the concentration of $NaIO_4$ increases the oxidation degree from 27 to 60%, which decreases the gelation time from 175 to 145 s [34]. From this study, the concentration of sodium periodate was optimized at 0.5% w/v for further studies.

Effect of concentration of oxidized neem gum on gelation time

Table 2 shows the effect of the concentration of oxidized NG on gelation time. Here, the oxidizing agent ($NaIO_4$) concentration was kept constant at 0.5% w/v. It is observed that with an increase in the concentration of NG-CHO-0.5 (1–5% w/v), a decrease in gelation time was observed (18 ± 1.3 to 3 ± 0.5 min). Furthermore, this behavior was explained using the simple unitary method. Table 2 shows the detailed analysis of the calculation, as the magnitude value of the ratio continuously increased from 1 to 5% w/v of NG-CHO-0.5, and the observed gelation time was also decreased (18 ± 1.3 to 3 ± 0.5 min). This is primarily because higher numbers of aldehyde functional groups are assessable to react with the amine groups of derivatized chitosan. For various types of polysaccharide hydrogels, it is necessary to adjust the ratio of the aldehyde group to the amino group to achieve an effective Schiff base reaction. Longer gelation times would result from an improper ratio. In a research work [35], an injectable hydrogel was fabricated using oxidized gum Arabic with the addition of 0.5 g of $NaIO_4$ and an oxidation degree of about 44% was achieved, also the amount of derivatized gum Arabic was changed from 20 to 40%, which resulted in reduced gelation time from 42 to 29 s.

Physical properties of injectable hydrogel

The physical properties in terms of gel content, porosity, syringeability, and network parameters are given in Table 3. The gel content and porosity of IH-0.5-1, IH-0.5-3, and IH-0.5-5 are found to be 58% ± 1%, 67% ± 1%, 76% ± 2%, 71% ± 2%, 59% ± 1%, and 47% ± 2%, respectively. Good crosslinks through imine bonds are responsible for injectable

Table 2 Effect of concentration of oxidized NG gum (NG-CHO-0.5) on gelation time

Sample code	Conc. of NG-CHO-0.5 (g) (B)	Aldehyde % of NG-CHO-0.5 (A)	Amine group % present in CMCh (C)	Ratio = $\frac{A \times B}{C}$	Gelation time (min)
IH-0.5-1	1	56.8 ± 1.4	67 ± 2.1	0.85	18 ± 1.3
IH-0.5-3	3	56.8 ± 1.4	67 ± 2.1	2.54	11 ± 1
IH-0.5-5	5	56.8 ± 1.4	67 ± 2.1	4.24	3 ± 0.5

Table 3 Physical properties of injectable hydrogels

Property	Value		
	IH-0.5-1	IH-0.5-3	IH-0.5-5
Gel content	58 ± 1%	67 ± 1%	76 ± 2%
Porosity	71 ± 2%	59 ± 1%	47 ± 2%
Syringeability	99 ± 0.5%	98.89 ± 0.5%	98.72 ± 0.7%
Average molecular weight between crosslinks	5.8×10^{2} g mol^{-1}	4.5×10^{2} g mol^{-1}	3.8×10^{2} g mol^{-1}
Crosslink density	0.7 mol cm^{3}	1.01 mol cm^{3}	1.2 mol cm^{3}
Mesh size	5.1×10^{-2} nm	4×10^{-2} nm	3.3×10^{-2} nm

hydrogels' higher gel content and lower porosity. Also, the higher the porosity, the higher the mesh size, the higher the average molecular weights between crosslinks, and the lower the crosslink density. This is because a more compact structure was formed as the concentration of NG-CHO-0.5 was increased (1–5% w/v). Yan et al. have fabricated an injectable hydrogel using Schiff base chemistry between poly(*L*-glutamic acid) and chitosan, in which the observed gel content was 79% [36]. Another group of researchers published a research article based on Schiff base crosslinking using oxidized galactomannan and *N*-succinyl chitosan, in which they evaluated a porosity determination of 69% [37].

Structural properties

Structural properties of the IH showing the least gelation time (IH-0.5-5) were characterized using FTIR spectroscopy and ^{1}H NMR. Figure 1a–c shows the FTIR spectra of the NG, NG-CHO-0.5, chitosan, CMCh, and IH-0.5-5. FTIR measurements of pure NG revealed absorption bands at 3404 cm^{-1} caused by –OH-stretching vibrations in the neem gum polysaccharide, 2930 cm^{-1} due to –C–H-stretching mode of -CH_3 groups in the neem gum, peaks at 1737 cm^{-1} caused by C = O-stretching vibrations of carboxylic acid present in the neem gum. Along with the peak at 1247 cm^{-1} due to C–O–C asymmetric stretching vibrations, a strong absorption band appeared at 1030 cm^{-1} in the FTIR spectrum of pure NG, which belongs to C–O–C stretching vibration (Fig. 1a) [5]. After NG reacts with sodium periodate, significant peaks are seen at 1727 and 874 cm^{-1} (Fig. 1a), mainly attributed to the production of symmetrical vibration of aldehydes and hemiacetal bonds, respectively. Another comparative research study shows similar outcomes for the oxidized product [38]. The FTIR spectra of pure chitosan (Fig. 1b) show the amino group's bending vibrational peak at 1570 cm^{-1}, the chitosan saccharide's C–O–C asymmetric structure at 1250 cm^{-1}, and the C–O-stretching vibration at 1024 cm^{-1}. The FTIR spectrum of CMCh (Fig. 1b) reveals the existence of two additional peaks at 1644 and 1420 cm^{-1}, which are attributable to symmetric and asymmetric stretching vibrational modes of the –COO group, respectively, and indicate the successful derivatization of chitosan and confirm the presence of carboxymethyl groups on CMCh [39].

IH-0.5-5 injectable hydrogel's FTIR spectrum (Fig. 1c) shows a peak at 1642 cm^{-1}, which demonstrates imine structure formation due to dynamic Schiff base crosslinking reaction between the amine group of CMCh and the aldehyde groups of NG-CHO. Further, the development of the Schiff base connection is shown by the loss of the aldehyde peak in the IH-0.5-5 spectrum [40].

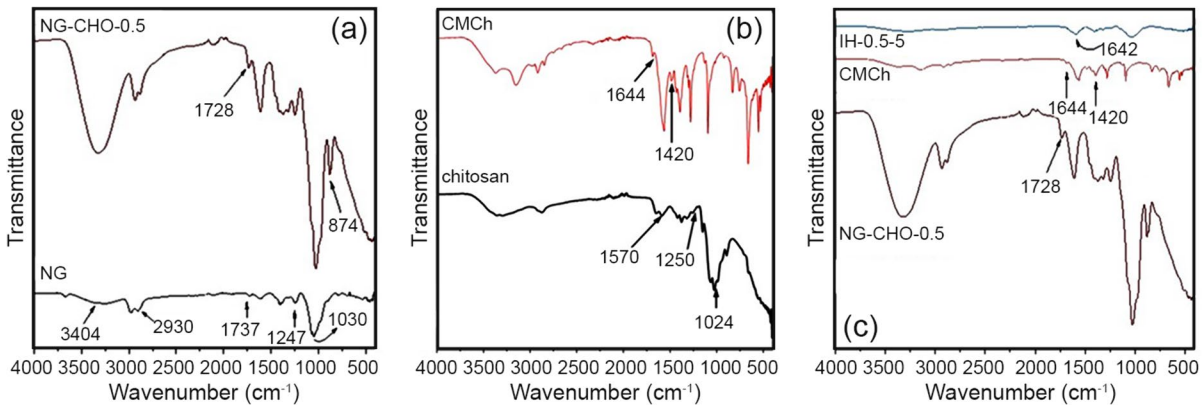


Fig. 1 FTIR spectra of: **a** NG, NG-CHO-0.5, **b** chitosan, CMCh, and **c** NG-CHO-0.5, CMCh, IH-0.5-5

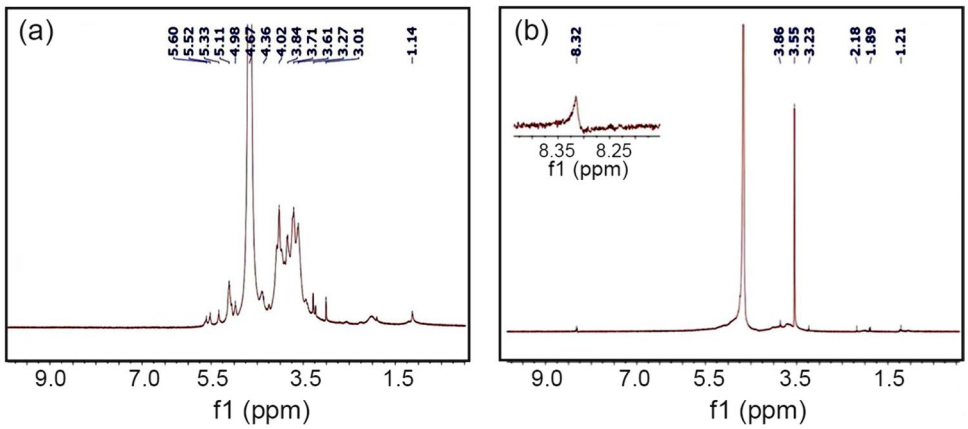


Fig. 2 NMR spectra of: **a** NG, and **b** NG-CHO-0.5

The structural properties also supported by proton NMR, and the spectra of chitosan and CMCh are described in Supplementary Materials (2.4). ^{1}H NMR was conducted for NG and oxidized NG using D_2O as solvent (Fig. 2). Due to the presence of sugar residues, the spectra displayed signals in the range of 3–6 ppm spectrum band, which correspond to the characteristic properties of polysaccharides sugar proton [41]. The peak at 3.13 ppm is due to the acetyl group ($COCH_3$) in neem gum. The peak at 1.3 ppm is due to the presence of methyl groups. The increase in signal strength between 3.3 and 3.8 ppm indicated evidence of an O-CH_3 group [42]. The peak at 4.36 to 5.36 was due to the –OH group. The central peak observed was at 4.67 ppm, designated using D_2O. The non-anomeric protons (H_2-H_6) contributed signals in the range of 3.3–4.5 ppm (Fig. 2a). A new distinctive peak of the aldehydic proton is observed at 8.32 ppm, which confirms the oxidation of NG (Fig. 2b).

Rheological properties

The time sweep study was evaluated using a rheometer at a fixed frequency (1 rad s^{-1}) and constant strain (1%). Precursors 1 and 2 were simultaneously extruded at the rheometer platform in a 1:1 ratio and allowed to crosslink on the platform at 37 °C. The changes in storage modulus (G′) and loss modulus (G") over time were determined in isothermal mode (37 °C). The G′ value rose after the first few seconds, because the Schiff base reaction continued slowly as the mixture's viscosity rose. After 3000 s, the G′ value reached the highest for IH-0.5-5 (584.75 ± 27 Pa) than IH-0.5-3 (514.7 ± 19 Pa), followed by IH-0.5-1 (122.99 ± 15 Pa). This is because when NG-CHO-0.5 concentration increased (1% to 5% w/v), resulting in tighter polymer networks and more excellent mechanical characteristics. A quick interaction between the aldehyde group of NG-CHO-0.5 and NH_2 groups of CMCh caused the storage modulus (G′) to grow dramatically in a very short period. Although the G" values in the IHs with different weight ratios of NG-CHO-0.5 (1%, 3%, and 5% w/v) showed no apparent differences (Fig. 3a). Furthermore, the time sweep study supports the gelation test (Fig. S3, in Supplementary Materials).

When measuring the amplitude sweep test, the IHs were subjected to several strain levels between 0.01% and 100% at 37 °C (Fig. 3b) to check the linear viscoelastic region (LVE). The LVE of the IH with a medium amount of NG-CHO-0.5 (IH-0.5-1) was measured with the help of the amplitude sweep test and was found to have strain values between 0.01% and 4%. A strain value of 1% was selected for all kinetic studies involving gelation. It was observed that the IHs exhibited higher values of G′ than G″ values at variable strain values, and the crossovers for the IH-0.5-5, IH-0.5-3, and IH-0.5-1 appeared at 30%, 27%, and 22%, respectively, which shows IHs changing physically from solid to liquid. Thus, IH possesses a higher amount of NG-CHO-0.5, resulting in more excellent elasticity, which is necessary for these IHs to resist the administration-linked strains during their injectable formulation, thus supporting syringeability. Balitaan et al. fabricated an injectable hydrogel using a Schiff base reaction between acrylamide-modified β-chitin and alginate dialdehyde. In the amplitude sweep test, strain value varied from (10^0% to 10^4%), and all the injectable hydrogels showed higher storage modulus values than loss modulus [25].

For performing a frequency sweep test, the variable angular frequency shear in the range of 0.1–100 rad s^{-1} was applied to the IHs at a constant amplitude (1%) (Fig. 3c). Across all the tested frequency ranges, these IHs exhibited mechanical robustness with higher G′ values than G″ values. The higher value of G′ than G″ indicates these injectable hydrogels' elastic and rigid nature. The IH-0.5-5 hydrogel showed higher values of G′ (509.03 ± 29 Pa) than the IH-0.5-1 (188.76 ± 16 Pa) and IH-0.5-3 (378.85 ± 20 Pa) due to the higher crosslinking density, which restricts the mobility of chains within the hydrogel matrix. The G′ values obtained from the frequency sweep test followed the same pattern as the amplitude sweep. A group of researchers

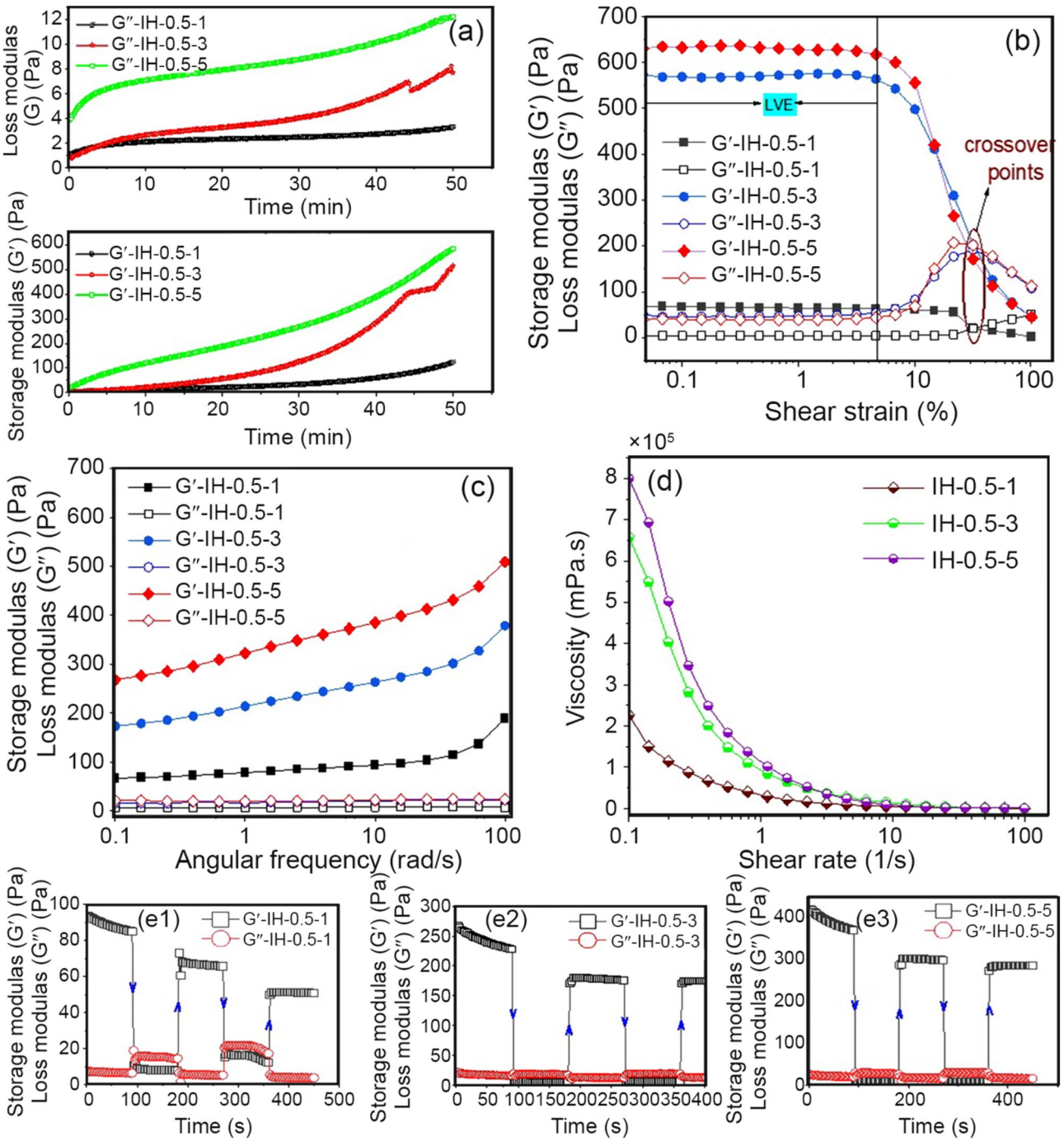


Fig. 3 **a** Isothermal time sweep study, **b** G' and G" versus shear strain showing LVE region and crossover point, **c** G' and G" versus angular frequency showing stiff and elastic nature, and **d** viscosity versus shear rate showing shear-thinning behavior. Oscillatory step-strain measurements of (e1) IH-0.5-1, (e2) IH-0.5-3, and (e3) IH-0.5-5

synthesized an injectable hydrogel using Schiff base chemistry and performed rheological studies. A frequency sweep test was performed (0.1 to 10 rad s^{-1}), and a higher value of G′ (120 Pa) was found than G" (9.8 Pa) (G′ > G″) [43].

To check the shear-thinning behavior, a continuous flow sweep test (1–100 s^{-1}, 1% constant strain) was performed at 37 °C, in which injectable hydrogels were sandwiched between the parallel plates of the instrument and permitted to squeeze at progressively higher shear rates. IH-0.5-5 has the highest concentration of NG-CHO-0.5, so the highest viscosity value at the beginning is compared to IH-0.5-1 and IH-0.5-3. It is observed that as the shear rate increases, the

viscosity reduced (800,680 ± 576 to 497.27 ± 31 m Pa. s for IH-0.5-5, 657,740 ± 482 to 329.96 ± 25 m Pa. s for IH-0.5-3, and 226,430 ± 354 to 131.83 ± 14 m Pa. s for IH-0.5-1), which shows shear-thinning behavior of injectable hydrogel. This behavior may be due to the breakage of Schiff base linkage at a high shear rate (Fig. 3d). The shear-thinning behavior of injectable hydrogel further supports the data on syringeability and injectability. The qualitative test of syringeability and injectability are given in Supplementary Materials (2.2). Wei and colleagues fabricated an injectable hydrogel using *N*-carboxyethyl chitosan and oxidized konjac glucomannan by Schiff base crosslinks. A flow sweep test was conducted, and the results depicted a decrease in viscosity with an increased shear rate [44].

Further, the self-healing properties of synthesized injectable hydrogels were quantitatively determined using a series of rheological recovery experiments. Oscillatory step-strain measurements were used to assess the injectable hydrogels' self-healing capacity. At a low strain value (1%), the storage modulus has a high magnitude value; as the strain shifted to a high value (200%), the storage modulus value was observed near the loss modulus in all injectable hydrogels. This was because the structure got distorted at a high strain value. The behaviors remained consistent across three cycles, demonstrating the dynamic Schiff base and dynamic coordination network's stability and self-healing ability. The percentage recovery after the third cycle was calculated using Eq. 5. The high crosslink density of IH-0.5-5 is responsible for the self-healing property. After all cycles of alternate strain (low to high), the highest recovery was found in the case of IH-0.5-5 (68% ± 5%, Fig. 3e3), followed by IH-0.5-3 (65% ± 4%, Fig. 3e2) and IH-0.5-1 (54% ± 2%, Fig. 3e1) for a given interval of time (Eq. 8)

$$\text{Recovery } \% \ = 100 - \frac{\text{initial } G' - \text{final } G'}{\text{initial } G'} \times 100 \tag{8}$$

Guo et al. synthesized an injectable hydrogel using Schiff base chemistry in which alternate strain sweep shows the self-healing behavior using high (500%) and low (1%) strain [45]. Self-healing property was also checked qualitatively (Supplementary Materials (2.1)).

Morphological evaluation by FESEM

The surface morphology of the IHs with three different concentrations of NG-CHO-0.5 (1%, 3%, and 5% w/v) was examined through FESEM (Fig. 4). The FESEM images show that there is a correlation between the IHs composition and the structure. The average pore diameter was calculated using histograms of Image J software. FESEM pictures reveal that the IHs display a continuous three-dimensional porous network at varying NG-CHO-0.5 concentrations (1% to 5% w/v). The IH-0.5-5 exhibits a tight network structure

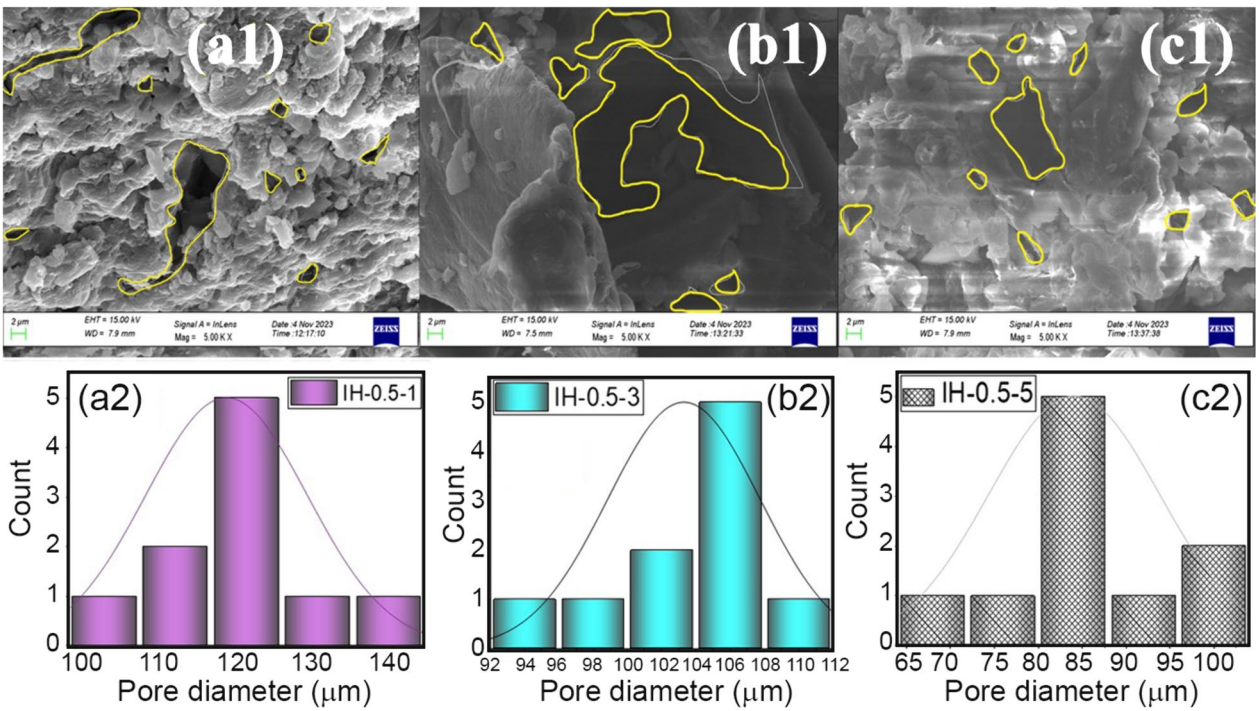


Fig. 4 FESEM images at different magnifications of: (a1) IH-0.5-1, (b1) IH-0.5-3, and (c1) IH-0.5-5. Histogram of (a2) IH-0.5-1, (b2) IH-0.5-3, and (c2) IH-0.5-5

and relatively smaller pore size (80 ± 10 µm) (Figs. 4c1 and 4c2) when compared to IH-0.5-3 (105 ± 4.45 µm, Figs. 4b1 and 4b2) and IH-0.5-1 (120 ± 10.22 µm, Figs. 4a1 and 4a2). This is because, when the NG-CHO-0.5 concentration in the IHs rose, the porosity rapidly reduced (71% ± 2% to 47% ± 2%, Table 3). The IHs network of interconnected pores may offer enough voids for the loading of drugs and also help with their release. Guan et al. fabricated an injectable hydrogel using Schiff base chemistry between oxidized dextran and gelatin, in which an SEM study revealed the surface pore size was 40 µm [50].

Swelling study of IHs

The swelling behavior of synthesized IHs is investigated in PBS at three pH levels (5.5, 7.4, and 8.5) for both dry and gel forms. In the case of the dry form of IHs, with decreasing concentrations of NG-CHO-0.5 and at a low pH (5.5), the swelling ratio rapidly increases. IH-0.5-1, which has a lower NG-CHO-0.5 concentration, shows a swelling ratio of roughly 611% ± 20%. While the swelling ratios of the hydrogels IH-0.5-3 and IH-0.5-5 were around 313% ± 19.42% and 162% ± 23%, respectively (Fig. 5a). In physiological pH (7.4), the swelling ratio values of IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 327% ± 12%, 198% ± 12%, and 158% ± 10%, respectively (Fig. 5b). Swelling ratios are also determined in alkaline pH (8.5) and the values for IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 394% ± 14%, 335% ± 12%, and 190% ± 10%, respectively (Fig. 5c). In the case of the freshly synthesized gel form of IHs, the swelling ratio increases quickly in an acidic environment (pH 5.5) with decreasing concentration of NG-CHO-0.5. The IH-0.5-1, which contains less NG-CHO-0.5, has a swelling ratio of about 476% ± 20%. The IH-0.5-3 and IH-0.5-5, on the other hand, have swelling ratios of about 250% ± 21.2% and 190% ± 27%, respectively (Fig. 5d). Figure 5e shows the swelling ratios for IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 253% ± 14%, 192% ± 12%, and 161% ± 10%, respectively, at physiological pH 7.4. The values for IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 306% ± 14%, 233% ± 10%, and 206% ± 12% for swelling ratios in alkaline pH of 8.5 (Fig. 5f). The IHs network structure becomes more compact as NG-CHO-0.5 concentration increases (1% to 5% w/v). As a result, the mobility of the polymer chains is more severely constrained, resulting in less exposure to water. The amino groups in CMCh are hydrophilic, making it simple to hydrate with water quickly. On the other hand, the imine group in the IHs is considered hydrophobic. Therefore, IHs may have more hydrophobic moieties due to the higher concentration of NG-CHO-0.5. As a result, the injectable hydrogel's capacity to make hydrogen bonds with water decreases, lowering its swelling rate. As a result, the hydrogel IH-0.5-5, which has a higher concentration of NG-CHO-0.5, displays less swelling ratio comparatively to IH-0.5-1 and IH-0.5-3 in both dry and gel forms in all pH ranges (5.5, 7.5 and 8.5). Additionally, swelling ratios of IH-0.5-1, IH-0.5-3, and IH-0.5-5 (dry and gel forms) decrease with increasing pH (5.5–7.4) and further increase when pH rises to 8.5. For an acidic pH of 5.5, the amino groups of CMCh undergo protonation, which causes electrostatic repulsion between inter- and intrapolymeric chains. Thus, the ruptured structure of the IHs network improves the water uptake. However, deprotonation within the amine group of CMCh reduces the swelling ratio in IHs at pH 7.4 (higher pH). Further, the carboxyl groups of CMCh become more ionized at pH 8.5, which increases the electrostatic repulsion and creates a loose network structure in IHs (dry and gel forms), resulting in higher swelling ratios. Thus, the above studies show that the swelling ratios of dry and gel forms of IHs depend on the pH of the solution and the concentration of NG-CHO-0.5 [21, 51]. Additionally, it was observed that the swelling ratio in the dried form is higher than that in the gel form of IHs across all pH ranges. This is because there is already some extent of water molecules available. Therefore, the gel form absorbs fewer water molecules than a dried form of IHs. Furthermore, the swelling percentage required for drug delivery varies depending on the application. A higher swelling percentage is often desired to prolong drug release and improve efficacy and adherence. The higher level of swelling allows the IH's matrix to expand and release the drug over a longer period of time, maintaining therapeutic levels in the body [52, 53]. Kumar et al. synthesized an injectable hydrogel based on *Moringa oleifera* gum and carboxymethyl chitosan using Schiff base crosslinks. A swelling study was conducted for dry and gel forms of IHs. The dry form shows the highest swelling in pH 5.5 (415%) than pH 7.4 (209%). In the case of gel form, the highest swelling was observed at pH 5.5 (830%) and then at pH 7.4 (430%) [31].

Drug release behavior

Using ciprofloxacin HCl as the model drug, the injectable hydrogel's drug release behavior was evaluated in PBS at pH 5.5, 7.4, and 8.5 at 37 °C. The antibiotic drug ciprofloxacin HCl is used to treat bacterial infections of the skin, urinary tract, and respiratory tract. According to the experimental section's description, two separate approaches were used. Figure 6 depicts the drug release pattern over time. In the first approach, i.e., drug release from a dried form of IHs, the Cipro release increases quickly at low pH (pH 5.5) with decreasing concentrations of NG-CHO-0.5. The IH-0.5-1 contains the least concentration of NG-CHO-0.5, giving a Cipro release of about 9621 ± 220 µg 100 mL^{-1} g^{-1}. On the other hand, the IH-0.5-3 and IH-0.5-5, give Cipro a release of about 6489 ± 353.21 and 5012 ± 523 µg 100 mL^{-1} g^{-1}, respectively (Fig. 6a). The Cipro release for IH-0.5-1, IH-0.5-3, and IH-0.5-5 is 5125 ± 144, 4662 ± 191, and

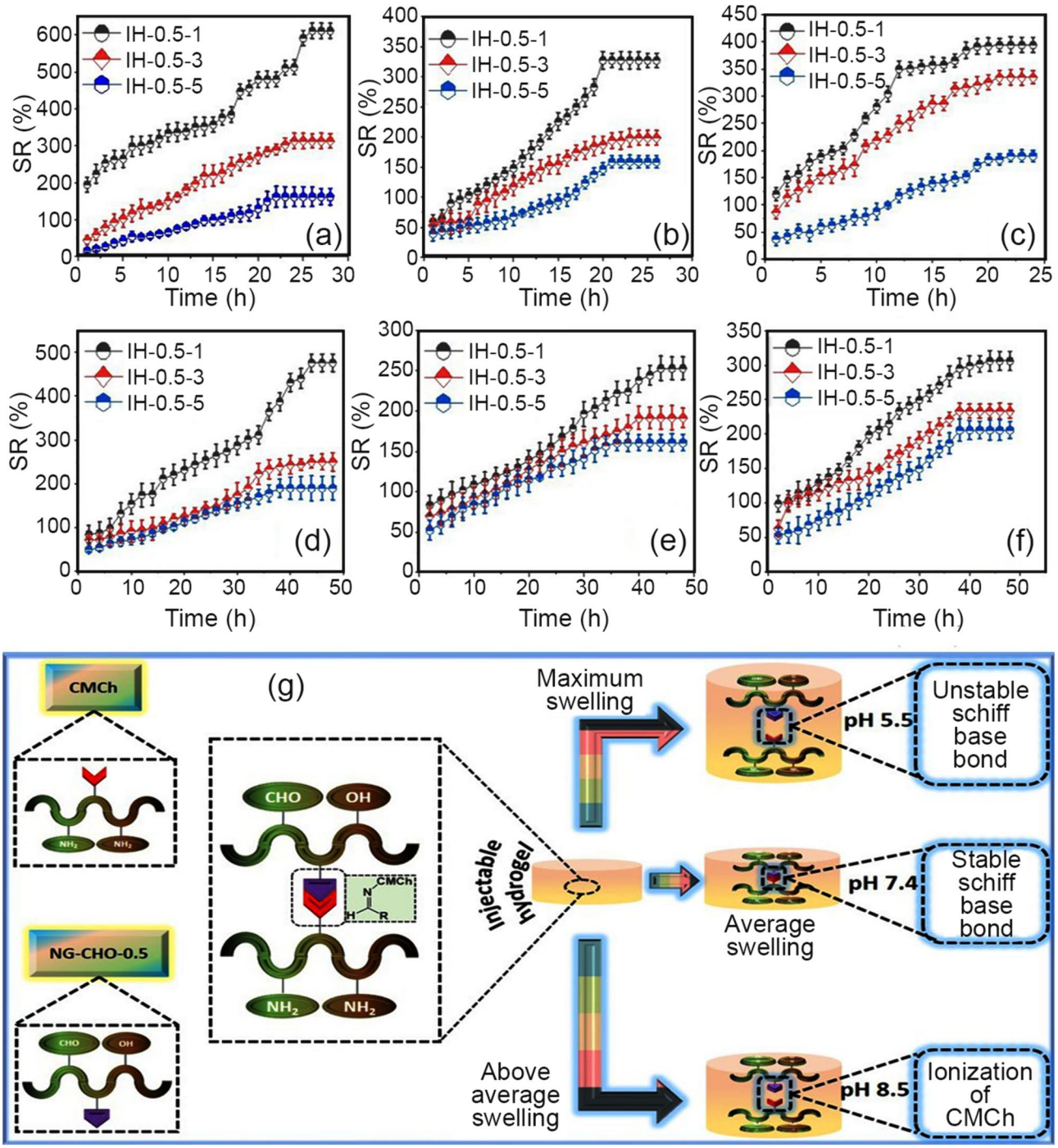


Fig. 5 Swelling behavior of IH in PBS in dry form: **a** at pH 5.5, **b** pH 7.4, and **c** pH 8.5, and in gel form: **d** at pH 5.5, **e** at pH 7.4, **f** at pH 8.5, and **g** mechanism of swelling behavior in pH 5.5, 7.4 and 8.5

3536 ± 170 µg 100 mL^{-1} g^{-1}, respectively, at physiological pH 7.4 (Fig. 6b). In alkaline media (pH 8.5), Cipro release values for IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 7165 ± 221, 6358 ± 219, and 5514 ± 148 µg 100 mL^{-1} g^{-1}, respectively (Fig. 6c). A related research article published by Wu et al., in which they fabricated an injectable hydrogel using Schiff base chemistry and 5-fluorouracil drug release studies, showed high swelling at pH 6.0 (58.9%) compared to that at pH 7.0 (80%) [54]. The second approach conducted the Cipro release through the gel form of IHs. The IH-0.5-1 includes the least concentration of NG-CHO-0.5, giving a Cipro release of $12{,}723 \pm 793$ µg 100 mL^{-1} g^{-1}.

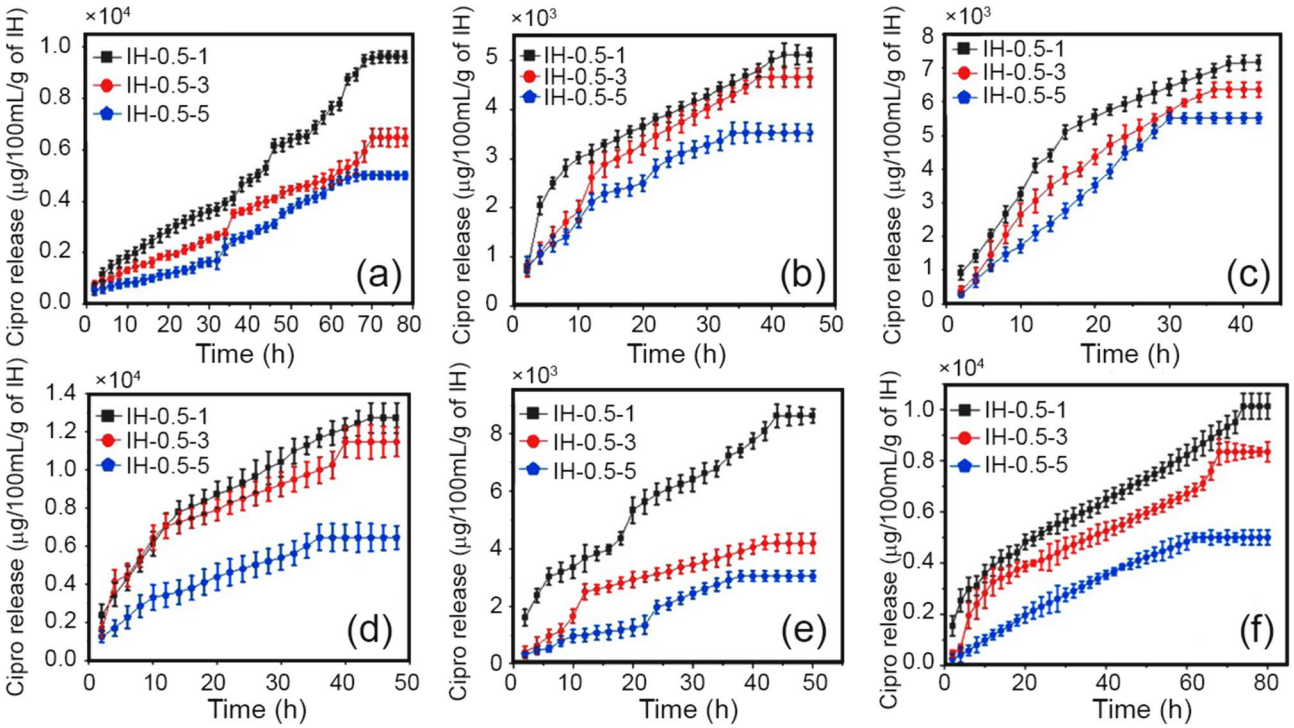


Fig. 6 Drug release behavior of IHs in PBS in dry form: **a** at pH 5.5, **b** pH 7.4, and **c** pH 8.5, and in gel form: **d** at pH 5.5, **e** at pH 7.4, and **f** at pH 8.5

Furthermore, the Cipro release values for IH-0.5-3 and IH-0.5-5 are 11,467 ± 770 and 6446 ± 70 μg 100 mL^{-1} g^{-1}, respectively (Fig. 6d). At physiological pH 7.4, the Cipro release values for IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 8626 ± 238, 4197 ± 346, and 3056 ± 190 μg 100 mL^{-1} g^{-1}, respectively (Fig. 6e). For Cipro release in the alkaline pH of 8.5, the values for IH-0.5-1, IH-0.5-3, and IH-0.5-5 are 10,127 ± 499, 8346 ± 388, and 5000 ± 289 μg 100 mL^{-1} g^{-1}, respectively (Fig. 6f).

This can be explained by the fact that as the concentration of NG-CHO-0.5 rises (from 1 to 5% w/v), the IHs network structure becomes more compact, which results in an increased crosslink density (0.7 to 1.2 g cm^{-3}). As a result, the mobility of the polymer chains is more severely limited, leading to reduced exposure to the release of Cipro. Since amino groups in CMCh are hydrophilic, it is simple to hydrate with water immediately. On the other hand, the IHs contain the imine group, which is regarded to have a hydrophobic tendency. Because there is a more significant concentration of NG-CHO-0.5 in IHs, they may contain more hydrophobic moieties. The ability of the injectable hydrogel to form hydrogen bonds with water is reduced, reducing the rate of release of Cipro. As a result, the IH-0.5-1 has a higher Cipro release than IH-0.5-3 and IH-0.5-5 in both dry and gel forms across all pH ranges (5.5, 7.5, and 8.5). Additionally, the Cipro release of IH-0.5-1, IH-0.5-3, and IH-0.5-5 (dry and gel forms) decreases as pH increases (5.5–7.4), indicating a pH-sensitive behavior. This was because the Schiff base bond was unstable at low pH (5.0) and easy to break, which caused the structure of the IHs to be destroyed, and the physically trapped drug Cipro was released along the concentration gradient. When the pH value was increased (pH 7.4), the Schiff base bond was relatively stable, and the release rate of Cipro was relatively slow due to the effect of physical interaction. Furthermore, at the pH of 8.5, CMCh carboxyl groups become more ionized, increasing electrostatic repulsion and forming a free network structure in IHs (dry and gel), increasing Cipro release. In the actual situation, when the IHs are used at the wound site, the second approach is more realistic than the first approach. The drug's slow release will help reduce microbial growth at the wound site during healing. Thus, the experiments demonstrated that the Cipro release of IHs in their dry and gel forms relies on the solution's pH and the amount of NG-CHO-0.5. The release of Cipro was found to be slower in gel form than in the dry form of IHs across all pH ranges. In dry form, PBS enters the IHs to release the drug, while in gel form, the drug follows a diffusion mechanism to release. This can be further explained using the Korsmeyer–Peppas (K–P) model (Table 4).

The K–P model provides the most accurate description of the antibiotic drug release mechanism through synthesized injectable hydrogels (Table 5). When release exponent value "n" is equal to or less than 1, it denotes non-Fickian or

Table 4 Comparison of the values of rheological behavior

References		Time sweep	Flow sweep (0.1%-100%)	Amplitude sweep (strain 0.1%-100%)	Frequency sweep (0.1–100 rad s^{-1}, strain 1%)	Step-strain analysis (low 1%, high 200%)
		G′ (up to 50 s)	Viscosity (mPa. s)	LVE region	G′ values	Recovery (%)
Neem gum and carboxymethyl chitosan injectable hydrogel (present study)	IH-0.5-1	122.99 ± 15 Pa	226,430 ± 354–13,183 ± 14	0.1–4%	188.76 ± 16 Pa	54% ± 2%
	IH-0.5-3	514.7 ± 19 Pa	657,740 ± 482–329.96 ± 25	0.1–4%	378.85 ± 20 Pa	65% ± 4%
	IH-0.5-5	584.75 ± 27 Pa	800,680 ± 576–497.27 ± 31	0.1–4%	509.03 ± 20 Pa	68% ± 5%
Hyaluronic acid and cystamine dihydrochloride [43]	–	–	–	–	120 Pa (0.1–10 rad s^{-1})	–
Carboxymethyl cellulose and poly(ethylene glycol) [46]	–	–	–	–	–	Self-healing behavior
Alginate and gelatin [47]	–	–	1890–4 Pa	–	–	–
Carboxymethyl chitosan [48]	–	–	–	0.1–1%	–	–
Pullulan and carboxymethyl chitosan [49]	–	~ 1100 Pa (up to 120 s)	–	–	–	Self-healing behavior

Table 5 Korsmeyer–Peppas (K–P) kinetic model parameters for IH-0.5-1, IH-0.5-3, and IH-0.5-5 in dry and gel forms in PBS at pH 5.5, 7.4, and 8.5

pH of PBS	Physical state of IH	Sample code	Parameters associated with the K–P model		
			kkp	n	R^2
5.5	Dry	IH-0.5-1	3.31	0.74	0.99
		IH-0.5-3	2.62	0.80	0.99
		IH-0.5-5	3.20	0.74	0.99
	Gel	IH-0.5-1	8.19	0.76	0.99
		IH-0.5-3	11.54	0.68	0.99
		IH-0.5-5	13.51	0.54	0.99
7.4	Dry	IH-0.5-1	14.48	0.61	0.99
		IH-0.5-3	7.45	0.80	0.99
		IH-0.5-5	12.83	0.60	0.99
	Gel	IH-0.5-1	14.32	0.61	0.99
		IH-0.5-3	7.45	0.80	0.99
		IH-0.5-5	12.83	0.60	0.99
8.5	Dry	IH-0.5-1	5.83	0.90	0.99
		IH-0.5-3	4.71	0.92	0.99
		IH-0.5-5	3.74	0.94	0.99
	Gel	IH-0.5-1	5.46	0.93	0.99
		IH-0.5-3	9.62	0.51	0.99
		IH-0.5-5	2.42	0.92	0.99

anomalous diffusion, a combination of diffusion- and erosion-controlled rate release. When $n = 0.5$, it denotes Fickian diffusion. Case II is indicated when $n = 1$, and super case transport 2 is shown when $n > 1$. For the release behavior of all the injectable hydrogels, the value for R^2 is found to be more than 0.99. The "n" value was more than 0.5 and less than 1, observed in all synthesized IHs. Therefore, the "n" value confirmed that all IHs showed non-Fickian or anomalous diffusion across all pH ranges in dry and gel forms. Additionally, the diffusion rate is higher than the relaxation rate of the hydrogel matrix in all cases. Li and researchers reported that the 5-fluorouracil drug release from injectable hydrogel generated by oxidized pectin and chitosan follows mostly Fickian diffusion at the pH of 6.8 and non-Fickian diffusion at pH of 7.4 [55].

In vitro cytotoxicity evaluation

The main requirement for injectable hydrogels to function as biomaterials for drug delivery applications is their biocompatibility with the host cells. Thus, the cytotoxicity of formulated IHs was investigated toward the L929 fibroblast cell line. The MTT assay of the IH-0.5-1, IH-0.5-3, and IH-0.5-5 exhibits insignificant toxicity with cell viability of 54% ± 1.5%, 60% ± 4.05%, and 67% ± 6.24% up to 500 µg mL^{-1}, respectively (Fig. 7a). This may be due to the presence of unreacted carbonyl groups present in the NG-CHO-0.5, which can interact with the cell proteins. A slight

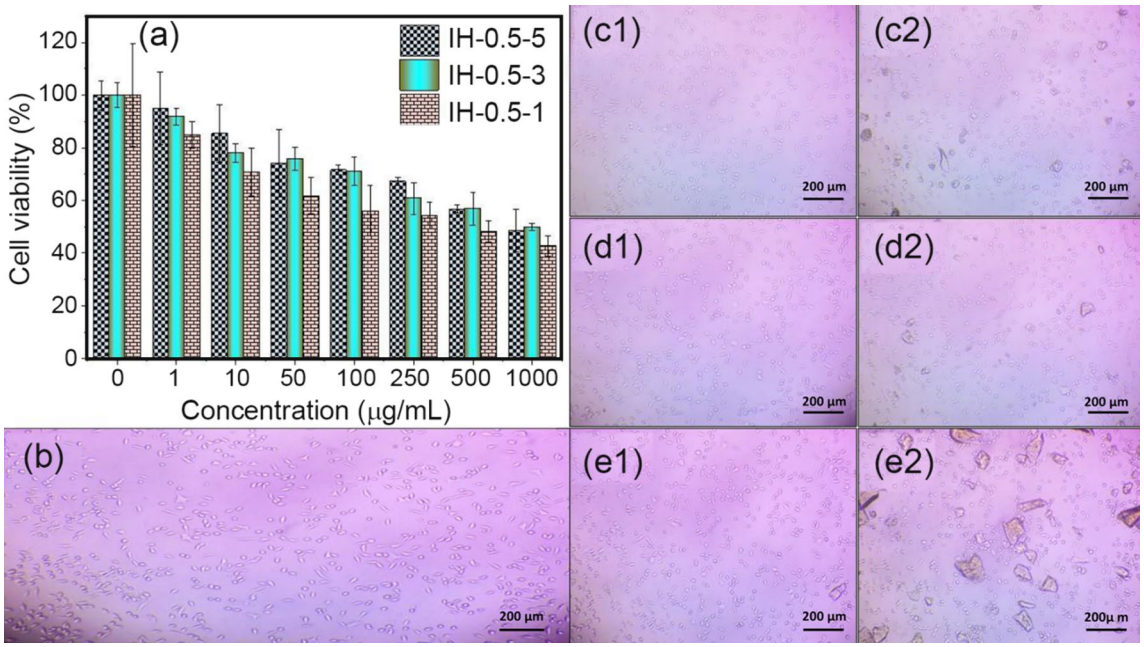


Fig. 7 **a** Cytotoxicity of injectable hydrogels. Inverted phase microscopic images of **b** control. IH-0.5-1 treated L-929 cell line at (c1) 1 μg mL^{-1}, and (c2) 500 μg mL^{-1}. IH-0.5-3 treated L-929 cell line at (d1)1 μg mL^{-1}, and (d2) 500 μg mL^{-1}. IH-0.5-5 treated L-929 cell line at (e1) 1 μg mL^{-1}, and (e2) 500 μg mL^{-1}

decrease in cell viability was observed as the concentration of NG-CHO-0.5 decreased [56]. Figures 7b-7e2 show synthesized inverted phase microscopic pictures of IHs. The results show that all injectable hydrogels are compatible with fibroblasts. Therefore, it can be said that these injectable hydrogels have exhibited good biocompatibility toward L-929 cell lines.

Antibacterial assay

Injectable hydrogels must be antibacterial to prevent bacteria from entering the wound, eliminate microbes, and reduce inflammation. The antibacterial properties of Schiff base IHs are pH-responsive, with the highest activity observed at acidic pH levels. This is due to the reversible nature of the Schiff base linkages, which are more stable under acidic conditions [57]. In this work, we assessed the IH's bactericidal properties using *S. aureus* and *E. coli* for 24 h at 37 °C. Figure 8 and Table S1 (in Supplementary Materials) display the antibacterial test results of positive control along with IH-0.5-1, IH-0.5-3, and IH-0.5-5 (50 to 1000 μg/disc). All IHs exhibited inhibitory effects against each bacterium across all concentrations up to 24 h at 37 °C. These inhibitory effects have come from two major approaches. First, chitosan has inherent antibacterial properties due to its polycationic nature and ability to disrupt bacterial cell walls. Since the amino group in CMCh exhibits a positive electrical charge, it traps bacteria with negative charges through electrostatic forces, thus effectively inhibiting bacterial growth. Additionally, it could induce the leakage of proteinaceous and other intracellular components from bacteria. Second, dynamic imine bonds have been proven to have certain antibacterial functions. Therefore, the combination of active Schiff base and protonated amino group in CMCh endows IHs with excellent antibacterial properties by defending the cell wall from the dispersion of active ingredients toward the cytoplasm. Thus, the imine bond in IHs further enhances the antibacterial effects. Li et al. synthesized an injectable hydrogel based on hyaluronic acid and chitosan through enzyme-catalyzed and Schiff-based crosslinks, showing excellent antibacterial results against *S. aureus* and *E. coli* [58].

Hydrolytic degradation in PBS at pH of 7.4

The imine bond undergoes hydrolytic cleavage in PBS (pH 7.4) at 37 °C. This hydrolytic degradation results in the breakdown of the polymer network and the formation of smaller molecular weight fragments. As the hydrolytic degradation progresses, the smaller molecular weight fragments produced become soluble in the PBS (pH 7.4) at 37 °C and dissolve or disperse into the liquid medium. It is important to note that degradation and dissolution can occur simultaneously or sequentially [59]. The hydrolytic biodegradability of synthesized IHs was studied in PBS (pH 7.4) at 37 °C for 30 days. Two approaches were used to quantify the result

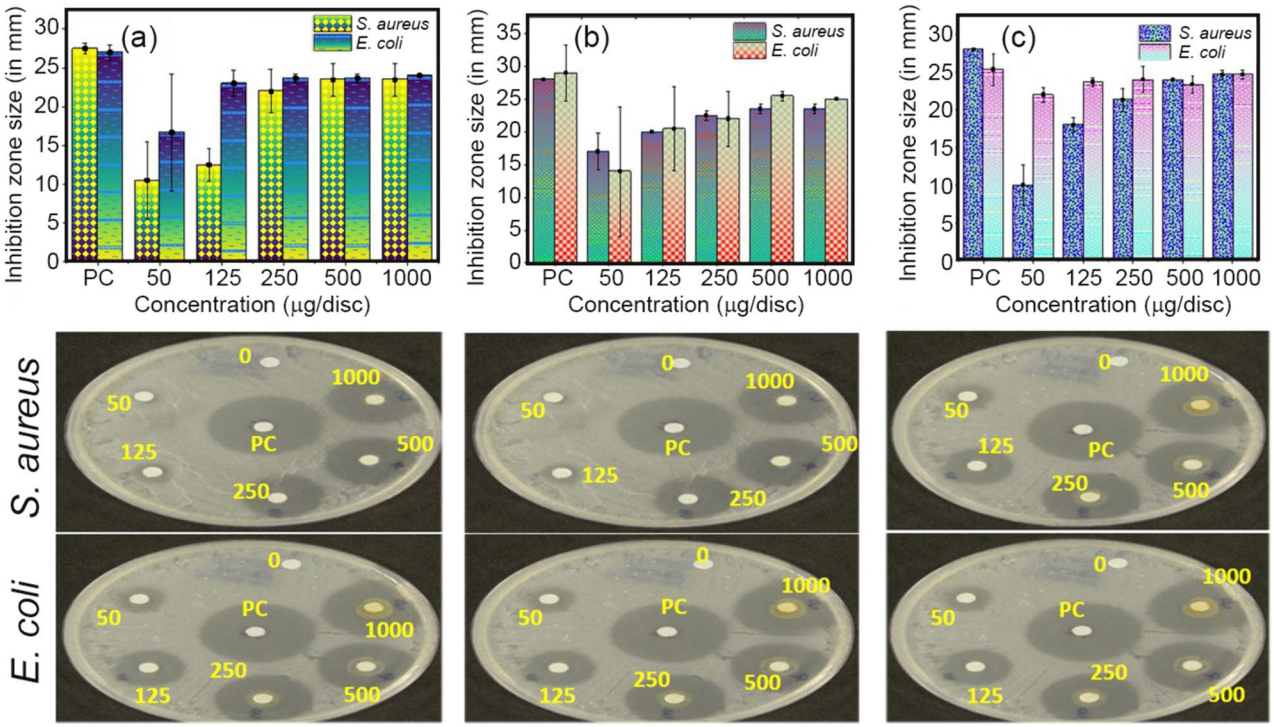


Fig. 8 Antibacterial activity of: **a** IH-0.5-1, **b** IH-0.5-3, and **c** IH-0.5-5

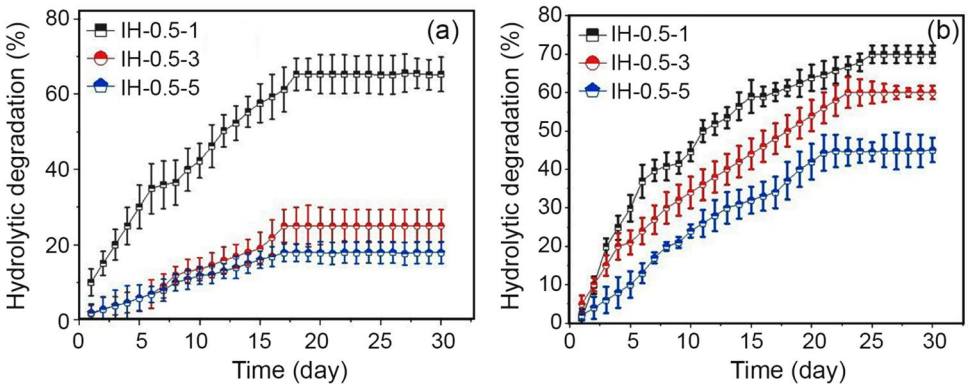


Fig. 9 Hydrolytic degradation behavior of IHs in PBS at pH 7.4: **a** dry form, and **b** gel form

(Fig. 9). The first approach conducted a hydrolytic degradation study using the dry form of IHs. The IH with the highest concentration of oxidized neem gum (IH-0.5-5) degraded the least (20% ± 4%). Meanwhile, the IH with the lowest concentration of oxidized neem gum (IH-0.5-1) was the most degraded (65% ± 5%). Additionally, IH-0.5-3 showed 25% ± 4% degradation (Fig. 9a). In the second approach, freshly synthesized gel forms of IHs were used to conduct the study. Here, the same type of degradation behavior was observed. The IH, having the highest concentration of oxidized neem gum (IH-0.5-5), observed 45% ± 7% degradation. At the same time, IH, which has the lowest concentration of oxidized neem gum (IH-0.5-1), observed 70% ± 3% degradation. Additionally, the concentration of oxidized neem gum between these two (IH-0.5-3) shows 60% ± 10% degradation (Fig. 9b). The correlation between the degradation rate and crosslinking density can explain this finding. The concentration of NG-CHO-0.5 and crosslinking density significantly impact the hydrolytic degradation of these IHs. As the concentration of NG-CHO-0.5 was increased, the crosslink density increased, making it more difficult for water molecules to penetrate and cleave the Schiff base bond of IHs. As a result, the rate of hydrolytic degradation was declined. Pandit et al. fabricated an injectable hydrogel using N, O-carboxymethyl chitosan and multialdehyde guar gum crosslink through the Schiff base mechanism; the IH having higher concentrations of oxidized guar gum degraded the least (85%) and the IH having most minuscule amount of

oxidized guar gum degraded the most (~99%) up to 24 days at 37 °C and at a pH of 7.4 [21].

Conclusion

This is the first study to report the use of multialdehyde neem gum in injectable hydrogel synthesis. The biodegradable, biocompatible, and self-healing IHs have been fabricated using dynamic Schiff base linkages between multialdehyde groups of oxidized neem gum and carboxymethyl chitosan. The optimized concentration of sodium periodate is 0.5% w/v, which converts the hydroxyl group to the aldehyde group of neem gum. The ideal ratio of the concentration of NG-CHO-0.5 to CMCh decides the least gelation time (18 ± 1.6 to 3 ± 0.5 min), which seems enough for administration. The synthesized IHs are ideal for syringeability and injectability evaluation because of their exceptional viscoelastic, shear-thinning, and self-healing properties. Because of the self-healing behavior, IHs rapidly recover their mechanical integrity to provide a protective, conformable matrix. The swellability and drug release behavior of IHs are pH-dependent and well correlated with crosslink density. The IHs showed drug release behavior using the model drug ciprofloxacin HCl, with a more significant release in PBS at pH 5.5 compared to pH 7.4 and 8.5. The prolonged drug release is found in the gel form of IHs, which supports wound healing. The drug release behavior follows the non-Fickian diffusion mechanism. In vitro, cytotoxicity evaluation toward L929 fibroblast cell lines and antibacterial assay favor the biocompatibility of synthesized injectable hydrogels. The IHs are observed to have better hydrolytic degradation in PBS (pH 7.4) at 37 °C.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13726-024-01386-7.

Acknowledgements The authors greatly acknowledge Delhi Technological University, Bawana Road, Delhi-110042, for providing financial assistance to conduct the research work.

Data availability The data are available upon reasonable request from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Kumar RN, Rao KG, Murthy RK (2022) Studies on the applicability of natural gums, neem gum and tamarind gum as release controlling polymers. J Pharm Negat Results 13:2022
2. Mankotia P, Choudhary S, Sharma K, Kumar V, Bhatia JK, Parmar A, Sharma S, Sharma V (2020) Neem gum based pH responsive hydrogel matrix: a new pharmaceutical excipient for the sustained release of anticancer drug. Int J Biol Macromol 142:742–755
3. Nasrine A, Narayana S, Ahmed GM, Sultana R, Noushida N, Salian TR, Almuqbil M, Almadani ME, Alshehri A, Alghamdi A, Alshehri S, Asdaq SMB (2023) Neem (Azadirachta Indica) and silk fibroin associated hydrogel: boon for wound healing treatment regimen. Saudi Pharm J 31:101749
4. Singh B, Ram K, Singh B (2022) Development and characterization of azadirachta indica gum-poly(2-hydroxyethyl methacrylate) crosslinked co-polymeric hydrogels for drug delivery applications. Chem Phys Lett 792:139401
5. Singh B, Singh B (2020) Graft copolymerization of polyvinylpyrollidone onto Azadirachta indica gum polysaccharide in the presence of crosslinker to develop hydrogels for drug delivery applications. Int J Biol Macromol 159:264–275
6. Singh B, Singh B (2021) Design of dietary fiber azadiracta indica gum based hydrogels for use in drug delivery. Food Hydrocoll Heal 1:100011
7. Sharma S, Bal T (2023) Exploring and characterizing the drug release potential of novel tissue compatible amphiphilic graft copolymer of polyvinyl acetate grafted neem gum (NG-*g*-PVAc). J Polym Environ 31:2364–2379
8. Singh B, Kumar R, Kumar S, Thakur N, Ram K (2023) Design of polysaccharide gum based network copolymeric hydrogels for drug delivery and wound dressing applications. Med Nov Technol Devices 20:100262
9. Singh B, Ram K (2023) Exploring neem gum in designing hydrogel dressings by covalent and supra-molecular interactions for better wound healing. Results Surf Interfaces 13:10016110
10. Hassan HT, Ismaeel GL, Jalil AT, Hadi WH, Jasim IK, Almulla AF, Radhea ZA (2023) Advanced injectable hydrogels for bone tissue regeneration. Biophys Rev 15:223–237
11. Tian W, Wang C, Chu R, Ge H, Sun X, Li M (2023) Injectable hydrogel nanoarchitectonics with near-infrared controlled drug delivery for in situ photothermal/endocrine synergistic endometriosis therapy. Biomater Res 27:1–16
12. Savetsakulanont J, Chalitangkoon J, Monvisade P (2021) Stimuli-responsive, self-healing, and injectable hydrogels with dual-crosslinked design from phenolphthalein-grafted *N*-carboxyethyl chitosan. Macromol Mater Eng 306:1–11
13. Shi Z, Yang F, Pang Q, Hu Y, Wu H, Yu X, Chen X, Shi L, Wen B, Xu R, Hou R, Liu D, Pang Q, Zhu Y (2023) The osteogenesis and the biologic mechanism of thermo-responsive injectable hydrogel containing carboxymethyl chitosan/sodium alginate nanoparticles towards promoting osteal wound healing. Int J Biol Macromol 224:533–543
14. Wang N, Ma J, Song W, Zhao C (2023) An injectable hydrogel to disrupt neutrophil extracellular traps for treating rheumatoid arthritis. Drug Deliv 30:2173332
15. Fazal T, Murtaza BN, Shah M, Iqbal S, Rehman M, Jaber F, Dera AA, Awwadi NS, Ibrahiumj HA (2023) Recent developments in natural biopolymer based drug delivery systems. RSC Adv 13:23087–23121
16. Shekhar S, Chaudhary V, Sharma B, Kumar A, Bhagi AK, Singh KP (2023) Sustainable polysaccharide hydrogels based on dynamic schiff base linkages as versatile building blocks for fabricating advanced functional materials. J Polym Environ 31:1257–1278
17. Dong X, Zhao Y, Zhao Z, Zhang X (2023) Injectable hydrogels for the repair and regeneration of articular cartilage. Polym Sci Ser B 65:555–577
18. Nguyen HTT, Do NHN, Lac HD, Nguyen PLN, Le PK (2023) Synthesis, properties, and applications of chitosan hydrogels as anti-inflammatory drug delivery system. J Porous Mater 30:655–670

19. Kou S, (Gabriel), Peters LM, Mucalo MR, (2021) Chitosan: a review of sources and preparation methods. Int J Biol Macromol 169:85–94
20. Li L, Xiang F, Wang F, Liu Y (2024) Preparation and antitumor study of intelligent injectable hydrogel: carboxymethyl chitosan–aldehyde gum Arabic composite graphene oxide hydrogel. Int J Biol Macromol 259:129429
21. Pandit AH, Nisar S, Imtiyaz K, Nadeem M, Mazumdar N, Rizvi MA, Ahmad S (2021) Injectable, self-healing, and biocompatible N, O -carboxymethyl chitosan/multialdehyde guar gum hydrogels for sustained anticancer drug delivery. Biomacromol 22:3731–3745
22. Lin X, Long H, Zhong Z, Ye Q, Duan B (2024) Biodegradable chitin nanofiber-alginate dialdehyde hydrogel: An injectable, self-healing scaffold for anti-tumor drug delivery. Int J Biol Macromol 270:132187
23. Amiryaghoubi N, Fathi M, Safary A, Javadzadeh Y, Omidi Y (2024) In situ forming alginate/gelatin hydrogel scaffold through Schiff base reaction embedded with curcumin-loaded chitosan microspheres for bone tissue regeneration. Int J Biol Macromol 256:128335
24. Zhang M, Chen G, Lei M, Lei J, Li D, Zheng H (2021) A pH-sensitive oxidized-dextran based double drug-loaded hydrogel with high antibacterial properties. Int J Biol Macromol 182:385–393
25. Balitaan JNI, Hsiao CD, Yeh JM, Santiago KS (2020) Innovation inspired by nature: biocompatible self-healing injectable hydrogels based on modified-β-chitin for wound healing. Int J Biol Macromol 162:723–736
26. Cheng F, Wu Y, Li H, Yan T, Wei X, Wu G, He J, Huang Y (2019) Biodegradable N, O-carboxymethyl chitosan/oxidized regenerated cellulose composite gauze as a barrier for preventing postoperative adhesion. Carbohydr Polym 207:180–190
27. Rasool N, Yasin T, Heng JYY, Akhter Z (2010) Synthesis and characterization of novel pH-, ionic strength and temperature- sensitive hydrogel for insulin delivery. Polymer 51:1687–1693
28. Liu C, Wu J, Gan D, Li Z, Shen J, Tang P, Luo S, Li P, Lu X, Zheng W (2020) The characteristics of mussel-inspired nHA/OSA injectable hydrogel and repaired bone defect in rabbit. J Biomed Mater Res Part B Appl Biomater 108:1814–1825
29. Sharma PK, Halder M, Srivastava U, Singh Y (2019) Antibacterial PEG-chitosan hydrogels for controlled antibiotic/protein delivery. ACS Appl Bio Mater 2:5313–5322
30. Pacelli S, Acosta F, Chakravarti AR, Samanta SG, Whitlow J, Modaresi S, Ahmed RPH, Rajasingh J, Paul A (2017) Nanodiamond-based injectable hydrogel for sustained growth factor release: preparation, characterization and in vitro analysis. Acta Biomater 58:479–491
31. Kumar J, Purwar R (2024) Self-healing, biocompatible injectable hydrogel based on multialdehyde moringa oleifera gum and carboxymethyl chitosan: a suitable platform for drug delivery in wound healing application. ChemistrySelect 9:e202400309
32. Hozumi T, Kageyama T, Ohta S, Fukuda J, Ito T (2018) Injectable hydrogel with slow degradability composed of gelatin and hyaluronic acid cross-linked by schiff's base formation. Biomacromol 19:288–297
33. Nair S, Remya NS, Remya S, Nair PD (2011) A biodegradable in situ injectable hydrogel based on chitosan and oxidized hyaluronic acid for tissue engineering applications. Carbohydr Polym 85:838–844
34. Su WY, Chen KH, Chen YC, Lee YH, Tseng CL, Lin FH (2011) An injectable oxidated hyaluronic acid/adipic acid dihydrazide hydrogel as a vitreous substitute. J Biomater Sci Polym Ed 22:1777–1797
35. Pandit AH, Mazumdar N, Imtiyaz K, Rizvi KMA, Ahmad S (2020) Self-Healing and injectable hydrogels for anticancer drug delivery: a study with multialdehyde gum arabic and succinic anhydride chitosan. ACS Appl Bio Mater 3:8460–8470
36. Yan S, Wang T, Li X, Jian Y, Zhang K, Lia G, Yin J (2017) Fabrication of injectable hydrogels based on poly(l-glutamic acid) and chitosan. RSC Adv 7:17005–17019
37. Lucas de Lima E, Fittipaldi Vasconcelos N, da Silva MJ, Andrade FK, Vieira RS, Feitosa JPA (2020) Injectable hydrogel based on dialdehyde galactomannan and *N*-succinyl chitosan: a suitable platform for cell culture. J Mater Sci Mater Med 31:5
38. Liang Y, Zhao X, Ma PX, Guo B, Du Y, Han X (2019) pH-Responsive injectable hydrogels with mucosal adhesiveness based on chitosan-grafted-dihydrocaffeic acid and oxidized pullulan for localized drug delivery. J Colloid Interface Sci 536:224–234
39. Ma L, Tan Y, Chen X, Ran Y, Tong Q, Tang L, Su W, Wang X, Li X (2022) Injectable oxidized alginate/carboxylmethyl chitosan hydrogels functionalized with nanoparticles for wound repair. Carbohydr Polym 293:119733
40. Yan Y, Guan S, Wang S, Xu J, Sun C (2022) Synthesis and characterization of protocatechuic acid grafted carboxymethyl chitosan with oxidized sodium alginate hydrogel through the Schiff's base reaction. Int J Biol Macromol 222:2581–2593
41. Manna PJ, Mitra T, Pramanik N, Kavitha V, Gnanamani A, Kundu PP (2015) Potential use of curcumin loaded carboxymethylated guar gum grafted gelatin film for biomedical applications. Int J Biol Macromol 75:437–446
42. Daoub RMA, Elmubarak AH, Misran M, Hassan EA, Osman ME (2018) Characterization and functional properties of some natural Acacia gums. J Saudi Soc Agric Sci 17:241–249
43. Li S, Pei M, Wan T, Yang H, Gu S, Tao Y, Liu X, Zhou Y, Xu W, Xiao P (2020) Self-healing hyaluronic acid hydrogels based on dynamic Schiff base linkages as biomaterials. Carbohydr Polym 250:116922
44. Wei X, Cui C, Fan C, Wu T, Li Y, Zhang X, Wang K, Pang Y, Yao P, Yang J (2022) Injectable hydrogel based on dodecyl-modified *N*-carboxyethyl chitosan/oxidized konjac glucomannan effectively prevents bleeding and postoperative adhesions after partial hepatectomy. Int J Biol Macromol 199:401–412
45. Guo H, Huang S, Yang X, Wu J, Kirk TB, Xu J, Xu A, Wei Xue W (2021) Injectable and self-healing hydrogels with double-dynamic bond tunable mechanical, gel-sol transition and drug delivery properties for promoting periodontium regeneration in periodontitis. ACS Appl Mater Interfaces 13:61638–61652
46. Li L, Wang L, Luan X, Pang Y, Zhang K, Cheng Y, Ji Z, Pang J (2023) Adhesive injectable cellulose-based hydrogels with rapid self-healing and sustained drug release capability for promoting wound healing. Carbohydr Polym 320:121235
47. Ma W, Dong W, Zhao S, Du T, Wang Y, Yao J, Liu Z, Sun D, Zhang M (2022) An injectable adhesive antibacterial hydrogel wound dressing for infected skin wounds. Biomater Adv 134:112584
48. Zhou T, Zhou H, Wang F, Zhang P, Shang J, Shi L (2024) An injectable carboxymethyl chitosan hydrogel scaffold formed via coordination bond for antibacterial and osteogenesis in osteomyelitis. Carbohydr Polym 324:121466
49. Zhao Y, Li R, Liu Y, Song L, Gao Z, Li Z, Peng X, Wang P (2023) An injectable, self-healable, antibacterial, and pro-healing oxidized pullulan polysaccharide/carboxymethyl chitosan hydrogel for early protection of open abdominal wounds. Int J Biol Macromol 250:126282
50. Guan S, Zhang K, Cui L, Liang J, Li J, Guan F (2021) Injectable gelatin/oxidized dextran hydrogel loaded with apocynin for skin tissue regeneration. Mater Sci Eng C 133:112604
51. Mohamed NA, El-Ghany NAA (2016) Swelling behavior of cross-linked terephthaloyl thiourea carboxymethyl chitosan hydrogels. Cellul Chem Technol 50:463–471

52. Chen Y-C, Ho H-O, Liu D-Z, Siow WS, Sheu MT (2015) Swelling/floating capability and drug release characterizations of gastroretentive drug delivery system based on a combination of hydroxyethyl cellulose and sodium carboxymethyl cellulose. PLoS ONE 10:e0116914
53. Arora G, Malik K, Singh I, Arora S, Rana V (2011) Formulation and evaluation of controlled release matrix mucoadhesive tablets of domperidone using Salvia plebeian gum. J Adv Pharm Technol Res 2:163
54. Wu X, He C, Wu Y, Chen X (2016) Synergistic therapeutic effects of Schiff's base cross-linked injectable hydrogels for local co-delivery of metformin and 5-fluorouracil in a mouse colon carcinoma model. Biomaterials 75:148–162
55. Li DQ, Wang SY, Meng YJ, Li JF, Li J (2020) An injectable, self-healing hydrogel system from oxidized pectin/chitosan/γ-Fe_2O_3. Int J Biol Macromol 164:4566–4574
56. Sharma PK, Taneja S, Singh Y (2018) Hydrazone-linkage-based self-healing and injectable xanthan–poly(ethylene glycol) hydrogels for controlled drug release and 3D cell culture. ACS Appl Mater Interfaces 10:30936–30945
57. Yeh Y, Tsai Y, Wu C, Tu LH, Bai MY, Yeh YC (2022) The Role of aldehyde-functionalized crosslinkers on the property of chitosan hydrogels. Macromol Biosci 22:2200366
58. Li J, Su J, Liang J, Zhang K, Xie M, Cai B, Li J (2024) A hyaluronic acid/chitosan composite functionalized hydrogel based on enzyme-catalyzed and Schiff base reaction for promoting wound healing. Int J Biol Macromol 255:128284
59. Efstathiou S, Wemyss AM, Patias G, Shok LA, Grypioti M, Coursari D, Ma C, Atkins CJ, Shegiwal A, Wan C, Haddleton DM (2021) Self-healing and mechanical performance of dynamic glycol chitosan hydrogel nanocomposites. J Mater Chem B 9:809–823

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

Jitendra Kumar[1] · Roli Purwar[1]

✉ Roli Purwar
roli.purwar@dtu.ac.in

[1] Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University, Delhi, India

arXiv:2409.09573v1 [cs.RO] 15 Sep 2024

Decentralized Safe and Scalable Multi-Agent Control under Limited Actuation

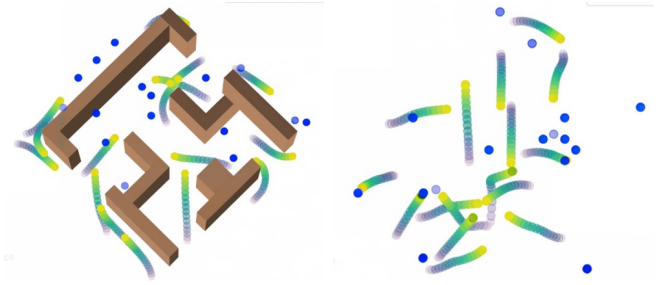
Vrushabh Zinage[1], Abhishek Jha [2], Rohan Chandra [3], Efstathios Bakolas [1]

***Abstract*— To deploy safe and agile robots in cluttered environments, there is a need to develop fully decentralized controllers that guarantee safety, respect actuation limits, prevent deadlocks, and scale to thousands of agents. Current approaches fall short of meeting all these goals: optimization-based methods ensure safety but lack scalability, while learning-based methods scale but do not guarantee safety. We propose a novel algorithm to achieve safe and scalable control for multiple agents under limited actuation. Specifically, our approach includes: (i) learning a decentralized neural Integral Control Barrier function (neural ICBF) for scalable, input-constrained control, (ii) embedding a lightweight decentralized Model Predictive Control-based Integral Control Barrier Function (MPC-ICBF) into the neural network policy to ensure safety while maintaining scalability, and (iii) introducing a novel method to minimize deadlocks based on gradient-based optimization techniques from machine learning to address local minima in deadlocks. Our numerical simulations show that this approach outperforms state-of-the-art multi-agent control algorithms in terms of safety, input constraint satisfaction, and minimizing deadlocks. Additionally, we demonstrate strong generalization across scenarios with varying agent counts, scaling up to 1000 agents. Videos and code are available at https://maicbf.github.io/.**

I. Introduction

We consider the problem of safe decentralized real-time multi-agent control in the presence of obstacles for a large number of agents under limited actuation. Such controllers are particularly important for robots that have to be deployed in cluttered environments with large number of agents such as in warehouse automation [1], autonomous driving [2], quadrotor swarms [3], [4], and automated intersection management [5] where the agents have to operate in real-time and remain safe at all times Furthermore, in these applications, real physical robots have limited actuation capabilities and cannot exceed certain control limits. Therefore, it becomes critical for the designed safe decentralized controller to satisfy input constraints to ensure the feasible deployment of robots. Lastly, as the environment gets more cluttered, resolving and preventing deadlocks becomes necessary as well.

There is a vast literature on this topic. Characteristic examples include references [6]–[20] which present either centralized or decentralized methods which have been popular choices due to their ability to guarantee safety, generalize to almost any nonlinear system and operate under limited actuation. Recently, control barrier functions (CBFs) [12] have been used increasingly used for guaranteed safety guarantees for nonlinear systems. For instance, CBF-based controllers have guaranteed safety for systems with higher relative degree [21], [22], hybrid systems [23], [24], sampled data systems [25], [26], unknown nonlinear systems [27]–[29], input delay systems [30], non-control affine systems [28], [31], [32] and systems with additive bounded disturbances. However, the main challenge in extending these approaches to multi-agent systems is that they are not scalable and deadlock-free in general. Furthermore, synthesizing a CBF for multi-agent scenarios is non-trivial in general. Learning-based methods [33]–[37] including multi-agent reinforcement learning (MARL) [38], [39], on the other hand, have demonstrated better scalability, with some methods scaling as much as up to 32 agents [37], [40], [41]. However, learning-based methods impose soft constraints (collision checking) on safety, as opposed to provable safety constraints imposed by CBFs, to generate collision-free trajectories, albeit without formal guarantees on safety, input constraint satisfaction, or deadlock resolution capabilities. Furthermore, some learning-based algorithms do not exploit the fact that governing nonlinear system equations are known either apriori or via system identification methods [41]. Consequently, these methods require large datasets comprising state-action pairs to train these MARL algorithms efficiently so that they can be competitive to control-theoretic approaches, which is impractical in many challenging real world scenarios.



(a) `Complex Maze` (b) `Empty environment`

Fig. 1: Demonstrating our decentralized safe and scalable multi-agent control in two environments. Blue and yellow denote goals and current positions, respectively.

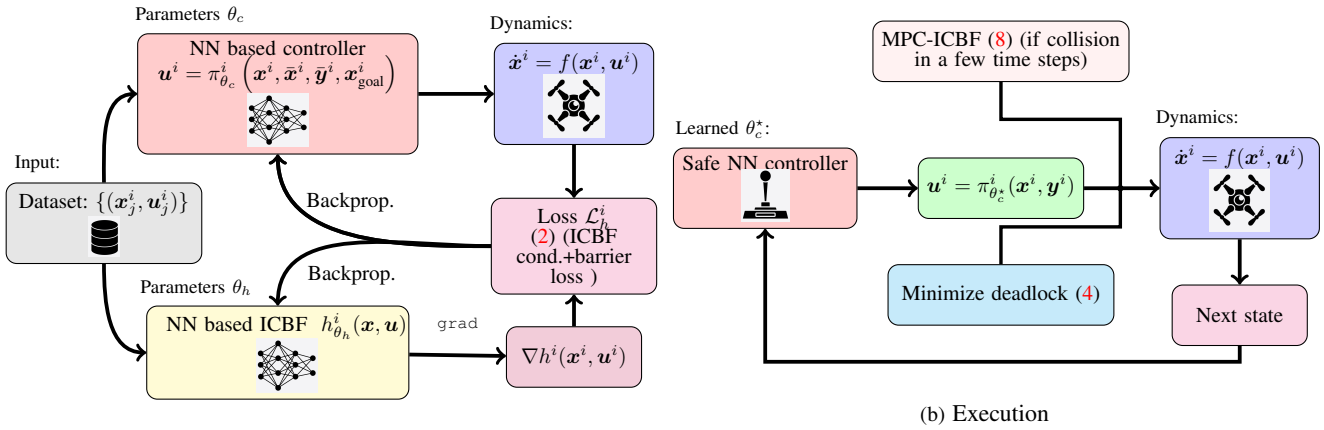
Recently, a line of research [42], [43] has combined the best of both classical optimization-based methods and learning-based methods to synthesize decentralized controllers that are scalable up to thousands of agents. The main idea of these approaches is to simultaneously learn safe feedback controllers and decentralized CBFs based neural certificates for each agent that ensure scalability and safety during planning. Although these approaches provide probabilistic guarantees on collision avoidance, they do not guarantee safety [42], [43], suffer from the problem of deadlock among these agents [42], cannot operate under limited actuation [43], and are restricted to control affine nonlinear systems [12], [43], [44]. Towards the goal of synthesizing decentralized controllers that are scalable and guarantee safety, our contributions are as follows:

1) *Safety and Scalability:* We propose a novel learning-based framework termed MA-ICBF based on Integral Control Barrier Functions (ICBFs) [28], [31] for safe multi-agent control that are scalable and are able to handle general nonlinear systems with input constraints.
2) *Limited Actuation:* As the proposed learning-based framework can only provide probabilistic guarantees, we embed a novel computationally light model predictive control (MPC) optimization-based framework termed MPC-ICBF into our approach that guarantees safety as well as input constraint satisfaction without

[1]Vrushabh Zinage and Efstathios Bakolas are with the Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin `vrushabh.zinage@utexas.edu`, `bakolas@austin.utexas.edu`

[2] Abhishek Jha is with the Department of Mechanical Engineering, Delhi Technological University, New Delhi, India `abhishekjha_me20a8_31@dtu.ac.in`

[3] Rohan Chandra is with the Department of Computer Science at the University of Virginia, USA `rohanchandra@virginia.edu`



(a) Training (b) Execution

Fig. 2: (a) Given the dataset of state-input pairs, the proposed learning framework first, jointly learns a decentralized safe controller and a decentralized ICBF for each agent $i \in V$. (b) Next, during implementation, MA-ICBF utilizes the learned safe decentralized controller to steer the agent to its goal position while at the same time guaranteeing safety (including input constraint satisfaction) and deadlock minimization via the MPC-ICBF (8) and the deadlock minimization modules (4) respectively.

compromising the overall computational time for our approach.

3) *Deadlock Resolution:* By drawing an analogy between well-established gradient-based methods for minimizing non-convex loss functions in ML and the problem of non-convexity (or the problem of local minima) that are common in deadlocks, we present a novel deadlock resolution technique to minimize deadlocks.

The overall structure of the paper is as follows. Section II discusses the preliminaries and the problem statement. Section III discusses the proposed approach followed by results in Section IV. Finally, we conclude in Section V.

II. Preliminaries and Problem Statement

A. Nomenclature

For integers a and $b(\geq a)$, the set $[a;b]_d$ denotes the set of integers $\{a, a+1, \ldots, b\}$. The superscript i on any variable indicates the variable for the i^{th} agent. The set $V = [1;N]_d$ denotes the set of indices for N agents. For a set $\mathcal{A}$, $\partial\mathcal{A}$ denotes its boundary.

B. Preliminaries and assumptions

We formulate the decentralized multi-agent control problem by $\langle N,\ M,\ \mathcal{X},\ \mathcal{X}_{\text{goal}}\ \mathcal{U}^i,\ \mathcal{O}_k,\ r,\ R,\ \mathcal{N}^i_a,\ \mathcal{N}^i_o\rangle$ where the superscript i denotes the i^{th} agent, N is the number of agents, M is the number of obstacles, $\mathcal{X} \subseteq \mathrm{R}^n$ is the state space, $\mathcal{X}_{\text{goal}} := \{\cup_{i=1}^N \boldsymbol{x}^i_{\text{goal}} \in \mathcal{X}\}$ is the set consisting of non-colliding goal states, $\mathcal{U}^i \in \mathbb{R}^m$ is the control space for the i^{th} agent, $\mathcal{O}_k \subset \mathbb{R}^n$ for $k \in [1;M]_d$ is the k^{th} obstacle space, $r > 0$ is such that the agents must keep a safe distance from the obstacles and also among themselves i.e. $\|\boldsymbol{x}^i - \boldsymbol{x}^j\| \geq 2r$, where $i \neq j$, $i, j \in V$, $R > 0$ is the limiting sensing radius for each agent, $\mathcal{N}^i_a = \{j \in V \mid \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \leq R, j \neq i\}$ is the set of neighboring agents for agent i and $\mathcal{N}^i_o$ is the set of neighboring obstacles for agent i. Each agent's motion is governed by the ordinary differential equation (ODE) $\dot{\boldsymbol{x}}^i = f(\boldsymbol{x}^i, \boldsymbol{u}^i)$, where $f : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is a continuously differentiable function, $\boldsymbol{x}^i \in \mathcal{X}$ and $\boldsymbol{u}^i \in \mathcal{U}^i$ the control input of the i^{th} agent.

The notion of Integral Control Barrier Function (ICBF) allows one to guarantee safety as well as input constraint satisfaction. For agent i, consider the integral feedback control defined as $\dot{\boldsymbol{u}} = \phi(\boldsymbol{x}, \boldsymbol{u})$, with $\boldsymbol{u}(0) := \boldsymbol{u}_0 \in \mathcal{U}$, where $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is a given smooth function. We define the safety set $\mathcal{S} := \{(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{S}_{\boldsymbol{x}} \times \mathcal{U}\}$, incorporating both state $\mathcal{S}_{\boldsymbol{x}}$ and input constraints $\mathcal{U}$. Now, consider a continuously differentiable scalar-valued function $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined such that $h(\boldsymbol{x}, \boldsymbol{u}) > 0$ if $(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{S}$, $h(\boldsymbol{x}, \boldsymbol{u}) = 0$ if $(\boldsymbol{x}, \boldsymbol{u}) \in \partial\mathcal{S}$ and $h(\boldsymbol{x}, \boldsymbol{u}) < 0$ if $(\boldsymbol{x}, \boldsymbol{u}) \in (\mathbb{R}^n \times \mathbb{R}^m) \setminus \mathcal{S}$. To guarantee that $(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{S}$, $h(\boldsymbol{x}, \boldsymbol{u})$ must satisfy the forward invariance condition i.e. $\dot{h}(\boldsymbol{x}, \boldsymbol{u}) + \alpha(h(\boldsymbol{x}, \boldsymbol{u})) \geq 0$ where $\dot{h}(\boldsymbol{x}, \boldsymbol{u})$ is computed along the system trajectories and α is a $\mathcal{K}_\infty$ function. Formally, ICBF is defined as follows:

Definition 1: (ICBF) [31] An Integral Control Barrier Function (ICBF) is defined as a function $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ that characterizes a safe set $\mathcal{S}$. Then, h is said to be an ICBF, for all $(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{S}$ if $\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{u}) = 0$, implies that $q(\boldsymbol{x}, \boldsymbol{u}) \leq 0$, where $q(\boldsymbol{x}, \boldsymbol{u}) := -\left(\nabla_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{u}) f(\boldsymbol{x}, \boldsymbol{u}) + \nabla_{\boldsymbol{u}} h(\boldsymbol{x}, \boldsymbol{u}) \phi(\boldsymbol{x}, \boldsymbol{u}) + \alpha(h(\boldsymbol{x}, \boldsymbol{u}))\right)$ and $\boldsymbol{p}(\boldsymbol{x}, \boldsymbol{u}) := \left(\nabla_{\boldsymbol{u}} h(\boldsymbol{x}, \boldsymbol{u})\right)^{\mathrm{T}}$.

The following theorem provides a method to synthesize safe controllers via ICBF's

Theorem 1: [31] If an integral feedback controller $\phi(\boldsymbol{x}, \boldsymbol{u})$ and a safety set $\mathcal{S}$, defined by an ICBF $h(\boldsymbol{x}, \boldsymbol{u})$, exist, then modifying the integral controller to $\dot{\boldsymbol{u}} = \phi(\boldsymbol{x}, \boldsymbol{u}) + \boldsymbol{v}^\star(\boldsymbol{x}, \boldsymbol{u})$, where $\boldsymbol{v}^\star(\boldsymbol{x}, \boldsymbol{u})$ is obtained by solving the following quadratic program (QP):

$$\boldsymbol{v}^\star(\boldsymbol{x}, \boldsymbol{u}) = \underset{\boldsymbol{v} \in \mathbb{R}^m}{\operatorname{argmin}} \|\boldsymbol{v}\|^2, \quad \text{s.t. } \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{u})^{\mathrm{T}} \boldsymbol{v} \geq q(\boldsymbol{x}, \boldsymbol{u}) \quad (1)$$

guarantees safety, i.e., forward invariance of the set $\mathcal{S}$.

C. Problem statement

Formally, in this paper, we consider the following problem

Problem 1: Given $< N,\ M,\ \mathcal{X},\ \mathcal{X}_{\text{goal}}\ \mathcal{U}^i,\ \mathcal{O}_k,\ r,\ R,\ \mathcal{N}^i_a,\ \mathcal{N}^i_o >$, for each agent $i \in V$, design a decentralized controller $\boldsymbol{u}^i(\boldsymbol{x}^i, \bar{\boldsymbol{x}}^i, \bar{\boldsymbol{y}}^i, \boldsymbol{x}^i_{\text{goal}})$ where $\bar{\boldsymbol{x}}^i$ represents the relative aggregated states of neighboring agents $\mathcal{N}^i_a$, $\bar{\boldsymbol{y}}^i$ the relative aggregated observations from neighboring obstacles $\mathcal{N}^i_o$ and $\boldsymbol{x}^i_{\text{goal}} \in \mathcal{X}_{\text{goal}}$ such that the following holds true:

- Input constraint satisfaction: Each agent's control policy satisfies the input constraints i.e. $\|\boldsymbol{u}^i\| \leq u_{\max}$ where $i \in V$ and $u_{\max} > 0$.
- Guaranteed safety: No collisions with obstacles (i.e., $\|\bar{\boldsymbol{y}}^i(t)\| > R$ as well as with other agents, meaning $\|\boldsymbol{x}^i(t) - \boldsymbol{x}^j(t)\| > 2r$ for $t \geq 0$, $r > 0$ and $j \neq i$.
- Minimizing the number of deadlocks i.e., the phenomenon when multiple agents stop before reaching their goal.

For the sake of brevity, we drop stating the arguments of $\boldsymbol{u}^i$ unless required. We are now in a position to discuss our proposed approach in the following section.

III. MA-ICBF : Proposed Framework

This section proposes our approach termed MA-ICBF that satisfies the objectives presented in Section II. Our approach is primarily divided into two steps. The first step involves the joint learning of the Neural Integral Control Barrier Function (NICBF) certificates and a decentralized safe control policy that enables the N agents to reach their respective goals. Since the control policies are parameterized by neural networks, they do not provide complete safety guarantees for the agents. Furthermore, as N increases, the likelihood of agents experiencing deadlock also increases. To address these two fundamental limitations, the second step involves developing theoretical frameworks that is augmented to the NN policy (in the first step) for avoiding collisions and minimizing deadlocks. To provide complete guarantees on the collision avoidance problem, we shift to a decentralized optimization-based control policy termed MPC-ICBF whenever we predict a collision in future time steps with its neighbors using the learned NN policy. Next, we analyze the deadlock problem by writing the multi-agent control problem as a QP. Then, we write the KKT conditions that provide sufficient conditions for the optimal control problem. Consequently, we draw an analogy between this deadlock problem and the problem of minimizing non-convex loss functions. Finally, we propose a method to modify the control input to ensure that the deadlocks are minimized.

A. Learning framework

Given the constants from Problem 1, the task for an agent $i \in V$ is to synthesize an ICBF $h^i_{\theta_h}(\boldsymbol{x}, \boldsymbol{u})$ and a decentralized feedback controller $\pi^i_{\theta_c}\left(\boldsymbol{x}^i, \bar{\boldsymbol{x}}^i, \bar{\boldsymbol{y}}^i, \boldsymbol{x}^i_{\text{goal}}\right)$ (where θ_h and θ_c are the parameters of the NN) that would take the agents towards their respective goal positions $\boldsymbol{x}^i_{\text{goal}} \in \mathcal{X}_{\text{goal}}$. Motivated by the fact that the neural network architecture for $h^i_{\theta_h}(\boldsymbol{x}, \boldsymbol{u})$ must adapt to changes in the size of $\bar{\boldsymbol{y}}^i$ and must be permutation invariant with respect to the neighboring agents $\mathcal{N}^i_a$, we use an architecture similar to that in [45]. Consider a time-varying observation matrix $\bar{\boldsymbol{y}}^i(t) \in \mathbb{R}^{n \times |\mathcal{N}^i_a|}$. Let $W \in \mathbb{R}^{p \times n}$ be the weight matrix and $\sigma(\cdot)$ the element-wise ReLU activation function. The mapping $\rho : \mathbb{R}^{n \times |\mathcal{N}^i_a|} \mapsto \mathbb{R}^p$ is given by $\rho\left(\bar{\boldsymbol{y}}^i\right) = \text{RowMax}\left(\sigma\left(W\bar{\boldsymbol{y}}^i\right)\right)$, where ρ maps a matrix $\bar{\boldsymbol{y}}^i$ with a dynamic column dimension and permutation to a fixed-length feature vector $\rho\left(\bar{\boldsymbol{y}}^i\right) \in \mathbb{R}^p$ and $\text{RowMax}(\cdot)$ as the row-wise max pooling operation. The dimension of $\rho\left(\bar{\boldsymbol{y}}^i\right)$ remains constant even if $\left|\mathcal{N}^i_a\right|$ changes over time. Similarly, the ICBF is parameterized, with the output being a scalar instead of a vector. The main objective is to ensure that the learned ICBF $h^i_\theta(\boldsymbol{x}, \boldsymbol{u})$ satisfies the ICBF based conditions given in Definition 1. Towards this goal, we consider minimizing the following loss function $\mathcal{L} = \sum_{j=1}^{N} \mathcal{L}^j_h(\theta)$ where $\mathcal{L}^j_h(\theta)$ for agent $j \in V$ $(\theta = \theta_c \cup \theta_h)$:

$$\mathcal{L}^j_h(\theta) = \sum_{i=1}^{N_h} \left(-p_\theta\left(\boldsymbol{x}^{j,\text{S}}_i, \boldsymbol{u}^{j,\text{S}}_i\right)^{\text{T}} \pi^j_{\theta_c} + q_\theta\left(\boldsymbol{x}^{j,\text{S}}_i, \boldsymbol{u}^{j,\text{S}}_i\right)\right)_+ + \sum_{i=1}^{N_s} \left(-h^j_{\theta_h}\left(\boldsymbol{x}^{j,\text{S}}_i, \boldsymbol{u}^{j,\text{S}}_i\right)\right)_+ + \sum_{i=1}^{N_h - N_s} \left(h^j_{\theta_h}\left(\boldsymbol{x}^{j,\text{US}}_i, \boldsymbol{u}^{j,\text{US}}_i\right)\right)_+ \tag{2}$$

where $x_+ = \max\{0, x\}$, N_h and $N_s (< N_h)$ denote the number of total state-input pairs and safe state-input pairs respectively, $\boldsymbol{x}^{j,\text{S}}_i \in \mathcal{S}^j_{\boldsymbol{x}}$, $\boldsymbol{x}^{j,\text{US}}_i \in \mathcal{X} \setminus \mathcal{S}^j_{\boldsymbol{x}}$, $\boldsymbol{u}^{j,\text{S}}_i \in \mathcal{U}$, $\boldsymbol{u}^{j,\text{US}}_i \notin \mathcal{U}$, and $p_\theta\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right)$ and $q_\theta\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right)$ are defined as:

$$p_\theta\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right) := \left(\nabla_{\pi^j_\theta} h^j_{\theta_h}\left(\boldsymbol{x}^j, \pi^j_\theta\right)\right)^{\text{T}},$$

$$q_\theta\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right) := -\Big(\nabla_{\boldsymbol{x}^j} h^j_{\theta_h}\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right) f\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right) + \nabla_{\pi^j_\theta} h^j_{\theta_h}\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right) \phi\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right) + \alpha\left(h^j_{\theta_h}\left(\boldsymbol{x}^j, \pi^j_{\theta_c}\right)\right)\Big)$$

where ϕ is an integral controller given by $\phi = \dot{k}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} k(\boldsymbol{x}) f(\boldsymbol{x}, \boldsymbol{u})$ and $k(\boldsymbol{x})$ is a nominal controller (for instance LQR or PID controller).

Data collection: Note that we do not use a fixed set of state-input pairs to train the decentralized ICBF $h^j_{\theta_h}$ and feedback controllers π^j_θ. Instead, we adopt an online training strategy, collecting data by running the current system. These pairs are temporarily stored to calculate loss and update the controllers via gradient descent. The updated controllers then generate new training data. This iterative process continues until the loss converges.

B. Deadlock analysis, deadlock resolution, and collision avoidance under limited actuation

In this section, we perform a deadlock analysis assuming that the governing dynamics is control-affine, i.e., $\dot{\boldsymbol{x}} = f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}$ with state constraints only. However, a similar analysis holds for non-control affine systems with input constraints as well. Note that each agent i computes its control input by solving the following QP (dropped the superscript i for brevity):

$$\boldsymbol{u}^\star = \underset{\boldsymbol{u} \in \mathbb{R}^m}{\text{argmin}} \ \|\boldsymbol{u} - k(\boldsymbol{x})\|_2^2 \qquad \text{s.t.} \ \ \bar{A}\boldsymbol{u} \leq \bar{\boldsymbol{q}} \tag{3}$$

where $\bar{A}$ and $\bar{\boldsymbol{q}}$ are given by

$$\bar{A} = [L_g h^1(\boldsymbol{x}), \ldots, L_g h^{i-1}(\boldsymbol{x}), L_g h^{i+1}(\boldsymbol{x}), \ldots L_g h^N(\boldsymbol{x})]$$
$$\bar{\boldsymbol{q}} = [L_f h^1(\boldsymbol{x}) + \alpha(h^1(\boldsymbol{x})), \ldots, L_f h^{i-1}(\boldsymbol{x}) + \alpha(h^{i-1}(\boldsymbol{x})), L_f h^{i+1}(\boldsymbol{x}) + \alpha(h^{i+1}(\boldsymbol{x})), \ldots, L_f h^N(\boldsymbol{x}) + \alpha(h^N(\boldsymbol{x}))]$$

The superscript in $h^i(\boldsymbol{x})$ denotes the CBF for the i^{th} agent. The notation $\bar{\boldsymbol{p}}_k$ represents the k^{th} row of $\bar{A}$, and $\tilde{q}_k$ the k^{th} component of $\bar{\boldsymbol{q}}$. The Lagrangian associated with (3) is $L(\boldsymbol{u}, \boldsymbol{\lambda}) = \|\boldsymbol{u} - k(\boldsymbol{x})\|_2^2 + \sum_{k=1}^{N} \lambda_k(\tilde{\boldsymbol{p}}_k^{\text{T}}\boldsymbol{u} - \tilde{q}_k)$. At optimality, the pair $(\boldsymbol{u}^\star, \boldsymbol{\lambda}^\star)$ satisfies the Karush-Kuhn-Tucker (KKT) conditions as follows:

1) **Stationarity**: The gradient of the Lagrangian L with respect to $\boldsymbol{u}$ at $(\boldsymbol{u}^\star, \boldsymbol{\lambda}^\star)$ is zero, leading to $\boldsymbol{u}^\star = k(\boldsymbol{x}) - \frac{1}{2}\sum_{k=1}^{N} \lambda^\star_k \tilde{\boldsymbol{p}}_k^{\text{T}}$.
2) **Primal Feasibility**: This requires that $\bar{\boldsymbol{p}}_k^{\text{T}}\boldsymbol{u}^\star \leq \tilde{q}_k$ for all $k \in V$.
3) **Dual Feasibility**: $\lambda^\star_k \geq 0$ for all $k \in V$.
4) **Complementary Slackness**: This ensures that for all constraints, $\lambda^\star_k(\tilde{\boldsymbol{p}}_k^{\text{T}}\boldsymbol{u}^\star - \tilde{q}_k) = 0$ for all $k \in V$. Mathematically, $\boldsymbol{u}^\star = k(\boldsymbol{x}) - \frac{1}{2}\sum_{k \in \mathcal{A}(\boldsymbol{u}^\star)} \lambda^\star_k \tilde{\boldsymbol{p}}_k^{\text{T}}$ where $\mathcal{A}(\boldsymbol{u}^\star) = \{k \in V \mid \tilde{\boldsymbol{p}}_k^{\text{T}}\boldsymbol{u}^\star = \tilde{q}_k\}$ is the set of active constraints .

The system is in a deadlock when $\boldsymbol{u}^\star$ is zero i.e. $k(\boldsymbol{x}) = \frac{1}{2}\sum_{k \in \mathcal{A}(\boldsymbol{u}^\star)} \lambda^\star_k \tilde{\boldsymbol{p}}_k^{\text{T}}$. This deadlock phenomenon of agents getting stuck in local minima (due to non-convexity of $k(\boldsymbol{x})$ wrt to $\boldsymbol{x}$ in (3)) is similar to the problem of minimizing non-convex functions in ML using well-established gradient methods. In the following, inspired by gradient based methods in ML that minimize a non-convex loss function, we propose a deadlock minimization method to minimize the deadlock present in large scale multi-agent control.

Deadlock resolution: Note that the cost function (3) is quadratic in $\boldsymbol{u}$, however as $\boldsymbol{u}$ in (3) is a function of the

state $\boldsymbol{x}$, the cost function can be non-convex with respect to $\boldsymbol{x}$. The deadlock occurs when the agents gets stuck in the local minima of this cost function i.e. when the cost becomes $\|k(\boldsymbol{x})\|_2^2$ (note that $\boldsymbol{x}$ is fixed when solving the QP (3)) instead of the global minima, before it reaches the goal. To avoid these local minima, we leverage tools from gradient-based methods in ML to escape the local minima. Particularly, in this paper, we leverage the momentum based gradient method to modify $\boldsymbol{u}_t^\star$ from (3) as follows (subscript t denotes the time instant at which (3) is solved):

$$\boldsymbol{u}_t^\star = \boldsymbol{u}_{t-1}^\star - \boldsymbol{w}_t + \Delta^\star, \quad \boldsymbol{w}_t = \gamma \boldsymbol{w}_{t-1} + \eta \nabla_{\boldsymbol{x}} c(\boldsymbol{x};\theta) \quad (4)$$

where η is the learning rate, $\gamma \in (0,1)$, $c(\boldsymbol{x};\theta) = \|\boldsymbol{u}(\boldsymbol{x};\theta_c) - k(\boldsymbol{x})\|_2^2$. As we know, the sequence of the states and the inputs of our neighboring agents at some previous time steps, it is possible to approximately compute the values of $h_{t-1}(\boldsymbol{x};\theta)$ for the agents. Please note that the input computed using the update scheme would be deadlock-free, however, there are no guarantees whether the agents are collision-free. Towards this goal, we choose the value of $\Delta^\star$ in (4) as follows:

$$\Delta^\star = \underset{\Delta \in \mathbb{R}^m}{\operatorname{argmin}} \ \|\Delta\|^2 \qquad \text{s.t. } \bar{A}\boldsymbol{u}_t^\star \leq \bar{\boldsymbol{q}} \quad (5)$$

where $\boldsymbol{u}_t^\star$ is from (4). Note that (5) is a QP as the optimization problem is quadratic with respect to the variable Δ. Updating the control input via (4) is only triggered when the magnitude of the control input $\boldsymbol{u}_t^\star$ of at least one of the agents is below a threshold i.e. $\|\boldsymbol{u}_t^\star\| \leq \epsilon$ for some small $\epsilon > 0$. To escape the local minima, state-of-the-art gradient-based update schemes can be used to update the control input at time t in (4).

Safety guarantee: One of the trivial methods to guarantee safety is to embed an optimization based framework such as MPC whenever a collision is anticipated by forward propagating the agent's dynamics for a few time steps via the NN based controller. The MPC problem when a collision is anticipated (via the NN control policy) is given by

$$\boldsymbol{u}_i^\star = \operatorname{argmin} c_{\text{MPC}}(\boldsymbol{x},\boldsymbol{u}) \quad (6a)$$

$$\text{s.t.} \quad h^j(\boldsymbol{x}) + \alpha(h^j(\boldsymbol{x})) \geq 0 \quad \forall \quad j \in \mathcal{N}_a^i \quad (6b)$$

$$h^j(\boldsymbol{u}) + \alpha(h^j(\boldsymbol{u})) \geq 0 \quad \forall \quad j \in \mathcal{N}_a^i \quad (6c)$$

where $c_{\text{MPC}}(\boldsymbol{x},\boldsymbol{u}) > 0$ is the cost function, usually taken to be a quadratic function, and $h^j(\boldsymbol{x})$ and $h^j(\boldsymbol{u})$ ($j \in \mathcal{N}_a^i$) encodes the state and the input constraints respectively. With a slight abuse in notation, we interchangeably change the inputs for h^j to $\boldsymbol{x}$ (only state constraints) or $\boldsymbol{u}$ (only input constraints) or both $(\boldsymbol{x},\boldsymbol{u})$. However, the optimization problem is computationally expensive especially when the number of agents are large (as the number of constraints are $N-1$), thereby resulting in poor scalability. To address this limitation, we transform the $2(N-1)$ constraints into a single constraint, thereby making the MPC problem devoid of the number of the agents. Particularly, we consider the combined ICBF $h_c^i(\boldsymbol{x},\boldsymbol{u})$ for an agent i as follows:

$$h_c^i(\boldsymbol{x},\boldsymbol{u}) = -\log\left(\sum_{j \in \mathcal{N}_a^i} e^{-\beta h^j(\boldsymbol{x})} + e^{-\beta h^j(\boldsymbol{u})}\right) \quad (7)$$

where $\beta > 0$. Consequently, the MPC-ICBF with a single constraint becomes

$$\boldsymbol{u}_i^\star = \underset{\boldsymbol{u} \in \mathcal{U}^i}{\operatorname{argmin}} \ c_{\text{MPC}}(\boldsymbol{x},\boldsymbol{u}) \ \text{ s.t. } \ \dot{h}_c^i(\boldsymbol{x},\boldsymbol{u}) + \alpha(h_c^i(\boldsymbol{x},\boldsymbol{u})) \geq 0 \quad (8)$$

Consider the following three constraint sets for agent i, i.e.,

$$\mathcal{S}_{\boldsymbol{x}}^i = \left\{\boldsymbol{x} : \cap_{j \in \mathcal{N}_a^i} h^j(\boldsymbol{x}) + \alpha(h^j(\boldsymbol{x})) \geq 0\right\} \quad (9a)$$

$$\mathcal{S}_{\boldsymbol{u}}^i = \left\{\boldsymbol{u} : \cap_{j \in \mathcal{N}_a^i} h^j(\boldsymbol{u}) + \alpha(h^j(\boldsymbol{u})) \geq 0\right\} \quad (9b)$$

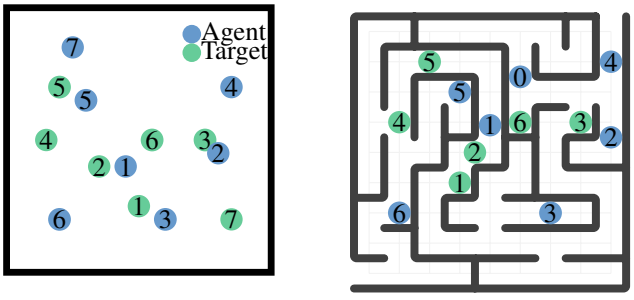
$$\mathcal{S}_c^i = \left\{(\boldsymbol{x},\boldsymbol{u}) : \ h_c^i(\boldsymbol{x},\boldsymbol{u}) + \alpha(h_c^i(\boldsymbol{x},\boldsymbol{u})) \geq 0\right\} \quad (9c)$$

It can be easily shown that the set $\mathcal{S}_{\boldsymbol{x}}^i \cup \mathcal{S}_{\boldsymbol{u}}^i \subset \mathcal{S}_c^i$. Furthermore, $\lim_{\beta \to \infty} \mathcal{S}_c^i = \mathcal{S}_{\boldsymbol{x}}^i \cup \mathcal{S}_{\boldsymbol{u}}^i$. For the sake of brevity, we do not present the proof in this paper.

Input constraint satisfaction: Based on the definition of ICBF, the combined ICBF (7) encodes both the state as well as the input constraints. Consequently, solving the proposed MPC-ICBF (8) ensures that the input constraints are satisfied.

IV. Evaluation and Discussion

In this section, we compare our proposed approach with two classes of methods. Through the experiments, we aim to answer the following questions (i) can MA-ICBF guarantee 100% collision avoidance as well as scale well with the number of agents (e.g. >1000 agents)? (ii) can MA-ICBF reduce the number of deadlocks among agents compared to baseline methods while synthesizing safe trajectories and respecting input constraints? (iii) does MA-ICBF take care of practical considerations such as input constraint satisfaction while maintaining the safety guarantee?, (iv) whether MA-ICBF provides significant computational benefits versus the approach that trivially combines the NN based controller and MPC with multiple ICBF constraints (6) and finally (v) whether MA-ICBF is generalizable. Particularly, does MA-ICBF perform well if they are trained on fewer agents and tested for a larger number of agents?

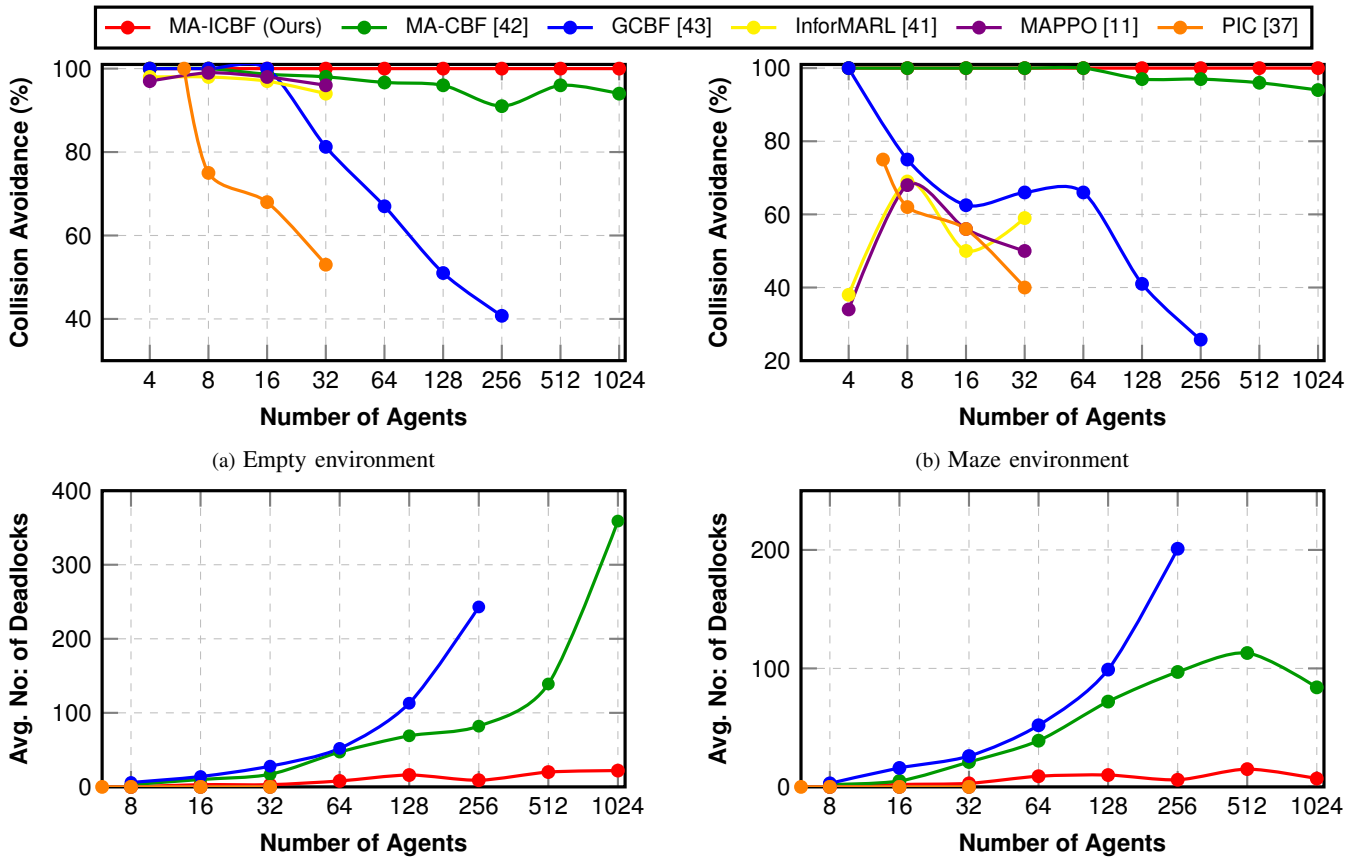


(a) Empty (b) Maze

Fig. 3: Representative testing environments used for experiments. (a) Empty environment. (b) Complex maze environment.

A. Experimental Setup

For the experiments, we consider an empty environment and a complex maze environment which is a $10m \times 10m$ grid as shown in Fig. 3. We consider a quadrotor system with dynamics similar to that in [42]. The training is performed with 70000 timesteps with a learning rate of 0.001. For each experiments 5 trials are conducted with each trial being randomized with respect to start and goal position in both of the environments. All experiments have been conducted on an AMD Ryzen machine @2.90GHz with 8 cores and 16.0 GB of RAM. The maximum number of agents tested in our experiments is 1024. First, we consider multi-agent reinforcement learning (MARL) algorithms [37], [40], [41]. The main limitation of these MARL algorithms is that they do not exploit the nonlinear dynamics of the system and henceforth require a large number of state-input pairs to train their algorithm. Due to computational intractability, we test the MARL algorithms only up to 32 agents similar to [42], [43] and up to 256 agents for method [43]. Second, we consider neural network-based (NN-based) approaches that leverage CBF-based safety certificates and exploit the



(a) Empty environment

(b) Maze environment

(c) Empty environment

(d) Maze environment

Fig. 4: (a)-(b): Collision avoidance percentage versus number of agents. (c)-(d): Average number of deadlocks.

nonlinear dynamics to design safe control inputs [42], [43]. We choose $\gamma = 0.8$ in (4) and $\beta = 50$ in (7).

B. 100% collision guarantee in MA-ICBF

As seen from Figs. 4a and 4b, our proposed approach (MA-ICBF) guarantees 100% collision avoidance, regardless of the number of agents in both empty as well as maze environments. This scalability feature is particularly significant, as it allows MA-ICBF to handle scenarios with more than $>$ 1000 agents effectively. Figs. 4a and 4b illustrate the performance of MA-ICBF as the number of agents increases, highlighting its ability to maintain collision-free trajectories in complex environments. From these experiments, we observe that in contrast to prior methods where the percentage of collision avoidance significantly decreases with the number of agents, MA-ICBF can provide safety guarantees even for a large number of agents. In addition, as seen from Fig. 5d, the computational load (runtime in ms) is not compromised as the number of agents increases, thereby showing good scalability. Our baselines, including GCBF, MA-CBF, and MARL methods are purely learning based with soft safety guarantees (collision-checking) and prioritize goal completion at the cost of safety as the number of agents scale and have comparatively low safety.

C. Minimizing deadlocks while generating safe trajectories:

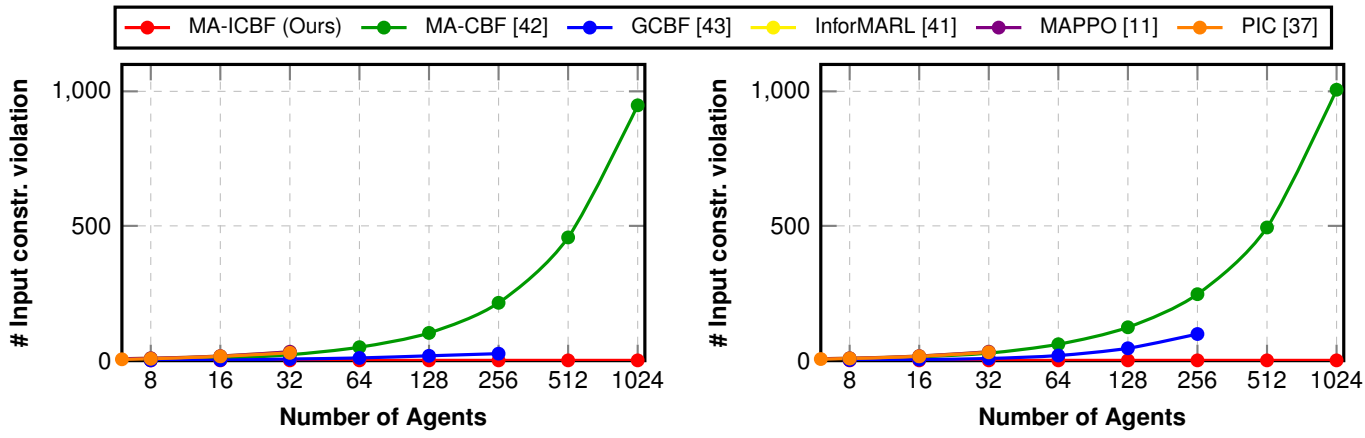
As observed from Figures 4c and 4d, there is a significant reduction in the number of deadlocks among agents in both empty and maze-like environments by leveraging the deadlock minimization mechanism described in III-B. For the experiments, the values of hyperparameters chosen were $\gamma = 0.99$, $\eta = 0.01$, and the threshold magnitude of control input for deadlock detection was taken as 0.01. In addition, from Fig. 4d, one can observe that MA-ICBF reduces the average number of deadlocks by up to $\approx 92\%$ with 1024 agents, compared to the baseline methods in a complex maze environment. Furthermore, the average number of deadlocks increases with the number of agents. However, the proposed method MA-ICBF successfully minimizes deadlocks while ensuring that the trajectories are safe. In addition, the safety and scalability of MA-ICBF are not compromised during deadlock minimization (see Figs. 4a, 4b, 5c, 5d).

D. Ensuring satisfaction of input constraints:

We set the input constraints in the range of $[-0.2, 0.2]$ meters per second for the linear velocity and $[-12, 12]$ degrees for the angular velocity. Figs. 5a and 5b show that the proposed approach MA-ICBF successfully constrains all agents within these bounds. In contrast, baselines GCBF, MA-CBF, and MARL methods show an exponential increase in the number of input constraint violations with a maximum of 1006 agents (out of 1024 agents) which means that $\approx 98\%$ agents have violated the input constraints.

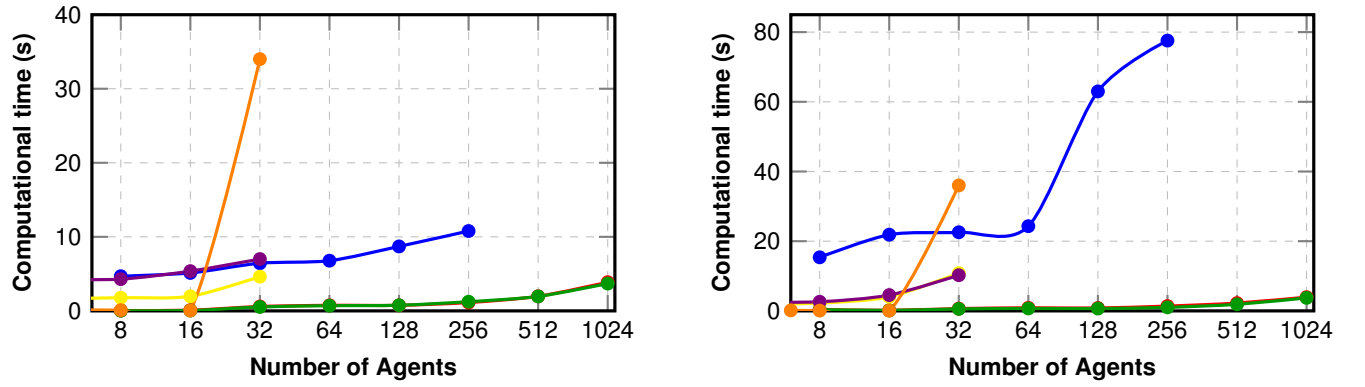
E. Computational benefits using log-sum-exp trick (7) in MPC-ICBF:

As discussed in the III-B, the log-sum-exp trick converts the multiple collision avoidance constraints (6) into a single constraint (8) for an agent. Furthermore, to accelerate the summation in the log-sum-exp expression in (7), we leverage parallelization using the `cupy` [46] package. We compare this approach with a baseline that trivially augments the learned NN control policy with the MPC based controller with ICBF constraints (NN+MPC). Fig. 6 shows the histogram plot comparing the computational time per agent in milliseconds for two methods— MA-ICBF using the



(a) Empty environment

(b) Maze environment



(c) Empty environment

(d) Maze environment

Fig. 5: (a)-(b): Average number of input constraint violations versus the number of agents in different settings. (c)-(d) Average computational time per agent in milliseconds in different environments

log-sum-exp trick and the baseline method (NN+MPC). As observed from Fig. 6, there is a significant computational advantage ($\approx 1.5 - 2\,\mathrm{x}$) between these two approaches. Furthermore, with the increase in the number of agents, the computational benefits become more apparent. This is mainly because, with the higher number of agents, the number of collisions increases, making the log-sum-exp trick's ability to reduce the computational load more noticeable.

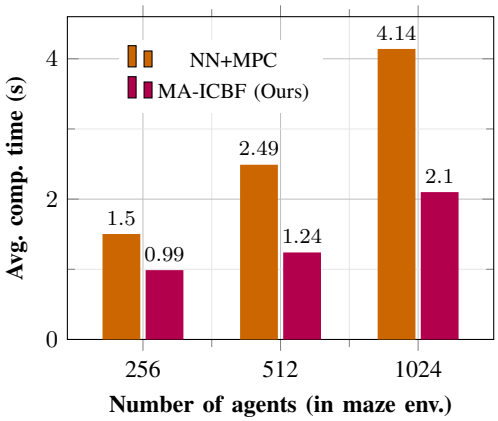


Fig. 6: The average computational time in seconds per agent of trivially encoding MPC and NN based controller versus MA-ICBF

F. Generalizibility:

Table I, shows how generalizable MA-ICBF is, when trained with fewer agents (4, 8 or 16) and then tested with a larger number of agents (256, 512, or 1024) in a maze environment. From table I, when trained with 16 agents, MA-ICBF performs well even with 1024 agents with no deadlocks, while at the same triggering the MPC-ICBF module twice to ensure complete collision avoidance and input constraint satisfaction.

Train \ Test		$n = 256$	$n = 512$	$n = 1024$
$m = 4$	#MPC-ICBF trig.	21	73	280
	#Deadlock	6	15	18
$m = 8$	#MPC-ICBF trig.	0	1	10
	#Deadlock	0	0	0
$m = 16$	#MPC-ICBF trig.	0	0	2
	#Deadlock	0	0	0

TABLE I: Test performance of MA-ICBF for the *maze* environment, when trained with m agents and tested with n agents.

V. Conclusion

In this paper, we have addressed the problem of multi-agent control under limited actuation, particularly focusing on the aspects of scalability, safety guarantees, and input constraint satisfaction. Our proposed approach first jointly learns a decentralized neural Integral Control Barrier (ICBF) certificate and a decentralized safe neural network (NN) controller. During the execution phase, the safe learned controller is augmented with a computationally light MPC-ICBF based framework as well as a deadlock minimization module to guarantee safety and minimize deadlocks respectively. The main limitation of our work is the assumption that the governing dynamics of each agent are known a priori. Future work includes extending this work to vision-control tasks, particularly focused on multi-agent control that is also able to handle dynamic obstacles.

REFERENCES

[1] X. Li, "Motion planning of warehouse automation robot based on recurrent neural network," in *2023 International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII)*. IEEE, 2023, pp. 385–389.

[2] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1826–1848, 2019.

[3] S. Batra, Z. Huang, A. Petrenko, T. Kumar, A. Molchanov, and G. S. Sukhatme, "Decentralized control of quadrotor swarms with end-to-end deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 576–586.

[4] J. Park, D. Kim, G. C. Kim, D. Oh, and H. J. Kim, "Online distributed trajectory planning for quadrotor swarm with feasibility guarantee using linear safe corridor," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4869–4876, 2022.

[5] T.-C. Au and P. Stone, "Motion planning algorithms for autonomous intersection management," in *Workshops at the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[6] G. Sharon, R. Stern, A. Felner, and N. R. Sturtevant, "Conflict-based search for optimal multi-agent pathfinding," *Artificial intelligence*, vol. 219, pp. 40–66, 2015.

[7] J. Kottinger, S. Almagor, and M. Lahijanian, "Conflict-based search for explainable multi-agent path finding," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 32, 2022, pp. 692–700.

[8] S. Dergachev and K. Yakovlev, "Distributed multi-agent navigation based on reciprocal collision avoidance and locally confined multi-agent path finding," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2021, pp. 1489–1494.

[9] J. Alonso-Mora, A. Breitenmoser, M. Rufli, P. Beardsley, and R. Siegwart, "Optimal reciprocal collision avoidance for multiple non-holonomic robots," in *Distributed autonomous robotic systems: The 10th international symposium*. Springer, 2013, pp. 203–216.

[10] R. Chandra, V. Zinage, E. Bakolas, J. Biswas, and P. Stone, "Decentralized multi-robot social navigation in constrained environments via game-theoretic control barrier functions," *arXiv preprint arXiv:2308.10966*, 2023.

[11] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 611–24 624, 2022.

[12] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*. IEEE, 2019, pp. 3420–3431.

[13] P. Glotfelter, J. Cortés, and M. Egerstedt, "Nonsmooth barrier functions with applications to multi-robot systems," *IEEE control systems letters*, vol. 1, no. 2, pp. 310–315, 2017.

[14] M. Jankovic and M. Santillo, "Collision avoidance and liveness of multi-agent systems with cbf-based controllers," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 6822–6828.

[15] R. Cheng, M. J. Khojasteh, A. D. Ames, and J. W. Burdick, "Safe multi-agent interaction through robust control barrier functions with learned uncertainties," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 777–783.

[16] K. Garg and D. Panagou, "Robust control barrier and control lyapunov functions with fixed-time convergence guarantees," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 2292–2297.

[17] M. Köhler, M. A. Müller, and F. Allgöwer, "Distributed mpc for self-organized cooperation of multi-agent systems," *IEEE Transactions on Automatic Control*, 2024.

[18] D. Saccani, L. Fagiano, M. N. Zeilinger, and A. Carron, "Model predictive control for multi-agent systems under limited communication and time-varying network topology," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 3764–3769.

[19] Y. Chen, A. Singletary, and A. D. Ames, "Guaranteed obstacle avoidance for multi-robot operations with limited actuation: A control barrier function approach," *IEEE Control Systems Letters*, vol. 5, no. 1, pp. 127–132, 2020.

[20] L. Wang, A. D. Ames, and M. Egerstedt, "Safety barrier certificates for collisions-free multirobot systems," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 661–674, 2017.

[21] W. Xiao and C. Belta, "Control barrier functions for systems with high relative degree," in *2019 IEEE 58th conference on decision and control (CDC)*. IEEE, 2019, pp. 474–479.

[22] C. Wang, Y. Meng, Y. Li, S. L. Smith, and J. Liu, "Learning control barrier functions with high relative degree for safety-critical control," in *2021 European Control Conference (ECC)*. IEEE, 2021, pp. 1459–1464.

[23] M. Marley, R. Skjetne, and A. R. Teel, "Hybrid control barrier functions for continuous-time systems," *IEEE Transactions on Automatic Control*, 2024.

[24] L. Lindemann, H. Hu, A. Robey, H. Zhang, D. Dimarogonas, S. Tu, and N. Matni, "Learning hybrid control barrier functions from data," in *Conference on Robot Learning*. PMLR, 2021, pp. 1351–1370.

[25] A. J. Taylor, V. D. Dorobantu, R. K. Cosner, Y. Yue, and A. D. Ames, "Safety of sampled-data systems with control barrier functions via approximate discrete time models," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 7127–7134.

[26] L. Niu, H. Zhang, and A. Clark, "Safety-critical control synthesis for unknown sampled-data systems via control barrier functions," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 6806–6813.

[27] V. Zinage and E. Bakolas, "Neural koopman control barrier functions for safety-critical control of unknown nonlinear systems," in *2023 American Control Conference (ACC)*. IEEE, 2023, pp. 3442–3447.

[28] V. Zinage, R. Chandra, and E. Bakolas, "Neural differentiable integral control barrier functions for unknown nonlinear systems with input constraints," *arXiv preprint arXiv:2312.07345*, 2023.

[29] V. Zinage and E. Bakolas, "Neural koopman lyapunov control," *Neurocomputing*, vol. 527, pp. 174–183, 2023.

[30] M. Jankovic, "Control barrier functions for constrained control of linear systems with input delay," in *2018 annual American control conference (ACC)*. IEEE, 2018, pp. 3316–3321.

[31] A. D. Ames, G. Notomista, Y. Wardi, and M. Egerstedt, "Integral control barrier functions for dynamically defined control laws," *IEEE control systems letters*, vol. 5, no. 3, pp. 887–892, 2020.

[32] V. Zinage, R. Chandra, and E. Bakolas, "Disturbance observer-based robust integral control barrier functions for nonlinear systems with high relative degree," *arXiv preprint arXiv:2309.16945*, 2023.

[33] P. Long, T. Fan, X. Liao, W. Liu, H. Zhang, and J. Pan, "Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6252–6259.

[34] Y. F. Chen, M. Liu, M. Everett, and J. P. How, "Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 285–292.

[35] M. Everett, Y. F. Chen, and J. P. How, "Collision avoidance in pedestrian-rich environments with deep reinforcement learning," *Ieee Access*, vol. 9, pp. 10 357–10 377, 2021.

[36] A. Kamenev, L. Wang, O. B. Bohan, I. Kulkarni, B. Kartal, A. Molchanov, S. Birchfield, D. Nistér, and N. Smolyanskiy, "Predictionnet: Real-time joint probabilistic traffic prediction for planning, control, and simulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8936–8942.

[37] I.-J. Liu, R. A. Yeh, and A. G. Schwing, "Pic: permutation invariant critic for multi-agent deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2020, pp. 590–602.

[38] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[39] X. Wang, Z. Zhang, and W. Zhang, "Model-based multi-agent reinforcement learning: Recent progress and prospects," *arXiv preprint arXiv:2203.10603*, 2022.

[40] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 611–24 624, 2022.

[41] S. Nayak, K. Choi, W. Ding, S. Dolan, K. Gopalakrishnan, and H. Balakrishnan, "Scalable multi-agent reinforcement learning through intelligent information aggregation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 25 817–25 833.

[42] Z. Qin, K. Zhang, Y. Chen, J. Chen, and C. Fan, "Learning safe multi-agent control with decentralized neural barrier certificates," *arXiv preprint arXiv:2101.05436*, 2021.

[43] S. Zhang, K. Garg, and C. Fan, "Neural graph control barrier functions guided distributed collision-avoidance multi-agent control," in *Conference on robot learning*. PMLR, 2023, pp. 2373–2392.

[44] K. P. Tee, S. S. Ge, and E. H. Tay, "Barrier lyapunov functions for the control of output-constrained nonlinear systems," *Automatica*, vol. 45, no. 4, pp. 918–927, 2009.

[45] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[46] R. Nishino and S. H. C. Loomis, "Cupy: A numpy-compatible library for nvidia gpu calculations," *31st confernce on neural information processing systems*, vol. 151, no. 7, 2017.

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-024-09493-6

RESEARCH ARTICLE-ELECTRICAL ENGINEERING

Derivation Analysis and Control of Multiport Flyback Converter with Lyapunov Function-Based Controller in Renewable Energy Systems Considering Circuit Parasitics

Shabana Urooj[1] · Dheeraj Joshi[2] · Ashutosh Gupta[2] · Amal BaQais[3]

Received: 3 November 2023 / Accepted: 8 August 2024
© King Fahd University of Petroleum & Minerals 2024

Abstract
This paper addresses the analytical modeling of multiport flyback converters, which are governed by linear differential–algebraic equations. The modeling of these converters poses a challenge due to the switching between multiple such equations that govern the circuit's behavior. The study emphasizes the importance of robust control strategies for addressing the non-idealities in flyback converters. Using MATLAB/Simulink, a dynamic model of the converter is developed. A comparison is made between a Lyapunov function-based controller, a linear proportional integrator controller, and an integral sliding mode controller. The controller consists of two main parts: a duty-ratio feedforward control unit for steady-state parameters and a Lyapunov function-based feedback control unit to handle disturbances. The feedforward control signal helps reduce the workload on the Lyapunov feedback controller. This control system ensures global exponential stability of the closed-loop system, enabling a swift transient response even under line and load disturbances. Mathematical simulations demonstrate its superior performance and stability, while experimental validations are conducted considering equivalent series resistances of each component. The proposed model and control scheme are further validated through a hardware prototype tested under various load and line disturbances. This research highlights the necessity of stress testing and performance evaluation for ensuring smooth and efficient operation of controlled multiport flyback converters across diverse conditions.

Keywords Sliding mode control · Lyapunov control · Multiport DC–DC converters · Nonlinear controller · Parasitics topology derivation · Renewable energy sources

1 Introduction

Over the past two decades, renewable energy sources such as photovoltaic (PV) and wind energy have become increasingly important due to their clean, pollution-free nature and abundance. Researchers and scientists are actively investigating PV energy generation systems, focusing on improving inverter designs and control strategies to boost efficiency. However, despite these technological advancements, the high costs have limited their widespread use, mainly residential applications. Flyback topology-based DC/DC converters, known for their simplicity, low cost, ease of isolation, and ability to provide multiple outputs, have found broad acceptance in automotive systems, telecommunications equipment, and electronic devices. These converters can function in both continuous conduction mode (CCM) and discontinuous conduction mode (DCM) [1, 2]. In [3], a grid-connected system for renewable energy source (RES) applications, comprising a modified switched-capacitor (SC)-based multilevel inverter paired with a DC–DC flyback converter, is presented.

When operating in the continuous conduction mode (CCM), the non-ideal flyback DC/DC converter exhibits higher efficiency, lower voltage and current stress, and a more straightforward filter design than the DCM mode. As a result, there is considerable interest in utilizing the flyback CCM converter for industrial applications [4–7].

However, the flyback CCM converter faces a challenge due to its nonminimum phase behavior, which is attributed

✉ Ashutosh Gupta
ashutoshgupta.nitm@gmail.com

[1] Department of Electrical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

[2] Department of Electrical Engineering, Delhi Technological University, Delhi, 110042, India

[3] Department of Chemistry, College of Science, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Published online: 16 September 2024

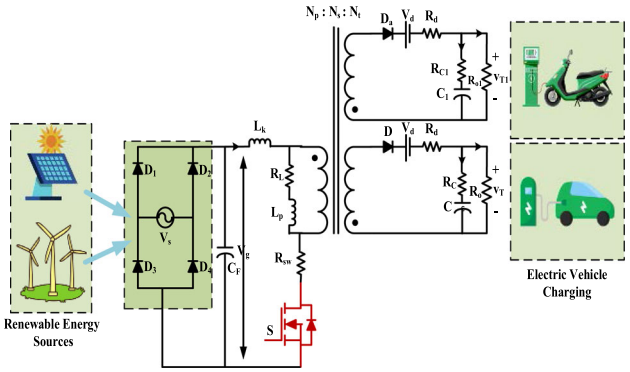


Fig. 1 Multiport flyback converter topology

to its transfer function from the control duty to the output voltage having a right-half-plane zero [8]. Consequently, achieving precise output voltage control for the flyback converter with equivalent series resistances poses a challenging issue. The MFC architecture of the renewable energy-based system is shown in Fig. 1, having an isolated flyback converter.

Control designs for these converters are usually implemented using small-signal analysis around the operating point to enable the application of linear control strategies [9, 10]. Linear control design techniques are popular for the flyback CCM converter around a fixed operating point, but their use requires a linearized model that does not capture the entire dynamic behavior, making it difficult to ensure stability under large perturbations. Researchers have proposed nonlinear control strategies derived directly from the converter's model, avoiding the need for linearization. Based on Lyapunov functions, backstepping control stabilizes the inductor current and output voltage, but it is sensitive to parameter variations and exhibits slow transient responses due to inherent singularities [11–13]. On the other hand, sliding mode control (SMC) ensures robustness against variations and disturbances with faster transient performance. However, SMC generates a bang–bang-type control input, leading to chattering and significant switching frequency variation with varying input voltage and output load [14]. In [15], it employs a control strategy grounded in Lyapunov functions to ensure the stability of a cascaded DC–DC converter.

Meanwhile, [16] introduced a technique for calculating the switching control signals for the Cuk DC–DC converter utilizing the Lyapunov theory. It is important to note that the concepts outlined in [17] have not yet been implemented in a practical experimental setup. A notable challenge with the Lyapunov controller is achieving fast convergence of the Lyapunov function while maintaining robustness against variations in the controller coefficients, as mentioned in [18]. Therefore, developing a practical approach for selecting these coefficients is essential to ensure rapid response and stability across a wide range of operating conditions.

In [19], a SISO flyback converter employing Lyapunov function control is introduced. The study accounts for parasitic elements in the diode, inductor, and switch but overlooks the inclusion of capacitor equivalent series resistances ESR. The paper digs into the design of the converter's controller but does not discuss the aspects of converter stability and the stress placed on the switches. In [19], the authors introduce a new seven-level inverter based on switched capacitors that operates with a single DC power supply. The design can produce a seven-level output voltage waveform while also boosting the voltage. Remarkably, the inverter achieves a voltage gain factor of 3 without the need for an additional DC–DC boost converter.

This paper proposes a CCM flyback converter considering ESR of all components, including the capacitor, having two outputs. This converter's nonminimum phase behavior, caused by a right-half-plane zero in its transfer function, leads to poor transient response under large signal perturbations [20–22]. The converter stability is analyzed by deriving the transfer function for one of the outputs. However, the Lyapunov function-based feedback controller makes the system's response globally exponentially stable during the transient response compared to the linear PID controller and the integral mode sliding mode controller. We strengthen it by adding a duty-ratio feedforward controller to lessen its load. By considering parasitic elements, we create a dynamic model for the flyback CCM converter, allowing precise control of the desired output voltage. Numerical simulations demonstrate better performance and stability, while practical tests confirm the effectiveness of our proposed control method in different applications. In [23], the authors propose a five-level boost inverter with a single DC source, featuring a DC–DC boost converter for voltage transfer, PCC method for switching pattern, component design, power loss analysis, performance comparison, and a laboratory prototype tested with the local grid for active and reactive power control.

This paper is structured as follows: Section 2 presents the dynamic modeling and small-signal analysis of the non-ideal flyback converter. Section 3 illustrates the proposed converter's controller design and stability analysis. Hardware design and fabrication are explained in Sect. 4. Section 5 showcases the simulation, hardware implementation and experimental results. Lastly, Sect. 6 concludes the paper with work and analysis of the proposed flyback converter.

2 Dynamic Modeling of Flyback Converter with Parasitic Components

The flyback CCM converter's modeling and controller design process can be challenging and complex, leading to the omission of parasitic components. This results in an ideal/lossless

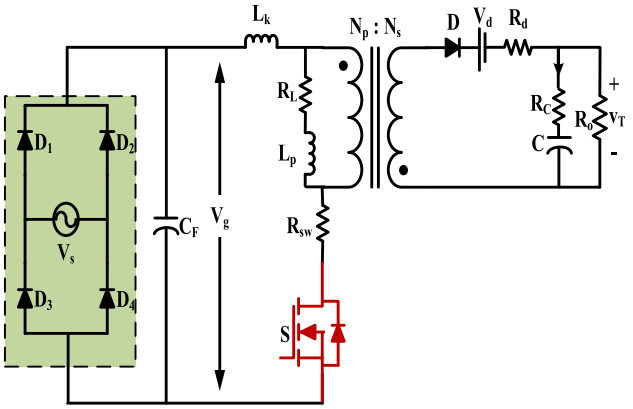


Fig. 2 Electrical model of Flyback converter with parasitic components

model that simplifies the development and aids in understanding the converter's main features. However, accounting for parasitic elements becomes crucial for enhancing model accuracy and studying dynamic performance. Including these components can lead to improved steady-state and transient performance. Therefore, we consider the parasitic components in our approach to construct a more comprehensive dynamic model of the flyback CCM converter.

Figure 2 describes the circuit diagram of the non-ideal flyback converter with ESRs. This electrical model presents the transformer as an equivalent inductor *Lp*. Within this model, the parameters V_g, S, R_{sw}, L_p, L_k, D, C, R_{o}, $n = \frac{N_p}{N_s}$, i_{Lp}, V_d, R_L, R_d, R_c, v_c, and v_T symbolize the input voltage, MOSFET switch, switch resistance, transformer inductance, leakage inductance, diode, output filter capacitor, load resistor, transformer turns ratio, inductor current, diode forward drop, inductor resistance, parasitic diode resistance, ESR of the capacitor, output capacitor voltage and output voltage of the non-ideal Flyback converter, respectively.

In any converter, two primary states of a switch are employed to formulate the voltage gain equation. The voltage and current waveforms for PWM switching of the Flyback converter are illustrated in Fig. 3 to depict this relationship. The working of the flyback converter is expressed as:

Switch $S(1 < t < \text{DT})$ is turned ON, as shown in Fig. 4a, the primary side of transformer T experiences an increasing current flow as the input DC voltage supplies power to it. Consequently, energy is stored in the transformer's magnetizing inductance L_p. Due to the polarity of the transformer's dot, diode D becomes reverse-biased, causing the secondary side to be disconnected and not conduct electricity. In this state, the load is powered by the current supplied from the output capacitor C.

When switch $S(\text{DT} < t < T)$ is toggled OFF, as shown in Fig. 4b, by the transformer's dot polarity, the inductor L_p Undergoes a transfer of energy to capacitor C in a unidirectional manner. This causes diode D on the secondary side to become forward-biased and initiate conduction based on Faraday's Law of electromagnetic induction. This cyclic process concludes, resulting in the load resistance receiving a direct current (DC) voltage.

Assuming minimal resistance in the windings and leakage inductance and acknowledging the presence of a nonzero core reluctance, the transformer can be simplified as two inductors with the magnetizing inductance L_p along with inductor resistor R_L and leakage inductance L_k on the primary side of an ideal transformer. This system involves two crucial variables: the current i_{LP} and the voltage v_T.

By utilizing Kirchhoff's voltage and current principles, we can deduce the state-space equations for both the cases when the switch is in the on and off positions, expressed as follows, respectively.

Switch turned ON $(1 < t < \text{DT})$:

$$\begin{bmatrix} \frac{\mathrm{d}i_{Lp}}{\mathrm{d}t} \\ \frac{\mathrm{d}v_c}{\mathrm{d}t} \end{bmatrix} = \begin{bmatrix} -\frac{(R_L+R_{\text{sw}})}{(L_P+L_k)} & 0 \\ 0 & -\frac{1}{C(R_{\text{o}}+R_c)} \end{bmatrix} \begin{bmatrix} i_{L_p} \\ v_c \end{bmatrix} + \begin{bmatrix} \frac{1}{(L_P+L_k)} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_g \\ V_d \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} V_o \end{bmatrix} = \begin{bmatrix} \frac{R_{\text{o}}}{(R_{\text{o}}+R_c)} & 0 \end{bmatrix} \begin{bmatrix} i_{Lp} \\ v_c \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_g \\ V_d \end{bmatrix} \tag{2}$$

Switch turned OFF $(\text{DT} < t < T)$:

$$\begin{bmatrix} \frac{\mathrm{d}i_{Lp}}{\mathrm{d}t} \\ \frac{\mathrm{d}v_c}{\mathrm{d}t} \end{bmatrix} = \begin{bmatrix} -\frac{\left(R_L+n^2R_d+\frac{n^2R_cR_{\text{o}}}{R_{\text{o}}+R_c}\right)}{L_p} & -\frac{nR_{\text{o}}}{L_p(R_{\text{o}}+R_c)} \\ \frac{nR_{\text{o}}}{C(R_{\text{o}}+R_c)} & -\frac{1}{C(R_{\text{o}}+R_c)} \end{bmatrix} \begin{bmatrix} i_{L_p} \\ v_c \end{bmatrix} + \begin{bmatrix} 0 & \frac{n}{L_p} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_g \\ V_d \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} V_o \end{bmatrix} = \begin{bmatrix} \frac{nR_{\text{o}}R_c}{(R_{\text{o}}+R_c)} & \frac{R_{\text{o}}}{(R_{\text{o}}+R_c)} \end{bmatrix} \begin{bmatrix} i_{Lp} \\ v_c \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_g \\ V_d \end{bmatrix} \tag{4}$$

In these equations, i_{L_k} represents the current flowing through the leakage inductor,i_{LP} is the current flowing through the magnetizing inductor, V_s stands for the input AC voltage, V_g represents the DC voltage after passing through the filter C_F, and v_T is the output voltage.

The basis of an optimized controller design lies in the small-signal model. Particularly in a flyback converter, a well-constructed model proves precious for achieving closed-loop control, fine-tuning, and optimizing the converter's dynamics. Equations (1)–(4) can be represented in the single switched model of the flyback converter considering both averaged and small-signal disturbances as

$$\frac{\mathrm{d}\bar{i}_{L_p}}{\mathrm{d}t} = (d-1)\left\{\frac{nV_d}{L_p} + \frac{nR_{\text{o}}\overline{v}_c}{L_p(R_{\text{o}}+R_c)} + \bar{i}_{L_p}\frac{(n^2R_d+R_L+\frac{n^2R_{\text{o}}R_c}{R_{\text{o}}+R_c})}{L_p}\right\}$$

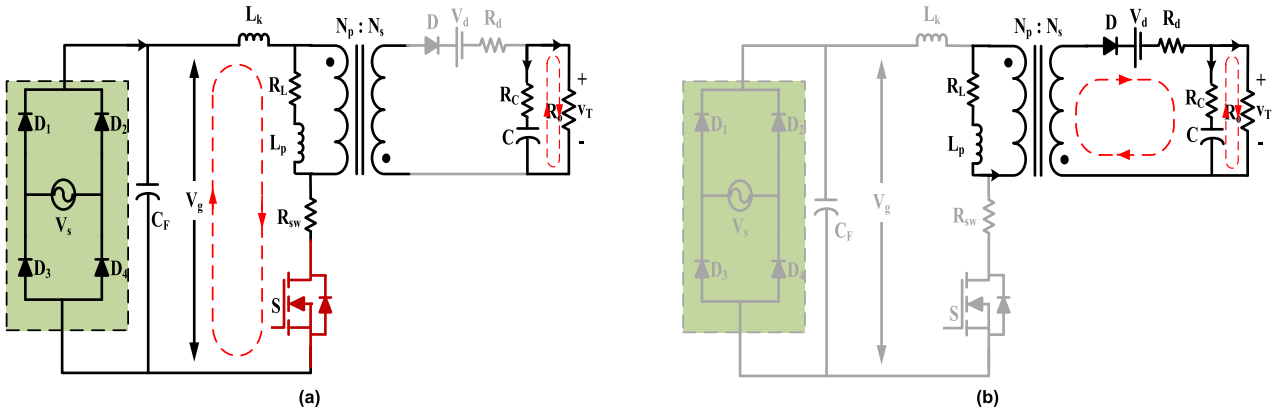


Fig. 3 Steady-state waveform of voltage and current during switching of Flyback converter

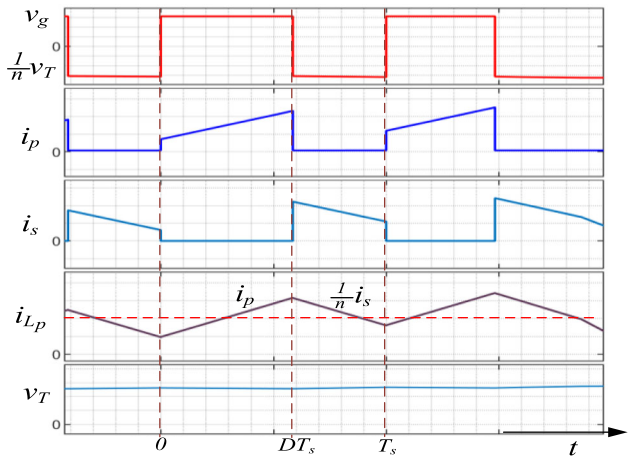


Fig. 4 Operation mode of flyback converter with parasitics: **a** State 1 ($0 < t \leq \mathrm{DT}$) and **b** State 2 ($\mathrm{DT} < t \leq T$)

$$+ d\left\{\frac{V_g}{(L_k + L_p)} - \bar{i}_{Lp}\frac{(R_L + R_{sw})}{(L_k + L_p)}\right\} \tag{5}$$

$$\frac{\mathrm{d}\bar{i}_{Lk}}{\mathrm{d}t} = d\left(\frac{V_g}{(L_p + L_k)} - \bar{i}_{Lp}\frac{(R_L + R_{sw})}{(L_p + L_k)}\right) \tag{6}$$

$$\frac{\mathrm{d}\overline{v}_c}{\mathrm{d}t} = -\left(\frac{\overline{v}_c}{(R_o + R_c)C} - \frac{nR_o\bar{i}_{Lp}}{(R_o + R_c)C} + \frac{dR_o n\bar{i}_{Lp}}{(R_o + R_c)C}\right) \tag{7}$$

where d is the duty ($0 \leq d \leq 1$) and $\bar{i}_{Lp}$, $\bar{i}_{Lk}$ and $\overline{v}_T$ represent the average values of i_{LP}, i_{Lk} and v_T over each switching cycle, respectively. The dynamic model will be subjected to averaging, followed by circuit averaging, during which small perturbations in variables will be introduced, finally resulting in the linearization of the model [24].

$$s\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s) + ((A_1 - A_2)X(s) + (B_1 - B_2)U(s))\hat{d}(s)$$

$$\hat{y}(s) = C\hat{x}(s) + D\hat{u}(s) + ((C_1 - C_2)X(s) + (D_1 - D_2)U(s))\hat{d}(s) \tag{8}$$

where

$$\begin{aligned} A &= A_1 d + A_2(1-d) = (A_1 - A_2)d + A_2 \\ \mathrm{B} &= \mathrm{B_1 d} + \mathrm{B_2}(1-\mathrm{d}) = (\mathrm{B_1} - \mathrm{B_2})\mathrm{d} + \mathrm{B_2} \end{aligned} \tag{9}$$

From Eq. (8), the transfer function corresponding outputs to the duty cycle can be expressed as

$$\frac{\widehat{V}_o(s)}{\widehat{d}(s)} = \frac{-4.9448(s + 7.779e04)(s - 2.848e04)}{(s + 7.692e04)(s + 2.963e04)} \tag{10}$$

$$\frac{\widehat{i}_{Lp}}{\widehat{d}} = \frac{3.8112e05(s + 2653)}{(s + 7.692e04)(s + 2.963e04)} \tag{11}$$

Utilizing MATLAB software, we constructed open-loop Bode diagrams to illustrate the relationship between the system's output and control input, employing the simulation parameters drawn in Table 1. Upon analysis of these Bode plots, the inductor current to duty shown in Fig. 5a is comparably stable, while output voltage to duty as shown in Fig. 5b, has become evident that the system exhibits an inadequate gain margin of − 13.9 dB and instability in the open-loop configuration. Consequently, to improve the gain margin and enhance system stability, developing a dedicated controller is considered necessary. A closed-loop controller is designed with different linear and nonlinear controllers for stability and performance analysis of the converter. PID controller is applied on the linearized and perturbated transfer function (10), whereas SMC and Lyapunov stability function is directly implemented on the differential equation as obtained in (1)–(4).

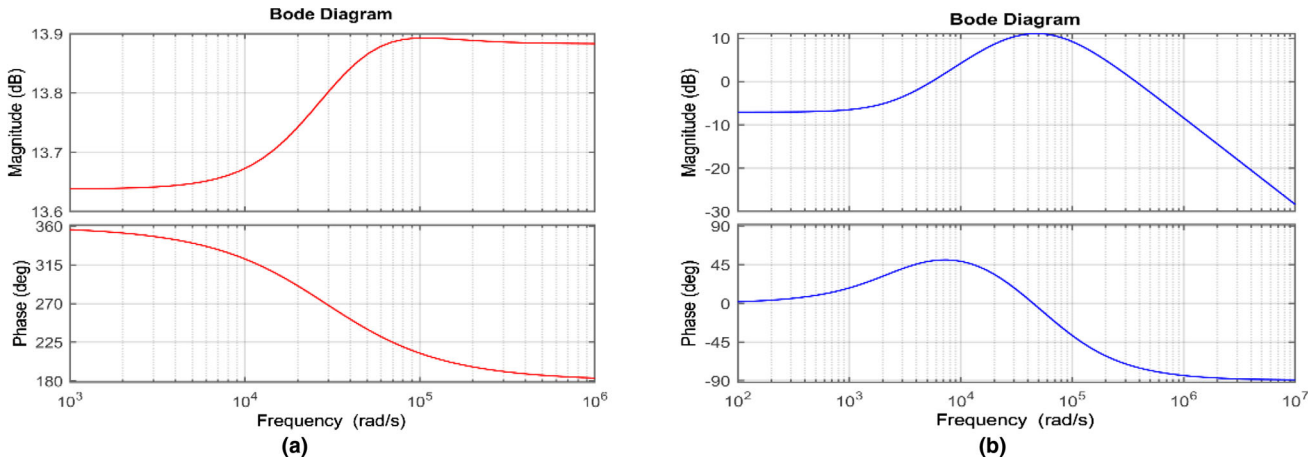


Fig. 5 Open-loop bode plot for flyback converter **a** output voltage to duty, **b** inductor current to duty

Table 1 Specification for flyback converter

Description	Parameter	Value
Power	P_o	10 W
Input voltage	V_s	220V (R.M.S.)
Output voltages	V_{o1}/V_T	180V/5V (DC.)
Frequency	f_s	100 kHz
Inductance	L_p	1.5 mH
Leakage inductance	L_k	1.3 μH
Capacitance	C	220 μF
Parasitic capacitor resistance	Rc	0.01 Ω
Parasitic inductor resistance	R_L	2.1 Ω
Parasitic diode resistance	R_d	0.1 Ω
Forward voltage drop	V_d	0.7 V
Switch resistance	R_{sw}	0.8Ω
Load resistance	R_{o}	2.5Ω
Turn ratio	n	320/6

3 Control Methodology and Stability Investigation

Controller designing is a critical aspect of system control, aimed at developing control strategies that ensure desired system behavior and performance [15]. The main objective of the non-ideal flyback converter controller is to keep the DC link voltage near the appropriate reference value as feasible. This voltage regulation is necessary for both operating modes, resulting in the same steady-state duty cycle for both modes. Traditionally, a proportional-integral control scheme based on a linearized model has been commonly employed to stabilize the converter. However, when subjected to significant disturbances, this technique produces a poor transient response because the accuracy of the small-signal model at each operational point is limited. To address these limitations, we propose a Lyapunov function-based controller that achieves quick dynamic response and global exponential stability even in the presence of notable disturbances. This controller consists of two units. The proposed control unit, the first, enables fast dynamic responses for the flyback converter and assures global exponential stability despite significant disturbances. The second component, referred to as the duty-ratio feedforward controller, reduces the burden for the feedback controller.

3.1 Duty-Ratio Feedforward Controller

The duty-ratio feedforward control element, which is essential for analyzing the error dynamics, is introduced in this section. The suggested feedback controller is designed using the findings from the dynamic error model. For further progress, we will establish the reference magnetizing inductor current as I_{Lp} and the reference output voltage as V_T. The flyback converter is assumed to function under stable, steady-state conditions. By substituting I_{Lp} and V_T. into the expressions for $\bar{i}_{Lp}$ and $\bar{v}_T$ We proceed with the analysis, equating (3) to (5) with zero.

$$I_{Lp} = -\frac{\frac{d_{\text{ff}} * V_g}{L_k+L_p} + \frac{(d_{\text{ff}}-1)}{L_p}\left\{nV_d + \frac{nV_cR_{\text{o}}}{R+R_c}\right\}}{\frac{\left(R_d n^2 + R_L + \frac{n^2 R_{\text{o}} R_c}{R_{\text{o}}+R_c}\right) * (d_{\text{ff}}-1)}{L_p} - \frac{d_{\text{ff}}(R_L+R_{\text{sw}})}{L_k+L_p}} \tag{12}$$

$$V_c = n(1-d_{\text{ff}})I_{Lp}R_{\text{o}} \tag{13}$$

Substituting (12) in (13), we then obtain, which in turn, can be arranged as

$$\propto_1 d_{\text{ff}}^2 + \propto_2 d_{\text{ff}} + \propto_3 = 0 \tag{14}$$

where

$$\begin{aligned}\propto_1 = &\left(-L_p R_o^2 V_g n - L_k R_o^2 V_c n^2 - L_p R_o^2 V_c n^2 - L_k R_o^2 V_d n^2\right.\\ &\left.-L_p R_o^2 V_d n^2 - L_p R_o R_c V_g n - L_k R_o R_c V_d n^2 - L_p R_o R_c V_d n^2\right)\end{aligned} \tag{15}$$

$$\begin{aligned}\propto_2 = &L_p R_o^2 V_g n + 2L_k R_o^2 V_c n^2 + 2L_p R_o^2 V_c n^2\\ &+ 2L_k R_o^2 V_d n^2 + 2L_p R_o^2 V_d n^2 + L_k R_o R_L V_c\\ &- L_p R_o R_{sw} V_c + L_k R_c R_L V_c - L_p R_c R_{sw} V_c\\ &+ L_p R_o R_c V_g n + L_k R_o R_c V_c n^2 + L_p R_o R_c V_c n^2\\ &+ 2L_k R_o R_c V_d n^2 + L_k R_o R_d V_c n^2 2L_p R_o R_c V_d n^2\\ &+ L_p R_o R_d V_c n^2 + L_k R_c R_d V_c n^2\\ &+ L_p R_c R_d V_c n^2\end{aligned} \tag{16}$$

$$\begin{aligned}\propto_3 = &- L_k R_o^2 V_c n^2 - L_p R_o^2 V_c n^2 - L_k R_o^2 V_d n^2\\ &- L_p R_o^2 V_d n^2 - L_k R_o R_L V_c - L_p R_o R_L V_c\\ &- Lk R_c R_L V_c - L_p R_c R_L V_c - L_k R_o R_c V_c n^2\\ &- L_p R_o R_c V_c n^2 - L_k R_o R_c V_d n^2 - L_k R_o R_d V_c n^2\\ &- L_p R_o R_c V_d n^2 - L_p R_o R_d V_c n^2 - L_k R_c R_d V_c n^2\\ &- L_p R_c R_d V_c n^2\end{aligned} \tag{17}$$

d_{ff} can be calculated from Eq. (14) as

$$d_{\text{ff}} = \frac{A_1 \pm \sqrt{A_2}}{A_3} \tag{18}$$

where $d_{\text{ff}} = \frac{A_1 - \sqrt{A_2}}{A_3}$ is considered since its value is less than one. The output voltage of the flyback CCM converter is not directly influenced by d_{ff} itself. Instead, d_{ff} serves as a feedforward input to closely track the reference output voltage. This approach improves the feedback controller's workload, allowing it to operate with lower gain since it no longer needs to regulate the output voltage precisely.

3.2 Feedback Control System Utilizing Lyapunov Functions

Apply the feedforward duty d_{ff} to extract the error dynamics to generate the proposed feedback controller for the converter. The error dynamics $ei_{Lp} = \bar{i}_{Lp} - I_{Lp}$, $ev_c = \overline{v}_c - V_c$, and feedback control input as $d_{\text{fb}} = d - d_{\text{ff}}$. Substituting ei_{Lp}, ev_c and d_{fb} into Eqs. (3) to (5) results to

$$\begin{aligned}\frac{\mathrm{d}ei_{L_p}}{\mathrm{d}t} = &\left(\frac{V_g - (R_L + R_{\text{sw}})(I_{Lp} + ei_{Lp})}{L_k + L_p}\right.\\ &+ \frac{(I_{Lp} + ei_{Lp})\left(R_L + R_d n^2 + \frac{R_o R_c n^2}{R_o + R_c}\right)}{L_p}\\ &\left.+ \frac{nV_d}{L_p} + \frac{nR_o(V_c + ev_c)}{L_p(R_o + R_c)}\right)d_{\text{fb}}\\ &+ (d_{\text{ff}} - 1)\\ &\left(\frac{(I_{Lp} + ei_{Lp})\left(R_L + R_d n^2 + \frac{R_o R_c n^2}{R_o + R_c}\right)}{L_p} + \frac{V_d n}{L_p} + \frac{R_o n(V_c + ev_c)}{L_p(R_o + R_c)}\right)\\ &+ \frac{d_{\text{ff}}(V_g - (R_L + R_{\text{sw}})(I_{Lp} + ei_{Lp}))}{L_k + L_p}\end{aligned} \tag{19}$$

$$\begin{aligned}\frac{\mathrm{d}ev_c}{\mathrm{d}t} = &\left(-\frac{R_o n(I_{Lp} + ei_{Lp})}{C(R_o + R_c)}\right)d_{\text{fb}}\\ &- \frac{V_c + ev_c - R_o n(I_{Lp} + ei_{Lp}) + R_o d_{\text{ff}} n(I_{Lp} + ei_{Lp})}{C(R_o + R_c)}\end{aligned} \tag{20}$$

The error dynamics (19) and (20) can be summarized as

$$\dot{e} = \mathrm{G}e + (\mathrm{H}e + h)d_{\text{fb}} \tag{21}$$

where

$$e = \left[ei_{Lp} ev_c\right]^T \tag{22}$$

$$G = \begin{bmatrix} -\frac{(R_L + R_{\text{sw}})}{L_k + L_p} d_{\text{ff}} - \frac{(R_d n^2 + R_L + \frac{R_o R_c n^2}{R_o + R_c})(1 - d_{\text{ff}})}{L_p} & -\frac{nR_o(1 - d_{\text{ff}})}{L_p(R_o + R_c)} \\ \frac{nR_o(1 - d_{\text{ff}})}{C(R_o + R_c)} & -\frac{1}{C(R_o + R_c)} \end{bmatrix} \tag{23}$$

$$H = \begin{bmatrix} -\frac{R_L + R_{\text{sw}}}{L_k + L_p} + \frac{(R_d n^2 + R_L + \frac{R_o R_c n^2}{R_o + R_c})}{L_p} & \frac{nR_o}{L_p(R_o + R_c)} \\ -\frac{nR_o}{C(R_o + R_c)} & 0 \end{bmatrix} \tag{24}$$

$$h = \begin{bmatrix} \left(-\frac{R_L + R_{\text{sw}}}{L_k + L_p} + \frac{\left(R_d n^2 + R_L + \frac{R_o R_c n^2}{R_o + R_c}\right)}{L_p}\right) I_{Lp} + \frac{V_g}{L_k + L_p} \\ + \frac{nR_o}{L_p(R_o + R_c)} V_c + \frac{n}{L_p} V_d - \frac{nR_o}{C(R_o + R_c)} I_{Lp} \end{bmatrix} \tag{25}$$

The Lyapunov stability theorem states that a system attains exponential stability when the specified condition is satisfied: The function $V(x)$ satisfies $c_1\|x\|^\alpha \le V(x) \le c_2\|x\|^\alpha$ and $\dot{V}(x) \le -c_3\|x\|^\alpha$ criteria, where $0 < c_1 < c_2 < c_3$, and α is a positive constant. The potential Lyapunov function represents the energy held in inductors and capacitors.

Using the Lyapunov function, we can depict the energy stored in the inductor and output capacitor:

$$V(e) = \frac{1}{2} L_p ei_{Lp}^2 + \frac{1}{2} C ev_c^2 = \frac{1}{2} e^T P e \tag{26}$$

where $P = \mathrm{diag}[L_p C]$, then it holds that

$$\begin{aligned}&\frac{1}{2}\min\left(L_p, C\right)\|e\|_2^2 = \frac{1}{2}\lambda_{\min}(P)\|e\|_2^2 \le V(e)\\ &\le \frac{1}{2}\lambda_{\max}(P)\|e\|_2^2 = \frac{1}{2}\max\left(L_p, C\right)\|e\|_2^2\end{aligned} \tag{27}$$

Taking the derivative of Eq. (27) with respect to time, we obtain the following result:

$$\dot{V}(e) = \frac{1}{2}\dot{e}^T P e + \frac{1}{2} e^T P \dot{e}$$

$$= -e^T Qe + d_{fb} e^T Ke + d_{fb} h^T Pe \tag{28}$$

where

$$Q = -\frac{1}{2}\left(G^T P + PG\right)$$

$$= \begin{bmatrix} L_p\left(\frac{d_{ff}(R_L+R_{sw})}{L_k+L_p} - \frac{(d_{ff}-1)\left(R_L+R_d n^2+\frac{R_o R_c n^2}{R_o+R_c}\right)}{L_p}\right) & 0 \\ 0 & \frac{1}{R_o+R_c} \end{bmatrix}$$

$$> 0 \tag{29}$$

$$K = \frac{1}{2}\left(H^T P + PH\right)$$

$$= \begin{bmatrix} L_p\left(\frac{R_L+R_d n^2+\frac{R_o R_c n^2}{R_o+R_c}}{L_p} - \frac{R_L+R_{sw}}{L_k+L_p}\right) & 0 \\ 0 & 0 \end{bmatrix} \tag{30}$$

By choosing feedback controller as

$$d_{fb} = -\alpha(e^T Ke + h^T Pe) \tag{31}$$

where $\alpha > 0$, substituting (31) in (28)

$$\dot{V}(e) = -e^T Qe - \alpha\left(e^T Ke + h^T Pe\right)^2$$

$$\leq \frac{1}{2}\lambda_{\min}(P)\|e\|_2^2$$

$$= -\min\left(\begin{bmatrix} L_p\left(\frac{d_{ff}(R_L+R_{sw})}{L_k+L_p} - \frac{(d_{ff}-1)\left(R_L+R_d n^2+\frac{R_o R_c n^2}{R_o+R_c}\right)}{L_p}\right) & 0 \\ 0 & \frac{1}{R_o+R_c} \end{bmatrix} \right)\|e\|_2^2$$

$$< 0 \tag{32}$$

From (27) and (32), the stability theorem $eiL_p \rightarrow 0 \& ev_T \rightarrow 0$

exponentially fast as $t \rightarrow 0$. Relating (27) and (32), we obtain

$$\dot{V}(e) \leq -\frac{2\lambda_{\min}(Q)}{\lambda_{\max}(P)} V(e) \tag{33}$$

and solving (33) gives

$$V(e(t)) \leq V(e(t_o))\exp\left\{-\frac{2\lambda_{\min}(Q)}{\lambda_{\max}(P)}(t-t_o)\right\} \tag{34}$$

For all $t \geq t_o$, then the error trajectory [25] is bounded as

$$\|e\|_2 \leq \sqrt{\left\{\frac{2V(e(t))}{\lambda_{\min}(P)}\right\}}$$

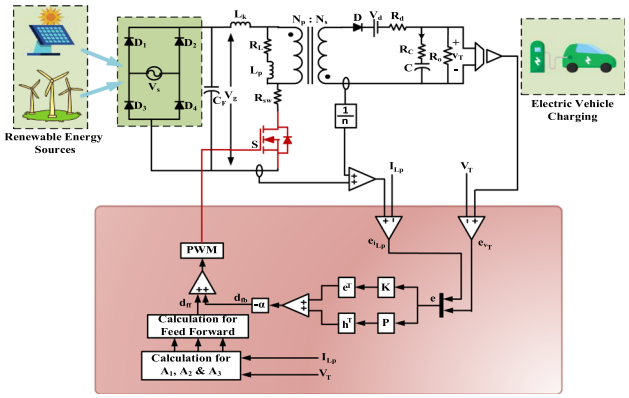


Fig. 6 Proposed Lyapunov-based flyback CCM converter control system with parasitics

$$\leq \left[\frac{\lambda_{\max}(P)\|e\|_2^2 \exp\left\{-\frac{2\lambda_{\min}(Q)}{\lambda_{\max}(P)}(t-t_o)\right\}}{\lambda_{\min}(P)}\right]^{\frac{1}{2}} = \left\{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}\right\}^{\frac{1}{2}} \|e(t_o)\|_2$$

$$\exp\left\{-\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}(t-t_o)\right\} \tag{35}$$

As a result, the error trajectories converge rapidly to zero with an exponential rate of $-\frac{\lambda_{\min}(Q)}{\lambda_{\max}(P)}$

The complete control input is subsequently determined using Eqs. (18) and (31) as follows:

$$d = d_{ff} + d_{fb} = \frac{A_1 - \sqrt{A_2}}{A_3} - \alpha\left(e^T \mathrm{K} e + h^T \mathrm{P} e\right) \tag{36}$$

Following Eq. (36), the complete control strategy for the flyback converter considering ESRs is shown in Fig. 6. This average value of inductor current I_{Lp} and capacitor voltage V_c is sensed and compared with their references, thereby generating errors. These errors are minimized using the proposed controller, whose derivation is discussed earlier, and a control law (Eq. 36) is developed.

4 Controller Implementation on Converter

To confirm the validity of the theoretical analysis, we establish the following design specifications. $V_{in} = 220\,\mathrm{VAC}$, $V_{o1} = 180\,\mathrm{V}$, $V_T = 5\,\mathrm{V}$, $P_o = 10\,\mathrm{W}$, and switching frequency $f_S = 100\,\mathrm{kHz}$ and the remaining parameters are shown in Table 1. The closed-loop system was simulated to evaluate how it reacts to changes in the reference output voltage, input voltage, and output load. These tests help us understand the system's transient behavior under various conditions.

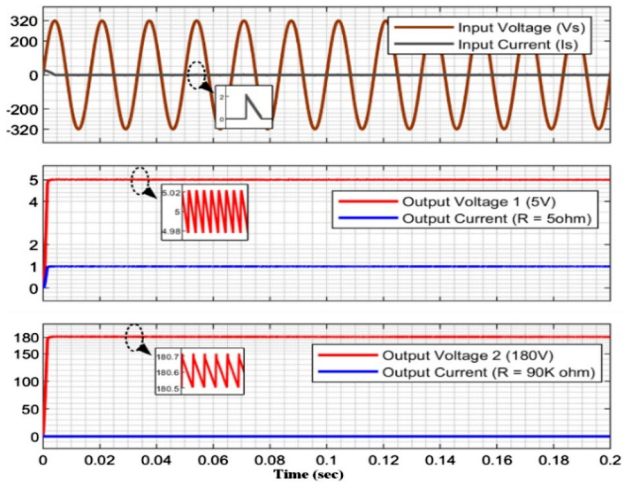


Fig. 7 Input and output profile of PI-controlled flyback converter

4.1 Linear Pi Controller

The paper presents the development of a simulation model for a Flyback Converter using PI control, where essential parameters are considered, and the calculated values of K_p and K_i are implemented.

$$u(t) = K_P e(t) + K_I \int e(t)\mathrm{d}t \tag{37}$$

where

$$e(t) = v_{\text{ref}} - v_c \tag{38}$$

Through successive trial and error, the values of K_p and K_i are determined as 10 and 3.2, respectively. Figure 7 illustrates the flyback converter's output performance under PI control for its voltage and current, respectively. The flyback converter, operating with the PI controller, achieves a peak output voltage of 5.042 V, with a rise time of 49.438 μs to reach the desired output voltage of 5 V. However, when subjected to PI control, the voltage and current output exhibit some ripple. Furthermore, the flyback converter demonstrates a steady-state error of 5.16%.

4.2 Sliding Mode Controller

The flyback converter undergoes simulation using a developed model based on the proposed design. The results of this simulation model are precisely analyzed. To ensure a regulated 5 V output voltage, a Simulink model for the flyback converter is established, including a sliding mode controller (SMC) for closed-loop control. Initially, a sliding surface is formulated to minimize voltage and current errors, adhering to the principles of sliding mode control, where the sliding surface $s(x)$ is set to 0. These are the SMC state variables (Fig. 8):

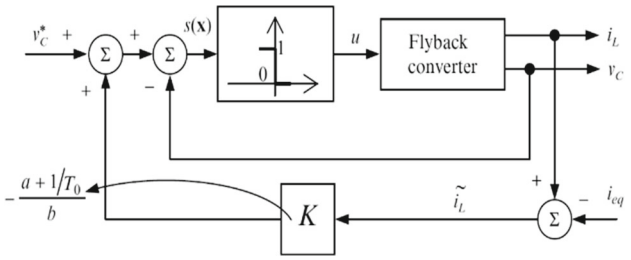


Fig. 8 Control schematic for ISMC for nonlinear flyback converter with voltage and current loop

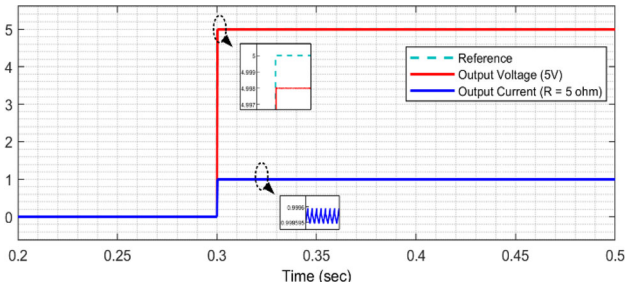


Fig. 9 Performance profile of ISMC-controlled flyback converter for 5 V output voltage

$$x_1 = i_{\text{ref}} - i_{Lp}$$

$$x_2 = v_{\text{ref}} - v_c$$

$$x_3 = \int (x_1 + x_2)\mathrm{d}t \tag{39}$$

The suggested controller sliding surface S is specified as follows:

$$S = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \tag{40}$$

Figure 9 shows the output voltage waveforms observed while utilizing the sliding mode controller, the schematic diagram of which is depicted in Fig. 8.

Before obtaining stability at the desired reference output voltage, the actual output voltage exhibits significant overshooting and minor oscillations. However, the settling time for this waveform is smaller than that depicted in Fig. 7.

The peak output voltage achieved by the flyback converter with the SMC controller is measured at 4.997 V. The rise time required to attain the desired state for the 5 V output voltage is determined to be 470.15 ns. Additionally, the steady-state error for the flyback converter under closed-loop control using sliding mode is calculated to be 0.4%.

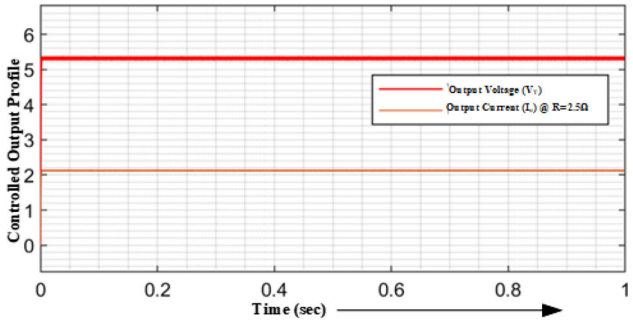


Fig. 10 Performance profile of Lyapunov-based controlled flyback converter for 5 V output voltage

4.3 Lyapunov Function-Based Controller

Figure 6 illustrates the control scheme employed for the multiport flyback converter (MFC) system. Lyapunov function-based controllers use mathematical functions to analyze and ensure the stability of dynamic systems. Simulation results are used to calculate the controller's efficiency and the converter's performance. The converter's responses are examined when both the input voltage and the load change. The parameter α within the designated controller is increased to evaluate the controller's response. Employing a parameter tuning approach, a gain of 0.0004 is achieved in the simulation.

Figure 10 shows the waveforms of both the actual and reference output voltages, considering variations in the output load. Lyapunov controller offers a rise time of 517.4 ns and a settling time of 0.4 ms. Notably, the Lyapunov controller outperforms the SMC in this scenario. This superiority can be attributed to the closed-loop system's output impedance, which is smaller under SMC than under PI control. The output voltage demonstrates reduced peak behavior when the sliding mode controller is used (Fig. 9). However, the proposed controller performs even better, consistently maintaining its output voltage, even when subjected to load variations (as evident in Fig. 10).

The load is adjusted between 0.25 A and 4 A within a duration of 1 s, while the output voltage remains constant at 5 V when the controller is implemented as shown in Fig. 12. In Fig. 13, the input voltage is depicted varying between 290 and 350 V, while the output voltage remains fixed at 5 V due to the Lyapunov controller. The graph demonstrates the controller's effective regulation, ensuring a constant output voltage despite changes in the input voltage range, highlighting its reliability and stability. The load is adjusted between 0.25 A and 4 A within a duration of 1 s, while the output voltage remains constant at 5 V when the controller is implemented. The graph demonstrates the controller's effective regulation, ensuring a constant output voltage despite changes in the input voltage range, highlighting its reliability and stability. The proposed converter stability analysis is performed for a closed loop for output voltage and inductor current to duty. The parasitic resistances of the inductors and the capacitors are included while predicting the magnitude and phase margin. The gain margin is improved to 89 dB for output voltage when the controller is applied, which can be observed from the bode plot, as shown in Fig. 11. The gain margin is positive, and the phase margin is sufficiently large, suggesting that the converter is stable under the specified operating conditions.

Figure 12 showcases the simulation output voltage waveforms of the Lyapunov-controlled flyback converter when subjected to load variations. This provides a visual representation of how the converter responds to changes in the load. In Fig. 13a, simulation graphs demonstrate the behavior of the non-ideal flyback converter when the proposed controller

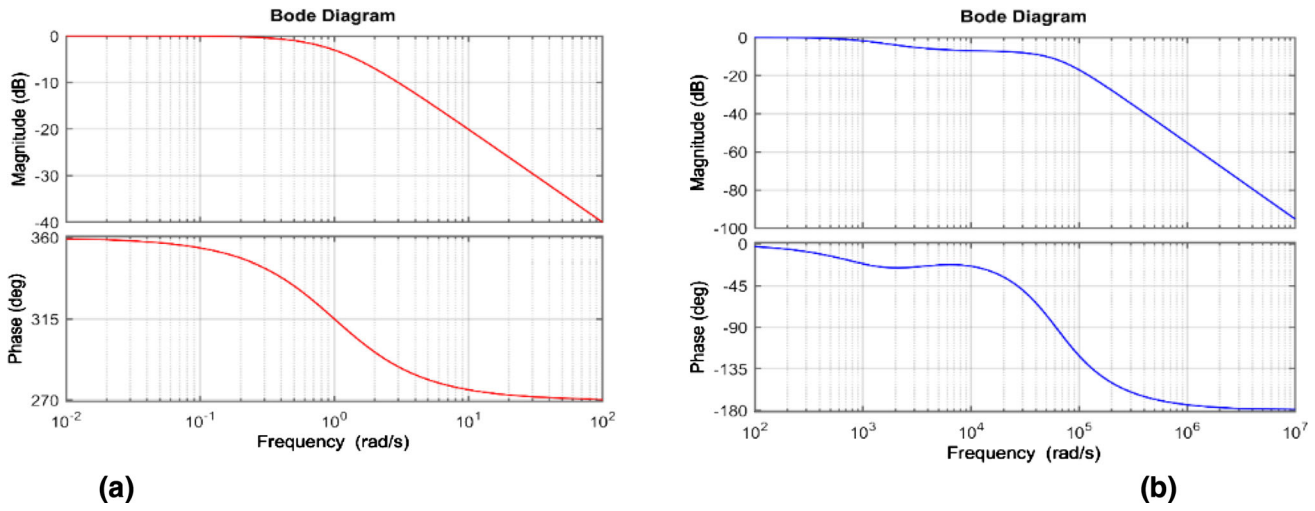


Fig. 11 Closed loop bode plot for flyback converter considering circuit parasitics **a** output voltage to duty, **b** inductor current to duty

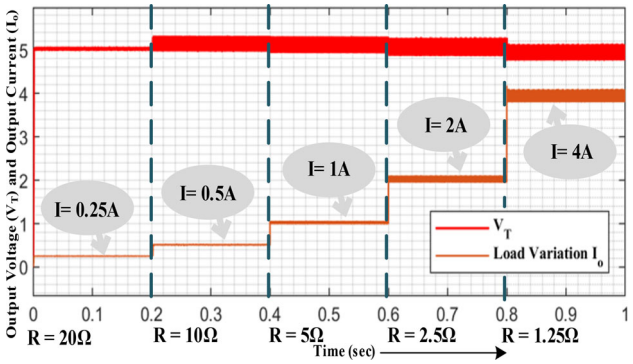


Fig. 12 Simulation output voltage waveforms of Lyapunov-controlled flyback converter under load variations

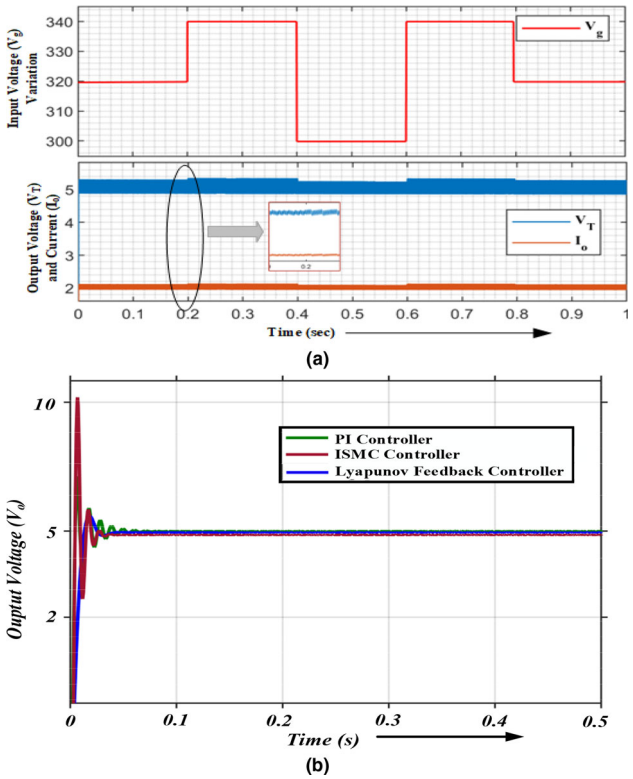


Fig. 13 **a** Simulation graphs demonstrate the behavior of the non-ideal flyback converter when the proposed controller is applied in response to changes in input voltage. **b** Comparative analysis of different controller applied on the converter

is applied in response to fluctuations in input voltage. This allows for an examination of how the converter behaves under varying input conditions with the proposed control strategy. Additionally, Fig. 13b presents a comparative analysis of the converter's performance with different controllers applied. This comparison offers insights into the effectiveness of the proposed controller in comparison with other control strategies in regulating the flyback converter.

Table 2 Time domain specifications

Controller	Rise time	%Peak undershoot (%)	Settling time
PI	49.43 µs	5	32 ms
SMC	0.47 µs	2.6	20 ms
Proposed controller	517.4 ns	1.3	0.4 ms

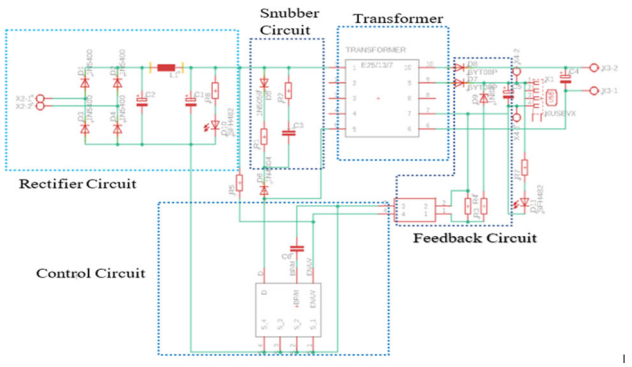


Fig. 14 Multi-output flyback converter circuit with the controller

In assessing the effectiveness of the proposed controller, we examined various parameters. These parameters included the rise time and settling time of transient responses when subjected to changes in the reference output voltage and the percentage of overshoot and settling time in response to variations in the input voltage and output load (refer to Table 2). Notably, when compared to other controllers, the proposed controller has the fastest and most reliable step response (Fig. 13b). The complete representation of the converter design can be explained by considering the converter's subparts, as shown in Fig. 14.

The comparison reveals that the proposed controller exhibits a significant reduction in settling time and oscillations compared to both PI and SMC controllers. Although there is a slight increase in undershooting, the substantial improvement in settling time outweighs this drawback. Consequently, the overall performance of the proposed controller surpasses that of the PI and SMC controllers. Figures 15 and 16 show the variation of V(e) and $\dot{V}$(e) , respectively, with respect to e_{iL_p} and $e(v_T)$ when the proposed controller is applied for various disturbances, and the surface shows the convergence of error to zero as the system reaches the steady state at fixed loading conditions without affecting the system stability, even with circuit parasitics.

5 Experimental Results

The converter was designed with four subsections, namely the rectifier circuit, snubber circuit, feedback circuit, and

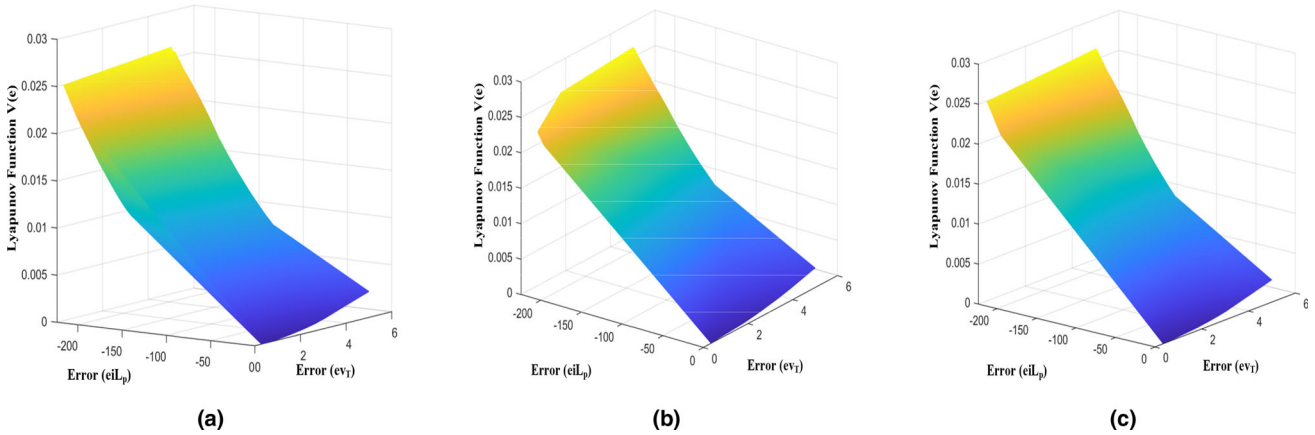


Fig. 15 Surface graph of the flyback converter for variations in $V(e)$ with respect to eiL_p and ev_T when the proposed controller is applied for **a** line variation, **b** load variation, **c** both line and load variation

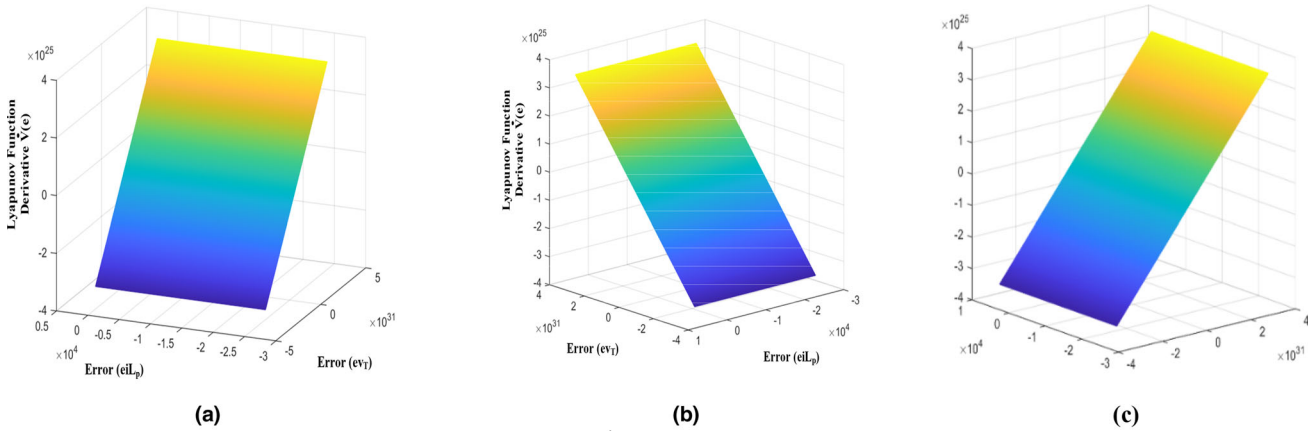


Fig. 16 Surface graph of the flyback converter for variations in $\dot{V}(e)$ with respect to eiL_p and ev_T when the proposed controller is applied for **a** line variation, **b** load variation, **c** both line and load variation

switch controller circuit. A full bridge rectifier, comprised of four diodes, was used to generate an unregulated DC supply, followed by a pi filter consisting of two capacitors and an inductor. The snubber circuit was designed to prevent voltage surges on the primary side of the transformer during transient conditions, thereby protecting the transformer from spikes. An auxiliary bias winding on the high-frequency Flyback transformer also provided overvoltage protection in case of open feedback loop faults in the circuit. An LED indicates the output state. The subsequent section will delve further into the design and implementation of the flyback converter using a customized flyback transformer. Parameters for the flyback converter are calculated in steps and initially analyzed on the Simulink model before designing. A flyback transformer is designed using the parameter given in Table 3. The crucial part was designing the air gap inductance for which an air gap of 0.12 mm was provided using paper. The complete representation of the converter design can be explained by considering the converter's subparts, as shown in Fig. 14. The IN4007 diode was selected for bridge

Table 3 Specification for flyback transformer

Description	Parameter	Value
Value of primary current	I_1	0.177 A (R.M.S.)
Value of secondary current	I_2	6.24A (R.M.S.)
Value of tertiary current	I_3	0.207 A (R.M.S.)
Packing factor	K_w	0.5
Frequency	f_s	100 kHz
Inductance	L_p	1.5 mH
Leakage inductance	L_k	1.3 µH
Current density	J	3 A/mm^2
Core window area	A_w	10.937 mm^2
Parasitic inductor resistance	R_L	2.1 Ω
Length of air gap	l_g	0.12 mm
Gauge of a_1 winding	a_1	SWG = 34
Gauge of a_2 winding	a_2	SWG = 15
Gauge of a_3 winding	a_3	SWG = 31

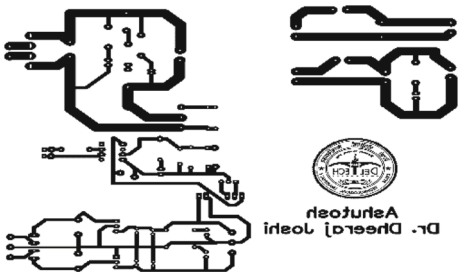


Fig. 17 PCB layout of hardware model for flyback converter with two outputs

rectification of AC input, ensuring that the maximum voltage across diodes must not exceed 700 V. The snubber circuit prevents voltage surge represented on the transformer's primary side throughout transient conditions, thus protecting the transformer from spikes. The use of parallel RC optimizes both energy efficiency and EMI. Further, additional bias winding on the flyback transformer offers overvoltage protection in the event of open feedback loop faults in the circuit; LED, which is used as an indicator, specifies the output state. A Flyback converter is a coupled inductor with a gapped core. During each cycle, energy is stored in the core's gap when primary winding is induced with input voltage. It is then transferred to the secondary winding to provide power to the load. Now, we can emphasize the transformer design as the necessary constraints for the converter design have been calculated. One of the prevalent methods in the design of high frequency is circuit isolation and voltage transformation in a flyback converter. The flyback transformer parameter is shown in Table 3.

By finding all the parameters for the flyback transformer, the next step is to design converter schematics. The core used here was a high-frequency ferrite EE core. Figure 17 shows the PCB layout of hardware model of a multi-output flyback converter with an output of 5 V and 180 V DC with 2 A and 0.002 A current, respectively. The hardware is designed to provide power to charge batteries and power the microcontroller.

The size of the developed prototype is $2.5^{"}x4^{"}x1.5^{"}$. Its power density is 0.67 W/inch^3. This flyback converter module can be incorporated into MMC to supply three-phase supplies. This design features a snubber circuit to prevent transient spikes. The transformer is designed to provide features such as protection against overvoltage and undervoltage. It can provide power to instruments with high sensitivity to spikes and transients. The PCB layout is designed using Fusion 360 software. The calculation for the design has been explained in an earlier section.

Figure 18 shows the experimental configuration to employ flyback converter hardware with a standalone PV-connected system. The ac power supply through an autotransformer, a LV25-P voltage sensor and an ACS712 current sensor

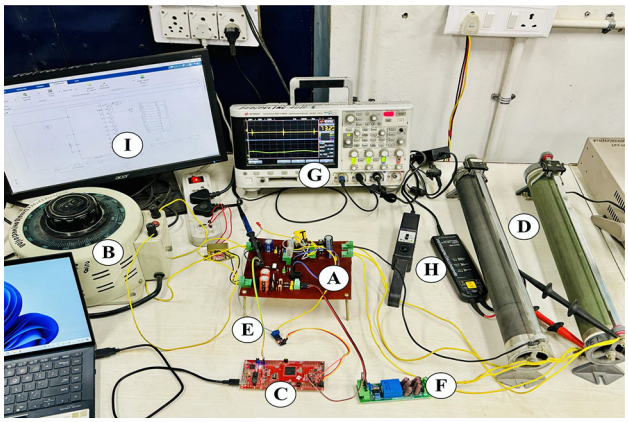


Fig. 18 Experimental setup for hardware implementation A. Multiport flyback converter prototype, B. Input Source, C. TMS320F28379D Microcontroller, D. Rheostat, E. Current Sensor, F. Voltage Sensor, G. DSO, H. Voltage and Current Probe and I. Computer

are used to detect voltage and current, respectively. A TMS320F28379D microcontroller is used for PWM pulse generation. A DSO of type KEYSIGHT DSO-X2024A is used to record the obtained results.

The control algorithm, which generates an adjustable duty cycle gate pulse through a controller, is presented and implemented. Throughout the period when power MOSFET is turned on, drain current keeps rising in the IC and MOSFET is switched off shortly as the drain-source current limit is reached by the current. Furthermore, a second switch attached to the load helps to highlight load variations and is controlled by its own gate pulse. The complete flyback converter model and all necessary calculations and circuitry are integrated onto a PCB and thoroughly tested in the laboratory. The hardware is subjected to various resistive loads to analyze its regulation characteristics throughout the testing process. A usability test is conducted to improve the hardware design. Furthermore, critical component tests are performed to evaluate their performance before finalizing the prototype. Each section of the setup is labeled with its respective name to facilitate clear identification and comprehension.

Figures 19, 20, 21 and 22 show the Flyback converter's DSO output. An input supply is taken from a regular 230 V single-phase supply to power the flyback converter to test its performance. This section examines how the proposed control strategy performs on the MFC when there are changes in the reference values of input and output voltages. Before that, it provide the steady-state response with a resistive load of R = 10 Ω, as illustrated in Fig. 19.

Similarly, AC input varies from 100 to 260 V using a variable in Fig. 20 on a closed-loop controlled flyback converter, which shows strong reluctance in output voltage to input variation. An error of less than 3% is observed while varying the source. The non-ideal flyback converter has a practical efficiency of 86.4% when it is loaded on rated conditions. The

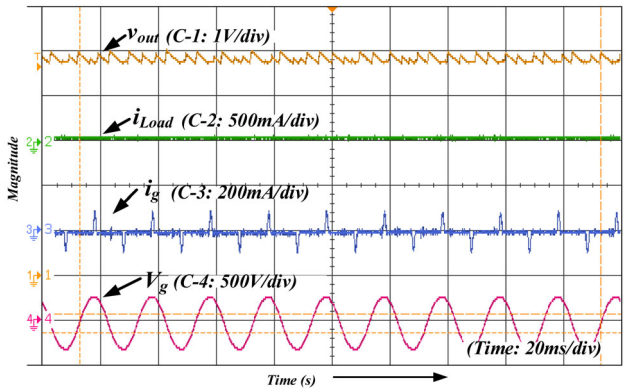


Fig. 19 Hardware steady-state waveforms of the flyback converter

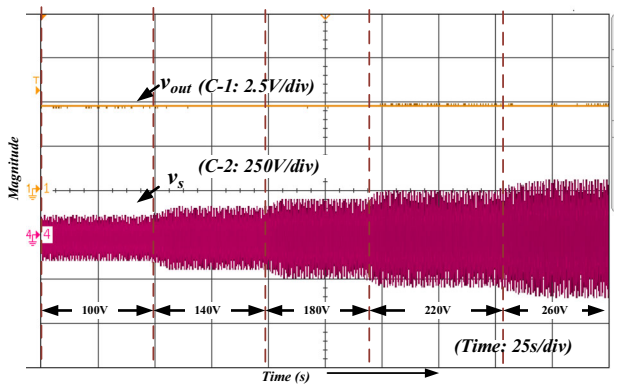


Fig. 20 Experimental waveforms of the flyback converter under input voltage variation with the steady-state error of 3%

variation in efficiency can be seen in Fig. 24. The drop in efficiency is observed due to the ESRs of the components.

Figure 21 also shows the MFC's transient responses when subjected to load fluctuations. As shown in this Fig. 21, the output voltage maintains stability with an excellent transient response time, allowing for a proportionate reduction in a

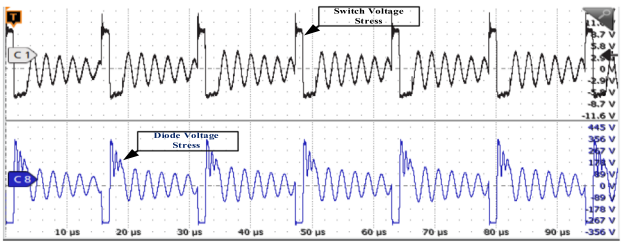


Fig. 22 Voltage stress on the switch and diode of the flyback converter

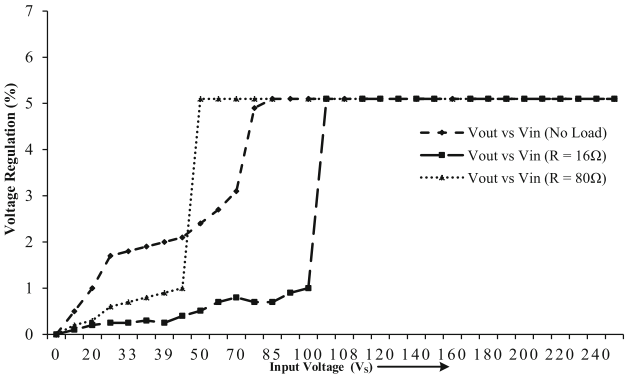


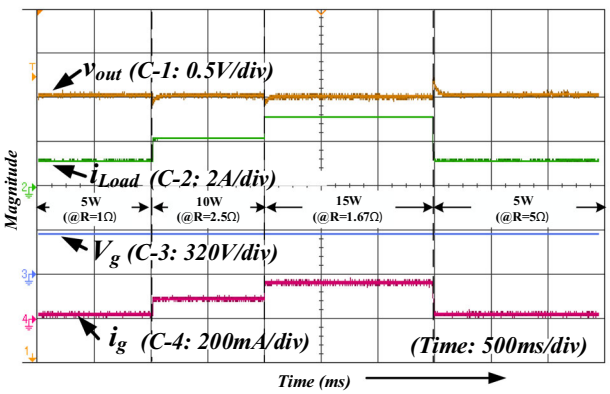
Fig. 23 Variation of output voltage with input voltage under no load condition

portion of the output voltage as needed. A less than 2% percentage error is observed within acceptable limits for resistive load variation from 1 to 10 Ω.

Figure 22 typically shows two separate curves representing the stress on the MOSFET and diode, respectively. The *x*-axis represents time, while the *y*-axis will show the magnitude of the stress.

The different plot for input voltage sweeps over output voltage for changed load is shown in Fig. 23. The plot shows output voltage regulation concerning the input voltage. The input voltage is varied from 0 to 230 V using a variac, and the load is changed from 2 A to very low values, showing no load

Fig. 21 Experimental waveforms of the flyback converter when utilizing the proposed controller while undergoing variations in the output load



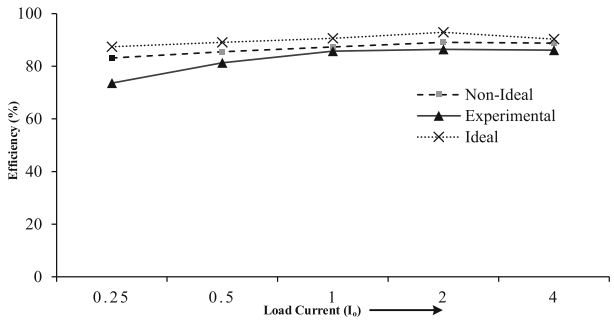


Fig. 24 Efficiency for non-ideal flyback converter for load variation

condition. A percentage error of less than 2% is observed within acceptable limits for the application in gate drivers such as IGBT and MOSFET. Figure 24 depicts the efficiency of the non-ideal flyback converter under load variation from below rated condition to above rated condition.

6 Conclusion

This paper presents a non-ideal flyback converter with dual outputs, controlled by a Lyapunov function-based approach, highlighting its simplicity, cost-effectiveness, and isolation. Extensive investigations were conducted to enhance the robustness and precision of the control method against load and voltage reference variations. The proposed system ensures global exponential stability, rapid transient response, and good gain and phase margins, even considering the effects of equivalent series resistances (ESRs). The inclusion of the duty ratio in the feedforward loop helps alleviate its burden. The average model, which accounts for parasitic components, improves control accuracy. Both numerical simulations and experimental tests confirm the superior performance of the converter in precisely regulating the reference output voltage. Additionally, this paper compares the effectiveness of the Lyapunov-based controller with that of a linear PI controller and a nonlinear integral sliding mode controller (ISMC) in managing an MFC for DC/DC conversion. Experimental results for closed-loop controlled flyback converters demonstrate the Lyapunov-based controller's advantages in handling various perturbations, reducing steady-state error to under 1%, and achieving quicker settling times. Notable stress on the switch and diode, within an acceptable range, will be addressed in future research.

Acknowledgements The author(s) extend their appreciation to the Deputyship for Research & Innovation. Ministry of Education in Saudi Arabia for funding this research work through the project number RI-44-0275.

References

1. Liu, W.-S.; Chen, J.-F.; Liang, T.-J.; Lin, R.-L.; Liu, C.-H.: Analysis, design, and control of bidirectional cascoded configuration for a fuel cell hybrid power system. IEEE Trans. Power Electron. **25**(6), 1565–1575 (2010). https://doi.org/10.1109/TPEL.2009.2037003
2. Imtiaz, A.M.; Khan, F.H.: "Time shared flyback converter" based regenerative cell balancing technique for series connected li-ion battery strings. IEEE Trans. Power Electron. **28**(12), 5960–5975 (2013). https://doi.org/10.1109/TPEL.2013.2257861
3. Marangalu, M.G.; Kurdkandi, N.V.; Hosseini, S.H.; Tarzamni, H.; Dahidah, M.; Sarhangzadeh, M.: A modified switched-capacitor based seventeen-level inverter with reduced capacitor charging spike for RES applications. IEEE Open J. Power Electron. **4**, 579–602 (2023). https://doi.org/10.1109/OJPEL.2023.3302282
4. Pesce, C.; Blasco, R.; Riedemann, J.; Andrade, I.; Peña, R.: A DC-DC converter based on modified flyback converter topology. IEEE Lat. Am. Trans. **14**(9), 3949–3956 (2016). https://doi.org/10.1109/TLA.2016.7785917
5. Papanikolaou, N.P.; Tatakis, E.C.: Active voltage clamp in flyback converters operating in CCM mode under wide load variation. IEEE Trans. Industr. Electron. **51**(3), 632–640 (2004). https://doi.org/10.1109/TIE.2004.825342
6. Marangalu, M.G.; Kurdkandi, N.V.; Babaei, E.: Single-source multilevel inverter based on flyback DC-DC converter. IET Power Electron. **14**, 1237–1255 (2021). https://doi.org/10.1049/pel2.12101
7. Park, J.; Roh, Y.-S.; Moon, Y.-J.; Yoo, C.: A CCM/DCM dual-mode synchronous rectification controller for a high-efficiency flyback converter. IEEE Trans. Power Electron. **29**(2), 768–774 (2014). https://doi.org/10.1109/TPEL.2013.2256371
8. Soltanzadeh, K.; Khalilian, H.; Dehghani, M.: Analysis, design, and implementation of a zero-voltage switching two-switch CCM flyback converter. IET Circuits, Devices System **10**(1), 20–28 (2016)
9. Chiu, H.-J., et al.: A Single-stage soft-switching flyback converter for power-factor-correction applications. IEEE Trans. Industr. Electron. **57**(6), 2187–2190 (2010). https://doi.org/10.1109/TIE.2009.2033622
10. Hamanah, W.M.: Performance analysis of SiC-based DC/DC converter for solar power tower heliostat application. Arab. J. Sci. Eng. **48**, 15207–15221 (2023). https://doi.org/10.1007/s13369-023-08044-9
11. Chetty, P.: Current injected equivalent circuit approach to modelling switching dc-dc converters. IEEE Trans. Aerosp. Electron. Syst. **AES-17**(6), 501–505 (1981)
12. Pandey, S.K.; Patil, S.L.; Phadke, S.B.: Comment on "PWM-Based Adaptive Sliding-Mode Control for Boost DC–DC Converters" [Aug 13 3291–3294]. IEEE Trans. Industr. Electron. **65**(6), 5078–5080 (2018). https://doi.org/10.1109/TIE.2017.2764872
13. Gupta, A,; Rana, K,; Joshi, D.: Design and analysis of multioutput flyback converter under continuous and discontinuous conduction mode using PID. Controller. In: 2020 First IEEE International Conference on Measurement, Instrumentation, Control and Automation (ICMICA), Kurukshetra, India, pp. 1–5. (2020). https://doi.org/10.1109/ICMICA48462.2020.9242853.
14. Rajamani, M.P.E.; Murugappan, M.; Prakash, N.B., et al.: Photovoltaic-based dual output DC–DC converter using gravitational search algorithm-tuned PI and sliding mode controllers. Arab. J. Sci. Eng. (2024). https://doi.org/10.1007/s13369-024-08962-2
15. Mukherjee, N.; Strickland, D.: Control of cascaded DC–DC converter-based hybrid battery energy storage systems—Part II: Lyapunov approach. IEEE Trans. Industr. Electron. **63**(5), 3050–3059 (2016). https://doi.org/10.1109/TIE.2015.2511159

16. Lekić, A.; Stipanović, D.; Petrović, N.: Controlling the Ćuk Converter using polytopic Lyapunov functions. IEEE Trans. Circuits Syst. II Express Briefs **65**(11), 1678–1682 (2018). https://doi.org/10.1109/TCSII.2017.2781621
17. Olm, J.M,; Fossas, E,; Repecho, V,; Dòria-Cerezo, A,; Griñó, R.: Adaptive control-based voltage regulation of a magnetically coupled multiport DC–DC converter for electrified vehicles applications. In: 2020 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. Seville, Spain, (2020). https://doi.org/10.1109/ISCAS45731.2020.9180489
18. Tang, W.; Daoutidis, P.: A bilevel programming approach to the convergence analysis of control-Lyapunov functions. IEEE Trans. Autom. Control **64**(10), 4174–4179 (2019). https://doi.org/10.1109/TAC.2019.2892386
19. Wi, S.-M.; Lee, J.S.; Kim, M.: Exponentially stable Lyapunov-function-based controller for a flyback CCM converter. IEEE Trans. Industr. Electron. **65**(2), 1213–1225 (2018). https://doi.org/10.1109/TIE.2017.2733459
20. Marangalu, M.G.; Kurdkandi, N.V.; Monfared, K.K.; Talebian, I.; Neyshabouri, Y.; Vahedi, H.: A new high step-up SC-based grid-tied inverter with limited charging spike for RES applications. IEEE Open J. Power Electron. **5**, 295–310 (2024). https://doi.org/10.1109/OJPEL.2024.3366165
21. Davoudi, A.; Jatskevich, J.: Parasitics realization in state-space average-value modeling of PWM DC–DC converters using an equal area method. IEEE Trans. Circuits Syst. I Regul. Pap. **54**(9), 1960–1967 (2007). https://doi.org/10.1109/TCSI.2007.904686
22. Kumar, V.; Behera, R.K.; Joshi, D.; Bansal, R.: Power Electronics, Drives, and Advanced Applications. CRC Press, United States (2020)
23. Marangalu, M.G.; Kurdkandi, N.V.; Alavi, P.; Khadem, S.; Tarzamni, H.; Mehrizi-Sani, A.: A new single DC Source five-level boost inverter applicable to grid-tied systems. IEEE Access **11**, 24112–24127 (2023)
24. Zhao, Q.; Fang, Y.; Wang, Z.; Xie, Y.: Small-signal model and control of PV grid-connected micro inverter based on interleaved parallel flyback converter. In: Sun, Z.; Deng, Z. (Eds.) 2013 Chinese Intelligent Automation Conference. Lecture Notes in Electrical Engineering, Vol. 255. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38460-8_66
25. Battula, S.; Garg, M.M.; Panda, A.K.; Senapati, L.: Bidirectional quasi-Z-source DC–DC converter with Lyapunov function-based controller in stand-alone photovoltaic-connected system. Int. J. Circ. Theor. Appl. **51**(5), 2327–2350 (2023). https://doi.org/10.1002/cta.3518

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

International Journal of Circuit Theory and Applications

WILEY

ORIGINAL ARTICLE

Design and Control of DVR Based on Adaptive Bateman Polynomial for Power Quality Improvement

Kanchan Bala Rai | Narendra Kumar | Alka Singh

Department of Electrical Engineering, Delhi Technological University, Delhi, India

Correspondence: Kanchan Bala Rai (kanchan_2k19phdee503@dtu.ac.in)

Received: 19 March 2024 | **Revised:** 5 August 2024 | **Accepted:** 20 August 2024

Keywords: Bateman polynomial (BMP) | dynamic voltage restorer (DVR) | power quality (PQ) issues | voltage source converter (VSC)

ABSTRACT

Power quality issues encompass a spectrum of disturbances that affect the stability and reliability of the power supply. This study explores voltage-related anomalies, including sags, surges, flickers, unbalances, harmonics, interruptions, and swells, examining their origins, characteristics, and repercussions. These issues arise from a myriad of sources, such as load fluctuations, equipment malfunctions, and grid dynamics. The ramifications of these disturbances extend to equipment malfunctions, reduced operational efficiency, and financial losses. Dynamic voltage restorer (DVRs) are series custom power devices for mitigating these voltage-related anomalies and use real-time monitoring and control to quickly inject exact voltages into the grid during disruptions, restoring voltage levels to acceptable levels. This study demonstrates the capacity of DVRs to successfully alleviate power quality concerns, hence improving equipment dependability and system stability by analyzing case studies and simulation findings. The control system used to supply switching signals to the voltage source converter (VSC) of the DVR is based on adaptive Bateman polynomial (ABMP). Traditional DVR control techniques often rely on fixed or linear control strategies, which may not adequately handle varying load conditions and disturbances. ABMP offers an adaptive approach where the control parameters can adjust dynamically based on real-time system conditions. This adaptive capability enhances the accuracy of voltage restoration, ensuring that the DVR responds optimally to varying loads and disturbances. During voltage sag and swell conditions at the grid side, the VSC injects the compensating voltage in series with the feeder with a constant switching frequency. A battery energy-supported system (BESS) based DVR is considered for the proposed system. The supply is connected to the critical and sensitive loads. The proposed control scheme is validated through extensive simulation and experimental results.

1 | Introduction

Power quality (PQ) refers to the reliability and stability of electrical power supply. It encompasses factors like voltage stability, frequency consistency, and absence of disturbances. Good power quality ensures equipment operates efficiently and reliably. Issues like voltage sags, harmonics, and interruptions can degrade power quality. Effective management and monitoring are crucial to maintaining optimal power quality standards. Power quality issues in a distribution system refer to various disturbances and deviations from the ideal sinusoidal waveform of electrical voltage and current. These issues can have adverse effects on electrical equipment and the overall performance of the distribution system. Some common power quality issues in distribution systems include voltage sags, voltage swells, voltage interruptions, harmonics, voltage unbalance, flickers, and so on [1]. Voltage sags are short-term reductions in voltage levels, typically caused by faults or the starting of large motors. They can lead to equipment malfunctions and process disruptions. Voltage swells are short-term increases in voltage levels, which can result from switching events or capacitor bank operations. They can stress equipment insulation and cause overheating. Voltage interruptions are complete failures of voltage supply, often due to faults or

© 2024 John Wiley & Sons Ltd.

switching operations. They can lead to downtime and equipment damage. Harmonics are non-sinusoidal waveform distortions that occur at multiples of the fundamental frequency (e.g., 60 Hz in the United States). Harmonics can cause overheating in motors, transformers, and capacitors and affect sensitive electronic equipment. Voltage unbalance occurs when the voltages in a three-phase system are not equal. It can lead to unequal loading and overheating in motors. Flicker is related to rapid voltage fluctuations that can cause visual discomfort, disrupt sensitive equipment, and affect lighting quality. It often occurs due to load fluctuations [2].

A dynamic voltage restorer (DVR), also known as a series active power filter, is a device that injects compensating voltage into the distribution line to manage the levels of voltage while protecting sensitive loads from power quality concerns such as voltage sag and swell [3–8]. DVRs are specifically designed to target frequent power quality issues in distribution networks, such as voltage sags and swells. These dips and surges can result in equipment damage especially for sensitive electronic loads resulting in malfunctions and data loss. DVRs boast a fast response time, typically in milliseconds. This rapid response ensures near-instantaneous voltage correction, minimizing the impact of voltage disturbances on critical loads. The advantages of DVR include focus on voltage sags and swells, active power injection, improved efficiency and isolation from grid faults. Unlike passive filter based techniques using capacitors and inductors, DVRs actively inject voltage into the line to regulate voltage levels. This allows for dynamic control and correction even during large voltage disturbances. Because DVRs only operate during voltage sags and swells, it minimizes energy losses compared with passive correction methods that continuously draw power. Some already available control schemes such as synchronous reference frame theory (SRFT), second-order generalized integrator (SOGI), and second-order second-order generalized integrator (SO-SOGI) for DVR help to mitigate the voltage related PQ issues; however, they lack in fast dynamic performance under phase and frequency jump. Most of the designed control schemes provide controllability based on voltage sag, swell, and voltage unbalancing conditions. A new adaptive Bateman polynomial (ABMP) controller for DVR is proposed in this research work, and its performance is evaluated under several PQ problems in the utility grid.

A standard three-phase DVR system consists of a three-phase VSC with a DC-link equipped with an energy storage device, like a battery, and a series transformer that isolates the VSC AC voltage from the grid AC voltage. The compensating voltage is then injected into the power distribution line using a controller to retrieve the error voltage signal and produce switching pulses for the voltage source inverter. The voltage fluctuations originating on the supply side are reduced on the load side connected at the point of common coupling (PCC) [9].

In many DVR systems, a battery energy storage system (BESS) is used on the DC side of the VSC. When the grid voltages are sagging or experiencing a disturbance, the DVR draws power from the battery to inject into the system. This injected power not only corrects the voltage but also supports the real power needs of the load. During normal operation, the DVR can recharge the battery when there is excess power available. In some cases, instead of a battery, a rectifier or an active front end can be used on the DC side of the VSC. This rectifier converts the incoming AC power from the grid into DC power, which is then used to supply real power to the load during disturbances. When the grid voltages are within acceptable limits, the DVR operates in a pass-through mode, allowing the AC power to flow directly to the load. These components (battery, rectifier, or active front end) are essential for ensuring that the DVR can provide both reactive power (voltage support) and real power (active power support) when needed. This capability makes the DVR a versatile and comprehensive solution for addressing power quality issues and maintaining the stability of the electrical grid [10, 11].

Previous works [12, 13] discuss synchronous reference frame theory implemented for the control of DVR. Naidu et al. [14] present an improved enhanced phase lock loop (PLL) technique for the DVR controller. The angle between the PCC and load voltages during disturbances is calculated using the dual P-Q theory-based energy-optimized DVR, which is discussed in Pradhan and Mishra [15]. A novel Instantaneous dq transform detection mechanism is proposed in Zheng et al. [16] for DVR with the rapid response superconducting magnetic energy storage (SMES) device, along with a pre-sag mitigation technique. The dynamic reaction of DVR is considerably improved by the SMES's quick response time [17]. Further, the superconducting fault current limiter is proposed to reduce the external fault current in addition to SMES device-based DVR. Most of the control schemes have been tested for voltage sag, swell, harmonics, and unbalancing, but very little research has been done on load unbalancing with linear and nonlinear loads. A mixed second- and third-order generalized integrator (MSTOGI) SRF-based control scheme can mitigate voltage-related PQ issues with balanced and unbalanced linear/nonlinear loads. Due to its straightforward structure, quick dynamic response, and simple software implementation, the synchronous reference frame PLL (SRF-PLL) [18, 19] has been utilized extensively in grid-connected systems. However, the output of SRF-PLL gives erroneous results if there are imbalances or harmonics in the grid voltages, making it impossible to precisely track the fundamental positive sequence component of the grid voltages. This may even compromise the stability of grid-connected systems. In addition to having a more straightforward structural design, the second-order generalized integrator PLL (SOGI-PLL) [20] also efficiently achieves accurate phase locking even in non-ideal grid voltages circumstances while also removing the delay caused by LPFs in the double decoupled SRF (DDSRF-PLL). Because of its improved performance, the SOGI-PLL has recently been the subject of intensive research [21]. When grid voltages contain harmonics and DC offset, SOGI PLL generates some error when extracting the fundamental positive grid voltages and fails to accurately determine PLL signals.

Bateman polynomials (BMPs) are a class of functions used to represent time-dependent responses in dynamic systems. In the context of power quality improvement, these polynomials can effectively model the behavior of voltage disturbances and load variations over time. The key innovation of the ABMP method lies in its adaptive capability. Unlike traditional fixed-parameter control methods, ABMP adjusts its control

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

parameters in real-time based on the actual conditions of the power system. This adaptability allows the DVR to respond optimally to changing load conditions, disturbances, and variations in the distribution network. The ABMP proposed in this work provides mitigation of problems such as voltage sag swell, unbalance, harmonics at PCC, and load unbalancing. The proposed strategy makes use of an adaptive Bateman's polynomial controller in conjunction with a dynamic voltage restorer (DVR). The most recent advances for addressing PQ problems usually depend on traditional control approaches that may not be flexible enough to adjust to the changing grid conditions. This work introduces adaptive Bateman's polynomial based controller for the DVR in order to overcome this constraint. Real-time variations to the compensation voltage based on dynamic changes in the grid voltage are possible by suitable design of the adaptive polynomials. This contribution has the potential to significantly enhance the effectiveness of the DVR in maintaining power quality under fluctuating grid conditions.

In this paper, an ABMP control scheme is developed for BESS supported DVR. Various conventional control schemes have been mentioned in literature to mitigate voltage related PQ issues, but orthogonal polynomial based control schemes have not been discussed. The generalized BMP is discussed [22]. This control scheme is designed to perform well under both balanced/unbalanced linear and nonlinear loads. The proposed control scheme can mitigate voltage-related PQ issues and load unbalancing effects.

2 | System Configuration, Design, and Principle of Operation of DVR

The proposed system consists of a three-phase, three-wire AC supply, IGBT switches, feeder impedance, and three single-phase transformers. It is shown in Figure 1 that the three-phase DVR with BESS at the DC side is connected to critical loads. The DVR is designed to regulate the voltage across the load (v_{Labc}) with rated magnitude even when PQ issues occur at the supply end. The proposed three-phase DVR is coupled to a three-phase line through a series transformer, which will inject compensating voltage $(v_{cabc(inj)})$ into the line. An additional ripple filter is connected across the series transformer (T_r) comprising capacitance and resistance (C_r, R_r), which will filter the ripples in injected voltage. The terminal voltage (v_{sa}, v_{sb}, v_{sc}) is shown in Figure 1 at the point of common coupling (PCC). The equivalent supply voltage of phase "a" v_{Ma} is connected to the point of common coupling (PCC), v_{sa} through short-circuit impedance z_{sa}. The voltage injected by the DVR in phase A is depicted as $v_{ca(inj)}$, which is computed such that the load voltage v_{La} is of rated magnitude and undistorted. A three-phase 400 V, 10 kVA, and 50 Hz system is considered. The voltage rating of VSC depends on the amount of maximum voltage required by the critical loads. The maximum voltage injected per phase is chosen as 20% of the phase voltage in case the BESS DVR system. The maximum sag in the source voltages per phase can be calculated as $230.9 \times 0.8 = 184.72$V. The series injection voltage required to be injected per phase is given by Equation (1)

$$V_{ca(inj)} = \sqrt{V_{sa}^2 - V_{La}^2} = 138.54V \tag{1}$$

The current rating of the VSC of the DVR depends on the connected load. The current rating of VSC is expressed as

$$\sqrt{3}V_{Sa}I_{Sa} = 10000, I_{Sa} = 25A \tag{2}$$

kVA rating of VSC is calculated as

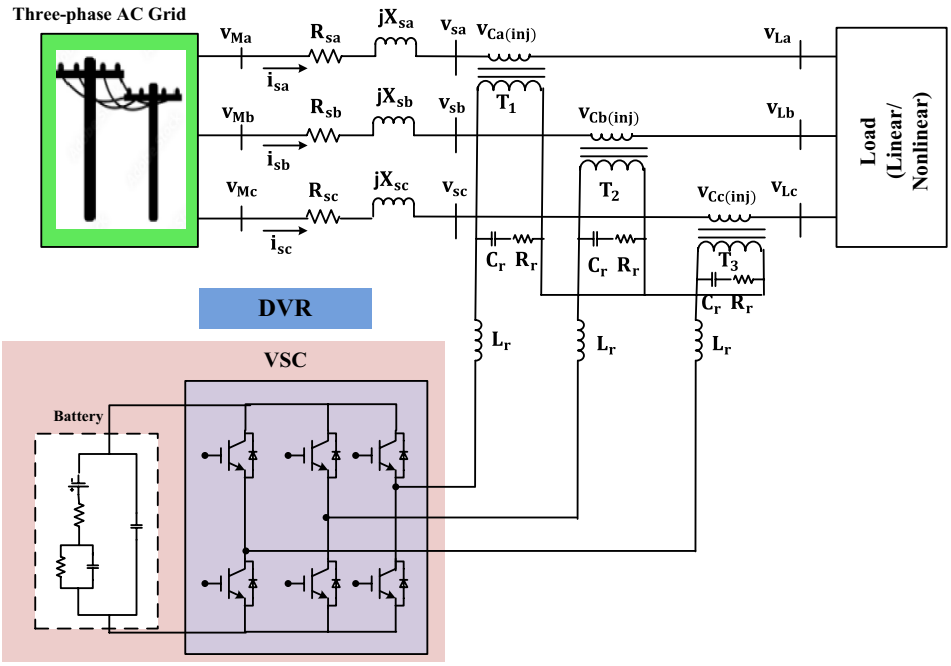


FIGURE 1 | Schematic diagram of a three-phase battery-supported DVR system.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

$$S = \frac{3 \times V_{Ca(inj)} I_{sa}}{1000} = 10.39\ \text{kVA} \tag{3}$$

The kVA rating of the series injection transformer will be the same as the kVA rating of VSC, that is, 10.39 kVA. For BESS DVR, the DC voltage considered is 300 V. The interfacing inductor L_r is calculated [3] as

$$L_r = \frac{n \times \left(\sqrt{3/2} m V_{DC}\right)}{6 a f_{sw} \Delta I_s} = 0.946\text{mH} \approx 1\text{mH} \tag{4}$$

where n is the transformation ratio of the series injection transformer, m is the modulation index, f_{sw} is the switching frequency, and ΔI_s is the ripple current of DVR. A ripple filter is designed to filter the switching frequency ripples that occur due to a series injection transformer. The expression for ripple filter which consists of series resistance (R_r) and series capacitor (C_r) is given as

$$\frac{f_s}{2} = \frac{1}{2\pi R_r C_r} \tag{5}$$

For $R_r = 1\Omega, C_r = 31.8\mu F$

When the voltage injected by a DVR is in-phase with the current, it means that the DVR is compensating not only for the voltage magnitude but also for the real power. In this scenario, the DVR does not only correct voltage distortions (sags, swells, or harmonics) but also contributes real power to the system. The series active power filter also known as DVR injects compensating voltage in series with the source voltages. The DVR will inject the required voltage to maintain the load voltages across critical loads when voltage-related PQ issues such as voltage sag, voltage swell, voltage unbalance, voltage flicker, and harmonics occur in the grid. The gating pulses for VSC's switches are obtained through PWM control. The proposed control scheme is shown in Section 3.

3 | Mathematical Analysis of BMP-Based Neural Network

The term "Bateman polynomial" was first coined by Pasternack in 1939. The BMPs are a family of orthogonal polynomials. BMP [22] can be defined as

$$F_n\left(\frac{d}{dx}\right)\text{sech}(x) = \text{sech}(x) P_n(\tanf(x)) \tag{6}$$

where P_n is a Legendre polynomial. Some dominant polynomial functions are represented as

$$\left.\begin{aligned} F_0(x) &= 1 \\ F_1(x) &= -x \\ F_3(x) &= \frac{1}{4} + \frac{3}{4}x^2 \\ F_4(x) &= \frac{-7}{12}x - \frac{5}{12}x^3 \end{aligned}\right\} \tag{7}$$

The BMPs satisfy the orthogonality and recurrence relation. This polynomial is also having the generating function defined as

$$\sum_{n=0}^{\infty} t^n F_n(z) = (1-t)^z F_1\left(\frac{1+z}{2}, \frac{1+z}{2}; 1; t^2\right) \tag{8}$$

The polynomial function F_n in Equation (7) is analogous to source voltage of each phase v_{sa}, v_{sb}, and v_{sc}. Thus, the polynomial function $(F_n(x))$ can be replaced by $(V_n(x))$. In the proposed research article, BMPs are designed for source voltage compensation. The source voltage $v_{sn}(x)$ is expanded into its intermediate components $V_0(x), V_1(x), V_2(x)$. The input signal is the source voltage v_{sn} where n = a, b, c and the output of the BMP is the extracted fundamental active source voltage component. The intermediate components are the first three terms of the BMPs. The expanded source voltage components for phase "a" source voltage is given by Equation (9).

$$\left.\begin{aligned} V_0(x) &= 1 \\ V_1(x) &= -v_{sa} \\ V_2(x) &= \frac{1}{4} + \frac{3}{4}{v_{sa}}^2 \end{aligned}\right\} \tag{9}$$

Similarly, the expansion terms are expressed for the other two phases of the three-phase grid-connected system. A single-layer neural network consists of three layers: an input layer, a hidden layer, and an output layer as shown in Figure 2. Input layer consists of every input. In the hidden layer, each input signal is expanded into its intermediate component at each perceptron unit. The output of each perceptron hidden layer unit is multiplied with training weights, and it is summed up together. The output of the hidden layer is passed through the activation function "tanh." The last layer is the output layer, where the actual and calculated output is subtracted to produce the error. This error is used to update the training weight using least mean squares (LMS) algorithm.

4 | Proposed ABMP Control Scheme

The proposed control scheme based on ABMP is shown in Figure 3. The BMNN is used to approximate fundamental source voltage. The active and reactive weights for the source of each phase source voltage are updated through the LMS scheme. The total weight of the proposed control scheme is utilized to generate the switching signals for VSC's switches through the PWM voltage controller. The control action is discussed in the subsection below.

4.1 | Estimation of ABMP Weights (W_{pa}, W_{pb}, W_{pc})

The phase "a" source voltage is processed through BMP, which approximates the source "a" phase voltage (v_{sa}) into three approximate components (V_0, V_1, V_2). These approximated components are multiplied by ANN weights of phase "a," $W_{pa1}, W_{pa2}, W_{pa3}$, respectively. The product output weights of BMP expansion (V_0, V_1, V_2) are summed and passed through the activation function "tanh." The active fundamental weight component (W_{Pa}) of source voltage (v_{sabc}) is obtained from activation function. The error between the actual or final weight and reference output W'_{Pa} is used to update the weights

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

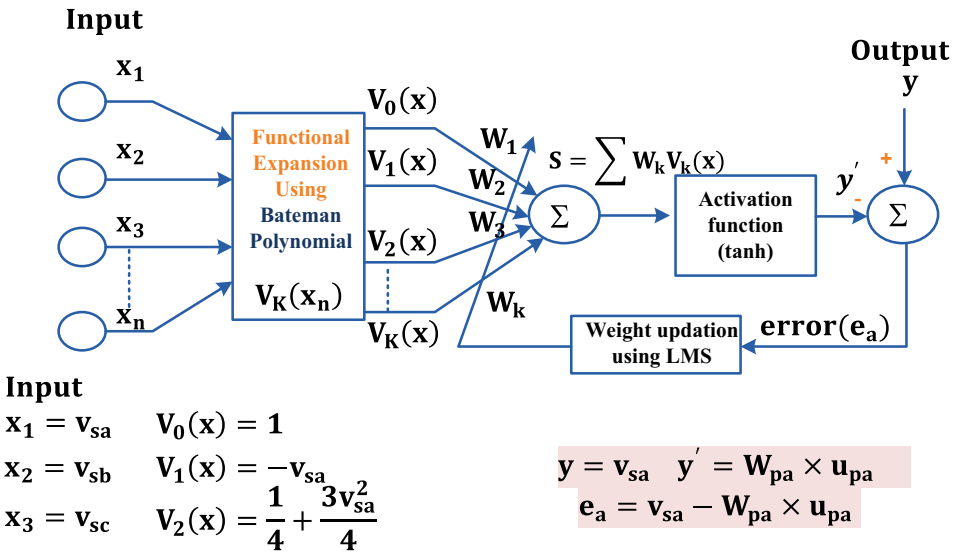


FIGURE 2 | Structure of single-layer neural network with ABMP functional expansion layer.

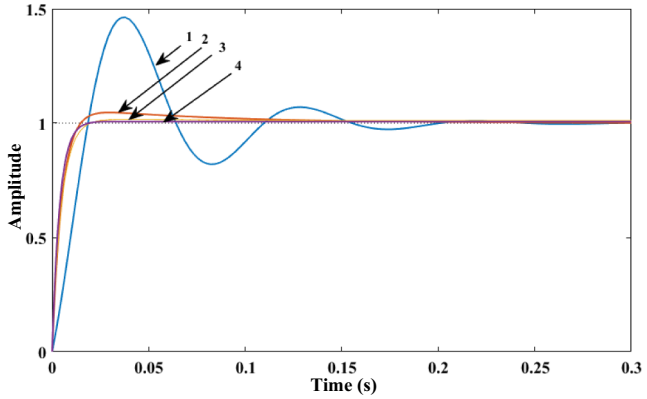


FIGURE 3 | Step response of PI controlled converter model.

$W_{pa1}, W_{pb1}, W_{pc1}$. Similarly, the final output is extracted for phase "b" and "c" source voltage.

$$W_{Pb} = \left[V_0 * W_{pb1} \;\; V_1 * W_{pb2} \;\; V_2 * W_{pb3}\right]^T \tag{10}$$

$$W_{Pc} = \left[\; V_0 * W_{pc1} \quad V_1 * W_{pc2} \quad V_2 * W_{pc3} \;\right]^T \tag{11}$$

where $W_{pa1}, W_{pa2}, W_{pa3}$ are the training weights of phase "a" source voltage which are updated using gradient descent method.

$$W_{pa1}(n+1) = W_{pa1}(n) + \frac{\mu}{2} \nabla W_{Pa} e_a(n) \tag{12}$$

where $\nabla W_{Pa} e_a(n) = \frac{\partial E(e_a{}^2)}{\partial W_{Pa}}$ denotes the gradient of the error.

$$MSE = E\left[e_a{}^2\right] = \left(v_{sa} - W_{Pa} u_{pa}\right)^2 \tag{13}$$

Differentiating Equation (13) gives

$$\frac{\partial E(e_a{}^2)}{\partial W_{Pa}} = -2u_{pa} \underbrace{\left(v_{sa} - W_{Pa} u_{pa}\right)}_{e_{pa}} = -2e_{pa} u_{pa} \tag{14}$$

$$W_{pa1}(n+1) = W_{pa1}(n) - \frac{\mu}{2}\left(-2e_{pa} u_{pa}\right) \tag{15}$$

$$W_{pa1}(n+1) = W_{pa1}(n) + \mu e_{pa} u_{pa} \tag{16}$$

Similarly, W_{pa2}, W_{pa3}

$$W_{pa2}(n+1) = W_{pa2}(n) + \mu e_{pa} u_{pa} \tag{17}$$

$$W_{pa3}(n+1) = W_{pa3}(n) + \mu e_{pa} u_{pa} \tag{18}$$

Equations (16), (17), and (18) show the training weights of phase "a" source voltage, where "μ" is the step size, "e_{pa}" is the error between the actual and reference source voltage of phase "a," and "u_{pa}" is the active unit template. The average magnitude of the weighted value of fundamental active components is expressed as

$$W_P = \frac{\left(W_{pa} + W_{pb} + W_{pc}\right)}{3} \tag{19}$$

Similarly, the fundamental reactive weight component of phase "a" is given by

$$W_{qa} = \left[\; V_{sqa0} \quad V_{sqa1} \quad V_{sqa2} \;\right]^T \tag{20}$$

$$W_{qa} = \left[\; V_0 * W_{qa1} \quad V_1 * W_{qa2} \quad V_2 * W_{qa3} \;\right]^T \tag{21}$$

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

$$W_{qb} = \begin{bmatrix} V_0 * W_{qb1} & V_1 * W_{qb2} & V_2 * W_{qb3} \end{bmatrix}^T \quad (22)$$

$$W_{qc} = \begin{bmatrix} V_0 * W_{qc1} & V_1 * W_{qc2} & V_2 * W_{qc3} \end{bmatrix}^T \quad (23)$$

The fundamental reactive weight components of phase "a" of source voltage are given as

$$W_{qa1}(n+1) = W_{qa1}(n) + \mu e_{qa} u_{qa} \quad (24)$$

Similarly, W_{qa2}, W_{qa3}

$$W_{qa2}(n+1) = W_{qa2}(n) + \mu e_{qa} u_{qa} \quad (25)$$

$$W_{qa3}(n+1) = W_{qa3}(n) + \mu e_{qa} u_{qa} \quad (26)$$

The average magnitude of the weighted value of fundamental reactive components is expressed as

$$W_q = \frac{\left(W_{qa} + W_{qb} + W_{qc}\right)}{3} \quad (27)$$

A PI controller is utilized to provide the reactive power component (W_{Lq}), which is injected by the DVR under voltage sag condition in source voltage. The output of the PI controller at *n*th instant of time is expressed as

$$W_{Lq}(n+1) = W_{Lq}(n) + K_{pq}\left(e_{AC}(n+1) - e_{AC}(n)\right) + K_{iq} e_{AC}(n+1) \quad (28)$$

where K_{pq} and K_{iq} are the gains of PI controller indicated in Figure 3; $e_{AC}(n)$ and $e_{AC}(n+1)$ are error between reference load voltage V^*_{Lq} and sensed load voltage V_{Lq} at *n*th and $(n+1)$th instant of time respectively. The error e_{AC} is expressed as

$$e_{AC} = V^*_{Lq} - V_{Lq} \quad (29)$$

The amplitude of load voltage V_{Lq} at the PCC is calculated from the ac voltages (V_{La}, V_{Lb}, and V_{Lc}) as $V_{Lq} = \sqrt{\frac{2}{3}\left(v^2_{La} + v^2_{Lb} + v^2_{Lc}\right)}$. The total reactive weight component of load voltage is expressed as

$$W_{q(total)} = W_q + W_{LP} \quad (30)$$

The function of PI controller is to regulate the load terminal voltage. The steady-state error in the load voltage is estimated, which is then regulated to zero using PI controller. The transfer function of VSC is given by

$$G_{VSC}(s) = \frac{V_{out}(s)}{V_{ref}(s)} = K_{VSC} \quad (31)$$

For an LC filter with inductance L_r and capacitance C_r, the transfer function is

$$G_{filter}(s) = \frac{1}{L_r C_r s^2 + R_r s + 1} \quad (32)$$

The PI controller is represented by

$$G_{PI}(s) = K_P(s) + \frac{K_i(s)}{s} \quad (33)$$

The open-loop transfer function is given by

$$G_{OL}(s) = G_{PI}(s) \cdot G_{VSC}(s) \cdot G_{filter}(s) \quad (34)$$

The closed loop transfer function is given by

$$G_{CL}(s) = \frac{G_{OL}(s)}{1 + G_{OL}(s)} \quad (35)$$

$$L_r = 1\text{mH}, C_r = R_r = 1\Omega, C_r = 31.8\mu F$$

For practical case, $K_{VSC} = \frac{m \cdot V_{dc}}{2 \cdot V_{ref}}$; $m = 1, V_{dc} = 300\text{V}, V_{ref} = \frac{V_{LL}}{\sqrt{3}} \cdot \sqrt{2} = 326.54\text{V}$

$$K_{VSC} = 0.45$$

$$G_{filter}(s) = \frac{1}{31.8 \times 10^{-9} s^2 + s + 1} \quad (36)$$

$$G_{OL}(s) = \left(K_P(s) + \frac{K_i(s)}{s}\right)\left(\frac{1}{31.8 \times 10^{-9} s^2 + s + 1}\right) \times 0.45 \quad (37)$$

$$G_{CL}(s) = \frac{0.45\left(K_P s + K_i\right)}{31.8 \times 10^{-9} s^3 + s^2 + 0.45K_P + \left(0.45K_i + 1\right)} \quad (38)$$

The step response of the transfer function of Equation (38) is shown in Figure 2 (Table 1).

Figure 3 shows the step response is shown, which shows that higher peak overshoot at lower K_P and higher K_i, higher will be peak overshoot, rise time, and settling time. It is better to select $K_P = 0.4$ and $K_i = 1.5$, which provides lower peak overshoot, rise time, and settling time.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 | Comparison table of K_P and K_i.

S. no.	K_P	K_i	Peak overshoot (%)	Rise time, s	Settling time, s
1.	0.08	10	46.30	0.0144	0.185
2.	0.4	5	4.63	0.0092	0.098
3.	0.5	1	1.6133	0.0101	0.0162
4.	**0.4**	**1.5**	**0.643**	**0.0083**	**0.0140**

4.2 | Generation of Active Unit Template (u_{pa}, u_{pb}, u_{pc}) and Reactive Unit Template (u_{qa}, u_{qb}, u_{qc})

The unit templates u_a, u_b, u_c have divided into active (u_{pa}, u_{pb}, u_{pc}) and reactive unit templates (u_{qa}, u_{qb}, u_{qc}) with the source currents (i_{sa}, i_{sb}, i_{sc}). The magnitude of source current at PCC is expressed as

$$I_t = \sqrt{\frac{2}{3}\left(i_{sa}^2 + i_{sb}^2 + i_{sc}^2\right)} \tag{39}$$

The active unit templates are calculated as

$$u_{pa} = i_{sa}/I_t, u_{pb} = i_{sb}/I_t, u_{pc} = i_{sc}/I_t \tag{40}$$

The reactive unit templates are calculated as

$$u_{qa} = \frac{-u_{pb}}{\sqrt{3}} + \frac{u_{pc}}{\sqrt{3}} \tag{41}$$

$$u_{qb} = \frac{-\sqrt{3}u_{pb}}{2} + \frac{(u_{pb} - u_{pc})}{2\sqrt{3}} \tag{42}$$

$$u_{qc} = \frac{-\sqrt{3}u_{pa}}{2} + \frac{(u_{pb} - u_{pc})}{2\sqrt{3}} \tag{43}$$

4.3 | Estimation of Reference Load Voltages

The estimated active component of reference load voltage for each phase are calculated using Equations (44)–(46):

$$v_{Lap}^* = W_p \times u_{pa} \tag{44}$$

$$v_{Lbp}^* = W_p \times u_{pb} \tag{45}$$

$$v_{Lcp}^* = W_p \times u_{pc} \tag{46}$$

Similarly, the reactive component of reference load voltage for each phase are calculated using Equation (47)

$$\left.\begin{aligned} v_{Laq}^* &= W_{q(total)} \times u_{qa} \\ v_{Lbq}^* &= W_{q(total)} \times u_{pb} \\ v_{Lcq}^* &= W_{q(total)} \times u_{pc} \end{aligned}\right\} \tag{47}$$

The total reference load voltages $V_{La}^*, V_{Lb}^*, V_{Lc}^*$ are calculated as the sum of active and reactive reference load voltage for each phase

$$\left.\begin{aligned} V_{La}^* &= v_{Lap}^* + v_{Laq}^* \\ V_{Lb}^* &= v_{Lbp}^* + v_{Lbq}^* \\ V_{Lc}^* &= v_{Lcp}^* + v_{Lcq}^* \end{aligned}\right\} \tag{48}$$

The error between sensed load voltages (V_{La}, V_{Lb}, V_{Lc}) and reference load voltages $(V_{La}^*, V_{Lb}^*, V_{Lc}^*)$ is used over a controller to generate gating pulses to the VSC of the DVR. The PWM controller is designed to generate the switching signals $(S_1 - S_6)$ for DVR. The PWM controller has a fixed switching frequency of 10 kHz. Figure 4 shows the details of the controller developed.

5 | Simulation Results

The performance of the proposed ABMP control scheme of a three-phase grid-connected DVR is investigated under various disturbances at grid supply or supply voltage. The various disturbances that are considered are voltage sag, swell, imbalance, and harmonics. The transient response of DVR is shown in this section with the following cases:

Case 1. Voltage sag of 20% in the source voltages

Case 2. Voltage swell of 20% in the source voltages

Case 3. Voltage harmonics (5th and 7th order harmonics)

Case 4. Unbalanced voltage with 20% sag in source voltage (phase "a")

Case 5. Load unbalancing

Case 6. Phase jump

Case 7. Frequency variation

5.1 | Voltage Sag in the Source Voltages

One of the most prominent voltage-related PQ issues is voltage sag. In this case scenario, a balanced voltage sag of 20% is applied in the source voltages (v_{sabc}). During t = 0.3 s to 0.4 s, a voltage sag is applied for five cycles in source voltages. The load voltages are observed to be balanced and sinusoidal with

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

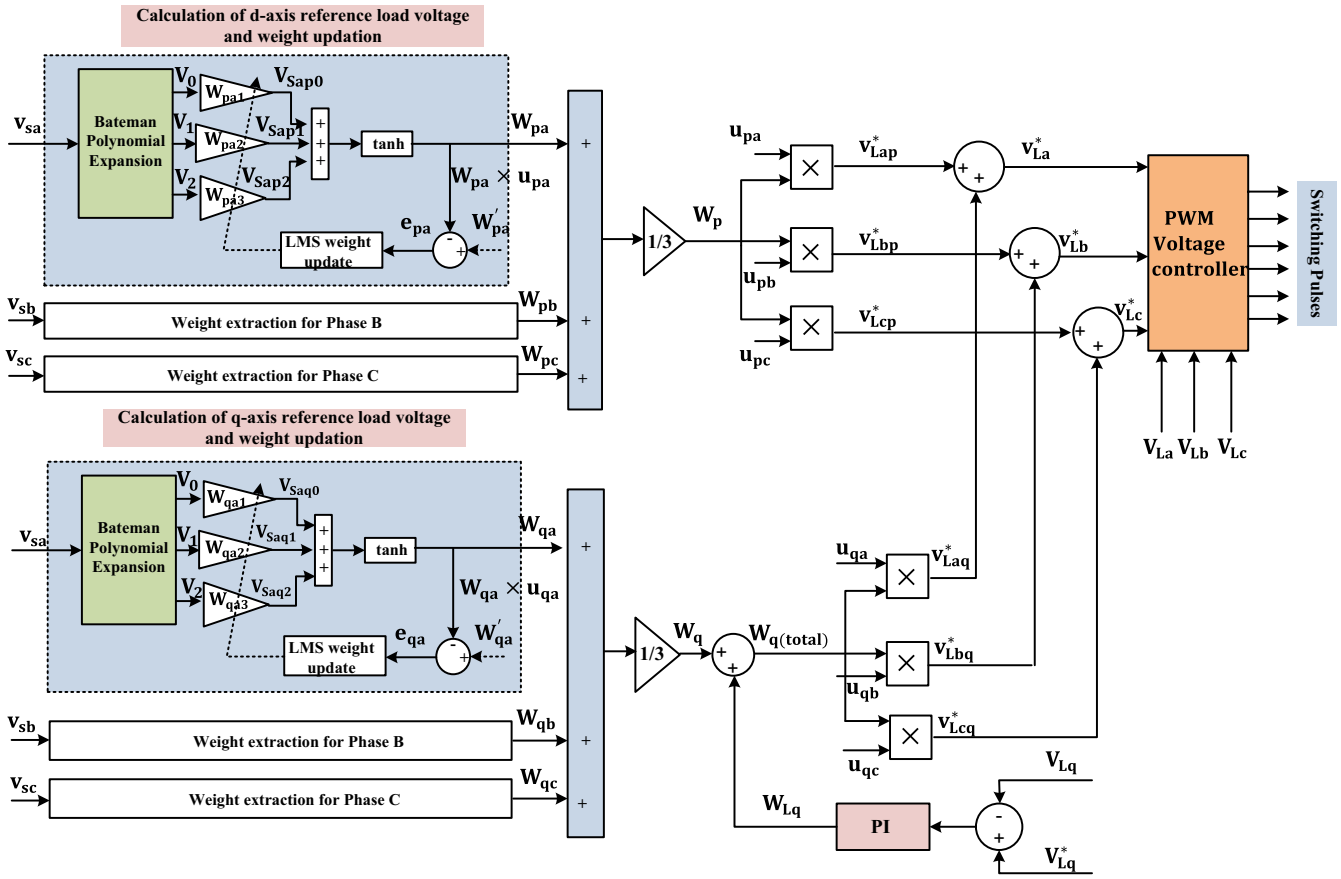


FIGURE 4 | Block diagram of the proposed adaptive Bateman polynomial (ABMP) based control scheme for BESS-DVR.

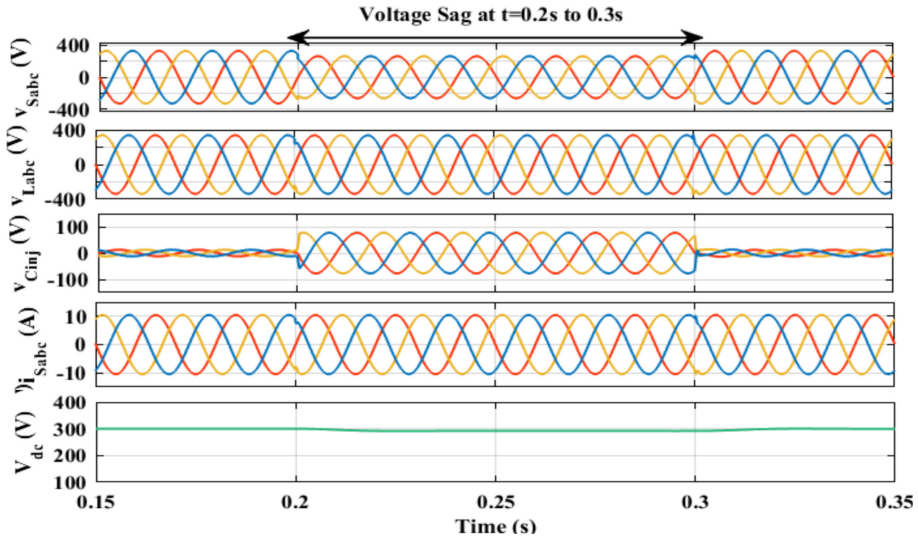


FIGURE 5 | Simulation result of the proposed DVR system under voltage sag condition.

the desired magnitude at voltage sag conditions. The PCC voltage (v_{sabc}), load voltages (v_{Labc}), series injecting compensating voltage (v_{Cinj}), source currents (i_{sabc}), and battery DC voltage (V_{dc}) is shown in Figure 5. The required load voltages are well regulated even under voltage sag condition due to DVR support. Requisite voltages are injected by the DVR. The source currents are sinusoidal under voltage sag condition due to the proposed control technique. The battery DC voltage shows a very small variation of 2 V under transient conditions.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

5.2 | Voltage Swell at the Source Side

Another common disturbance observed in source voltages is voltage swell. At t = 0.45 to 0.55 s, voltage swell is simulated in source voltages. The considered swell is balanced with a 20% increase in source voltages for five cycles. Before t = 0.45 s, the system is operating at steady-state. The voltage swell is cleared in five cycles. The battery DC voltage is slightly decreased due to the transient effect but gets stable within one cycle. The response of the proposed system under voltage swell conditions is depicted in Figure 6, which also shows well regulated load voltage as desired.

5.3 | Voltage Harmonics in the Source Supply

To investigate the performance of DVR during voltage harmonics, fifth- and seventh-order harmonics are incorporated into source voltages. The transient response of the system is noticed from t = 0.5 s to t = 0.6 s (five cycles). It is depicted in Figure 7 that load voltages are regulated at its rated value with constant magnitude. The effect of harmonics is cleared after five cycles. The required voltage for load voltage compensation is fed by a series injection transformer. The battery DC voltage is regulated at 300 V with slight variation under transient conditions. The harmonic spectra of phase "a" of load voltages and source voltages are shown in Figure 8. The source voltages are significantly polluted with 25% harmonics, while load voltages have only 0.96%, which is acceptable as per IEEE-519 standards.

5.4 | Voltage Unbalancing in the Source Supply

In this case study, voltage unbalancing is created in source voltages. All the three-phase voltages are observed to be unbalanced

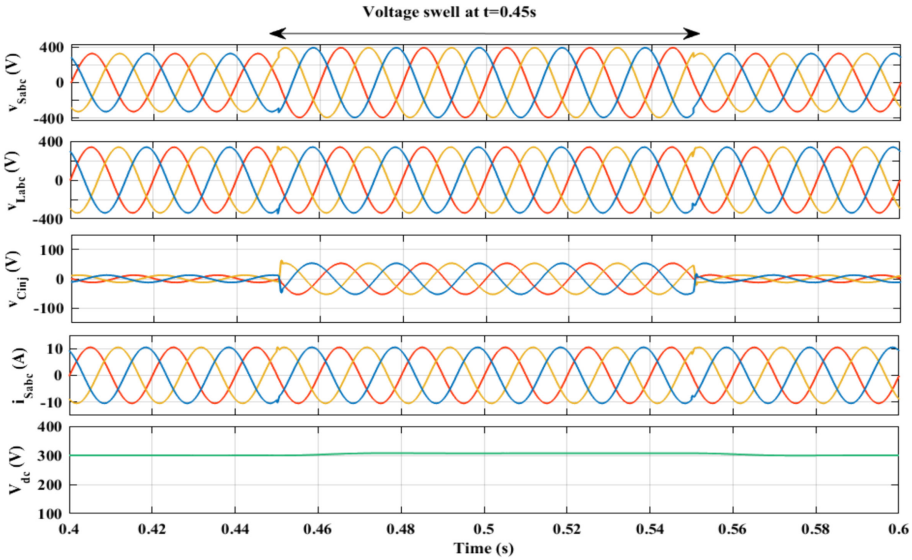


FIGURE 6 | Simulation results of the proposed DVR system under voltage swell condition.

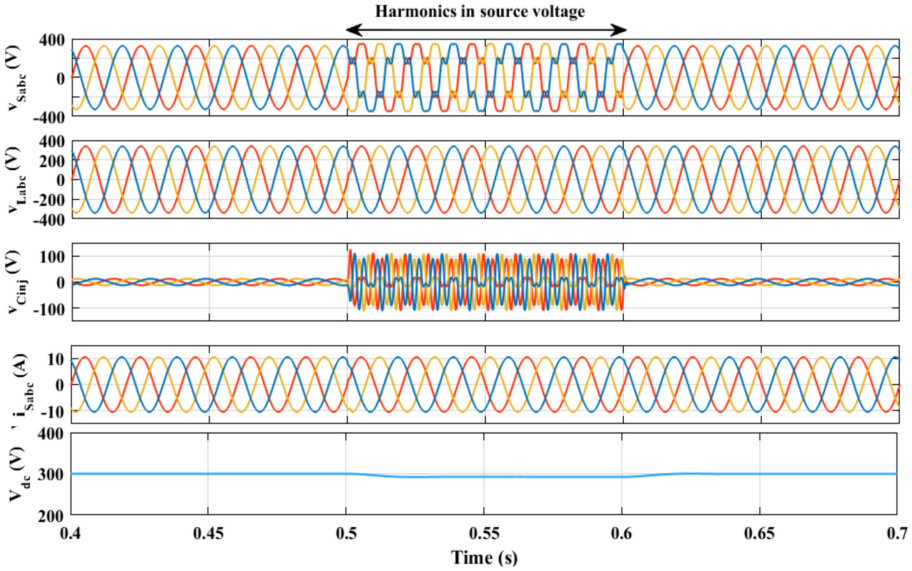


FIGURE 7 | Simulation result of the proposed DVR system under voltage harmonics condition.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

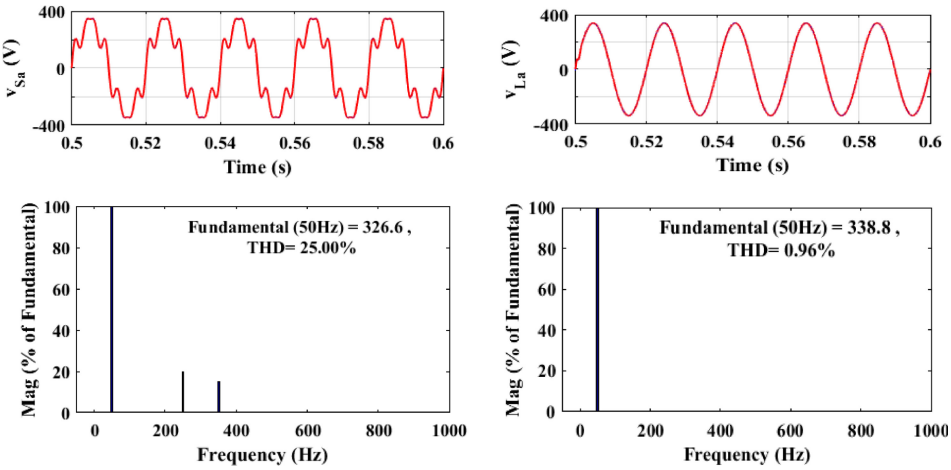


FIGURE 8 | THD spectrum under source voltages harmonics (a) source voltages (b) load voltages.

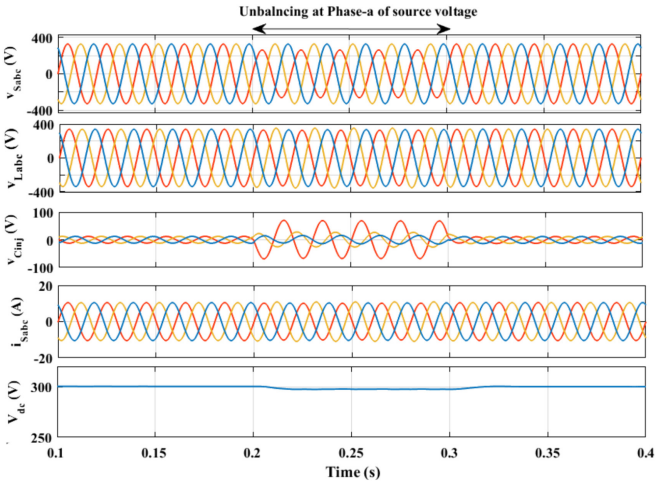


FIGURE 9 | Simulation result of the proposed DVR system under unbalancing of source voltages.

from t = 0.2 s to t = 0.3 s. Due to this unbalancing in voltages, a slight variation in battery DC voltage is observed in Figure 9. Despite voltage imbalance, the load voltages are balanced, sinusoidal and of constant peak magnitude. The voltage unbalancing is simulated for five cycles (t = 0.2 s to t = 0.3 s). The source current is sinusoidal under linear load conditions.

5.5 | Performance Under Load Unbalancing

An unbalanced nonlinear load is connected to the DVR. Phase "c" of the nonlinear load is disconnected at t = 0.6 s to 0.7 s. In Figure 10, it is seen that there is no effect of load unbalancing at critical load voltages. The load voltages remain constant even with a harmonic at PCC voltage and phase "c" load is disconnected at t = 0.6 s.

5.6 | Performance Under Frequency Change of +5 Hz and Phase Shift of 15°

The performance of the proposed DVR is tested under frequency change and phase shift of +5 Hz, and 15°, respectively. The source voltages are simulated to experience a phase jump of 15° from t = 0.3 s to t = 0.4 s as shown in Figure 11a. The source voltages are simulated to have frequency change of +5 Hz (i.e., transition from 50 to 55 Hz) from t = 0.2 s to t = 0.3 s as shown in Figure 11b. It can be observed that even under frequency change, the load voltages are balanced and sinusoidal with the required magnitude in both phase jump and frequency shift. The series injection transformers inject the required voltage into the system.

The control signals of PWM voltage controller are shown in Figure 12 under the grid adverse condition such as voltage sag and voltage swell. The fundamental active weight component of source voltage (W_{pa}, W_{pb}, W_{pc}) of each phase is decreasing under the condition of voltage sag and increases under voltage swell condition. The total averaged active fundamental voltage component of source voltage (W_p) is also decreased under voltage sag condition and increased under voltage swell condition. Similarly, the fundamental reactive weight components of phase a, b, c of source voltage are also shown in Figure 12 under voltage sag and swell condition. The reference load voltage (V^*_{Labc}) is constant throughout the grid disturbances.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

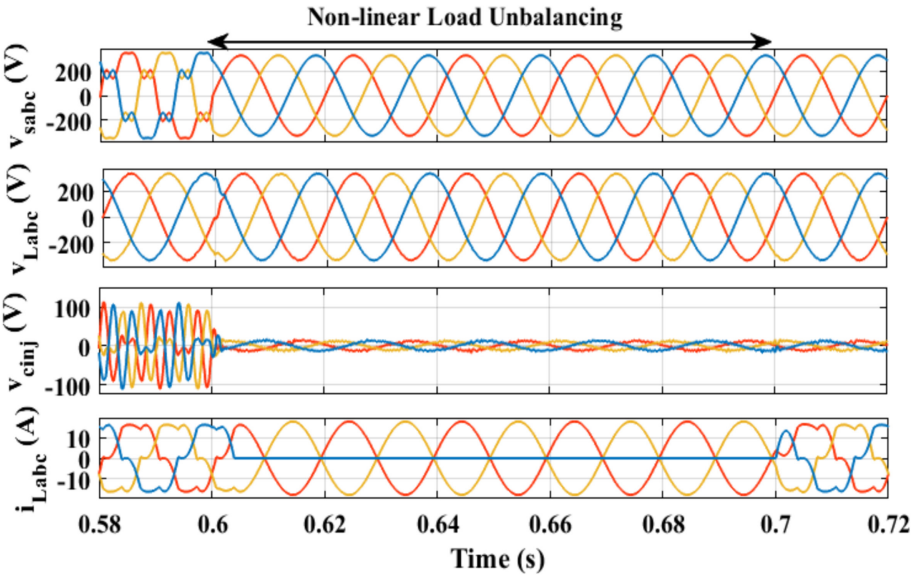


FIGURE 10 | Simulation result of the proposed DVR system under unbalanced load condition.

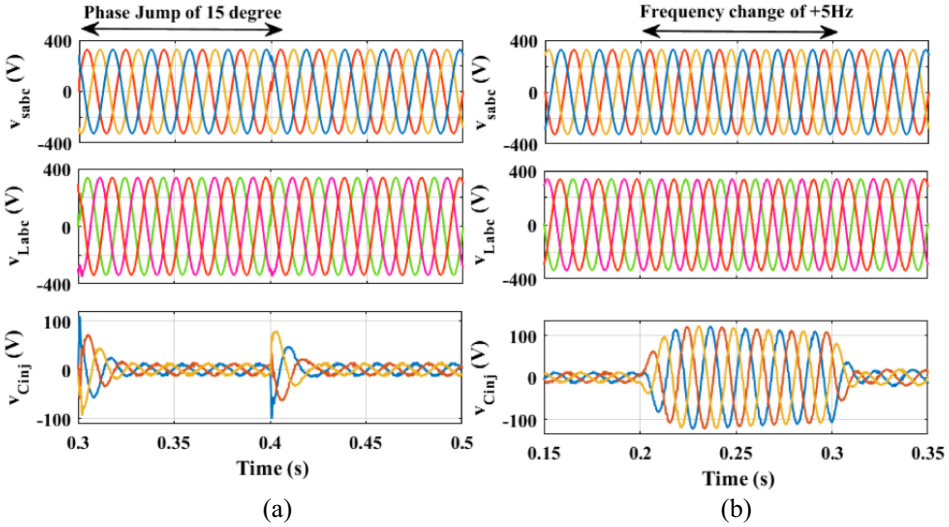


FIGURE 11 | Simulation result of the proposed DVR system under (a) phase jump of 15° (b) frequency change +5 Hz.

6 | Experimental Results

A real-time hardware-in-the-loop (HIL) system is created to evaluate the viability of the proposed approach for DVR control. Figure 13 depicts a schematic of the designed laboratory experimental setup. The real-time HIL system comprises OPAL-RT real-time digital simulator with FPGA Xilinx Kintex-7 410T. The OPAL-RT is a real-time simulation platform powered by Intel Xeon Four-Core 3.7 GHz processors and operates in the RT-LAB software environment. It features 32 digital inputs and 32 digital outputs. The proposed system is simulated in MATLAB Simulink and the results are verified using OPAL-RT.

A voltage sag of 20% is created in the source voltage of phase "a" as shown in Figure 14a. The source voltage under voltage sag condition is reduced by 20%, and the respective intermediate components $(V_1, V_2, V_3,)$ of ABMP are decrease with the source voltage of phase "a." As the intermediate signals are decreasing, the fundamental active voltage component is also decreased with the ABMP learning method. The ABMP will update its active and reactive weights under sag conditions. The compensating voltage is computed, and $(V_{Ca(inj)})$ is fed through the series injection transformer. These active and reactive fundamental components are added to generate the reference load current. It is validated from Figure 14a that the load voltages are always maintained at pre-sag condition. The DVR injects required voltages to regulate the load voltages under voltage sag conditions. These voltages are fed by a series injection transformer $(v_{ca(inj)})$. A small change in battery DC voltage is visible at the start of the voltage sag condition for less than one cycle. The series injection transformer injects very low voltage before the voltage sag condition and higher magnitude during sag conditions.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

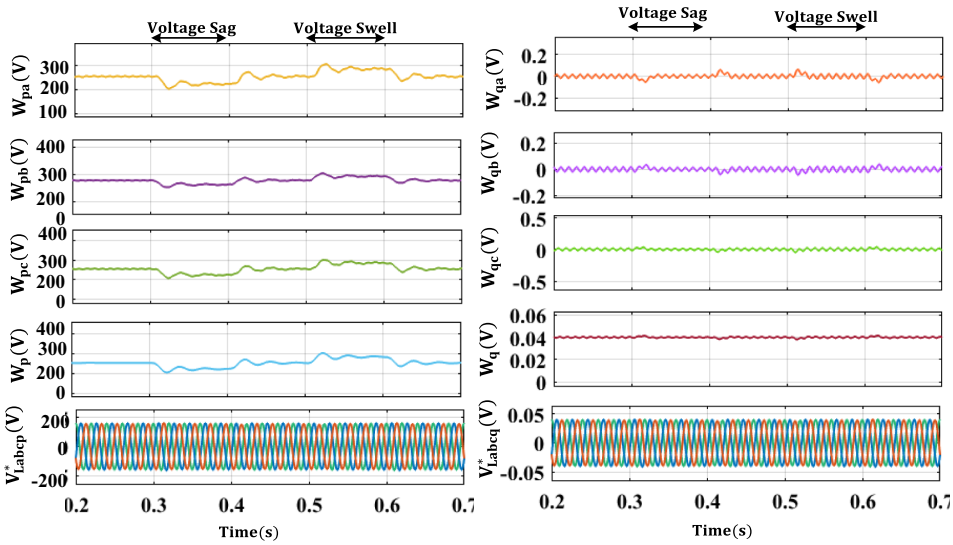


FIGURE 12 | Control signals of DVR under voltage sag and swell condition.

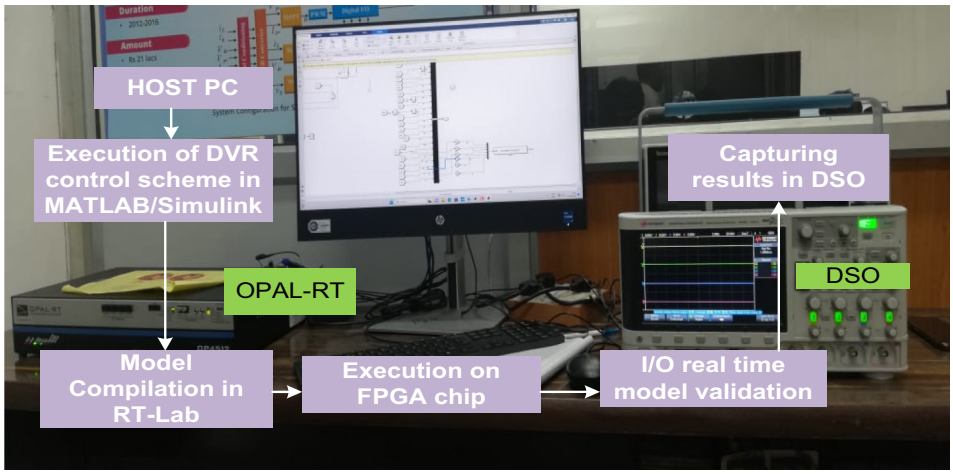


FIGURE 13 | Experimental setup of the proposed ABMP control based DVR implementation in OPAL-RT.

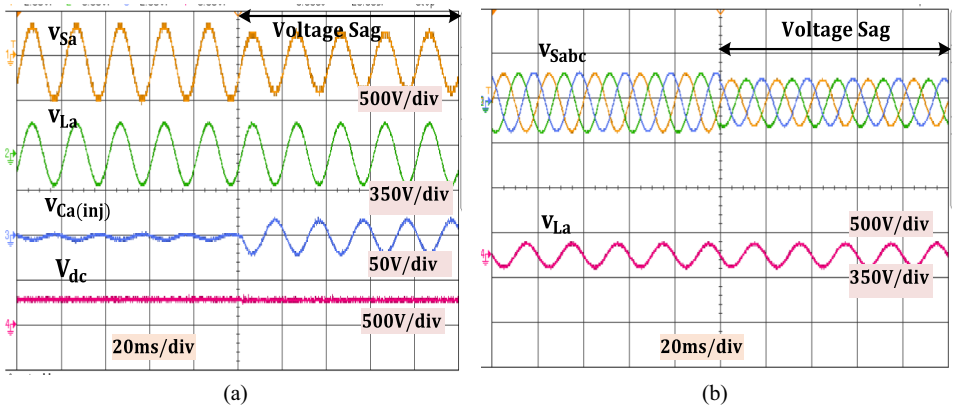


FIGURE 14 | Experimental result of the proposed DVR under voltage sag of 20%. (a) Phase "a" source voltage (v_{sa}), phase "a" load voltage (v_{La}), phase "a" series injected voltage $(v_{ca(inj)})$, and DC voltage (V_{dc}). (b) Three-phase source voltages (v_{sabc}) and phase "a" load voltage (v_{La}).

Another important voltage-related PQ issue tested here is the voltage swell. In Figure 15a, a voltage swell of 20% of source voltage has been created at some instant of time to investigate the performance of the proposed control scheme under voltage swell conditions. The source voltage under voltage swell condition is increased by 20%, and the respective intermediate

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

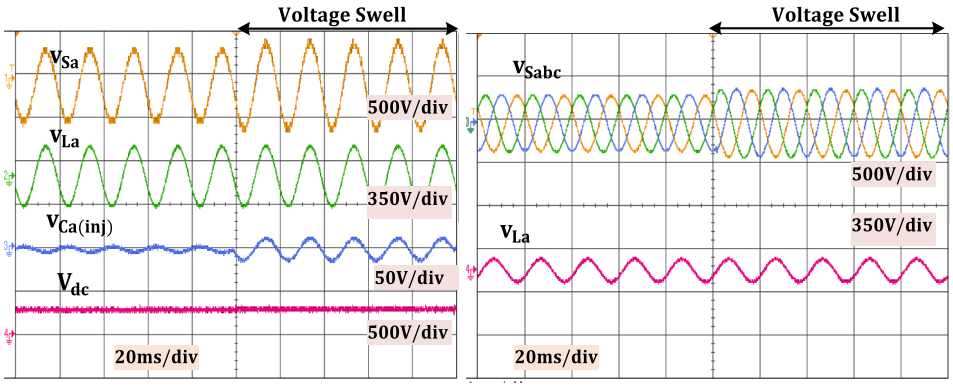


FIGURE 15 | Experimental result of the proposed DVR under voltage swell of 20% with signals. (a) Phase "a" source voltages (v_{sa}), phase "a" load voltages (v_{La}), phase "a" series injected voltage $(v_{ca(inj)})$, and battery DC voltage (V_{dc}). (b) Three-phase source voltages (v_{sabc}) and phase "a" load voltages (v_{La}).

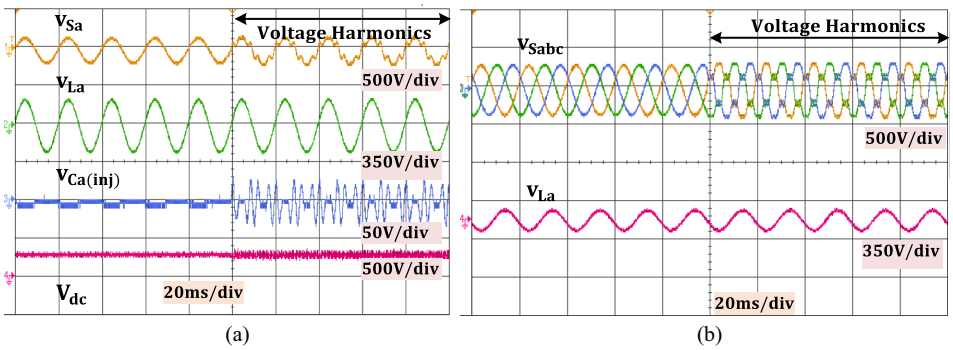


FIGURE 16 | Experimental results of the proposed DVR under voltage harmonics having signals. (a) Phase "a" source voltage (v_{sa}), phase "a" load voltage (v_{La}), phase "a" series injected voltage $(v_{ca(inj)})$, and battery DC voltage (V_{dc}). (b) Three-phase source voltages (v_{sabc}) and phase "a" load voltage (v_{La}).

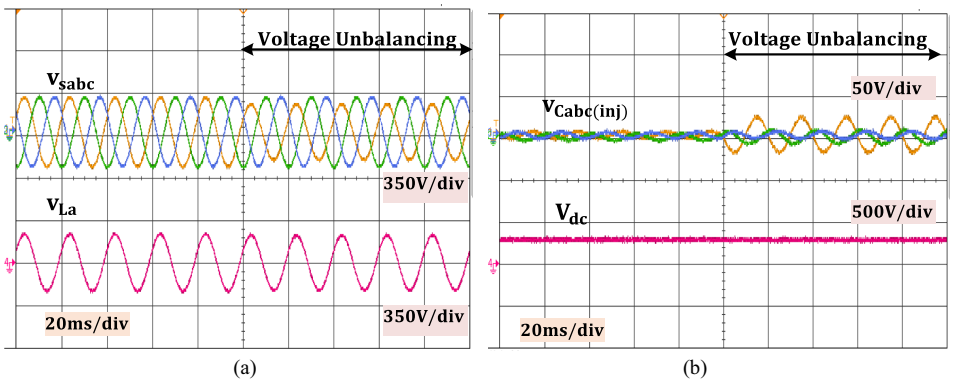


FIGURE 17 | Experimental result of the proposed DVR under voltage unbalancing with signals. (a) Phase "a" source voltage (v_{sa}), phase "a" load voltage (v_{La}), phase "a" series injected voltage $(v_{ca(inj)})$, and battery DC voltage (V_{dc}). (b) Three-phase source voltages (v_{sabc}) and phase "a" load voltage (v_{La}).

components $(V_1, V_2, V_3,)$ of ABMP are increasing with the source voltage of phase "a." As the intermediate signal increases, the fundamental active voltage component is also increased with the proposed method. The ABMP will update its active and reactive weights under swell conditions. The compensating voltage $(V_{Ca(inj)})$ is fed through the series injection transformer. The compensating voltage is added with the active weight component of source voltage to maintain the load voltage

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

constant. The load voltages of phase "a" (v_{La}) is regulated to its required magnitude. A negligible change in battery DC voltage is observed. Figure 15b shows the three-phase source voltages under voltage swell conditions.

Harmonic voltage is created in three-phase source voltages (v_{sabc}) having a THD of 25%. It is validated through the experimental results as shown in Figure 16a that load voltages are pure sinusoidal with reduced harmonics component. This also proves that the proposed control algorithm is working effectively in all adverse grid conditions. The series injection transformer injects the required voltage to maintain the load voltages of phase "a" constant. A small ripple is seen in the battery DC voltage across the battery. In Figure 16b, three-phase source voltages with harmonics and the load voltages across phase "a" are shown.

On the same lines, voltage unbalancing is created at some instant of time in three-phase source voltages (v_{sabc}). The performance of the proposed controller even under unbalanced conditions with sag, the controller is acting very fast and compensation is done less than one cycle. However, the DVR action regulates the load terminal voltages. There is no adverse effect on load voltages as shown in Figure 17a. In Figure 17b, series injection voltage $(v_{ca(inj)})$ and battery DC voltage (V_{dc}) are shown, which show the necessary compensating voltages injected by the DVR. The dc link voltage the BESS-supported DVR is well regulated.

The performance of a three-phase grid-connected PV-based DVR is performed on the same system parameters as conventional SRFT and the proposed ABMP control scheme. The voltage-related PQ issues considered are voltage sag, voltage swell, voltage unbalancing, and voltage harmonics, which are mitigated with the effect of designed control schemes.

7 | Result Analysis

Table 2 shows the comparative performance of control schemes under all the mentioned PQ issues. It is seen from this table that the proposed ABMP gives source current THD under voltage sag, swell, unbalancing, and voltage harmonics of 2.1%, 2.3%, 1.5%, and 0.96%, respectively.

This table compares the performance of two systems—conventional SRFT and the proposed ABMP—in terms of total harmonic distortion (THD) under different system conditions: voltage sag, voltage swell, voltage unbalancing, and voltage harmonics. The THD values are given for different phases (A, B, C) and for both current and voltage.

Voltage sag: The THD for current is consistently lower with the proposed ABMP algorithm compared with the conventional SRFT across all phases (A, B, C). The THD for voltage remains the same in both systems.

Voltage swell: Similar to voltage sag, the proposed ABMP system shows lower current THD compared with the conventional SRFT. However, the voltage THD remains unchanged across both systems.

Voltage unbalancing: The proposed ABMP algorithm has significantly lower current THD compared with the conventional SRFT, indicating better performance in managing unbalanced voltages. Voltage THD remains unchanged.

Voltage harmonics: For voltage harmonics, the THD for voltage remains at 25% in both systems. However, the proposed ABMP system achieves a significantly lower THD for current compared with the conventional SRFT, which indicates improved harmonic performance in terms of current.

TABLE 2 | Comparative performance analysis of SRFT and proposed ABMP under various operating conditions.

		Conventional SRFT		Proposed ABMP	
System condition	**Phase**	**Current**	**Voltage**	**Current**	**Voltage**
Voltage sag THD (%)	A	3.05%	4.63%	**2.1%**	4.63%
	B	3.4%	4.5%	**2.81%**	4.5%
	C	3.3%	5.41%	**2.6%**	5.41%
Voltage swell THD (%)	A	3.8%	3.30%	**2.3%**	3.30%
	B	3.3%	5.29%	**2.14%**	5.29%
	C	3.5%	5.30%	**2.6%**	5.30%
Voltage unbalancing THD (%)	A	3.15%	0.7%	**1.5%**	0.7%
	B	3.25%	0.6%	**1.6%**	0.6%
	C	3.12%	0.7%	1.32%	0.7%
Voltage harmonics THD (%)	A	3.5%	25%	**0.96%**	25%
	B	3.1%	25%	1.02%	25%
	C	3.3%	25%	1.2%	25%

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 | Conclusion

ABMP control-based DVR system is proposed in this paper, and it is considered as an effective solution for enhancing power quality in electrical distribution network. The paper shows effectively controlled DVR can mitigate voltage sags and swells, harmonics, and other power quality issues, thereby improving the reliability of the power supply. The successful implementation of ABMP control-based DVR systems relies on robust control strategy. The designed control technique is observed to be capable of detecting and responding to voltage disturbances quickly and accurately. It is verified that the proposed control scheme is well suited for PQ mitigation and results in well-regulated load voltages supply under all voltage-related PQ issues such as voltage sags, swells, unbalancing, harmonics, frequency change, and phase angle change.

The proposed ABMP control scheme has some drawbacks such as computational complexity and overfitting. ABMP method relies on polynomial equations, which can become computationally expensive for complex grid configurations, especially during real-time operation. There is a risk of overfitting the polynomial to the training data, which can lead to poor performance on unseen data. The simulation results are verified with OPAL-RT experimental results.

DVRs can play a crucial role in integrating renewable energy sources like wind and solar into the grid. They can mitigate voltage fluctuations caused by intermittent renewable power generation. Research is ongoing to develop effective control algorithms that can address the PQ problems. These algorithms could incorporate real-time system data for improved accuracy and dynamic response.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. S. Mittal, A. Singh, and P. Chittora, "Power Quality Enhancement in Single Phase Two Level/Five Level Converters Using Adaptive-RBFNN Algorithm," *Electrical Engineering* (2024): 1–14, https://doi.org/10.1007/s00202-024-02432-6.

2. A. Farooqi, M. M. Othman, A. F. Abidin, S. I. Sulaiman, and M. A. M. Radzi, "Mitigation of Power Quality Problems Using Series Active Filter in a Microgrid System," *International Journal of Power Electronics and Drive Systems* 10, no. 4 (2019): 2245.

3. B. Singh, A. Chandra, and K. Al-Haddad, *Power Quality: Problems and Mitigation Techniques* (The Atrium, Sourthen Gate, UK: John Wiley & Sons, 2014).

4. S. Mittal, A. Singh, and P. Chittora, "Design and Development of Leaky Least Mean Fourth Control Algorithm for Single-Phase Grid-Connected Multilevel Inverter," *International Journal of Circuit Theory and Applications* 52, no. 1 (2024): 328–345.

5. K. B. Rai, N. Kumar, and A. Singh, "Design and Analysis of Hermite Function-Based Artificial Neural Network Controller for Performance Enhancement of Photovoltaic-Integrated Grid System," *International Journal of Circuit Theory and Applications* 51, no. 3 (2023): 1440–1459.

6. K. B. Rai, N. Kumar, and A. Singh, "Design and Analysis of the Shunt Active Power Filter With the ε-NSRLMMN Adaptive Algorithm for Power Quality Improvement in the Distribution System," *IETE Journal of Research* 70 (2022): 1–15.

7. K. B. Rai, N. Kumar, and A. Singh, "Performance Analysis of UPQC Using Mittag Leffler Polynomial Based Neural Network Control," in *2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)* (USA: IEEE, 2022).

8. K. B. Rai, A. Singh, and N. Kumar, "Bernoulli Polynomial-Based Control Technique for PV-Integrated Grid System Under Distorted Supply," *International Journal of Circuit Theory and Applications* 51 (2023): 3204–3225.

9. T. L. Ilamkar and V. Joshi, "Voltage Sag Compensation Using Synchronously Reference Frame Theory Based Dynamic Voltage Restorer," in *2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT)* (USA: IEEE, 2018).

10. S. G. Reddy, S. Ganapathy, and M. Manikandan, "Three Phase Four Switch Inverter Based DVR for Power Quality Improvement With Optimized CSA Approach," *IEEE Access* 10 (2022): 72263–72278.

11. V. Babu, K. S. Ahmed, Y. M. Shuaib, and M. Mani, "A Novel Intrinsic Space Vector Transformation Based Solar Fed Dynamic Voltage Restorer for Power Quality Improvement in Distribution System," *Journal of Ambient Intelligence and Humanized Computing* 12 (2021): 1–16.

12. P. Kanjiya, B. Singh, A. Chandra, and K. Al-Haddad, ""SRF Theory Revisited" to Control Self-Supported Dynamic Voltage Restorer (DVR) for Unbalanced and Nonlinear Loads," *IEEE Transactions on Industry Applications* 49, no. 5 (2013): 2330–2340.

13. P. Jayaprakash, B. Singh, D. P. Kothari, A. Chandra, and K. Al-Haddad, "Control of Reduced Rating Dynamic Voltage Restorer With Battery Energy Storage System," *IEEE Transactions on Industry Applications* 50, no. 2 (2014): 1295–1303.

14. T. A. Naidu, S. Raj Arya, and R. Maurya, "Multiobjective Dynamic Voltage Restorer With Modified EPLL Control and Optimized PI-Controller Gains," *IEEE Transactions on Power Electronics* 34, no. 3 (2019): 2181–2192.

15. M. Pradhan and M. K. Mishra, "Dual P-Q Theory Based Energy-Optimized Dynamic Voltage Restorer for Power Quality Improvement in a Distribution System," *IEEE Transactions on Industrial Electronics* 66, no. 4 (2019): 2946–2955.

16. Z. Zheng, X. Xiao, X. Y. Chen, C. Huang, and J. Xu, "Performance Evaluation of a MW-Class SMES-Based DVR System for Enhancing Transient Voltage Quality by Using d–q Transform Control," *IEEE Transactions on Applied Superconductivity* 28, no. 4 (2018): 1–5.

17. Z. Zheng, X. Xiao, C. Huang, and C. Li, "Enhancing Transient Voltage Quality in a Distribution Power System With SMES-Based DVR and SFCL," *IEEE Transactions on Applied Superconductivity* 29, no. 2 (2019): 1–5.

18. L. Tong, X. Zou, S. Feng, et al., "An SRF-PLL-Based Sensorless Vector Control Using the Predictive Deadbeat Algorithm for the Direct-Driven Permanent Magnet Synchronous Generator," *IEEE Transactions on Power Electronics* 29, no. 6 (2014): 2837–2849.

19. S. Golestan and J. M. Guerrero, "Conventional Synchronous Reference Frame Phase-Locked Loop Is an Adaptive Complex Filter," *IEEE Transactions on Industrial Electronics* 62, no. 3 (2015): 1679–1682.

20. J. Matas, M. Castilla, J. Miret, L. Garcia de Vicuna, and R. Guzman, "An Adaptive Prefiltering Method to Improve the Speed/Accuracy Trade-Off of Voltage Sequence Detection Methods Under Adverse Grid Conditions," *IEEE Transactions on Industrial Electronics* 61, no. 5 (2014): 2139–2151.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

21. S. Mittal, A. Singh, and P. Chittora, "Design and Development of Reduced Switch Five Level Inverter for Interfacing Renewable Energy Sources and EV Charger," *Electrical Engineering* (2024): 1–15, https://doi.org/10.1007/s00202-024-02373-0.

22. M. Qasim, P. Kanjiya, and V. Khadkikar, "Artificial-Neural-Network-Based Phase-Locking Scheme for Active Power Filters," *IEEE Transactions on Industrial Electronics* 61 (2014): 3857–3866, https://doi.org/10.1109/TIE.2013.2284132.

Appendix A

The parameters of the proposed DVR system are as follows:

a. Source voltage = 400 V, 50 Hz

b. Interfacing inductor = 1 mH

c. DVR ripple filter, series resistance $\left(R_r\right) = 1\Omega$, and series capacitor $\left(C_r\right) = 31.8\ \mu F$

d. Battery specification: nominal voltage = 260 V, rated capacity = 1.2 Ah, initial state of charging = 100%, battery response time = 30 s

e. Series injection transformer rating: three single phase transformers with rating of each: 10.39 kVA, 300 V/300 V.

1097007x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cta.4270 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) | 979-8-3503-9156-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSTSN61422.2024.10670853

Design and Implementation of BIST logic for High Speed and Energy Efficient Carry Select Adder(CSLA)

Garima Solanki
Department of Electrical Engineering
Delhi Technological University
Delhi, India
garimasolanki@dtu.ac.in

Sourav Agarwal
Department of Electrical Engineering
Delhi Technological University
Delhi, India
souravagarwalcr7skr@gmail.com

Tushar Mishra
Department of Electrical Engineering
Delhi Technological University
Delhi, India
tusharmishra263@gmail.com

Vansh Khandelwal
Department of Electrical Engineering
Delhi Technological University
Delhi, India
vansh.khandelwal06@gmail.com

***Abstract*—This paper investigates the implementation of Built-In Self-Testing (BIST) techniques tailored specifically for high-speed and energy-efficient Carry Select Adders (CSLA). The study proposes a novel CSLA design dominated by carry generation logic, aiming to optimize performance while minimizing power consumption and chip area. By integrating BIST circuitry into the CSLA architecture, the research contributes to enhancing fault detection and ensuring system integrity in high-speed computing applications.The proposed CSLA design employs carry generation logic to improve computational efficiency while mitigating power consumption and propagation delay. Through simulation using Xilinx Vivado, the study evaluates the effectiveness of the integrated BIST circuit in detecting and diagnosing faults within the CSLA. The results demonstrate the practical viability of the BIST-enhanced CSLA design in maintaining arithmetic computation integrity and reliability in high-speed computing environments.This research underscores the importance of integrating BIST techniques into CSLA designs to enhance fault detection and diagnosis capabilities, thereby improving overall system reliability and performance. The findings offer valuable insights into the implementation of BIST methodologies in high-speed and energy-efficient computing architectures, with implications for enhancing fault tolerance and system robustness in real-world applications.**

***Index Terms*—Carry Select Adder(CSLA), Built-In Self Test(BIST), Multiple Input Signature Register(MISR), Linear Feedback Shift Register(LFSR), Circuit Under Test(CUT)**

I. Introduction

Adders, crucial for performing arithmetic computations, are commonly employed in Multiplier-Accumulator (MAC) operations within signal processing systems [1]. The performance of the adder significantly impacts the overall system efficiency. While the Ripple Carry Adder (RCA) is widely used, it suffers from a longer Carry Propagation Delay (CPD). Conversely, the Carry Select Adder (CSLA) can improve CPD but at the cost of increased chip area and power consumption. This paper presents a BIST implementation on a high-speed, energy-efficient CSLA design, focusing on carry generation logic as the Circuit Under Test (CUT). In the proposed CSLA, all input variables are carry functions, and each functional block within the CSLA generates carry functions as outputs [2]. CSLA when optimized and manufactured, defects or faults may occur. Upon optimization and fabrication, the possibility of defects or faults arises. To address this concern, the implementation of a built-in methodology or supplementary circuit becomes imperative. This method or circuit should have the capability to autonomously generate test patterns and conduct comprehensive circuit testing. This can be achieved by using the Built-In Self-Testing (BIST) technique [3], [4]. Built-In Self-Test (BIST) is a method in which test patterns are generated to test the circuit. The circuit being tested is referred to as the Circuit Under Test (CUT). Manual testing of circuits can be highly time-consuming and economically impractical, especially during the fabrication of devices or integrated circuits (ICs). To address this issue, BIST is employed, automating the testing phase and making it more efficient and cost-effective [5]. In this study, BIST logic is employed on 4-bit CSLA with and without fault, implemented using Xilinx Vivado. The subsequent sections provide an overview of prior related research (Section 2), detail the BIST logic (Section 3), present the proposed methodology and outcomes (Section 4), and conclude the paper (Section 5).

II. Literature Review

In the available literature, several related contributions have been identified [2], [5]–[10].Minho Nam et al. [2] proposed a design for a high-speed and energy-efficient carry select adder (CSLA) emphasizing carry generation logic dominance. Their findings suggest a reduction in computational delay by 24%, a decrease in power consumption by 17%, and a 37% decrease in the power-delay product compared to conventional implementations.

Patil et al. [5] introduced a research paper focusing on the design and implementation of a Built-In Self-Test (BIST) system, capable of testing any combinational circuit for

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:55 UTC from IEEE Xplore. Restrictions apply.

faults. The paper elaborates on the operation of BIST and its constituent units, including LFSR, comparator, signature analyzer, and Circuit Under Test (CUT).

Yashwanth et al. [6] proposed and compared five designs of an 8-bit Vedic multiplier utilizing various adder architectures, including the ripple carry adder, carry select adder, advanced carry select adder, Brent-Kung adder, and Ling adder. The study utilized Verilog HDL and the Xilinx tool to evaluate these designs based on delay and area metrics.

Supraja et al. [7] introduced a built-in self-test (BIST) technique enabling the circuit to autonomously detect errors using additional hardware. This technique reduces test complexity, duration, and power consumption compared to conventional methods. The authors implemented the proposed design using VHDL and compared its performance with existing models, demonstrating high efficiency, low area, and high speed.

Mishra et al. [8] presented a modified test pattern generator (TPG) utilizing a linear feedback shift register (LFSR) to reduce the number of registers and consequent power consumption. Their study concluded that the proposed design achieved a balance between low power consumption and high-speed performance suitable for Built-In Self-Test (BIST) applications.

Jamuna et al. [9] described a design and implementation of BIST logic for an Arithmetic Logic Unit (ALU) on a Field-Programmable Gate Array (FPGA). The authors applied BIST to a 4-bit ALU comprising a ripple carry adder and a Wallace tree multiplier, assessing fault coverage, code coverage, power consumption, and resource utilization. Xilinx Vivado and QuestaSim were utilized for synthesis, implementation, and simulation, revealing a high 94.2

Gizopoulos et al. [10] proposed a novel testable arithmetic logic unit (ALU) capable of complex arithmetic and logic operations. The ALU is testable for the cell fault model (CFM) using a deterministic BIST scheme based on a 7-bit counter and an accumulator-based compactor.

III. BIST

BIST incorporates the test functions into the Circuit Under Test (CUT) and the effect is incorporated into the design and implementation of the circuit. Typically, BIST incorporates three additional blocks into a digital circuit: A Test Pattern Generator (TPG) [11], [12], a test controller, and a response analyzer. It signifies that the TPG creates test vectors for the circuit under test, where in this context, an LFSR acts as the TPG. CUT is used to process and check if the circuit behaves in a manner that corresponds to the desired operational model using Output Response Analyzer (ORA) to compress the responses [9].

During normal operation, the tested circuit receives inputs from the TPG and operates as usual. However, during the testing phase, the TPG generates test patterns which are applied to the circuit, and the resulting responses are analyzed by the ORA. In typical BIST setups, the responses obtained from testing are optimized to generate fault signatures, which are subsequently compared with golden signature values. An error signal is then employed to determine whether the chip is faulty or fault-free. Critical parameters to consider during the development of BIST techniques encompass fault coverage [13], test set size, hardware overhead, and performance overhead [9].

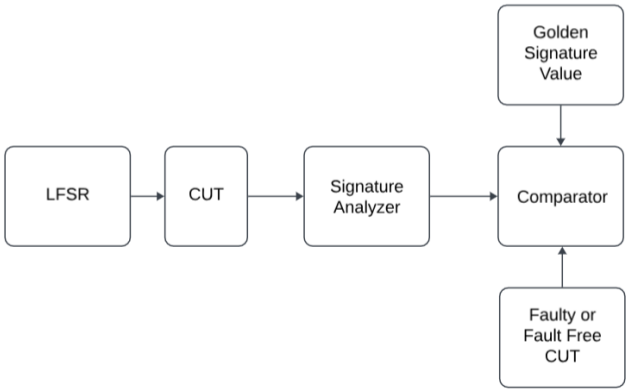


Fig. 1. BIST Block Diagram

A. Linear Feedback Shift Register (LFSR)

Linear Feedback Shift Register, a pattern generator that uses a linear relationship between current and previous input bits. LFSR design involves employing a characteristic polynomial equation to define tap bits, which are then XORed with the leftmost input bit. Two types of LFSR exist: Fibonacci and Galois. LFSR can generate $2^n - 1$ number of test patterns, where n refers to the number of bits. The test patterns are administered to the circuit under test at the negative edge of the clock signal. The starting value of LFSR is called seed. LFSR is small and fast and produces maximum number of test patterns with a proper choice of tap bits. LFSR can also reduce power consumption and test time for BIST [5].

The LFSR equation implemented is M(x)=$x^{11} + x^2 + 1$ and its block diagram is Fig. 2.

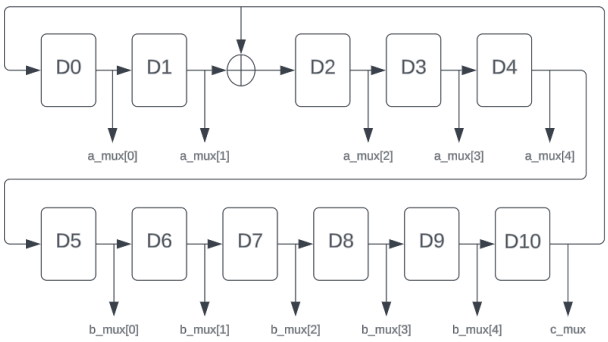


Fig. 2. Implemented LFSR Model

B. Signature Analyzer and Multiple Input Signature Register (MISR)

Signature Analyzer simplifies digital circuit testing by converting raw pulse waveforms into recognizable signatures. These signatures are compared with reference signatures to detect faults. The process involves capturing data, timing

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:55 UTC from IEEE Xplore. Restrictions apply.

control, and signature verification. Signature analyzers are compact, fast, and useful for BIST applications.

MISR enhances testing efficiency in multiple-output digital systems. It combines several input data streams into a single signature using parallel linear feedback shift registers (LFSRs). MISR compresses test response data, efficiently verifies system functionality, and detects faults.

In summary, signature analysis simplifies fault detection, while MISR efficiently combines input data streams for testing. The MISR equation implemented is M(x)=x^6+x+1 and its block diagram is Fig. 3.

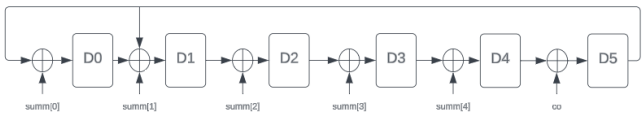


Fig. 3. Implemented MISR Model

C. Circuit Under Test (CUT)

Carry Select Adder uses combinational logic to reduce the power consumption, delay and PDP of binary addition. It contains three blocks: PCU, WCU, and FSU, which are connected in a cascaded manner. It can handle 5-bit binary input data.

1) Primary Carry Unit (PCU): The input stage originates primary carry using XOR and NAND gates. It produces two outputs, Cp0 and Cp1, for each bit of input. The delay remains constant regardless of the input bit-width.

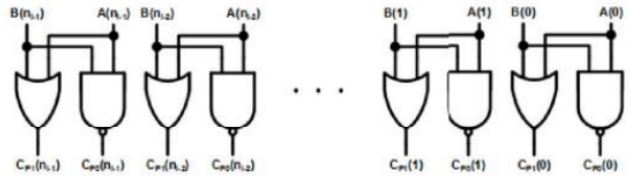


Fig. 4. Circuit Diagram of Primary Carry Unit

2) Wave Carry Unit (WCU): The second stage performs the carry propagation functions. It has two sub-blocks those are: Wave Carry Unit0 and Wave Carry Unit 1, that are activated by the carry input Cin. They produce two outputs, Cw0 and Cw1, for each bit of input. They are designed with two NAND gates in series and have a shorter delay than conventional full adders. The Least Significant Bit is given a free pass.

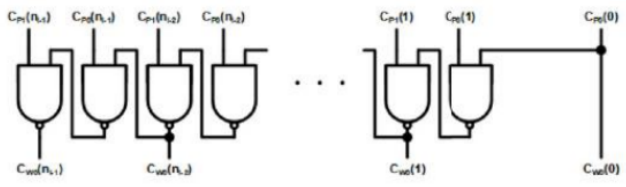


Fig. 5. Circuit Diagram of Wave Carry Unit 0

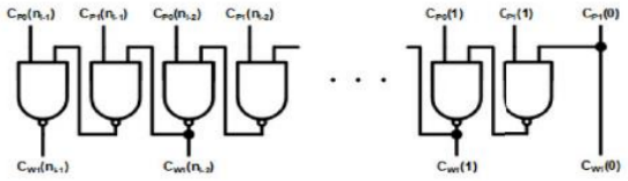


Fig. 6. Circuit Diagram of Wave Carry Unit 1

3) Final Selection Unit (FSU): The final stage computes the Sum and Cout outputs. It uses the PCU and WCU outputs, along with the carry input Cin. It is designed using three NAND gates and an XNOR gate per bit of input, it sustains a uniform delay, irrespective of the input bit-width. The Least Significant Bit and Most Significant Bit are handled differently.

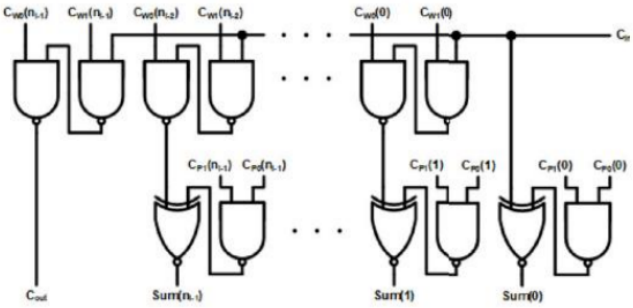


Fig. 7. Circuit Diagram of Final Selection Unit

IV. Methodology

The principal objective of this design proposal is to integrate BIST logic into a 5-bit High-Speed CSLA with a carry input. The synthesized RTL schematic of this proposed design, achieved through Xilinx Vivado, is shown in Fig. 8.

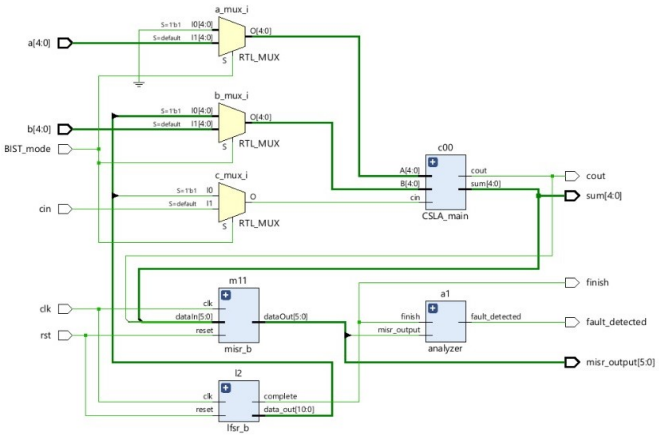


Fig. 8. RTL schematic of the synthesized design

Fig.8 shows the functional block diagram which consists of an LFSR circuit to generate input patterns for A, B, and cin for the high-speed CSLA, an MISR circuit that takes the sum and carry output to generate the signature, and an

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:55 UTC from IEEE Xplore. Restrictions apply.

analyzer to compare the signature of a non-faulty circuit and the Circuit Under Test(CUT), in our case which is the High-Speed CSLA.

Figure 9 also provides an overview of Circuit Under Test functional block diagram and more specifically of the High-Speed CSLA circuit. The circuit comprises of Primary Carry Unit block, the 2-wave Carry Unit block, and the Final Selection Unit block, which together are used to produce the sum and carry within this circuit.

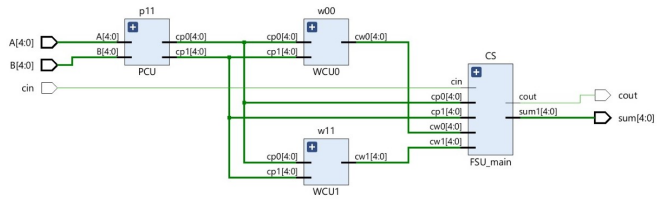


Fig. 9. CSLA RTL schematic

Fig. 10 presents the functional block diagram of BIST Implementation of the non-faulty circuit of the CUT to generate the signature. It consists of two test pattern generators LFSR and the MISR. The signature value for the non-faulty circuit came out to be 01011. This signature is used in the functional block diagram shown in Fig. 8 to detect whether the circuit is faulty or not.

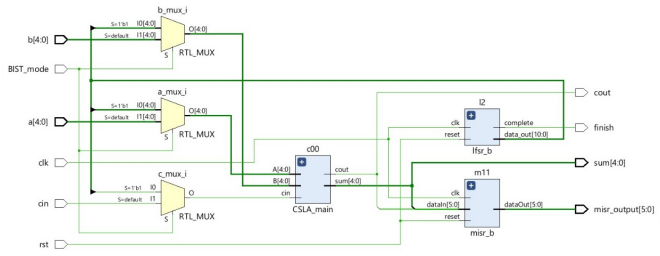


Fig. 10. RTL schematic of the Non-Faulty Circuit

V. Result

To evaluate the design on this subject, Xilinx Vivado was employed in simulating the proposed design. Figure 11 presents the simulation results of different aspects of the High-Speed CSLA circuit that aids in testing a Circuit Under Test.



Fig. 11. Output of CUT

Fig. 12 presents the simulation Non-Faulty Circuit output and Fig. 13 shows the Faulty circuit output.

We can see in Fig. 12 when the simulation was complete i.e when all the patterns were generated for a, b, and cin as

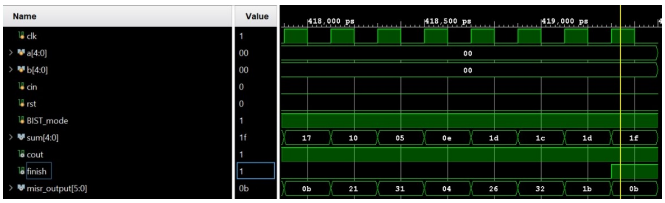


Fig. 12. Result without fault injection

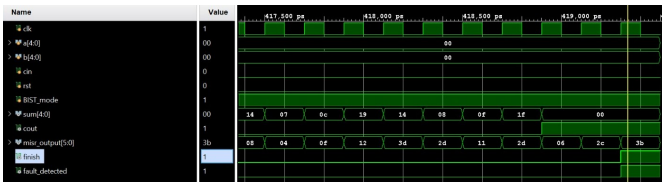


Fig. 13. Result with fault injection

inputs, the finish signal got high and the corresponding MISR output was 0b (1011) which is the signature value.

In Figure 13 when the simulation ends as evidenced by the high status of the finish signal having a value of '1', the signal analyzer values were 3b (0011 1011) which is different from the golden signature value. Therefore, this divergent accounts for an actual fault identification, and this leads to the fault-detected signal being set to 1.

VI. Conclusion

In this paper, the discussion on how the BIST techniques can be applied on a High Speed Carry Select Adder on an Xilinx Vivado platform using Verilog language has been described. The BIST circuit is used and integrated in the CSLA which not only increases the current reliability and solidity of the system to some extent but also helps in providing deeper diagnosis of faults.

The achieved results show the design of the BIST circuit is efficient in identifying and isolating any fault that may be existent in the carry-select adder and engulfs any arithmetical flaw in high-speed applications. First, the incorporation of BIST makes it easier to test a given digital circuit and at the same time, the reliability as well as the life cycle of the designed circuit is also boosted as a result.

Future work in this area may involve exploring optimization techniques for further enhancing the performance of the BIST circuit, as well as extending the methodology to other arithmetic components.

VII. Acknowledgment

We extend our sincere gratitude to Delhi Technological University (DTU) and the Department of Electrical Engineering for their invaluable support in our research endeavors. The facilities provided by DTU and the conducive academic environment of the Department of Electrical Engineering have been instrumental. We also appreciate the collaborative efforts of our colleagues. Thank you to all individuals and institutions for their support in our academic pursuits.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:55 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] H. Weste, Neil, E. Weste, and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective*. Reading: Addison-Wesley Publishing Company, 1993.

[2] M. Nam, Y. Choi, and K. Cho, "High-speed and energy efficient carry select adder (csla) dominated by carry generation logic," *Microelectronics Journal*, vol. 79, pp. 70–78, 2018.

[3] J. Rajski and J. Tyszer, *Arithmetic Built-In Self-Test for Embedded Systems*, 1st ed. Publisher, Year.

[4] L. T. Wang, C. W. Wu, and X. Wen, *VLSI Test Principles and Architectures*. Elsevier, 2006.

[5] M. Patil and H. B. Sharanabasaveshwar, "Design and implementation of bist," *2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEEC-COT)*, 2018.

[6] D. Yaswanth, S. Nagaraj, and R. V. Vijeth, "Design and analysis of high-speed and low-area vedic multiplier using carry-select adder," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2020.

[7] T. Supraja *et al.*, "Design of low power based carry look ahead adder using domino logic for high speed alu application," *Journal Name*, 2021.

[8] B. Mishra, R. Jain, and R. Saraswat, "Low power BIST based multiplier design and simulation using FPGA," in *2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE, 2016.

[9] S. Jamuna, P. Dinesha, and K. Shashikala, "Design and implementation of BIST logic for ALU on FPGA," in *2018 International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*. IEEE, 2018.

[10] D. Gizopoulos *et al.*, "An effective BIST scheme for arithmetic logic units," in *Proceedings International Test Conference 1997*. IEEE, 1997.

[11] S. Wang, "A BIST TPG for low power dissipation and high fault coverage," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 7, pp. 777–789, July 2007.

[12] A. Crouch, M. Pressly, and J. Circello, "Testability features of the mc68060 microprocessor," in *Proc. Int'l Test Conf.*, 1994, pp. 60–69.

[13] Oyiatzis, "An ALU-based BIST scheme for word-organized RAMs," *IEEE Transactions on Computers*, vol. 57, no. 5, pp. 577–590, May 2008.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:55 UTC from IEEE Xplore. Restrictions apply.

Indian Journal of Chemical Technology
Vol. 31, September 2024, pp. 681-690
DOI: 10.56042/ijct.v31i5.10942

Development and evaluation of gelatin/polyacrylamide/carboxymethyl tamarind kernel gum hydrogel for delivery of ampicillin sodium

Tanuja Kumari, Nitin, Priyanka Meena* & Sudhir G. Warkar*
Department of Applied Chemistry, Delhi Technological University, Delhi, India
*E-mail: priyankameena.du2016@gmail.com (PM), sudhirwarkar@gmail.com (SGW)

Received 6 May 2024; accepted 8 July 2024

Gelatin-based hydrogels, despite their excellent biocompatibility, face limitations in their utility for targeted drug release due to low mechanical strength. Thus, this research is directed towards the fabrication of pH-responsive hydrogels based on gelatin, polyacrylamide (PAM) and carboxymethyl tamarind kernel gum (CMTKG), followed by their loading with ampicillin sodium drug. The hydrogel has been optimized by varying crosslinker, and initiator amounts to observe their effect on swelling. The swelling is increased with an increase ininitiator, reaching a maximum of 1371% at pH 7.4 and 1218% at pH 1.2. The hydrogels are analyzed through Scanning Electron Microscopy (SEM), Powder X-ray Diffraction (PXRD), and Attenuated Total Reflection-Fourier Transform Infrared Spectroscopy (ATR-FTIR) techniques. Various parameters such as drug loading, porosity, and gel fraction are evaluated. The *in vitro* drug release is evaluated in both pH 1.2 and 7.4 buffer, with higher drug release (63.3%) observed under alkaline pH. The kinetic modeling data validates the Korsmeyer-Peppas model (R^2= 0.9871) to be the suitable model for the explanation of the drug release mechanism, suggesting the Fickian diffusion (n<0.5, pH 1.2) and non-Fickian diffusion (n>0.5, pH 7.4). Therefore, the Gelatin/PAM/CMTKG hydrogel shows potential for targeted release of ampicillin sodium in response to varying pH.

Keywords: Biopolymer, Drug Release, Gelatin, Hydrogel, Kinetic modelling

Introduction

Hydrogel is a three-dimensional framework composed of hydrophilic polymer chains, physically or chemically crosslinked. The hydrophilic groups such as -COOH, -SO_3H, -OH and -$CONH_2$ etc., facilitate the absorption of fluid into the matrix[1]. The surface tension and capillary forces facilitate the swelling characteristics of the hydrogel. Hydrogels are renowned for their biocompatibility and environmental friendliness and are widely used in industrial, pharmaceutical, and agricultural fields[2]. Various synthetic and natural polymers, including polyacrylamide[3], collagen[4], gelatin[5–7], hyaluronic acid etc., have been employed to fabricate hydrogel networks[8]. The cross-linking within the hydrogel structure serves as a crucial factor in preventing their dissolution in aqueous environments[9]. These polymeric networks exhibit a distinctive characteristic similar to living tissues, depicted by their soft and rubbery nature. This feature renders them well-suited for various biomedical applications. Hydrogels, also known as intelligent or smart hydrogels, exhibit responsiveness to different chemical and physical stimuli, such as magnetic field, pH, ionic strength, etc.[10].

In biomedical applications, pH plays a very crucial role in the design and implementation of targeted drug delivery systems. The pH-responsive hydrogels are categorized as either anionic or cationic based on their response to changes in pH. Anionic hydrogels exhibit higher swelling in alkaline pH as they undergo ionization at a pH greater than their pKa. However, cationic hydrogels show a greater swelling in acidic pH, as it ionizes at a pH lower than their pKa[10]. Polyacrylamide (PAM), is a non-toxic, pH-sensitive polymer known for its excellent biocompatibility [11,12]. It contains the –$CONH_2$ group, which fosters swelling and drug-release behavior at various pH conditions[13]. Nowadays, PAM have been combined with biopolymers such as chitosan[12], carboxymethyl cellulose[11], carboxymethyl tamarind kernel gum[14] etc. to formulate hybrid hydrogels with improved biocompatibility, swelling and mechanical properties.

Tamarind Kernel Gum (TKG) is an economical polysaccharide, procured from the Tamarindus indica L. plant seeds. It's carboxymethylated derivative, carboxymethyl tamarind kernel gum (CMTKG) is a biopolymer with molar proportions of xylose, glucose and galactose as 1:3:2(Ref.14). It is an anionic polymer with high swelling capacity and pH-dependent release properties, making it suitable for drug delivery applications[15].

Gelatin is another biopolymer, derived from the partial hydrolysis of collagen, obtained from animal tissues i.e., skin, ligament, bones, blood vessels, and heart. Its primary components include glycine, proline, and 4-hydroxyproline[16]. Gelatin emerges as a remarkable preference for drug delivery applications owing to its non-toxic nature, excellent gel-forming ability, good biodegradability and highly biocompatible[17].

Ampicillin sodium, a model drug, is a β-lactam antibiotic, that belongs to the class of penicillin. It is widely employed in addressing infections arising from gram-negative and gram-positive bacteria, in both humans and animals. This versatile drug is efficient in treating a range of infections such as gastrointestinal, meningitis, and urinary tract infections[18,19].

Numerous studies have explored hydrogel-based delivery systems for hydrophilic drugs. However, these approaches often encounter limitations such as challenges in achieving release at targeted pH in the gastrointestinal tract and having low mechanical strength, leading to burst drug release. For instance, gelatin/ß-cyclodextrin/dextran hydrogels were investigated for the release of 5-fluorouracil, but it exhibited approximately 50% drug release at *p*H 1.2 and around 80% at pH 7.4(Ref.20). Hence to minimize the release at untargeted sites, this study explored the fabrication of a series of gelatin, PAM and CMTKG based hydrogels for pH-dependent release of model drug ampicillin sodium. The novelty of this work is that the Gelatin/PAM/CMTKG hydrogel has been fabricated and tailored for the first time for the controlled release of ampicillin sodium. The synthesized hydrogels were analyzed by PXRD, ATR-FTIR, and SEM techniques. In addition, the impact of initiator, and cross-linker amount on the swelling of synthesized hydrogels was studied in pH 1.2 and 7.4 buffer. The evaluation of the drug release kinetics was done with different mathematical models comprising Hixson-Crowell, Korsmeyer-Peppas, Higuchi, Zero order and First order.

Experimental Section

Materials

Gelatin (Fischer Scientific), Potassium persulphate [KPS, MW=270.30 g/mol, (Fischer Scientific, Mumbai, India)], Acrylamide [AM, MW=71.08 g/mol, (CDH, New Delhi, India)] and N, N′-methylene-bis(acrylamide) [MBA, MW=154.17 g/mol (Merck, Germany)] were used as provided. CMTKG [0.20º substitution, MW=9.14 $\times 10^5$ g/mol] was generously given by Hindustan Gum and Chemicals Ltd., Haryana, India and Ampicillin sodium [MW=371.39 g/mol], was gifted by Unicure India Pvt Ltd, Noida, India.

Synthesis of Gelatin/PAM/CMTKG hydrogels

The series of Gelatin/PAM/CMTKG hydrogels (A-1 to A-9) were fabricated with different amounts of initiator KPS (14-38 mg), and crosslinker MBA (10-30 mg), as detailed in Table 1. A required amount of CMTKG, AM and gelatin were dissolved in distilled water while continuously stirring for 1 h. Besides this, KPS and MBA were added to the solution and stirred for an hour. Then poured into the test tubes and heated for 1h in a water bath at 60°C. The synthesized hydrogels were removed from the test tube, sliced, and oven-dried at 60°C. Fig. 1 shows the illustration of the fabrication of hydrogel[21].

Table 1 — Composition of Gelatin/PAM/CMTKG hydrogels

Sample code	CMTKG (g)	AM (g)	MBA (mg)	KPS (mg)	Gelatin (g)
A-1	0.2	0.5	10	32	0.5
A-2	0.2	0.5	15	32	0.5
A-3	0.2	0.5	20	32	0.5
A-4	0.2	0.5	25	32	0.5
A-5	0.2	0.5	30	32	0.5
A-6	0.2	0.5	10	14	0.5
A-7	0.2	0.5	10	20	0.5
A-8	0.2	0.5	10	26	0.5
A-9	0.2	0.5	10	38	0.5

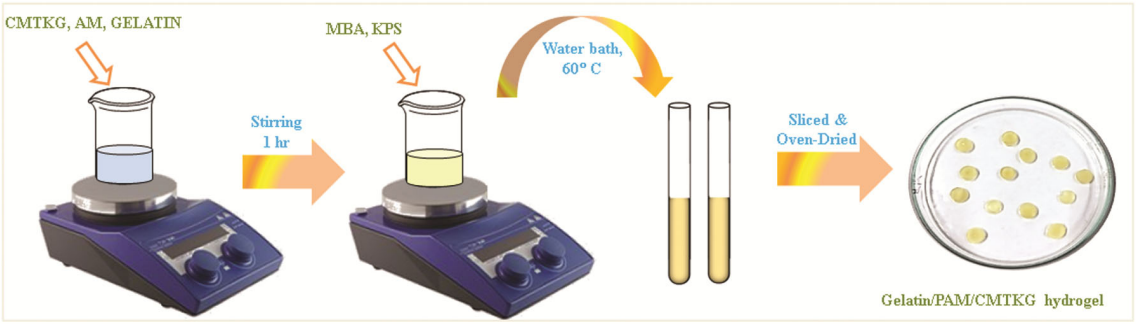


Fig. 1 — Synthesis of Gelatin/PAM/CMTKG hydrogel

Swelling studies

The swelling behavior of fabricated hydrogels was assessed in pH 7.4 and 1.2 buffer over 24 h. Dried hydrogel discs were accurately weighed (W_d) and subsequently immersed in the appropriate solutions. After 1 h, the hydrogel disc was extracted, gently blotted with filter paper, weighed (W_S), and subsequently re-immersed in the solution[22]. Swelling was conducted in triplicate and assessed using the given equation[23].

$$\text{Equilibrium Swelling (\%)} = \frac{(W_s - W_d)}{W_d} \times 100 \quad (1)$$

Gel fraction

To calculate the gel fraction (crosslinked portion) of all synthesized hydrogels, the weighed hydrogel disc (W_i) was dipped in distilled water for 24 h and then taken out, dried in an oven at 60°C and weighed (W_e) again. The gel fraction (%) was calculated using the following equation[24].

$$\text{Gel fraction (\%)} = \frac{W_e}{W_i} \times 100 \quad (2)$$

Porosity

The porosity of the fabricated hydrogels was evaluated using hexane as the solvent. The dried hydrogel discs (M_i) were dipped in 10 mL of hexane for 2 h. After removing the hydrogel disc from the hexane and wiping it with filter paper, the final weight (M_f) was recorded. The porosity of the hydrogel disc was calculated using the following equation:

$$\text{Porosity (\%)} = \frac{M_f - M_i}{\rho V} \times 10 \quad (3)$$

Here, ρ and V represent the density of hexane (0.659 g/mL), and volume of the hydrogel[25,26].

Drug loading and drug entrapment efficiency

The drug loading was carried out for formulation (A-1) as it exhibited the highest swelling among all the hydrogels. Ampicillin sodium-loaded Gelatin/ PAM/CMTKG hydrogel was formulated as mentioned in the synthesis procedure with the minor variation that drug ampicillin sodium (35 mg) was also added during the addition of MBA and KPS in the solution and the rest process remains the same. The calculation of drug loading and drug entrapment efficiency was performed for ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel. A hydrogel was immersed in a pH 7.4 buffer for 24 h, then the 3 ml solution was withdrawn, and the absorbance at λ_{max} 203 nm was measured through a UV-vis spectrophotometer (Model: Cary 300 UV–Vis). The amount of drug loaded was calculated following the Beer-Lambert law, using the calibration curve drawn with the known concentration of the drug solution. The equations employed for the determination of drug loading (DL%) and drug encapsulation efficiency (DEE%) were as follows[27,28]

$$DL(\%) = \frac{Amount\ of\ drug\ loaded\ in\ hydrogel}{Weight\ of\ hydrogel} \times 100 \quad (4)$$

$$DEE(\%) = \frac{Amount\ of\ drug\ loaded\ in\ hydrogel}{Initial\ amount\ of\ drug\ added\ in\ hydrogel} \times 100 \quad (5)$$

***In vitro* ampicillin sodium release**

The *in vitro* studyfor ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel was conducted in an orbital incubator shaker under pH 1.2 and pH 7.4buffer solutions. A precise quantity (0.1 g) of the ampicillin sodium-loaded hydrogel disc was immersed in 100 mL of pH 7.4 and 1.2solutions at 37 °C. At the interval of1 hour, 3 mL of the solution was taken out, and an equal volume of fresh buffer was added to it. Subsequently, the absorbance of the solution was measured employing a UV spectrophotometer (Model: Cary 300 UV-Vis) at the λ_{max} 203 nm. The drug amount was determined through the calibration curve drawn with the known concentration of the drug solution[29].

Kinetic modeling of ampicillin sodium

Various mathematical models have been employed to analyze drug release mechanisms, including the Korsmeyer-Peppas, Hixson-Crowell, Higuchi, Zero-order and First-order models. The regression coefficient (R^2) values were correlated and the model with an R^2 value closest to 1 was considered the most fitted one[30].

Characterization

Powder X-ray diffraction (PXRD) was performed for ampicillin sodium, Gelatin/PAM/CMTKG and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel, using a high-resolution Bruker D8 diffractometer at 2θ range of 10-70° with CuKα radiation. Attenuated Total Reflection-Fourier Transform Infrared Spectroscopy (ATR-FTIR) spectra were recorded for MBA, Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/ PAM/CMTKG hydrogel within a wavenumber range of 4000-450 cm^{-1}, using the ATR-FTIR spectrophotometer (Model: Nicolet iS50 FTIR Tri-detector). Scanning Electron Microscopy (SEM) images of Gelatin/ PAM/CMTKG and ampicillin sodium-loaded Gelatin/ PAM/CMTKG hydrogels were captured using an SEM (JEOL Japan Mode: JSM 6610LV).

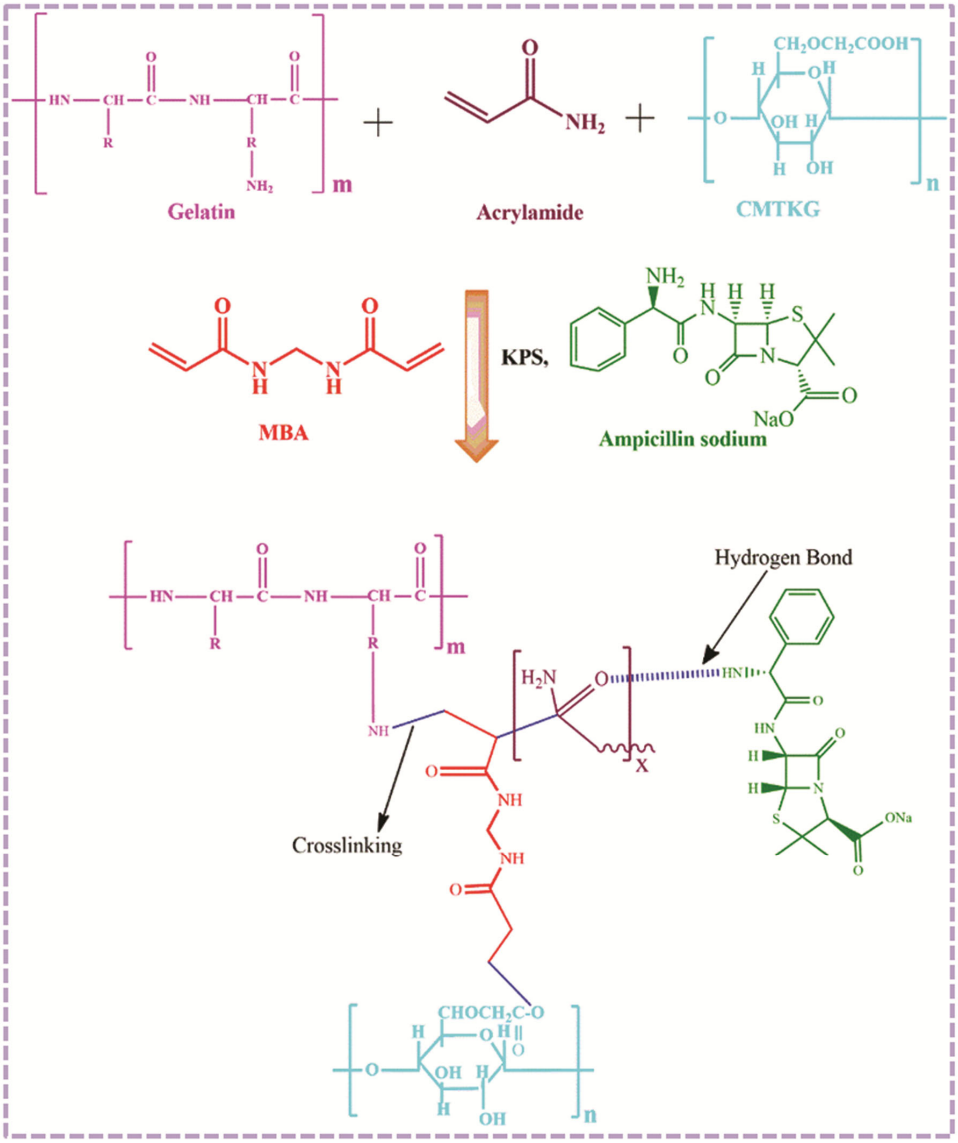
Statistical analysis

Each experiment was carried out thrice (n=3), and the results are reported as mean ± standard deviation. Statistical analysis was performed using ANOVA (One-way analysis of variance), with significance set at $p < 0.05$ for all studies.

Results and Discussion

Mechanism of Gelatin/PAM/CMTKG hydrogels and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel formation

Various hydrogel formulations are fabricated by varying amounts of KPS and MBA using free radical polymerization (as displayed in Table 1). The initiator KPS, decomposes to generate sulfate radicals at 60°C. The decomposition of the vinyl bond of AM in the existence of sulfate radicals leads to the formation of acrylamide radicals. These radicals-initiated chain formation, propagation, and finally termination, lead to the formation of PAM. The sulphate radical also abstracts the H atom from the COOH group of CMTKG and the -NH_2 group of Gelatin, generating radicals on polymeric chains. The MBA acted as a cross-linker, connecting the free radicals present on the Gelatin, CMTKG, and PAM chain, thus leading to the formation of a three-dimensional polymeric hydrogel network[31]. Moreover, in ampicillin, sodium-loaded Gelatin/PAM/CMTKG hydrogel, the hydrogen bonding among ampicillin sodium (drug) and PAM contributed to the development of ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel as displayed in Scheme 1.



Scheme 1 — Mechanism of synthesis of ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel

Swelling studies

Swelling analysis was conducted in pH 7.4 and pH 1.2 buffers for all fabricated hydrogels. Fig. 2 illustrates the impact of the initiator (KPS) and crosslinker (MBA) quantities on the equilibrium swelling (%) of the hydrogel.

Impact of cross-linker

The swelling of synthesized hydrogels was significantly affected by the concentration of MBA (Fig. 2a). The synthesized hydrogel (A-1) displayed the highest swelling of about 1371±0.75% at pH 7.4 and 1218±0.76%at pH 1.2 with 10 mg MBA, compared to all other formulations, as listed in Table 1. The swelling capacity decreased when the amount of MBA was increased from 10 mg to 30 mg. This reduction in swelling ($p < 0.05$) was attributed to higher concentrations of cross-linker, which minimized the free volume within the polymeric network, thereby decreasing the capacity for fluid absorption. Moreover, decreasing the MBA by 10 mg resulted in a gel-like structure due to insufficient cross-linking in the polymer network [32].

Impact of Initiator

The concentration of MBA is held constant at 10 mg, while variations in initiator (KPS) content (14-38 mg) are examined for their swelling effects, as illustrated in Fig. 2 (b). Initially, an increase in initiator content leads to a rise in the swelling (%), having a maximum swelling of 1371±0.75% at pH 7.4 and 1218±0.76% at pH 1.2 with 32 mg of KPS. However, a decrease occurs below 32 mg, attributed to unreacted monomers, resulting in a less compact polymer network and a subsequent reduction in swelling ($p < 0.05$). Additionally, initiator content exceeding 32 mg causes swelling to decrease again, due to the formation of a soluble oligomer chain instead of a polymeric chain which leads to a reduction in swelling[14]. For all the formulations, higher swelling ($p < 0.05$) was observed in alkaline pH 7.4 compared to acidic conditions (pH 1.2). In an

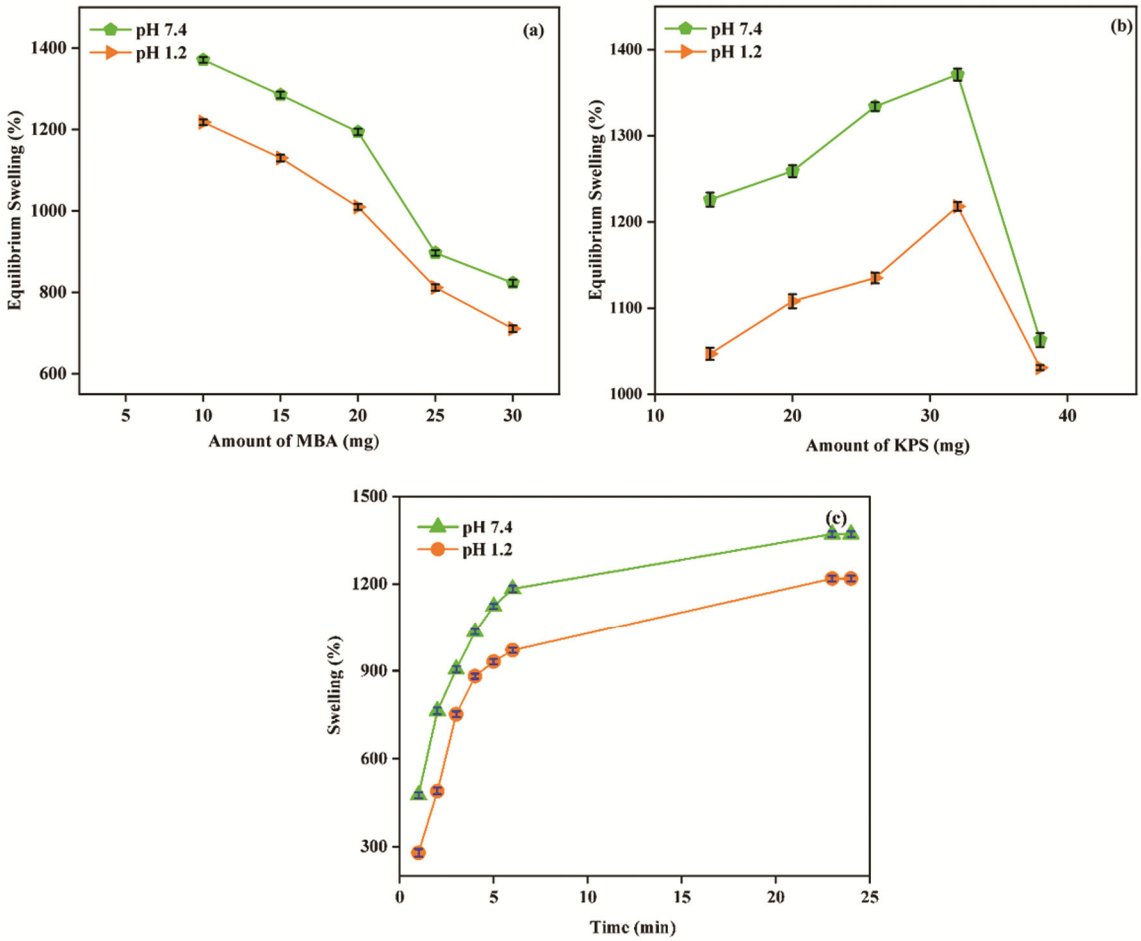


Fig. 2 — Impact of (a) cross-linker, (b) initiator on equilibrium swelling (%), and (c) swelling (%) vs time plot for A-1 formulation

alkaline pH, deprotonation of the COOH groups of polymeric chains happens, and the electrostatic repulsion among the COO^- ions leads to the polymeric chain relaxation and adopt an elongated conformation. This elongation generates extra space, allowing biofluid to penetrate the hydrogel matrix and increase swelling (%). Moreover, at pH 1.2, protonation of COOH groups occurs, leading to the hydrogen bonding among CMTKG, PAM, and gelatin, resulting in the contraction of the hydrogel network, thus reducing the swelling (%) of the hydrogel[31].

Gel fraction analysis

The gel fraction plot revealed that increasing the quantities of crosslinker MBA (A-1 to A-5) and initiator KPS (A-6 to A-8) significantly increases chain entanglement and cross-linking density, leading to a more compact structure, hence higher gel fraction ($p < 0.05$) (Fig. 3). However, exceeding 32 mg of initiator (A-9) results in decreased gel fraction ($p < 0.05$) due to oligomer formation. The presence of these uncross-linked soluble oligomer chains reduces the overall crosslinked gel content of the hydrogel. The highest gel fraction observed was 89.28±0.65% with 32 mg of KPS and 30 mg of MBA in the Gelatin/PAM/CMTKG hydrogel[33].

Porosity

The porosity analysis revealed an increase in the quantity of the crosslinker MBA (A-1 to A-5) results in a decrease in porosity ($p < 0.05$), due to increased crosslinking between Gelatin, PAM, and CMTKG chains, which leads to a denser structure with fewer voids. Moreover, as the amount of initiator KPS (A-6 to A-8) increases, porosity also increases ($p < 0.05$), due to more reaction, which provides more voids between chains. However, beyond 32 mg of KPS (A-9), excessive polymer chain formation, resulting in a reduction in porosity, as illustrated in Fig. 4[34].

Characterization

PXRD analysis was carried out for the ampicillin sodium, Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel to gain insights into the crystalline nature (Fig. 5). The sharp crystalline peak of ampicillin sodium was observed at 2θ value of 12.7, 15.5, 17.3, 18,19.3, 19.8, 21.8, 23.9, 25.8, 26.3, 26.7° (Ref.35). Moreover, the broad band observed in the PXRD spectrum of Gelatin/PAM/CMTKG and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel indicates that the synthesized hydrogel exhibits amorphous nature. Additionally, in the PXRD of the ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel, there is an absence of peaks related to ampicillin sodium, revealing that ampicillin sodium is evenly distributed within the cross-linked network of polymer[30].

ATR-FTIR spectra for the Gelatin/PAM/CMTKG hydrogel, ampicillin loaded-Gelatin/PAM/CMTKG hydrogel, and MBA are shown in Fig. 6. In both Gelatin/PAM/CMTKG and ampicillin loaded-Gelatin/PAM/CMTKG hydrogel, the band at 3321 cm^{-1} & 3196 cm^{-1} and 3317 cm^{-1} & 3194 cm^{-1} respectively, correspond to -NH and -OH overlapping stretching vibration. The peak at 1649 cm^{-1} and 1647 cm^{-1}

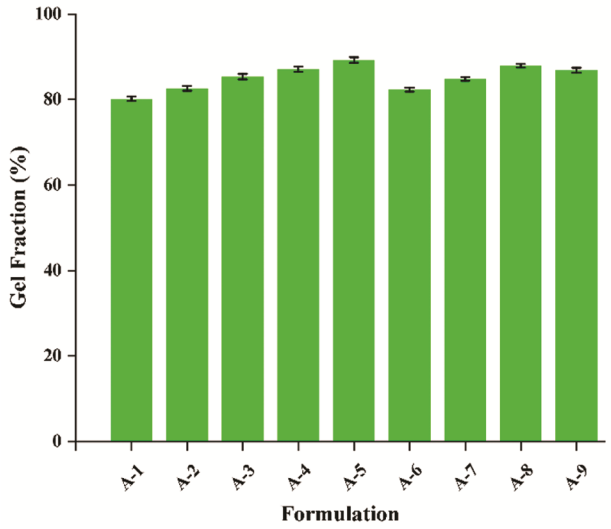


Fig. 3 — Gel Fraction analysis of all Gelatin/PAM/CMTKG hydrogels

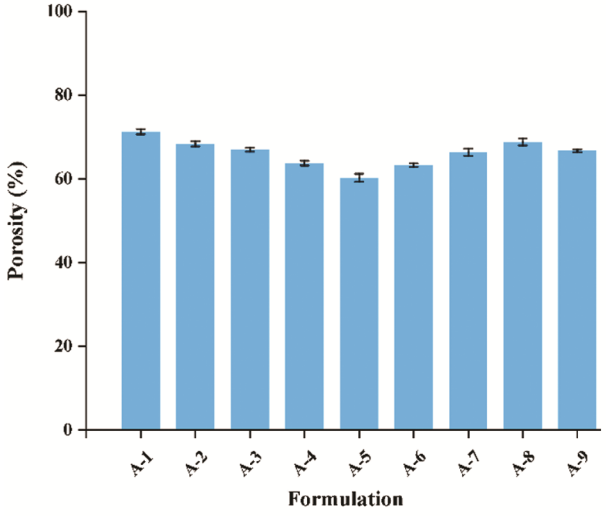


Fig. 4 — Porosity studies of all fabricated Gelatin/PAM/CMTKG hydrogels

signifies the C=O stretch in Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel[36]. The peaks at 2926 cm^{-1} and 2931 cm^{-1} are related to the C-H stretch in Gelatin/PAM/CMTKG and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel. Additionally, the peak assigned to the C-O-C group is sighted at 1094 cm^{-1} and 1106 cm^{-1} in Gelatin/PAM/CMTKG and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel[15].

In the ATR-FTIR spectrum of MBA, the peak at 1307 cm^{-1} ascribed to C-N stretch, is moved to 1323 cm^{-1} and 1321 cm^{-1} in the Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel, shows the decline in conjugation due to cross-linking and successful development of cross-linked hydrogel network[14]. Additionally, in the ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel, the absence of any additional peaks as compared to Gelatin/PAM/CMTKG hydrogel, indicates that there is likely only a weak or physical interaction between the drug and hydrogel.

SEM images of the Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel are presented in Fig. 7. The Gelatin/PAM/CMTKG hydrogel features a highly uneven and porous surface, facilitating the capacity of effective encapsulation of ampicillin sodium crystals within the polymer matrix. However, an SEM micrograph of ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel shows a smoother and less porous surface, as the entrapped drug captures the pores of the hydrogel matrix[10].

Drug loading and drug encapsulation efficiency

The drug loading and drug encapsulation efficiency for the Gelatin/PAM/CMTKG hydrogel (A-1) were examined. The drug encapsulation efficiency (%) was determined to be 61.16% and drug loading (%) was 19.54% for the Gelatin/PAM/CMTKG hydrogel.

***In vitro* Ampicillin sodium release**

An *in vitro* study was conducted for ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel, as

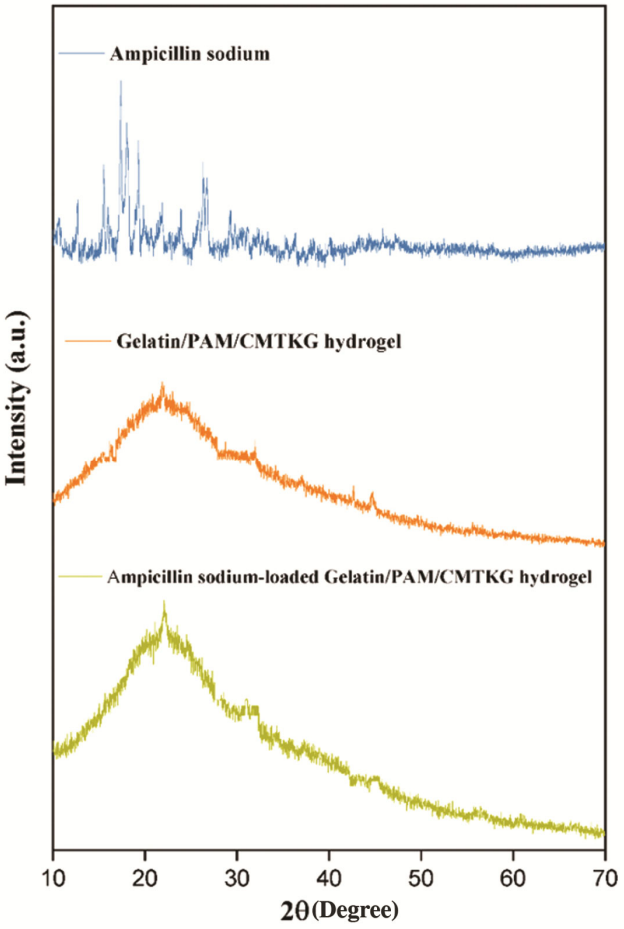


Fig. 5 — PXRD patterns of ampicillin sodium, Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel

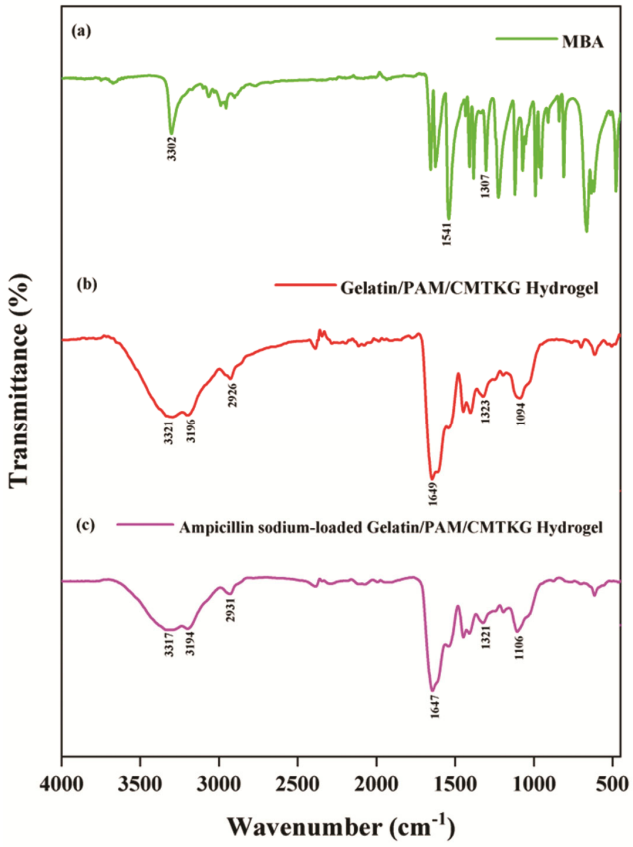


Fig. 6 — FTIR spectra of MBA, Gelatin/PAM/CMTKG hydrogel and ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel

Table 2 — Kinetic modelling data of ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel

Model	Equation	Coefficient of Regression (R^2)		Ref.
		pH 7.4	pH 1.2	
Zero Order	$W_t = W_\infty + k_0 t$ k_0 = zero-order constant	0.9012	0.9327	38
Higuchi	$W_t/W_\infty = k_{HG} t^{1/2}$ k_{HG} = Higuchi constant	0.9762	0.9809	24
First Order	$\text{Log}W_t = \text{Log}W_\infty + \frac{kt}{2.303}$ k = first order constant	0.7268	0.7831	39
Hixson-Crowell	$(W_t)^{1/3} - (W_\infty)^{1/3} = k_{HX} t$ k_{HX} = Hixson Crowell constant	0.8172	0.8239	39
Korsmeyer–Peppas	$W_t/W_\infty = kt^n$ k = kinetic constant n = diffusion exponent	0.9871	0.9892	22

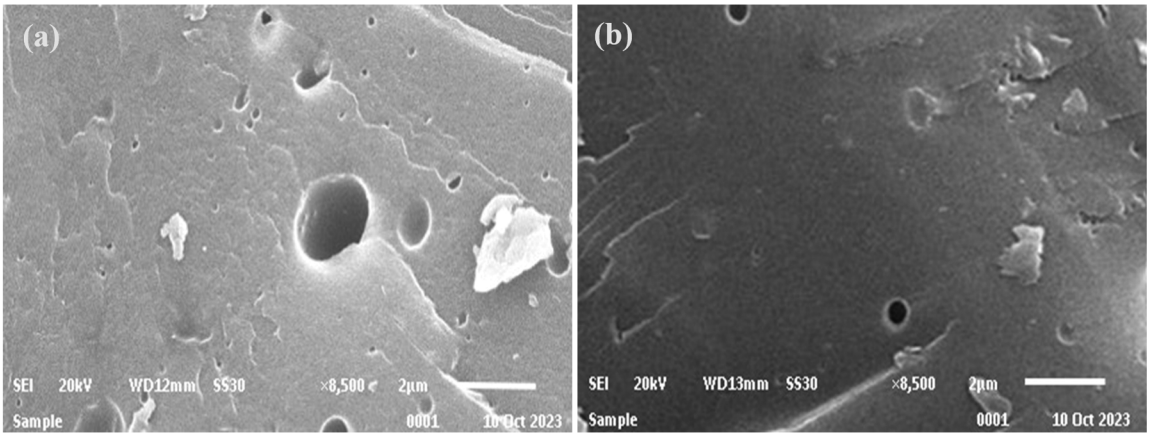


Fig.7 — SEM images of (a) Gelatin/PAM/CMTKG hydrogel and (b) ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel

shown in Fig. 8. The study revealed a higher drug release under pH 7.4 compared to pH 1.2, due to deprotonated COO^- ion present in hydrogel. At pH 7.4, these COO^- ions exhibit repulsion, causing the polymer chain to expand, leading to an increased drug diffusion ($p < 0.05$) from the hydrogel matrix. Under pH 1.2, the COOH group undergoes protonation, which facilitates the formation of hydrogen bonds among the CMTKG, Gelatin, and PAM chains, resulting in a contracted polymeric network, and lower drug release from the hydrogel. The release profile of ampicillin sodium from the hydrogels showed a gradual and prolonged release pattern without any burst release, ensuring optimal drug levels over extended periods and reducing the need for frequent dosing. The maximum ampicillin sodium release was observed to be 63.3±0.69% at pH 7.4, and 14.6 ±0.75% at pH 1.2 from the synthesized ampicillin sodium loaded Gelatin/PAM/CMTKG hydrogel. Hence, the synthesized ampicillin sodium loaded-Gelatin/PAM/CMTKG hydrogel can be utilized for pH-dependent controlled release of ampicillin sodium[30].

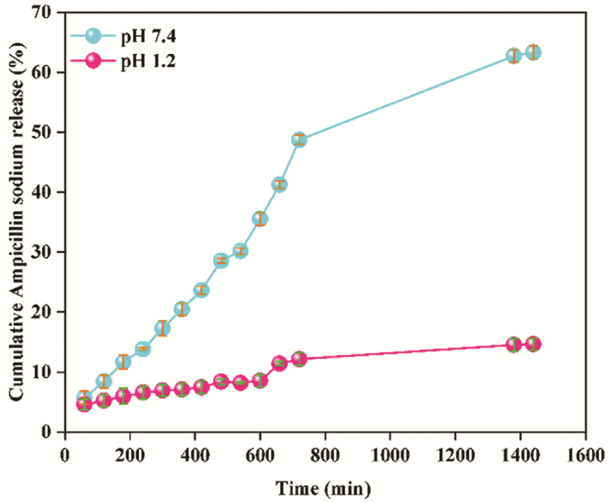


Fig. 8 — Ampicillin sodium release from Gelatin/PAM/CMTKG hydrogel in pH 7.4 and 1.2

Kinetic modelling of ampicillin sodium

The kinetic modelling was done for ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel using various kinetic models, as detailed in Table 2. The results demonstrated that the data closely fit the Korsmeyer-Peppas model under both pH conditions as illustrated in Fig. 9. At pH 7.4, an R^2 value of 0.9871 and an n value of 0.7847 indicated a non-Fickian diffusion mechanism, which revealed that drug release due to the diffusion and polymeric chain relaxation. However, at pH 1.2, the R^2 value was observed as 0.9892, with an n of 0.2557, indicating

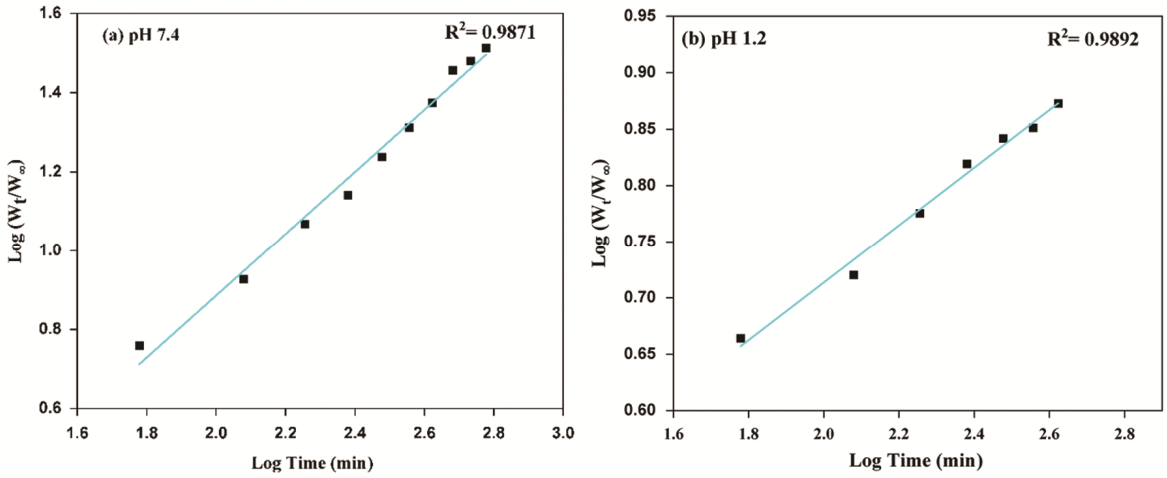


Fig. 9 — Kinetic modelling plot of ampicillin sodium as per Korsmeyer-Peppas model in (a) pH 7.4 and (b) 1.2

the Fickian diffusion mechanism, suggesting that drug release occurred primarily through the diffusion process alone[37].

Conclusion

In this study, the synthesis of pH-responsive ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel was successfully achieved using MBA as a crosslinker and KPS as the initiator. The swelling analysis revealed that an increase in KPS leads to an increase in swelling ($p < 0.05$), but increases in MBA result in decreased swelling values. The porosity of the hydrogels showed an increasing trend ($p < 0.05$) with higher KPS content and a decreasing trend with higher MBA content. Moreover, the gel fraction data showed an increase in gel fraction with higher KPS and MBA content. The swelling and *in vitro* drug release analysis were conducted in pH 1.2 and 7.4 buffer, revealing higher swelling and drug release at pH 7.4, making it suitable for targeted drug delivery. Moreover, the kinetic modelling data of drug release aligned well with the Korsmeyer–Peppas model with R^2 value of 0.9871 at pH 7.4 and 0.9892 at pH 1.2, indicating a Fickian diffusion mechanism ($n<0.5$) at pH 1.2 and a non-Fickian diffusion mechanism ($n>0.5$) at pH 7.4. Therefore, the synthesized ampicillin sodium-loaded Gelatin/PAM/CMTKG hydrogel shows promise for the pH-dependent release of ampicillin sodium.

Acknowledgement

The authors are grateful to the Delhi Technological University, New Delhi, India and the Council of Scientific & Industrial Research (CSIR), India.

References

1 Ghauri Z H, Islam A, Qadir M A, Gull N, Haider B, Khan R U & Riaz T, Development and evaluation of pH-sensitive biodegradable ternary blended hydrogel films (chitosan/guar gum/PVP) for drug delivery application, *Sci Rep*, 11 (2021) 1.

2 Zainal S H, Mohd N H, Suhaili N, Anuar F H, Lazim A M & Othaman R, Preparation of cellulose-based hydrogel: A review, *J Mater Res Technol*, 10 (2021) 935.

3 Maitra J & Shukla V K, Cross-linking in hydrogels: A Review, *Am J Polym Sci*, 4 (2014) 25.

4 Shoulders M D & Raines R T, Collagen structure and stability, *Annu Rev Biochem*, 78 (2009) 929.

5 Karim A A & Bhat R, Gelatin alternatives for the food industry: Recent developments, challenges and prospects, *Trends Food Sci Technol*, 19 (2008) 644.

6 Varghese S A, Rangappa S M, Siengchin S & Parameswaranpillai J, Natural polymers and the hydrogels prepared from them, Elsevier Inc, (2019) 17.

7 Jaipan P, Nguyen A & Narayan R J, Gelatin-based hydrogels for biomedical applications, *MRS Commun*, 7 (2017) 416.

8 Abatangelo G, Vindigni V, Avruscio G, Pandis L & Brun P, Hyaluronic acid: Redefining its role, *Cells*, 9 (2020) 1.

9 Malik N S, Ahmad M, Minhas M U, Tulain R, Barkat K, Khalid I & Khalid Q, Chitosan/xanthan gum based hydrogels as potential carrier for an antiviral drug: Fabrication, characterization, and safety evaluation, *Front Chem*, 8 (2020) 1.

10 Shoukat H, Pervaiz F, Noreen S, Nawaz M, Qaiser R & Anwar M, Fabrication and evaluation studies of novel polyvinylpyrrolidone and 2-acrylamido-2-methylpropane sulphonic acid-based crosslinked matrices for controlled release of acyclovir, *Polym Bull*, 77 (2020) 1869.

11 Godiya C B, Cheng X, Li D, Chen Z & Lu X, Carboxymethyl cellulose/polyacrylamide composite hydrogel for cascaded treatment/reuse of heavy metal ions in wastewater, *J Hazard Mater*, 364 (2019) 28.

12 Zhou C & Wu Q, A novel polyacrylamide nanocomposite hydrogel reinforced with natural chitosan nanofibers, *Colloids Surf B: Biointerfaces*, 84 (2011) 155.

13 Meena P, Singh P & Warkar S G, Fabrication and evaluation of stimuli-sensitive xanthan gum-based hydrogel as a

potential carrier for a hydrophobic drug ibuprofen, *Colloid Polym Sci*, 302 (2023) 377.
14 Meena P, Singh P & Warkar S G, Development and assessment of carboxymethyl tamarind kernel gum-based pH-responsive hydrogel for release of diclofenac sodium, *Eur Polym J*, 197 (2023) 112340.
15 Mali K K, Dhawale S C & Dias R J, Synthesis and characterization of hydrogel films of carboxymethyl tamarind gum using citric acid, *Int J Biol Macromol*, 105 (2017) 463.
16 Pal K, Banthia A K & Majumdar D K, Preparation and characterization of polyvinyl alcohol-gelatin hydrogel membranes for biomedical applications, *AAPS Pharm Sci Tech*, 8 (2007) 21.
17 Treesuppharat W, Rojanapanthu P, Siangsanoh C, Manuspiya H & Ummartyotin S, Synthesis and characterization of bacterial cellulose and gelatin-based hydrogel composites for drug-delivery systems, *Biotechnol Rep*, 15 (2017) 84.
18 Peechakara B V, Basit H & Gupta M, Stat Pearls Publishing, Treasure Island (FL), (2022).
19 Yuan X, Tu Y, Chen W, Xu Z, Wei Y, Qin K, Zhang Q, Xiang Y, Hongcai Zhang & Ji X, Facile synthesis of carbon dots derived from ampicillin sodium for live/dead microbe differentiation, bioimaging and high selectivity detection of 2,4-dinitrophenol and Hg(II), *Dyes Pigm*, 175 (2020) 108187.
20 Anirudhan T S & Mohan A M, Novel pH switchable gelatin based hydrogel for the controlled delivery of the anti cancer drug 5-fluorouracil, *RSC Adv*, 4 (2014) 12109.
21 Wei Q B, Fu F, Zhang Y Q, Wang Q & Ren Y X, PH-responsive CMC/PAM/PVP semi-IPN hydrogels for theophylline drug release, *J Polym Res*, 21 (2014) 453.
22 Rafique N, Ahmad M, Minhas M U, Badshah S F, Malik N S & Khan K, Designing gelatin-based swellable hydrogels system for controlled delivery of salbutamol sulphate: Characterization and toxicity evaluation, *Polym Bull*, 79 (2022) 4535.
23 Li S, Xiaowen Y, Yang Y, Liu L, Sun Y, Liu Y, Yin L & Chen Y, Osteogenic and anti-inflammatory effect of the multifunctional bionic hydrogel scaffold loaded with aspirin and nano-hydroxyapatite, *Front Bioeng Biotechnol*, 11 (2023) 1.
24 Rani I, Warkar S G & Kumar A, Nano ZnO embedded poly (ethylene glycol) diacrylate cross-linked carboxymethyl tamarind kernel gum (CMTKG)/poly (sodium acrylate) composite hydrogels for oral delivery of ciprofloxacin drug and their antibacterial properties, *Mater Today Commun*, 35 (2023) 105635.
25 Kalantari K, Mostafavi E, Saleh B, Soltantabar P & Webster T J, Chitosan/PVA hydrogels incorporated with green synthesized cerium oxide nanoparticles for wound healing applications, *Eur Polym J*, 134 (2020) 109853.
26 Chavda H V & Patel C N, Effects of solvent treatments on characteristics of superporous hydrogels, *Ethiop Pharm J*, 27 (2009) 16.
27 Khan R, Zaman M, Salawi A, Khan M A, Iqbal M O, Riaz R, Ahmed M M, Butt M H, Alvi M N, Almoshari Y & Alshamrani M, Synthesis of chemically cross-linked pH-sensitive hydrogels for the sustained delivery of ezetimibe, *Gels*, 8 (2022) 281.
28 Shahid N, Erum A, Zaman M, Tulain U R, Shoaib Q, Majeed A, Rasool M F, Imran I, Alshehri S, Noorani B & Alqahtani F, pH-responsive nanocomposite based hydrogels for the controlled delivery of ticagrelor; in vitro and in vivo approaches, *Int J Nanomed*, 16 (2021) 6345.
29 Jana S, Sharma R, Maiti S & Sen K K, Interpenetrating hydrogels of O-carboxymethyl Tamarind gum and alginate for monitoring delivery of acyclovir, *Int J Biol Macromol*, 92 (2016) 1034.
30 Ashames A, Ullah K, Al-Tabakha M, Khan S A, Hassan N, Mannan A, Ikram M, Buabeid M & Murtaza G, Development, characterization and In-vitro evaluation of guar gum based new polymeric matrices for controlled delivery using metformin HCl as model drug, *PLOS One*, 17 (2022) e0271623.
31 Sabbagh F & Muhamad I I, Acrylamide-based hydrogel drug delivery systems: Release of acyclovir from MgO nanocomposite hydrogel, *J Taiwan Inst Chem Eng*, 72 (2017) 182.
32 Ghumman S A, Noreen S, Hameed H, Elsherif M A, Shabbir R, Rana M, Junaid K & Bukhari S N A, Synthesis of pH-sensitive cross-linked basil seed gum/acrylic acid hydrogels by free radical copolymerization technique for sustained delivery of captopril, *Gels*, 8 (2022) 291.
33 Suhail M, Khan A, Rosenholm J M, Minhas M U & Wu P C, Fabrication and characterization of diclofenac sodium loaded hydrogels of sodium alginate as sustained release carrier, *Gels*, 7 (2021) 1.
34 Nawaz S, Khan S, Farooq U, Haider M S, Ranjha N M, Rasul A, Nawaz A, Arshad N & Hameed R, Biocompatible hydrogels for the controlled delivery of anti-hypertensive agent: Development, characterization and in vitro evaluation, *Des Monomers Polym*, 21, (2018) 18.
35 Özakar R S & Özakar E, The effect of polymer amount and crosslinker ratio in polymeric hydrogel beads on characterization, *J Res Pharm*, 25 (2021) 653.
36 Trivedi J H, Synthesis, characterization, and swelling behavior of superabsorbent hydrogel from sodium salt of partially carboxymethylated tamarind kernel powder-g-PAN, *J Appl Polym Sci*, 129 (2013) 1992.
37 Suhail M, Hsieh Y H, Khan A, Minhas M U & Wu P C, Preparation and in vitro evaluation of aspartic/alginic acid based semi-interpenetrating network hydrogels for controlled release of ibuprofen, *Gels*, 7 (2021) 68.
38 Goonoo N, Bhaw-Luximon A, Ujoodha R, Jhugroo A, Hulse G K & Jhurry D, Naltrexone: A review of existing sustained drug delivery systems and emerging nano-based systems, *J Control Release*, 183 (2014) 154.
39 Das D, Das R, Mandal J, Ghosh A & Pal S, Dextrin crosslinked with poly(lactic acid): A novel hydrogel for controlled drug release application, *J Appl Polym Sci*, 131 (2014) 1.

J Mater Sci: Mater Electron (2024) 35:1732

Development of flexible piezoelectric nanogenerator based on PVDF/KNN/ZnO nanocomposite film for energy harvesting application

Komal Verma[1], Aman Kumar[1], and Richa Sharma[1,*]

[1] *Department of Applied Physics, Delhi Technological University, Delhi 110042, India*

Received: 29 January 2024
Accepted: 31 July 2024

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

ABSTRACT

For solving energy shortages and environmental problems one of the key strategies is renewable energy generation. Among the various techniques available for energy generation and harvesting, piezoelectric energy harvesting has grabbed significant attention owing to the huge demand for self-powered electronic devices. Herein, we report the Potassium Sodium Niobate (KNN) and Zinc Oxide (ZnO) @ Poly (vinylidene fluoride) (PVDF) nanocomposite film-based flexible piezoelectric nanogenerator (PENG) and studied the effect of KNN and ZnO concentration on the piezoelectric performance of the device. In addition, the real-life application of the device was also investigated. The finding showed that PVDF/KNN/0.45%ZnO nanocomposite film possesses the highest electroactive (β) phase (~59.81%), remnant polarization (P_r) (2.87 μC/cm^2), maximum polarisation (P_s) (32.83 μC/cm^2) and output voltage (39.5 V). The synergistic effect of KNN and ZnO nanoparticles in PVDF polymer matrix considerably improves the resultant piezoelectric performance of the nanogenerator, enabling it to glow four red light-emitting diodes (LEDs) and power a digital wrist watch. Moreover, the developed nanogenerator was also able to generate output voltages of 3.08 V (for pen tapping), 1.72 V (for elbow bending), 1.46 V (for finger moving) and 660 mV (for wrist stretching). Therefore, this fabricated device is a promising candidate for harvesting green and renewable energy and for use in wearable and portable self-powered electronic devices, where a small amount of energy is the basic requirement for their function.

1 Introduction

With the advancement in the microelectronics and the semiconductor industry, the developed functional devices for instance, wearable electronics, the internet of things (IoT), and implantable medical devices have reached every corner and are playing a very important role in our daily life [1, 2]. Presently, for powering these devices, rechargeable batteries are required. However, the limited lifetime and capacity of these batteries has led to the widespread research interest in harvesting energy from renewable sources available in the surrounding environment which can be collected and utilized, such as thermal, solar,

Address correspondence to E-mail: richasharma@dtu.ac.in

https://doi.org/10.1007/s10854-024-13402-w
Published online: 16 September 2024

vibration, wind, etc. It is therefore of great significance to harvest these energies into useful electrical energy to increase the life time of the devices or to even get rid of batteries by developing self-powered systems [3–5]. Among these, harvesting energy from the abundant mechanical vibrations that are present in mechanical and human motions has become a promising and attractive area of research. A piezoelectric nanogenerator (PENG) which is a device capable of converting mechanical energy into useful electrical energy based on Zinc Oxide nanowire arrays was introduced by Wang et al. [6]. In piezoelectric materials, on the application of stress, vibration, tension, and so on, the central symmetry in the crystal structure is disturbed, resulting in the generation of electric potential [7]. The harvested energy can then be used for various purposes such as medical research, consumer devices, industry, defence to name a few [8, 9].

In general, piezoelectric materials can be divided into two categories: organic and inorganic piezoelectric materials. The first category includes poly (vinylidene fluoride) PVDF [10] with its copolymers: poly (vinylidene fluoride-co-hexafluoropropylene) P(VDF-HFP) [11] and (poly (vinylidene fluoride-trifluoroethylene) (P(VDF-TrFE) [12]. The latter category includes $BaTiO_3$ [13], (K, Na) NbO_3 [14–16], ZnO [17–21], PZT [10], etc. Certainly, the types of materials in these two categories have their advantages and disadvantages. Initially, the materials used for nanogenerator applications were inorganic piezoelectric ceramic materials due to their advantages like high sensitivity and electrical activity [22]. However, the inherent defects of these materials such as brittleness, low toughness, and high temperature fabrication processes, have led to the efforts for the development of polymer-based piezoelectric nanogenerators. Among the various available polymers, PVDF and its copolymer have grabbed significant research interest in the field of energy harvesting due to their light weight, easy synthesis technique, flexibility, and long-term stability under high electric field [23, 24]. PVDF is a semicrystalline polymer that has five polymorphs: α, β, γ, δ, and ε [25]. The piezoelectric behaviour of the PVDF is mainly determined by the dipole orientation in the β and γ which possess all T3GT3G′ trans chain conformation and TTTT trans-planar zigzag phases [26, 27]. Therefore, for enhancing the piezoelectric output performance of the PVDF-based PENGs, it is crucial to increase the β or γ phase contents.

One simple and effective way to increase the content of the β phase and improve the energy conversion efficiency of the PVDF-based PENGs is by dispersing fillers or additives in various concentrations in the PVDF matrix. The positive effects of the addition of additives in the PVDF polymer matrix are mainly attributed to the following aspects: (1) The addition of nanofillers can serve as nucleation and crystallization sites and improve the formation of β phase, crystal orientation, and consequently, the piezoelectric response [28]; (2) The addition of piezoelectric filler material in the PVDF composite can improve the piezoelectric properties [29]; (3) The addition of some conducting fillers in the polymer matrix can enhance the nucleation and also helps in transfer of the generated charges [30]. It is the combined effect of the filler materials and polymers in the PVDF-based composites that enhance the piezoelectric performance of the devices for wider applications. Presently, the fillers that have been incorporated in the PVDF matrix are lead-based perovskite ceramics [10], lead-free piezoelectric ceramics [31, 32], carbon-based additives [30, 33], metal oxides [34, 35], and other materials [35, 36].

Among the various fillers used, Potassium Sodium Niobate (KNN) is a promising candidate to be used as a filler in the PVDF matrix owing to its good ferroelectric, piezoelectric, high curie temperature [29]. For instance, Bairagi et al. prepared the piezoelectric nanocomposite film based on KNN nanorods embedded into the PVDF matrix. The PENG with 10 wt% KNN nanorods loading produced an output voltage of 3.4 V and a current of 0.100 μA. The high piezoelectric performance of the prepared device was attributed to the significant enhancement in the fraction of the β phase of the PVDF and its improved piezoelectricity which was strengthened by the addition of KNN nanorods [37]. Abdullah et al. fabricated the KNN- based piezotriboelectric hybrid energy films by incorporating the KNN particles in a PVDF polymer matrix and investigated the piezoelectric response of the synthesized energy film under the varying concentrations of KNN in the energy film and varying tapping frequency [38]. The maximum output voltage and current observed are 35.3 V and 15.8 μA, respectively.

Another nanofiller in the PVDF matrix that has shown promising results is Zinc Oxide (ZnO). Takhur et al. observed an increase in the fraction of β phase up to 84% with the addition of 0.85 vol% ZnO nanoparticles in the PVDF polymer. Ma et al. reported that a single hybrid PVDF/ZnO nanofiber generated

an output voltage which is 300% more than that of a single PVDF nanofiber [39]. Juan Yi et al. [40] fabricated PVDF/Y-doped ZnO electrospun films and investigated the piezoelectric output performance as a function of Y-doped ZnO content in the PVDF. The fabricated device with 15 wt% ZnO generated output voltages of ~20 V and ~35 V when walking or running.

Looking at the promising results demonstrated by the KNN and ZnO-based PENGs, we have fabricated an improved hybrid piezoelectric nanogenerator from three materials: KNN nanoparticles, ZnO nanorods, and PVDF to explore the synergistic effect of KNN and ZnO on the ferroelectric and piezoelectric performance of the KNN/ZnO/PVDF nanogenerator. It is reported in the literature that the addition of two or more nanofillers in the PVDF polymer matrix has resulted in profound influences on the piezoelectric output performance of PENG [41, 42]. In the present research work, we have synthesized six sets of nanocomposite film of PVDF, PVDF/KNN, PVDF/ZnO, and PVDF/KNN/ZnO with varying concentrations of KNN and ZnO in the PVDF matrix by drop cast method. ZnO will act as a nucleating/conducting nanofiller in the prepared nanocomposite film. The PVDF/KNN/ZnO based PENGs demonstrated superior energy harvesting response, and higher content of β phase in PVDF. To obtain the surface morphology of the KNN nanoparticles, ZnO nanorods, and prepared composite films, scanning electron microscopy (SEM) was done. X-ray diffraction and Fourier Infrared Transform Spectroscopy (FTIR) were carried out to analyse the presence of β phase in prepared samples. The ferroelectric properties of the prepared nanocomposite films were also studied. The maximum generated output voltage of 39.5 V was recorded for the PVDF/KNN/ZnO-based PENGs when subjected to force using the dynamic shaker.

2 Experimental details

2.1 Synthesis of KNN nanoparticles

KNN nanoparticles have been synthesized by the hydrothermal method as illustrated in Fig. 1. For the synthesis of KNN nanoparticles, 0.006 mol of Niobium Oxide (Nb_2O_5, Alfa Aesar, 99.5% metal basis) was mixed into a solution of KOH (Thermo Fischer Scientific India Pvt. Ltd. AR, 92%) and NaOH (Avantor Performance Material India Limited, AR, 96%) with a K:Na mol ratio of 1:5 mol by continuous stirring in magnetic stirrer for 2 h. Then, the properly mixed solution

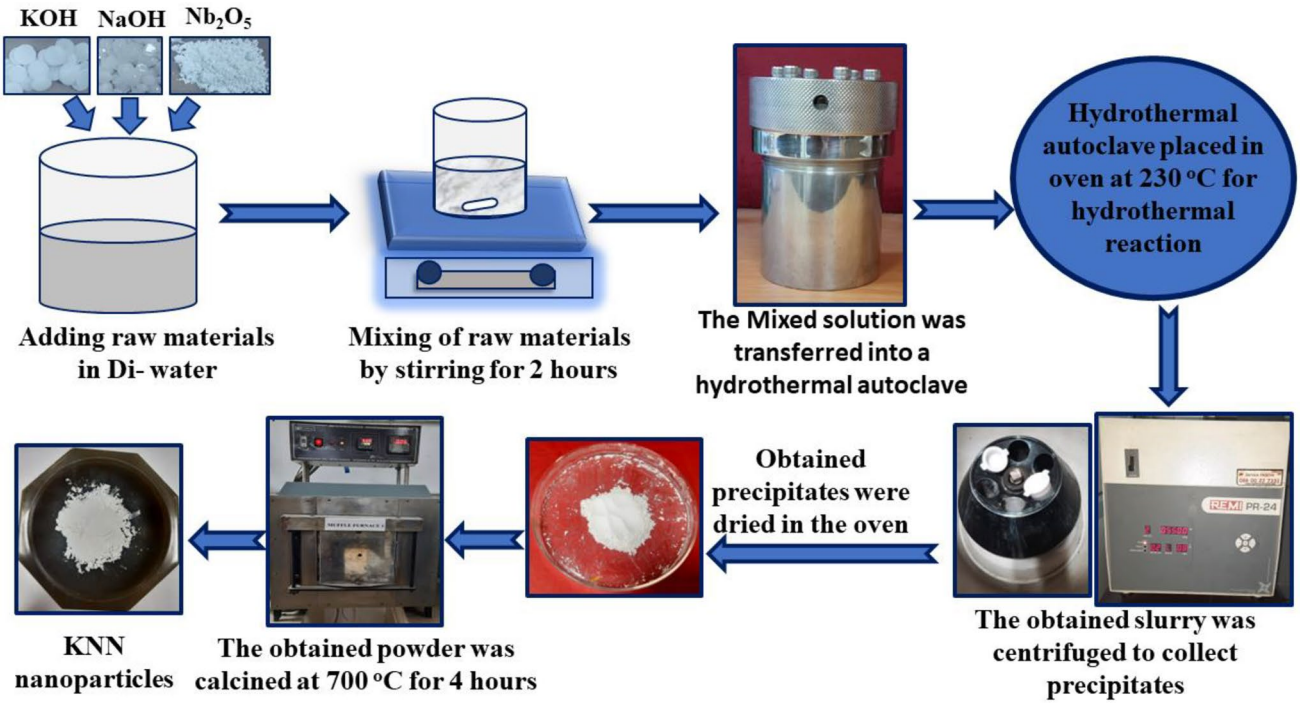


Fig. 1 Synthesis of KNN nanoparticles by hydrothermal method

was poured into a 200 ml Hydrothermal autoclave and kept in a hot air oven at 230 °C for 18 h. Thereafter, the precipitates from the autoclave were collected and washed multiple times with deionized water and ethanol to get rid of the residual alkali followed by drying in oven at 80 °C for 10 h. The obtained dried powder was then calcined at 700 °C for 4 h to obtain the KNN nanoparticles.

2.2 Synthesis of ZnO nanoparticles

The co-precipitation method has been used for the synthesis of ZnO nanoparticles, as presented in Fig. 2. $ZnCl_2$ (Thermo Fischer Scientific India Pvt. Ltd., 96%) was dissolved in 50 ml of deionized water and stirred for 30 min at 40 °C to make a precursor solution. After that, NaOH (Thermo Fischer Scientific India Pvt. Ltd., 92%) solution was prepared with deionized water and then added dropwise to the $ZnCl_2$ solution dropwise while being continuously stirred to keep the reactant's PH level at 12. The resulting precipitates are washed with ethanol and deionized water multiple times, and then dried for 6 h at 80 °C in oven. The resulting powder was then calcined at 500 °C for 2 h in a muffle furnace to obtain the ZnO nanoparticles.

2.3 Fabrication of PVDF/KNN/ZnO flexible nanocomposite films

For the fabrication of PVDF/KNN/ZnO flexible nanocomposite films, Polyvinylidene fluoride (PVDF, Alfa Aesar) is dissolved in N, N-Dimethylformamide (DMF, Fisher Scientific) by using a magnetic stirrer at 50 °C until a uniform transparent solution was obtained. Then KNN and ZnO nanoparticles are added into the DMF in varying wt. % as listed in Table 1 and ultrasonicated for 40 min at room temperature. Finally, the two solutions are mixed by magnetic stirrer for 5 h. The obtained uniform solutions with

Table 1 Details of weight percentages of KNN, ZnO, and PVDF in the fabricated flexible nanocomposite films

S. No	Polymer/Ceramic	Amount (wt %)		
		PVDF	KNN	ZnO
1	Pure PVDF	100	0	0
2	PVDF/KNN/0%ZnO	85	15	0
3	PVDF/KNN/0.45%ZnO	85	14.55	0.45
4	PVDF/KNN/1%ZnO	85	14	1
5	PVDF/KNN/7.5%ZnO	85	7.5	7.5
6	PVDF/KNN/15%ZnO	85	0	15

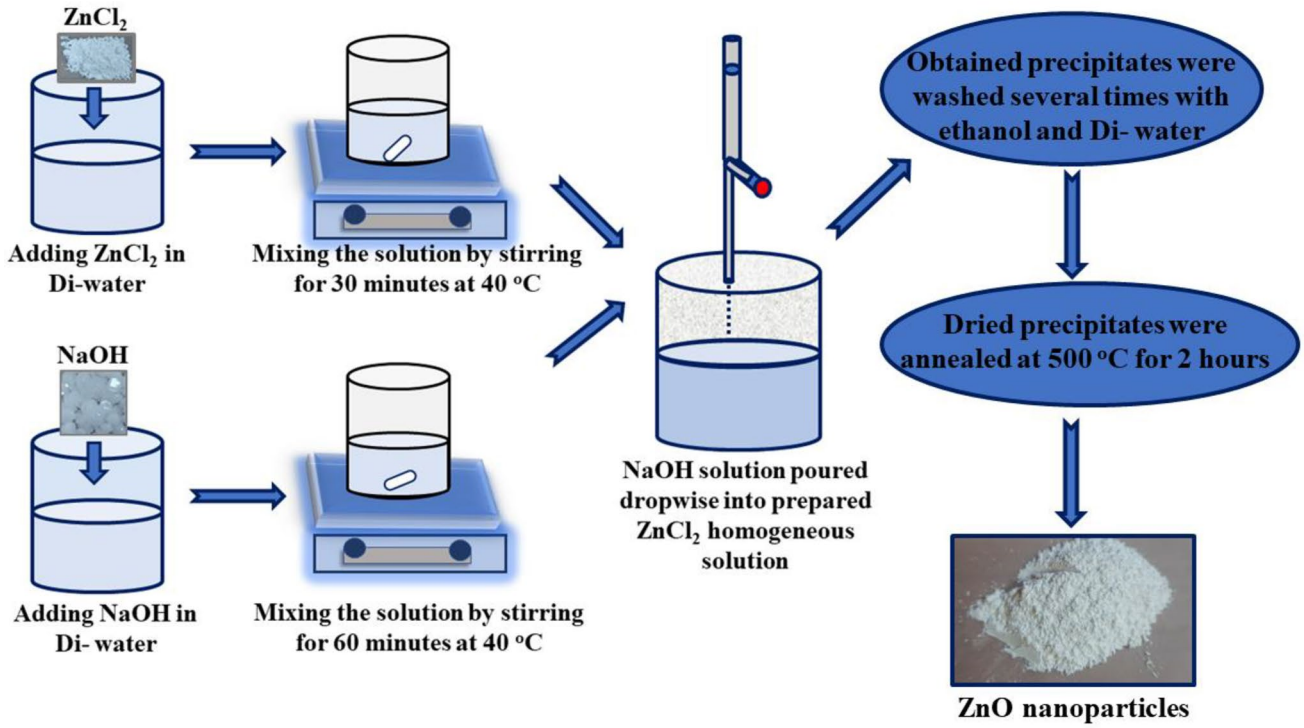


Fig. 2 Synthesis of ZnO nanoparticles by Co-precipitation method

varying KNN/ZnO ratios are poured onto the petri dish and kept in the oven for drying at 80 °C. After this process, a flexible PVDF/KNN/ZnO nanocomposite films was peeled off from the petri dish. For comparative study, films of pure PVDF, PVDF/KNN, and PVDF/ZnO are also prepared using a similar method. The detailed schematic representation of the synthesis process of flexible nanocomposite films is shown in Fig. 3.

2.4 Fabrication of PENG device

A lightweight, flexible, and cost-effective nanogenerator is fabricated by cutting the nanocomposite films into 2×2 cm^2 and then attaching aluminium tape on both sides of the nanocomposite films to make the electrodes. Then copper wires are attached to the aluminium tape followed by wrapping the complete device in the Kapton tape to pack and fix the nanocomposite film. The schematic representation of the fabricated PENG and the photograph of the prepared device is presented in the Fig. 4.

2.5 Characterizations techniques

To analyse the crystalline structure of the prepared KNN nanoparticles, ZnO nanoparticles, and nanocomposite films, X-ray Diffraction (XRD; Rigaku, Ultima-IV) with Cu-Kα radiation source (1.54 Å) was used. Field emission scanning electron microscopy (FE-SEM) (MIRA3LMH, TESCAN instrument) was used for the morphological study of KNN nanoparticles, ZnO nanoparticles, and prepared nanocomposite films. Fourier Transform Infrared spectroscopy (FTIR) (Perkin Elmer FTIR spectrum-II) was used to analyse the formation of the β phase in the nanocomposite films. The polarisation hysteresis loops of the prepared nanocomposite films are recorded at room temperature using the PE loop tracer. The generated output voltages of the fabricated PENG devices on application of force on the surface of the devices using a vibrator shaker are measured by using a digital storage oscilloscope (DSO; Tektronix, MDO500).

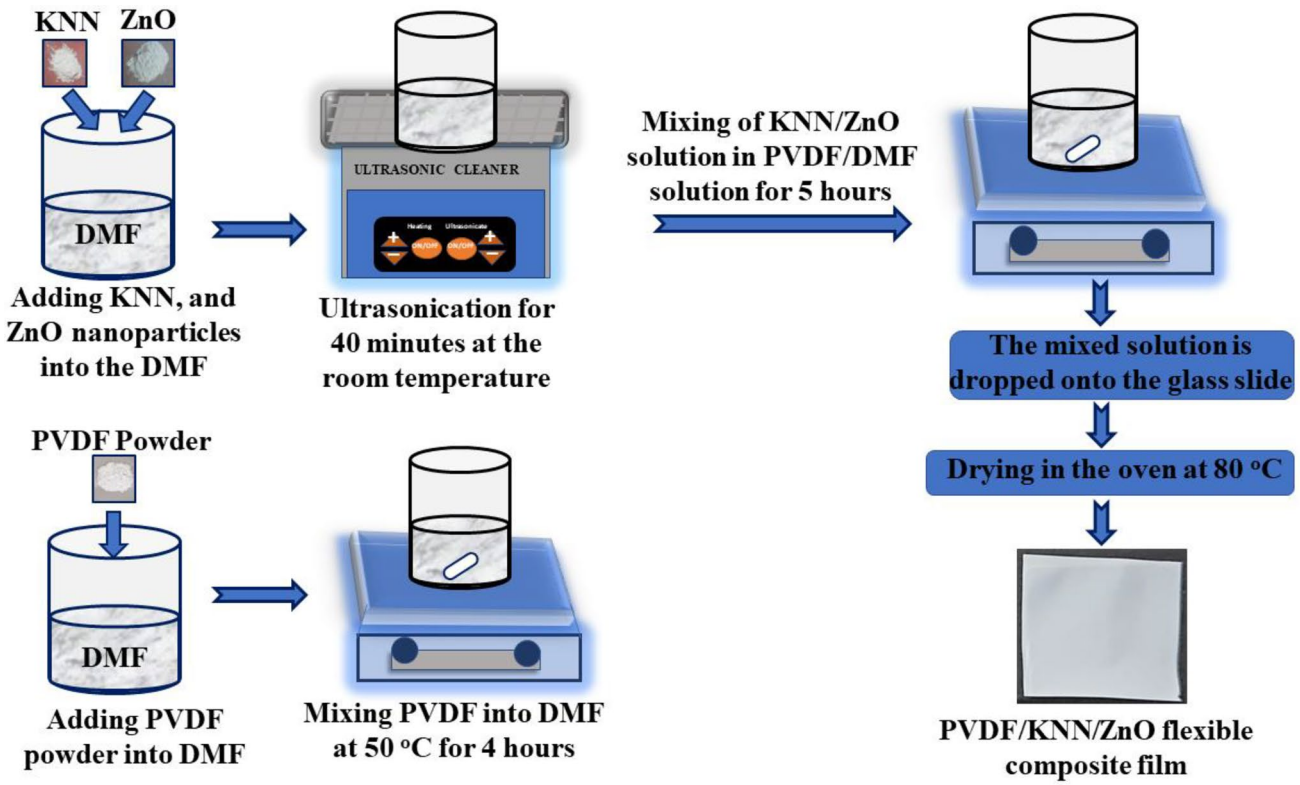


Fig. 3 Schematic representation of the preparation of PVDF/KNN/ZnO flexible nanocomposite films

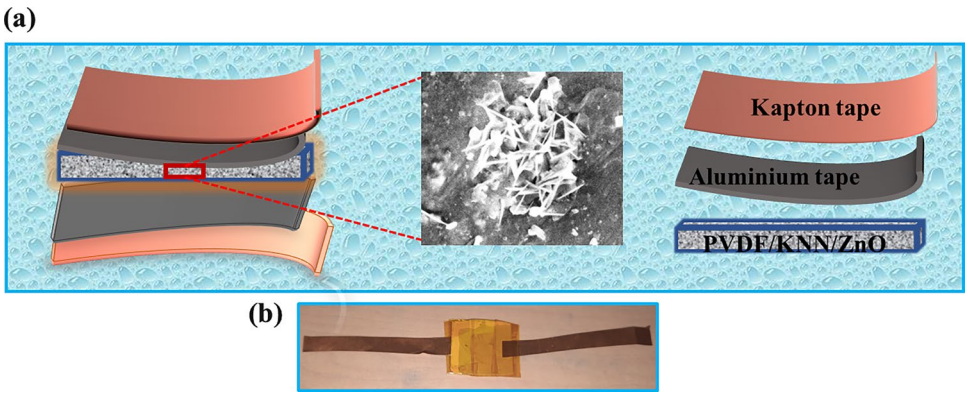


Fig. 4 **a** Schematic diagram of the fabricated PENG device and **b** Photograph of the prepared PENG device

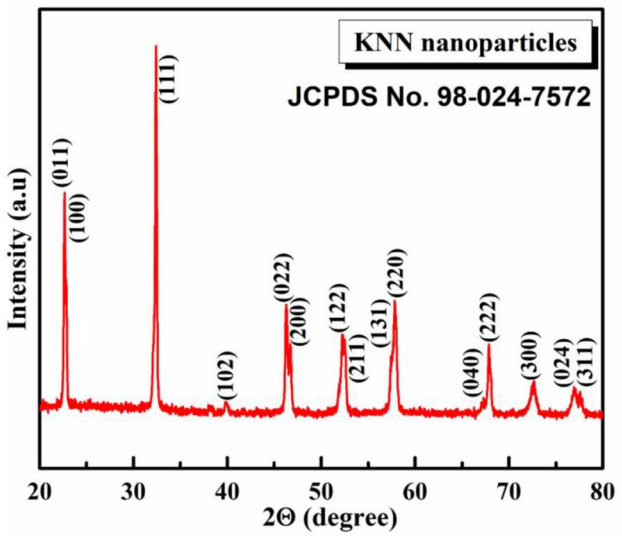


Fig. 5 XRD pattern of the calcined KNN nanoparticles

3 Result and discussion

3.1 Structural and microstructural analysis of KNN nanoparticles

The crystal structure and morphology of the synthesized KNN nanoparticles are investigated by XRD and FESEM as presented in Fig. 5 and Fig. 6, respectively. The XRD pattern shown in Fig. 5 reveals that a pure perovskite structure without the presence of any secondary phase has been successfully synthesized using the hydrothermal method. The formation of the orthorhombic structure for the KNN nanoparticles can be further confirmed by the splitting of the peak at 45°. From the FESEM micrograph presented in Fig. 6 a shows that synthesized KNN consisted of regular cubic nanoparticles. The average particle size

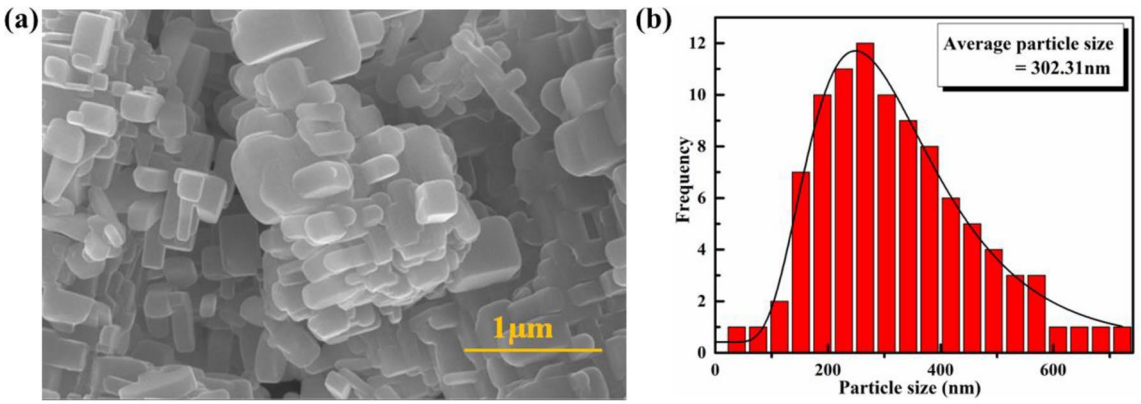


Fig. 6 **a** FESEM image of the calcined KNN nanoparticles and **b** Histogram showing the distribution of particle size

of the KNN nanoparticles is estimated to be ~303 nm as illustrated in Fig. 6 b.

3.2 Structural and microstructural analysis of ZnO nanoparticles

The XRD pattern of the synthesized ZnO nanoparticles is presented in Fig. 7. From the XRD pattern it can be seen that observed diffraction peaks for ZnO nanoparticles correspond to hexagonal wurtzite structure. Furthermore, the morphology and the particle size of the synthesized ZnO nanoparticles are investigated using FESEM and are presented in the Fig. 8a and b, respectively. The FESEM image of the ZnO nanoparticles in Fig. 8a depicts the formation of nanorods with a hexagonal shaped cross-section. The average diameter of the ZnO nanorods is estimated to be ~183 nm.

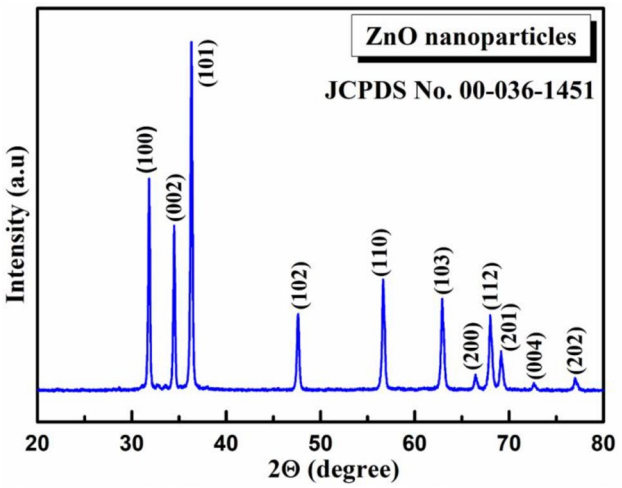


Fig. 7 XRD plot of calcined ZnO nanoparticles

3.3 Structural analysis of the nanocomposite films

3.3.1 XRD analysis of the PVDF/KNN/ZnO flexible nanocomposite films

To analyse the structure of the prepared nanocomposite films, XRD analysis is done and the corresponding plot is depicted in Fig. 9. The peaks which correspond to α and β phases of the PVDF are obtained at 18.6° and 20.2°, respectively in all the prepared films. The intensity of the diffraction peak corresponding to β phase is higher in comparison to α phase, indicating that β phase dominates over the phase in all the prepared nanocomposite films. In addition to this, other peaks that can be seen in the XRD plot are attributed to ZnO and KNN nanoparticles. The signs "♦" and "♠" are used to mark the presence of KNN and ZnO nanoparticle diffraction peaks, respectively in the XRD plot presented in Fig. 9.

3.3.2 FTIR analysis of the nanocomposite films

The FTIR plot of the prepared nanocomposite films is presented in Fig. 10. The presence of α and β phases of the PVDF in the prepared nanocomposite films can be identified by the presence of peaks at 766 cm^{-1} and

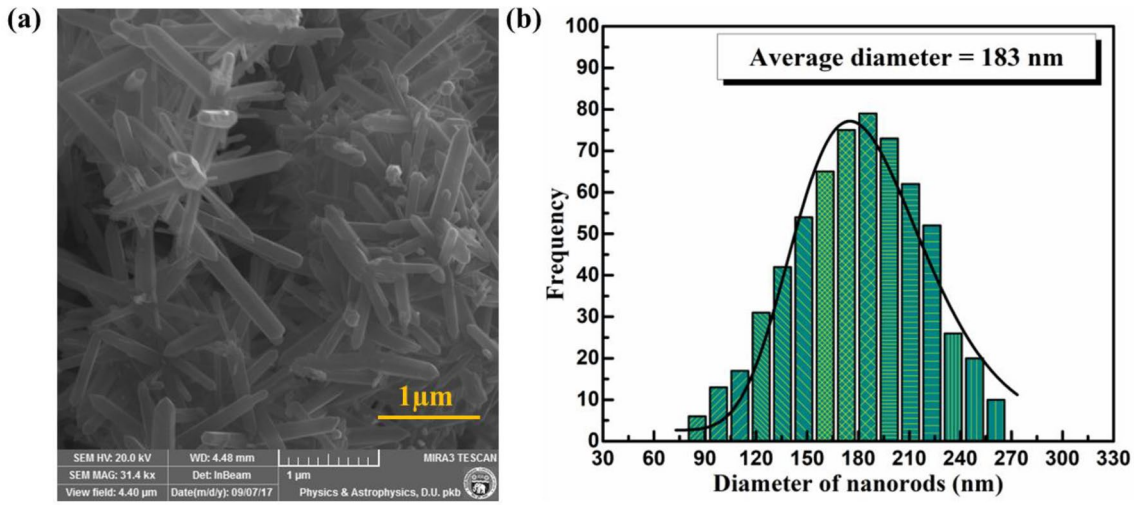


Fig. 8 **a** FE-SEM image of calcined ZnO nanorods and **b** Histogram showing the distribution of diameter of ZnO nanorods

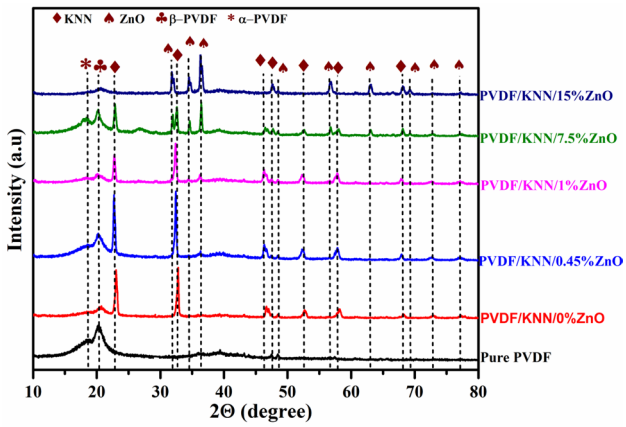


Fig. 9 X-Ray Diffraction patterns of the flexible nanocomposite films produced with different content of KNN and ZnO nanoparticles in the PVDF matrix

840 cm^{-1}, respectively. All the peaks corresponding to α and β phases of the PVDF are marked in the plot shown in Fig. 10a [43–46]. The content of the β phase in the prepared flexible films have been calculated using the Beer-Lambert Law [47] given below:

$$F(\beta) = \frac{A_\beta}{\left(\frac{K_\beta}{K_\alpha}\right)A_\alpha + A_\beta} \quad (1)$$

where $K_\alpha = 6.1 \times 10^4$ cm^2mol^{-1} and $K_\beta = 7.7 \times 10^4$ cm^2mol^{-1} are the absorption coefficient at 840 cm^{-1} and 766 cm^{-1}, respectively; A_α and A_β, are the absorbance intensities at 766 cm^{-1} and 840 cm^{-1}, respectively [48]. The variation of the β phase content in the films as a function of KNN and ZnO nanoparticle concentration is illustrated in Fig. 10c. It can be seen that with the introduction of KNN and ZnO nanoparticles, there is

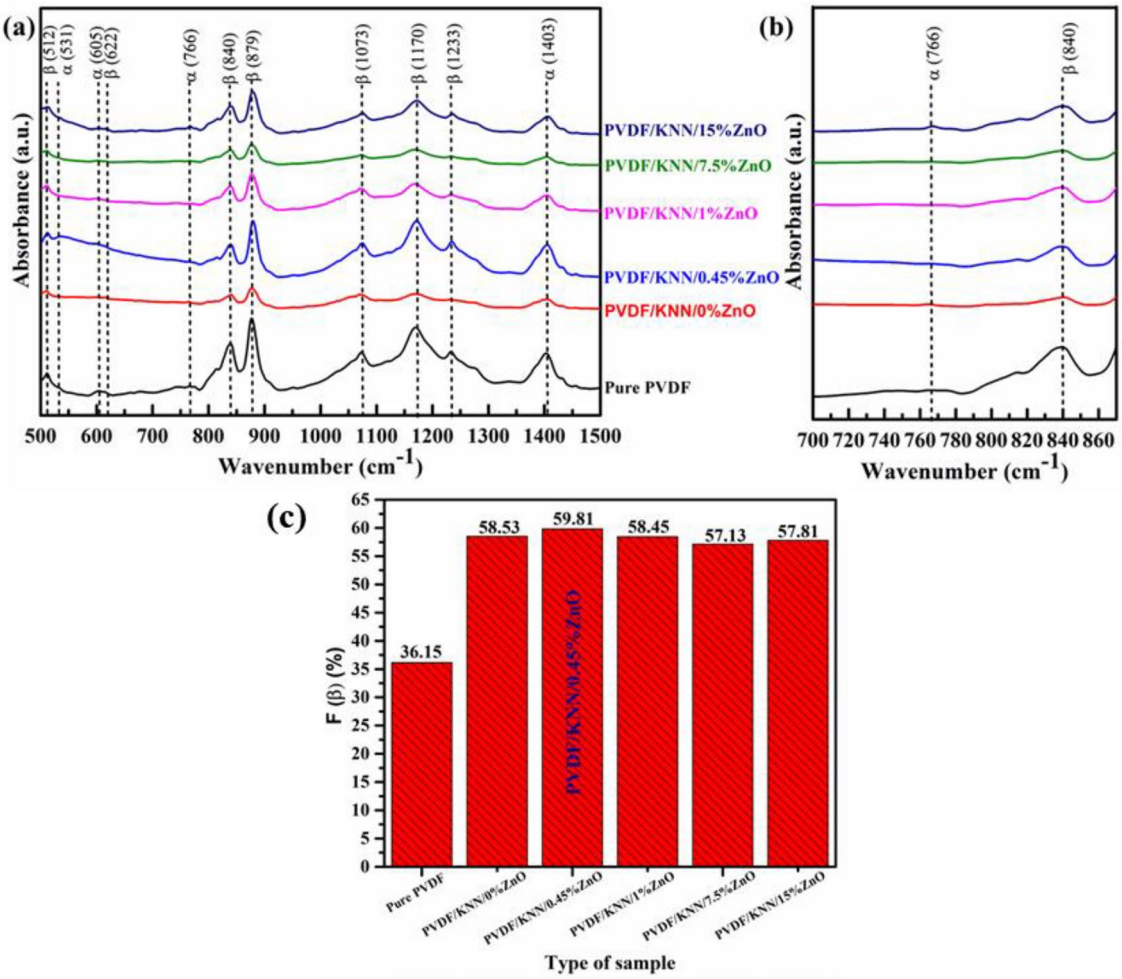


Fig. 10 **a** FTIR spectra of pure PVDF film and nanocomposite films, **b** Enlarged view of the FTIR spectra in the range 700 cm^{-1} to 870 cm^{-1} of Fig. **a** and **c** Variation of β- phase content as a function of KNN and ZnO nanoparticles concentration in the PVDF matrix

an enhancement in the β phase fraction in the nanocomposite films in comparison to the pure PVDF film. Furthermore, the highest β phase content is obtained for the nanocomposite film containing 0.45% ZnO nanorods in the PVDF matrix with further increasing the amount of ZnO nanorods above 0.45%, the fraction of β phase has decreased in the nanocomposite films. Therefore, it is expected that the device fabricated with PVDF/KNN/0.45%ZnO nanocomposite film will demonstrate superior piezoelectric performance in comparison to the other nanocomposite films as discussed in the next sections of the paper. Furthermore, the obtained FTIR results are in agreement with the XRD result confirming the presence of the β phase in the prepared nanocomposite films. Based on XRD and FTIR analysis, it is suggested that the crystallinity of the PVDF have been improved with the simultaneous addition of KNN nanoparticles and ZnO nanorods in the PVDF matrix.

3.4 Surface morphology studies of PVDF/KNN/ZnO flexible nanocomposite films

The surface morphology of fabricated pure PVDF and nanocomposite films are examined using FESEM and corresponding images are presented in Fig. 11. The pure PVDF film has a smooth surface (Fig. 11a). It can be seen from the Fig. 11b–f that KNN and ZnO nanoparticles are distributed uniformly in the PVDF polymer. Slight roughness on the surface of the nanocomposite films can be seen in the films when both KNN and ZnO nanoparticles are added together in the PVDF polymer.

3.5 Polarisation – Electric Field (P-E) hysteresis loops of the nanocomposite films

The room temperature P-E hysteresis plots of the prepared PVDF/KNN/ZnO flexible nanocomposite films are presented in Fig. 12a–f. For all the prepared films, non-saturated hysteresis loops have been observed. It can be seen from the hysteresis loops that films containing KNN and ZnO nanoparticles demonstrated larger ferroelectric behaviour in comparison to the pure PVDF film. Furthermore, the variation of the remnant polarization (P_r) and maximum polarisation (P_s) are also given in Fig. 12g and h, respectively. With the increase in the concentration of ZnO nanoparticles in the nanocomposite films, there is a decrease in both P_r and P_s values. In particular, the maximum values of P_r and P_s obtained are 2.87 μC/cm^2 and 32.83 μC/cm^2 respectively, for the sample containing the highest content of KNN and lowest content of ZnO in the nanocomposite film. This may be attributed to the larger polarizability of KNN in comparison to ZnO and PVDF [32]. The observed increase in P_r will also

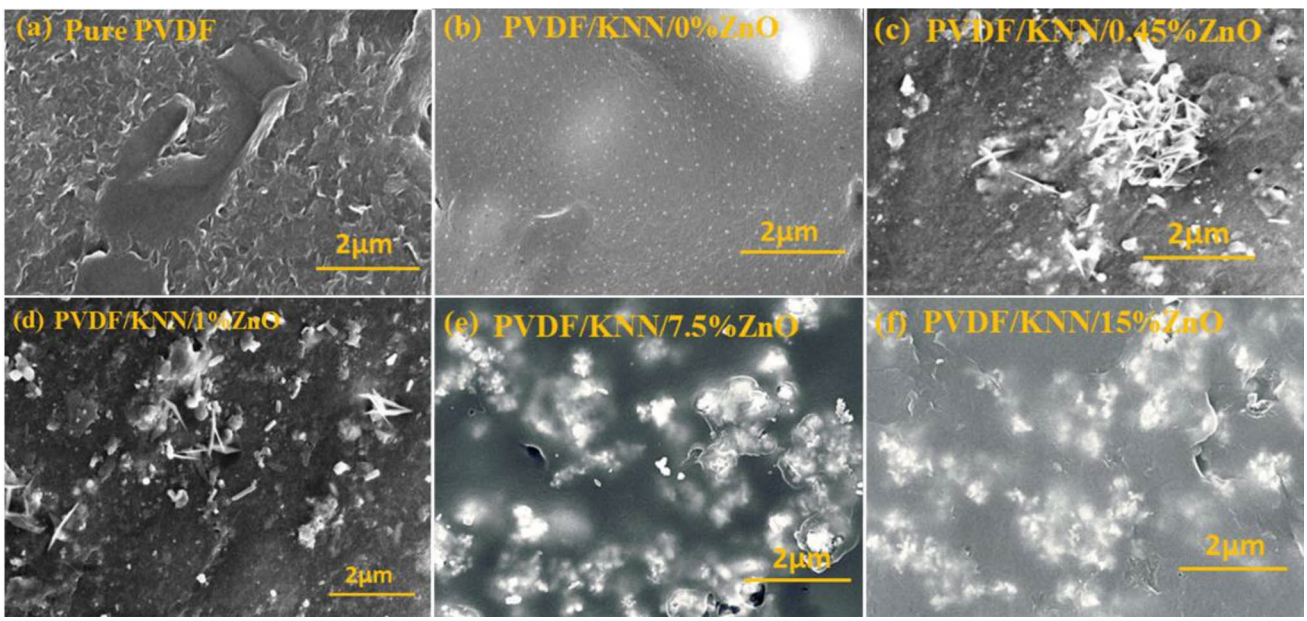


Fig.11 FESEM surface micrographs of **a** Pure PVDF, **b** PVDF/KNN/0%ZnO, **c** PVDF/KNN/0.45%ZnO, **d** PVDF/KNN/1%ZnO, **e** PVDF/KNN/7.5%ZnO, and **f** PVDF/KNN/15%ZnO

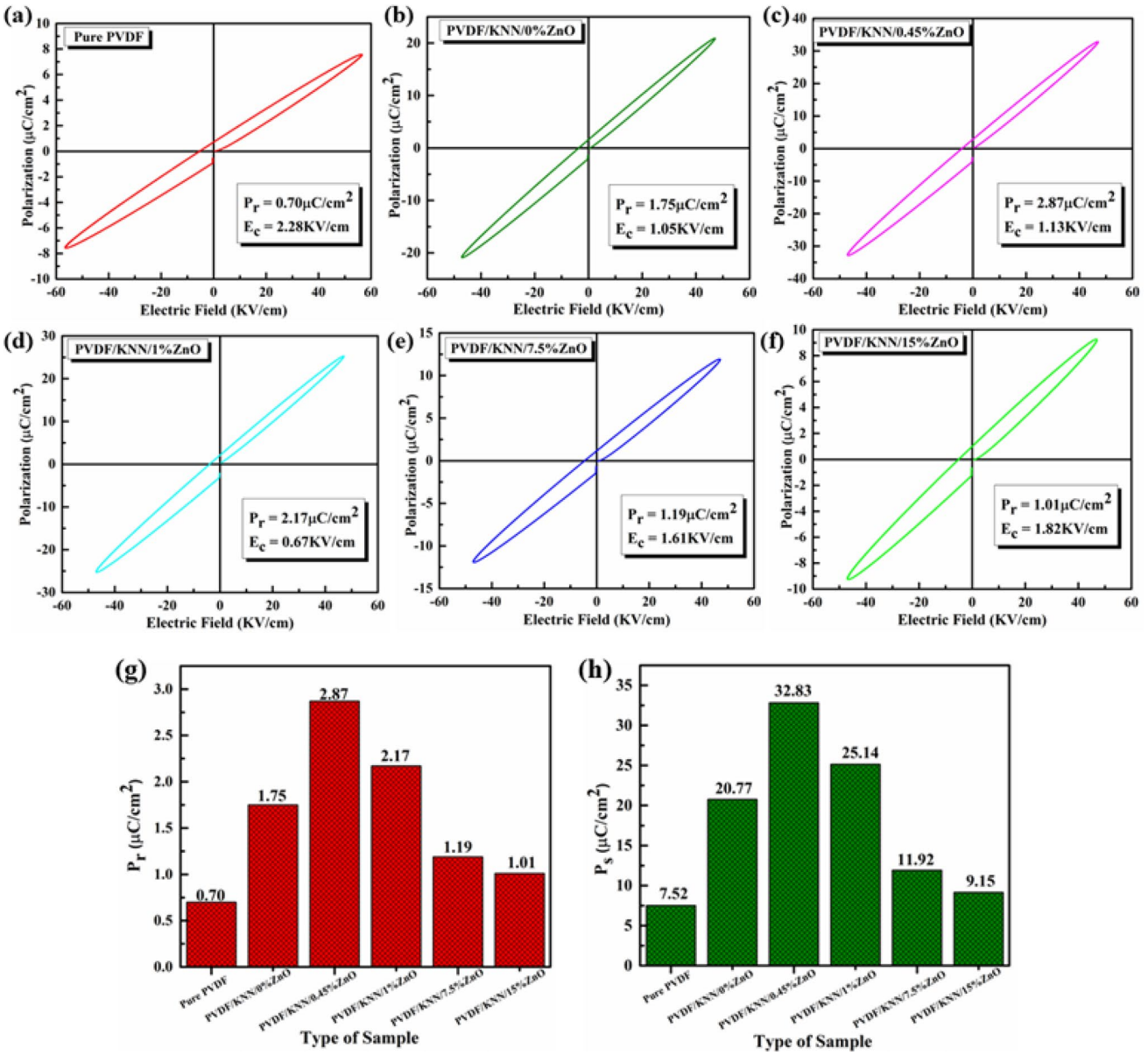


Fig. 12 *P-E* hysteresis Loops of **a** pure PVDF, **b** PVDF/KNN/0%ZnO, **c** PVDF/KNN/0.45%ZnO, **d** PVDF/KNN/1%ZnO, **e** PVDF/KNN/7.5%ZnO, **f** PVDF/KNN/15%ZnO **g** and **h** Variation of P_r and P_s as a function of KNN/ZnO concentration in PVDF matrix

increase of the piezoelectric behaviour of the nanocomposite films [49].

3.6 Piezoelectric performance of PVDF/KNN/ZnO PENG device

To test the piezoelectric performance of the fabricated devices, shaker was used to apply mechanical force on the surface of the PENG. Figure 13a–f depicts the piezoelectric performance of the fabricated PENG with different concentrations of KNN and ZnO in the PVDF matrix on application of repeated force. The peak-to-peak open circuit output voltage obtained for different PENG is presented in Fig. 13g. It can be noticed from the Fig. 14 that the output response for the device PVDF/KNN/0.45%ZnO is enhanced by almost 4.5 times for open circuit voltage in comparison to pure PVDF and 1.5 and 1.33 times in comparison to the device containing only KNN and ZnO in the PVDF matrix, respectively. The obtained better performance of the PVDF/KNN/0.45%ZnO PENG device in comparison

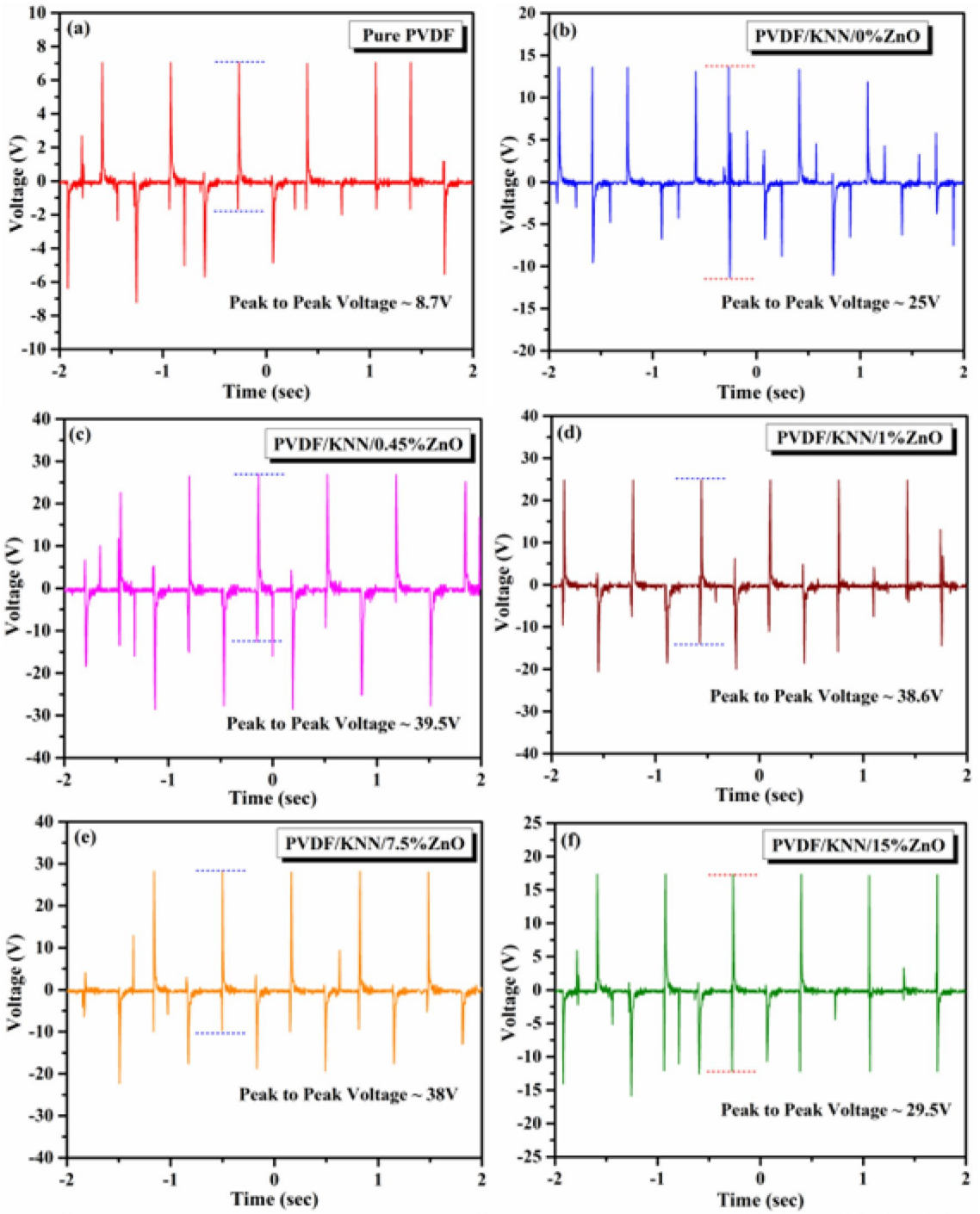


Fig. 13 The generated open circuit piezoelectric output voltage plots of the PENG **a** pure PVDF, **b** PVDF/KNN/0%ZnO, **c** PVDF/KNN/0.45%ZnO, **d** PVDF/KNN/1%ZnO, **e** PVDF/KNN/7.5%ZnO, and **f** PVDF/KNN/15%ZnO

to the other fabricated PENGs is mainly attributed to its higher β phase content and better ferroelectric properties. This generated output voltage using PVDF/KNN/0.45%ZnO PENG was enough to light up four red light emitting diodes (LEDs) and power a digital wrist watch as shown in Fig. 15f, g (Supplementary Video1), (Supplementary Video 2) and the corresponding circuit is shown in Fig. 15e.

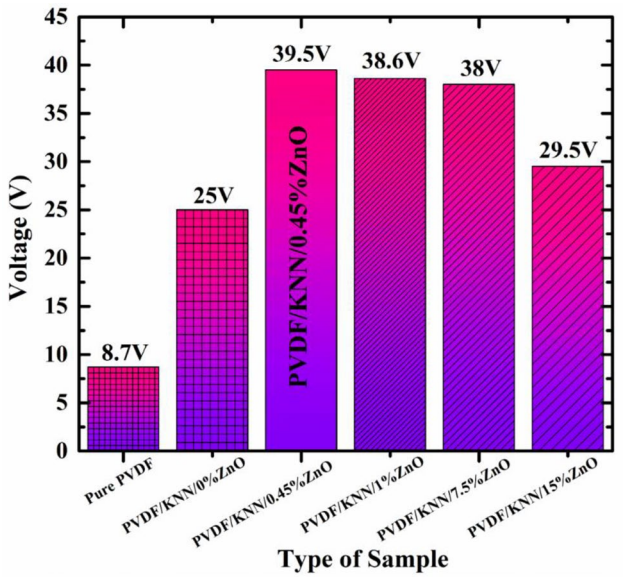


Fig. 14 Variation of output Voltage of the PENG as a function of KNN and ZnO nanoparticles concentration in the PVDF matrix

Further, to explore the mechanical energy which is wasted in our real-life fabricated PENG device with PVDF/KNN/0.45%ZnO nanocomposite film has been attached on the elbow, finger and wrist of the hand for harvesting energy from human body motions. The generated output voltages for pen tapping (Supplementary Video 3), elbow bending (Supplementary Video 4), finger moving (Supplementary Video 5) and wrist stretching (Supplementary Video 6) were 3.080 V, 1.720 V, 1.460 V and 660 mV, respectively. The corresponding output voltage plots are resented in Fig. 15a–d. This shows that the piezoelectric nanogenerator can be utilised as a great solution in numerous applications of flexible wearable and portable small electronic gadgets where, a small amount of power is required for self-operation.

From the obtained data it can be summarized that the PENG fabricated with incorporating KNN and ZnO nanoparticles together in the PVDF matrix has resulted in improved piezoelectric performance in comparison to the PENG fabricated by incorporating only KNN or ZnO in the PVDF matrix. The fabricated device can thus have the potential to be used in applications in which a small amount of energy is needed for self-operation. The comparison of the results obtained in the present work with the reported work in the literature is presented in Table 2.

4 Conclusion

In the present research, lead-free environmentally friendly flexible piezoelectric PVDF/KNN/ZnO nanocomposite films were successfully fabricated by using the drop casting method. The PENG prepared with PVDF/KNN/0.45%ZnO nanocomposite film has showed the excellent piezoelectric performance in terms of output voltage which is attributed to higher β phase formation and better ferroelectric properties in comparison to other prepared PENGs in the present work. The maximum value of the generated output voltage was 39.5 V. Further, to demonstrate the practical applicability of the developed nanogenerator, the performance of the device was examined by powering LEDs, a digital wrist watch, and harvesting energy from human body motion such as elbow bending, finger moving, and wrist stretching. Finally, it can be concluded that the versatile, durable, and flexible PVDF/KNN/0.45%ZnO based PENG studied in the present work demonstrate their excellent potential for future applications in energy harvesting, wearable and portable self-powered electronic devices.

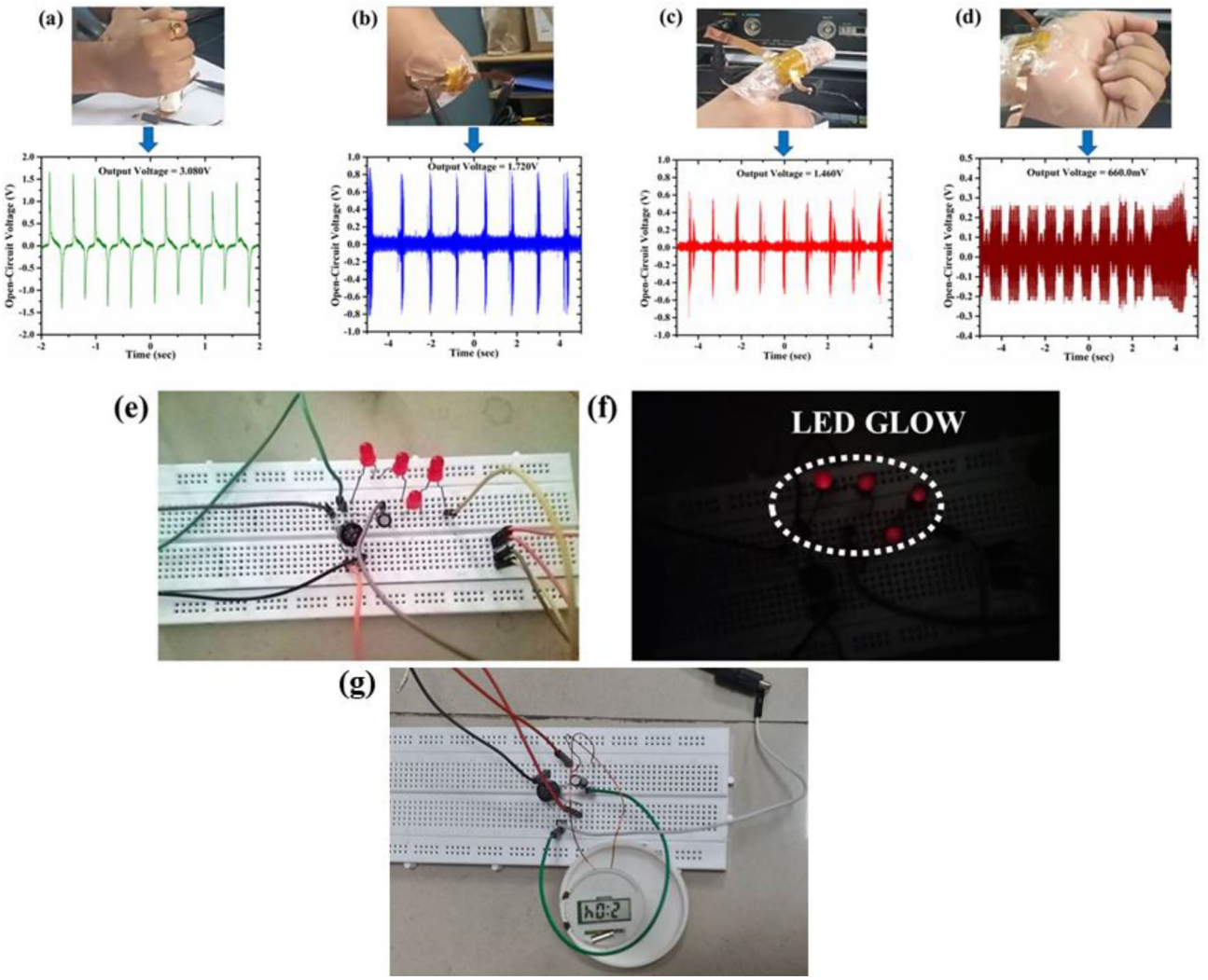


Fig. 15 The generated voltage with PENG fabricated with PVDF/KNN/0.45%ZnO nanocomposite film by **a** pen tapping **b** elbow bending **c** finger moving **d** wrist stretching and **e** circuit with full wave bridge rectifier for glowing LEDs and **f** showing glowing LED **g** power a digital wrist watch

Table 2 Comparison of the piezoelectric performance between PVDF/KNN/ZnO PENG nanogenerator fabricated in the present work and various nanogenerator reported in the literature

S. No	Filler material in the Polymer Matrix	Polymer used	Film Synthesis Technique	Open Circuit Voltage	References
1	KNN nanorods and ZnO nanorods	PVDF	electrospinning	25 V	[50]
2	KNN nanorods	PVDF	Twin screw melt mixing	3.7 V	[51]
3	ZnO nanofibers	PVDF	electrospinning	0.064 V	[29]
4	KNN nanorods and CNT	PVDF	electrospinning	23.24 V	[52]
5	BTO nanoparticles	PVDF	Solvent casting method	7.2 V	[53]
6	ZnO microrods	PVDF	Supersonic spraying	15.2 V	[54]
7	BTO nanoparticles	PVDF	Solvent casting	9.3 V	[55]
8	ZnO nanoparticles and MWCNT	PVDF	Drop casting	22 V	[56]
9	KNN nanofibers	PVDF	Twin screw melt mixing	17.5 V	[57]
10	KNN ceramic particles	PVDF	Drop casting	11.2 V	[16]
11	KNN nanoparticles and ZnO nanorods	PVDF	Drop casting	39.5 V	Present Study

Acknowledgements

The authors wish to thank the DTU Project Research Grant (F.NO.DTU/IRD/619/2019/2113) for providing the research grant to perform the research work. Komal Verma expresses her gratitude to Delhi Technological University (DTU), India for providing the DTU fellowship (2K19/PHDAP/06).

Author contributions

Komal Verma: Conceptualization, Writing-original Draft, methodology, investigation, Formal analysis. Aman Kumar: Data collection, investigation, methodology. Richa Sharma: Conceptualization, writing-review & editing, supervision of the manuscript at different stages, data curation.

Data policy and data availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no known conflict of interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10854-024-13402-w.

References

1. S.M.R. Islam, D. Kwak, M.H. Kabir, M. Hossain, K.-S. Kwak, The internet of things for health care: a comprehensive survey. IEEE Access **3**, 678–708 (2015)
2. A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of things for smart cities. IEEE Internet Things J. **1**, 22–32 (2014)
3. Y. Jie, N. Wang, X. Cao, Y. Xu, T. Li, X. Zhang, Z.L. Wang, Self-powered triboelectric nanosensor with poly(tetrafluoroethylene) nanoparticle arrays for dopamine detection. ACS Nano **9**, 8376–8383 (2015)
4. S. Lee, R. Hinchet, Y. Lee, Y. Yang, Z.H. Lin, G. Ardila, L. Montes, M. Mouis, Z.L. Wang, Ultrathin nanogenerators as self-powered/active skin sensors for tracking eye ball motion. Adv. Funct. Mater. **24**, 1163–1168 (2014)
5. G.-T. Hwang, H. Park, J.-H. Lee, S. Oh, K.-I. Park, M. Byun, H. Park, G. Ahn, C.K. Jeong, K. No, H. Kwon, S.-G. Lee, B. Joung, K.J. Lee, Self-powered cardiac pacemaker enabled by flexible single crystalline PMN-PT piezoelectric energy harvester. Adv. Mater. **26**, 4880–4887 (2014)
6. Z.L. Wang, J. Song, Piezoelectric nanogenerators based on zinc oxide nanowire arrays. Science **312**, 242–246 (2006)

7. H. Jaffe, Piezoelectric ceramics. J. Am. Ceram. Soc. **41**, 494–498 (1958)
8. H.S. Kim, J.H. Kim, J. Kim, A review of piezoelectric energy harvesting based on vibration. Int. J. Precis. Eng. Manuf. **12**, 1129–1141 (2011)
9. M.-G. Kang, W.-S. Jung, C.-Y. Kang, S.-J. Yoon, Recent progress on PZT based piezoelectric energy harvesting technologies. Actuators **5**(1), 5 (2016)
10. L. Paralı, M. Koç, E. Akça, Fabrication and Characterization of High Performance PVDF-based flexible piezoelectric nanogenerators using PMN-xPT (x: 30, 32.5, and 35) particles. Ceram. Int. **49**, 18388–18396 (2023)
11. X. Chen, X. Li, J. Shao, N. An, H. Tian, C. Wang, T. Han, L. Wang, B. Lu, High performance piezoelectric nanogenerators with imprinted P(VDF-TrFE)/BaTiO3 nanocomposite micropillars for self-powered flexible sensors. Small **13**(23), 1604245 (2017)
12. H.M.A. Yaseen, S. Park, P(VDF-TrFE)/BaTiO3 nanocomposite Langmuir-Schaefer thin film for piezoelectric nanogenerator. J. Alloys Compd. **952**, 169940 (2023)
13. F. Zamanpour, L. Shooshtari, M. Gholami, R. Mohammadpour, P. Sasanpour, N. Taghavinia, Transparent and flexible touch on/off switch based on BaTiO3/silicone polymer triboelectric nanogenerator. Nano Energy **103**, 107796 (2022)
14. S. Banerjee, S. Bairagi, S.W. Ali, A lead-free flexible piezoelectric-triboelectric hybrid nanogenerator composed of uniquely designed PVDF/KNN-ZS nanofibrous web. Energy **244**, 123102 (2022)
15. K. Verma, S. Goel, R. Sharma, Influence of calcination and sintering temperature on the microstructure, dielectric, ferroelectric and piezoelectric properties of the lead-free KNN ceramics. J. Mater. Sci. Mater. Electron. **33**, 26067–26085 (2022)
16. K. Verma, R. Sharma, A flexible piezoelectric generator based on KNN/PVDF composite films: role of KNN concentration on the piezoelectric performance of generator. Chin. J. Phys. **84**, 98–215 (2023)
17. M. Kumar, P. Kumari, Design and fabrication of self-powered flexible P(VDF-TrFE)/ZnO/TiO2 fiber mats as nanogenerator for wearable applications. Mater. Sci. Semicond. Process. **160**, 107429 (2023)
18. S. Goel, B. Kumar, A review on piezo-/ferro- electric properties of morphologically diverse ZnO nanostructures. J. Alloys Compd. **816**, 152491 (2020)
19. S. Goel, B. Kumar, Lead-free high T_c ferroelectric material: hierarchial Dy-doped ZnO architectures co- assembled by 1D nanorods and 2D nanosheets. J. Alloys Compd. **801**, 626–639 (2019)
20. N. Sinha, S. Goel, A.J. Joseph, H. Yadav, K. Batra, M.K. Gupta, B. Kumar, Y-doped ZnO nanosheets: gigantic piezoelectric response for an ultra-sensitive flexible piezoelectric nanogenerator. Ceram. Int. **44**, 8582–8590 (2018)
21. S. Goel, N. Sinha, H. Yadav, A.J. Joseph, B. Kumar, Experimental investigation on the structural, dielectric, ferroelectric and piezoelectric properties of La doped ZnO nanoparticles and their application in dye-sensitized solar cells. Physica E: Low-dimens. Syst. Nanostruct. **91**, 72–81 (2017)
22. D. Zhang, X. Zhang, X. Li, H. Wang, X. Sang, G. Zhu, Y. Yeung, Enhanced piezoelectric performance of PVDF/BiCl3/ZnO nanofiber-based piezoelectric nanogenerator. Eur. Polym. J. **166**, 110956 (2022)
23. H. Sun, H. Tian, Y. Yang, D. Xie, Y.-C. Zhang, X. Liu, S. Ma, H.-M. Zhao, T.-L. Ren, A novel flexible nanogenerator made of ZnO nanoparticles and multiwall carbon nanotube. Nanoscale **5**(13), 6117–6123 (2013)
24. S. Egusa, Z. Wang, N. Chocat, Z. Ruff, A. Stolyarov, D. Shemuly, F. Sorin, P. Rakich, J. Joannopoulos, Y. Fink, Multimaterial piezoelectric fibres. Nat. Mater. **9**(8), 643–648 (2010)
25. J. Yan, M. Liu, Y.G. Jeong, W. Kang, L. Li, Y.X. Zhao, N. Deng, B. Cheng, G. Yang, Performance enhancements in poly(vinylidene fluoride)-based piezoelectric nanogenerators for efficient energy harvesting. Nano Energy **56**, 662–692 (2019)
26. G.D. Zhu, Z.G. Zeng, L. Zhang, X.J. Yan, Piezoelectricity in beta-phase PVDF crystals: a molecular simulation study. Comput. Mater. Sci. **44**, 224–229 (2008)
27. X.M. Wang, F.Z. Sun, G.C. Yin, Y.T. Wang, B. Liu, M.D. Dong, Tactile-sensing based on flexible PVDF nanofibers via electrospinning: a review. Sensors **18**, 16 (2018)
28. N. Sezer, M. Koç, A comprehensive review on the state-of-the-art of piezoelectric energy harvesting. Nano Energy **80**, 105567 (2021)
29. S. Bairagi, S.W. Ali, A unique piezoelectric nanogenerator composed of melt-spun PVDF/KNN nanorod-based nanocomposite fibre. Eur. Polym. J. **116**, 554–561 (2019)
30. J. Yang, Y. Zhang, Y. Li, Z. Wang, W. Wang, Q. An, W. Tong, Piezoelectric nanogenerators based on graphene oxide/PVDF electrospun nanofiber with enhanced performances by in-situ reduction. Mater. Today Commun. **26**, 101629 (2021)
31. A. Patra, A. Pal, S. Sen, Polyvinylpyrrolidone modified barium zirconate titanate /polyvinylidene fluoride nanocomposites as self-powered sensor. Ceram. Int. **44**, 11196–11203 (2018)
32. K.S. Nair, H. Varghese, A. Chandran, U.N.S. Hareesh, A. Chouprik, M. Spiridonov, K.P. Surendran, Synthesis of KNN nanoblocks through surfactant-assisted hot injection method and fabrication of flexible piezoelectric

nanogenerator based on KNN-PVDF nanocomposite. Mater. Today Commun. **31**, 103291 (2022)
33. J.H. Eun, S.M. Sung, M.S. Kim, B.K. Choi, J.S. Lee, Effect of MWCNT content on the mechanical and piezoelectric properties of PVDF nanofibers. Mater. Des. **206**, 109785 (2021)
34. M.M. Alam, A. Sultana, D. Mandal, Biomechanical and acoustic energy harvesting from TiO_2 nanoparticle modulated PVDF nanofiber made high performance nanogenerator. ACS Appl. Energy Mater. **1**(7), 3103–3112 (2018)
35. C. Jin, N. Hao, Z. Xu, I. Trase, Y. Nie, L. Dong, A. Closson, Z. Chen, J.X.J. Zhang, Flexible piezoelectric nanogenerators using metal-doped ZnO-PVDF films. Sens. Actuators, A **305**, 111912 (2020)
36. C. Chen, Z. Bai, Y. Cao, M. Dong, K. Jiang, Y. Zhou, Y. Tao, S. Gu, Xu. Jie, X. Yin, W. Xu, Enhanced piezoelectric performance of BiCl3/PVDF nanofibers-based nanogenerators. Compos. Sci. Technol. **192**, 108100 (2020)
37. Y.H. Lee, D.H. Kim, Y. Kim, I. Shabbir, M. Li, K.H. Yoo, T.W. Kim, Significant enhancement of the output voltage of piezoelectric/triboelectric hybrid nanogenerators based on MAPbBr 3 single crystals embedded into a porous PVDF matrix. Nano Energy **102**, 107676 (2022)
38. S. Bairagi, S.W. Ali, Poly (vinylidene fluoride) (PVDF)/ Potassium Sodium Niobate (KNN) nanorods based flexible nanocomposite film: Influence of KNN concentration in the performance of nanogenerator. Org. Electron. **78**, 105547 (2020)
39. A.M. Abdullah, M.U.K. Sadaf, F. Tasnim, H. Vasquez, K. Lozano, M.J. Uddin, KNN based piezo-triboelectric lead-free hybrid energy films. Nano Energy **86**, 106133 (2021)
40. J. Ma, Q. Zhang, K. Lin, L. Zhou, Z. Ni, Piezoelectric and optoelectronic properties of electrospinning hybrid PVDF and ZnO nanofibers. Mater. Res. Exp. **5**, 035057 (2018)
41. J. Yi, Y. Song, Z. Cao, C. Li, C. Xiong, Gram-scale Y-doped ZnO and PVDF electrospun film for piezoelectric nanogenerators. Compos. Sci. Technol. **215**, 109011 (2021)
42. N. Yaacob, S.G. Pei, A.F. Ismail, N. Nazri, L.T. Yogarathinam, B.C. Ng, M.N.Z. Abidin, ZrO2–TiO2 incorporated PVDF dual-layer hollow fiber membrane for oily waste water treatment: effect of air gap. Membranes **10**, 124–142 (2020)
43. Y. Huan, X. Zhang, Y. Jianan Song, T.W. Zhao, G. Zhang, X. Wang, High-performance piezoelectric composite nanogenerator based on Ag/(K, Na)NbO3 heterostructure. Nano Energy **50**, 62–69 (2018)
44. P.K. Sarkar, S. Maji, G.S. Kumar, K.C. Sahoo, D. Mandalb, S. Acharya, Triboelectric generator composed of bulk poly(vinylidene fluoride) and polyethylene polymers for mechanical energy conversion. RSC **6**, 910–917 (2016)
45. H.T.T. Nong, A.N. Nguyen, J. Solard, A. Gomez, S. Mercone, Robust piezoelectric coefficient recovery by nano-inclusions dispersion in un-poled PVDF–Ni0.5Zn0.5Fe2O4 ultra-thin films. Appl. Sci. **12**, 1589 (2022)
46. J.Y. Yin, C. Boaretti, A. Lorenzetti, A. Martucci, M. Roso, M. Modesti, Effects of solvent and electrospinning parameters on the morphology and piezoelectric properties of PVDF nanofibrous membrane. Nanomater. **12**, 962 (2022)
47. Y. Hou, Y. Deng, Y. Wang, H.L. Gao, Uniform distribution of low content BaTiO3 nanoparticles in poly(vinylidene fluoride) nanocomposite: toward high dielectric breakdown strength and energy storage density. RSC Adv. **5**, 72090–72098 (2015)
48. P. Martins, A.C. Lopes, S. Lanceros-Mendez, Electroactive phases of poly(vinylidene fluoride): determination, processing and applications. Prog. Polym. Sci. **39**(4), 683–706 (2014)
49. R.K. Singh, S.W. Lye, J. Miao, Holistic investigation of the electrospinning parameters for high percentage of β-phase in PVDF nanofibers. Polym. **214**, 123366 (2021)
50. V. Jella, S. Ippili, J.H. Eom, J. Choi, S.G. Yoon, Enhanced output performance of a flexible piezoelectric energy harvester based on stable MAPbI3-PVDF composite films. Nano Energy **53**, 46–56 (2018)
51. S. Bairagi, S.Z. Ali, A hybrid piezoelectric nanogenerator comprising of KNN/ZnO nanorods incorporated PVDF electrospun nanocomposite webs. Int. J. Energy Res. **44**, 1–19 (2020)
52. M. Kim, Y.S. Wu, E.C. Kan, J. Fan, Breathable and flexible piezoelectric ZnO @ PVDF fibrous nanogenerator for wearable applications. Polym **10**, 10070745 (2018)
53. S. Bairagi, S.W. Ali, Investigating the role of carbon nanotubes (CNTs) in the piezoelectric performance of PVDF/ KNN- based electrospun nanogenerator. Soft Matter **16**, 4876 (2020)
54. M. Sahu, S. Hajra, K. Lae, P.L. Deepti, K. Mistewicz, H.J. Kim, Piezoelectric nanogenerator based on lead-free flexible pvdf-barium titanate composite films for driving low power electronics. Crystals **11**, 85 (2021)
55. D.D. Ojha, B. Joshi, E. Samuel, A. Khadka, A. Aldalbahi, G. Periyasami, D. Choi, S. An, S.S. Yoon, Int. J. Energy Res. **2023**, 3074782 (2023)
56. Y. Yang, H. Pan, G. Xie, Y. Jiang, C. Chen, Y. Su, Y. Wang, H. Tai, Flexible Piezoelectric Pressure Sensor based on Polydopamine- modified $BaTiO_3$/PVDF composite film for human motion monitoring. Sens. Actuators: A Phys. **301**, 111789 (2019)
57. S. Pratihar, A. Patra, A. Sasmal, S.K. Medda, S. Sen, Enhanced dielectric, ferroelectric, energy storage and

mechanical energy harvesting performance of ZnO-PVDF composite induced by MWCNT as an additive third phase. Soft Matter **17**, 8483–8495 (2021)

58. S. Bairagi, S.W. Ali, Influence of high aspect ratio lead-free piezoelectric fillers in designing flexible fibrous nanogenerators: demonstration of significant high output voltage. Energy Technol. **7**, 1900538 (2019)

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.

Electrical Engineering
https://doi.org/10.1007/s00202-024-02719-8

ORIGINAL PAPER

Development of improved functional neural network based autoregression models for power quality improvement

Alka Singh[1] · **Srishti Singh**[1]

Received: 6 September 2023 / Accepted: 31 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
This paper presents two improved and adaptive models based on functional neural network and Autoregression (FNNAR) analysis. These models have been developed for estimating the fundamental component of nonlinear and varying load current and computing the exact compensation required in a power distribution system. The proposed FNNAR analysis involves two steps: The first step is designed to estimate the fundamental current in terms of polynomial or trigonometric functional expansion terms; while, the second step involves computations based on the weighted sum of the delayed output terms. An activation function is additionally incorporated to account for the nonlinearity and sudden variations of load current. Both the FNNAR models are developed and their parameters computed in an adaptive manner from the input–output data. The simulation results on a single-phase 110 V, 50 Hz system power distribution system are validated by a scaled down experimental model showing hardware results depicting load compensation. Adequate comparison of the two developed models is also discussed in the paper with two advanced variants of conventional algorithms viz. Least means square algorithm and second order generalized integrator based filtering technique.

Keywords Compensation · Distribution network · Functional expansion · Nonlinear system · Power quality

1 Introduction

Power Quality issues have proliferated in the last few decades resulting into search for appropriate solutions. Power factor correction, load balancing, voltage regulation, harmonic mitigation etc. are some important challenges encountering distribution systems [1, 2]. Moreover, nonlinear loads have shown a phenomenal increase in power distribution networks. Such loads inject harmonic pollution in the electric grid and distort the load currents. The estimation of fundamental quantity from these nonlinear distorted currents becomes a challenging exercise especially when the consumer demand varies and is unpredictable. Load compensation involves estimation of the exact compensating current required. Conventional techniques such as LMS algorithm and SOGI based filtering techniques have been utilized for this purpose. However, recent literature studies [3–5] have also focused on improved versions of these conventional techniques such as cascaded non-identical SOGI (CNISOGI), Variable step size modified clipped least mean square (VSSCLMS) [3]. Zero Attracting Normalized LMS [4], quantum calculus based Least Means Forth (Q-LMF) algorithm [5]. All these algorithms aim to extract the fundamental converged weight with least error and in minimum time. However, these improved versions result in increased order and complexity of filter. Moreover, a basic issue with these higher order filters is the need to tune and adjust several parameters for optimum performance. These computations can be significantly improved if good modeling techniques are available and utilized. This paper discusses two such improved FNNAR techniques based on functional expansion and feedback terms which also overcome the limitations of conventional time series and autoregression moving average models (ARMA). Both the techniques are designed, simulated and tested with respect to each other as well as VSSCLMS and CNISOGI algorithms for achieving shunt compensation. The proposed techniques are self-adaptive in nature and require weight initialization once and no additional parameter tuning.

✉ Alka Singh
alkasingh.dr@gmail.com

Srishti Singh
srishhtisingh@gmail.com

[1] Electrical Engineering Department, Delhi Technological University, Delhi, India

Published online: 12 September 2024

Artificial neural networks (ANNs) and functional neural networks (FLNN) have found increased applications in the last decade [6–15]. Paper [6] discusses one such application for wastewater treatment and the authors have integrated ANN and ARMA model to predict multi step ahead prediction of faults. In paper [7] the authors have shown equivalence between the ANNs and linear/ nonlinear ARMA models. The coefficients of the simulated ARMA model were updated using the least squares method. Paper [8] discusses mid-term load forecasting on a power system in terms of daily peak prediction and one-week ahead prediction. On the same lines, paper [9] discusses the energy consumption demand based on a genetic algorithm-based ANN model. This model has been utilized since the conventional methods do not yield good results under temperature inaccuracies in forecasting. Day ahead short term demand and electricity price is predicted in smart grids using ARMA-FLNN network and discussed in [10]. Paper [11] discusses a Hammerstein model with nonlinear function followed by a linear dynamic system. Application of a hybrid model developed using ARMA, particle swarm optimization and support vector machines has been studied to estimate wind speed in wind power production [12]. An interesting application is shown in [13] for the dc arc modeling. In this paper, polynomial models have been designed to calculate the arc voltage and currents [13] since the arc phenomenon is difficult to predict. Newer and recent applications include development of models for machine health monitoring [14] and prediction of stock market trends [15]. A comprehensive review of methods based on optimization and uncertainty in microgrids is listed in [16]. Further, methods used to depict uncertainty in load, markets and generation and its impact on the outputs are detailed in [16, 17]. Thus, literature review mentions a number of methods and conventional techniques for modeling and estimation of different electrical as well as non-electrical quantities.

Several functional and neural network based expansions are reported using Chebyshev polynomials, Laguerre polynomials, Legendre polynomials, trigonometric terms, Spline functions, Bernoulli polynomials [18–23], etc., for grid synchronization and shunt compensation. Legendre polynomials and trigonometric polynomials are simple, accurate and commonly employed FLNN structures [18–25]. Some work is reported on ANN based models for synchronization [24], harmonic estimation [25] and control of Distribution static compensator (DSTATCOM) using back-propagation [26] and single neuron based Adaline model [27]. The literature studies reveal that it is normal for the artificial neural network (ANN) to comprise multiple hidden layers but FLNN is single layered in contrast. These are basically single layered structures as compared to the complex multilayer perceptron model of the ANNs employing several hidden layers. Moreover, the inherent advantages of FLNN involve expanding the input space into higher dimensional space having higher orders without the need of introducing hidden layers. Hence, it is possible to replace complex ANN by FLNN network having nonlinear functional expansion terms introduced in place of hidden layers. Hence, the computational effort also gets reduced exponentially since there is no hidden layer. The performance aspects of these FLNN structures can be further enhanced by considering the output along with a few time delayed terms, which constitutes the FNNAR series. These models can be highly useful in system identification of nonlinear systems. The development of adaptive FNNAR models is the objective of this paper because such models for estimating and computing compensation have not been utilized till date. Some alternatives to FNNAR-based system models include the use of conventional time series models but this gives very slow and unsatisfactory performance.

The focus of this paper is to present a real time application of FNNAR modeling techniques for power quality improvement in power systems. The proposed hybrid models are combinations of functional neural networks and ARMA models and designed to estimate the exact shunt compensator current injection requirement when a single-phase power system feeds varying nonlinear loads. Due to the dynamic variation of loads, the compensation requirement also varies and needs to be suitably predicted. This compensation requirement can then be met using well designed compensators so that effective system performance is ensured with good power quality and reliability. Most of the models developed in literature focus only on simulation and estimating quantities in the near future. This paper verifies the developed models on a test bench in the laboratory. The main highlights of the paper are:

(1) Development of two new and improved FNNAR models viz. trigonometric FNNAR and polynomial FNNAR for the estimation of fundamental component of nonlinear load current.
(2) Adaptive training of the developed models using the Least Means Square (LMS) technique.
(3) Application of the developed models for mitigation of power quality problems and achieving load compensation.
(4) Suitable comparison of both the developed models wrt new algorithms such as VSSLMS and CNISOGI to investigate the steady state and dynamic behavior of the system.

This paper is divided into six Sections. Section-2 discusses the configuration of the system, Section-3 discusses the mathematical analysis of the proposed FNNAR techniques and stability analysis. Section-4 and Section-5 deal with simulation and experimental results, respectively, and Sect. 6 lists the main conclusions of the paper.

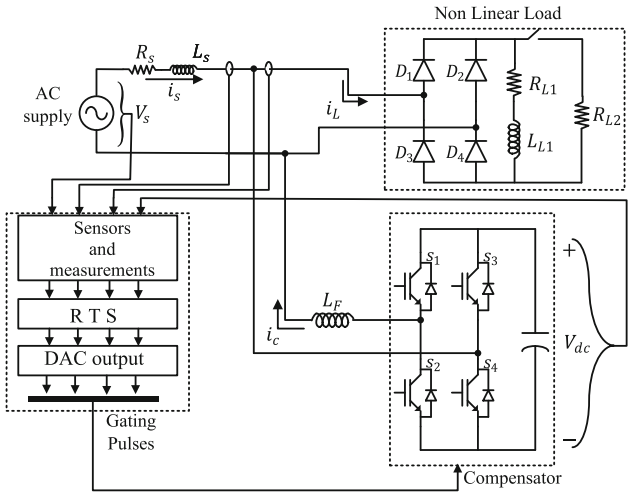


Fig. 1 Schematic diagram of single-phase grid connected system with shunt compensator

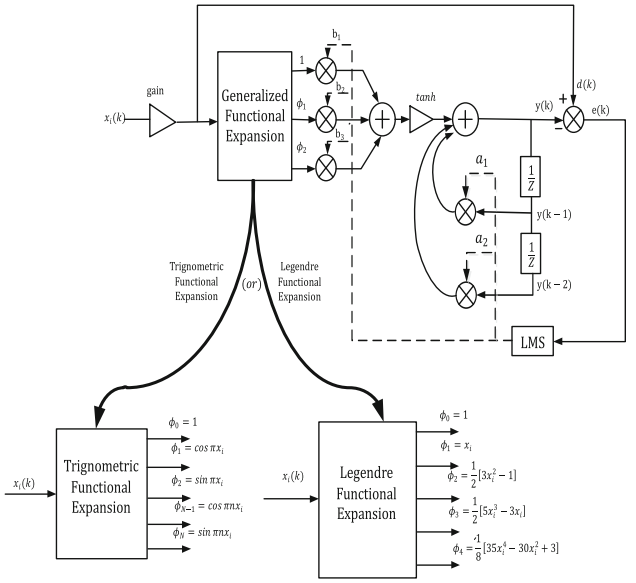


Fig. 2 Schematic diagram of two proposed improved FNNAR models for compensation

2 System configuration

Figure 1 shows a single-phase grid connected system (110 V, 50 Hz) feeding nonlinear loads. This type of load is depicted by a resistive-inductive (RL) branch connected at the end of diode rectifier ($R_{\mathrm{L1}} = 30\Omega$, $L_{\mathrm{L1}} = 80\mathrm{mH}$). All the parameters are listed in the Appendix. The problem with such loads is the injection of harmonics back to the grid. Further, the grid supply feeds multiple loads at the point of common coupling (PCC). However, the PCC voltage itself gets distorted and affects the performance of these multiple loads interconnected there. Several other power quality problems like poor power factor, reactive power burden can be mitigated by connecting a compensator (H bridge converter) as shown in Fig. 1. This figure also shows the measurement and sensing unit, real time simulator (RTS) and digital to analog (DAC) conversion blocks required for experimental setup. The system shown in Fig. 1 is also tested experimentally under load variations for validation of improved models.

3 Mathematical analysis

Figure 2 explains the improved FNNAR models for estimating the fundamental current component ($y(k)$) from the load current input ($x_{\mathrm{i}}(k)$) for the kth sample. The novelty of the paper is that both the models are self-adaptive in nature. The input is first normalized using the gain block and passed through the functional expansion block. The first part of the FNNAR models comprises functional neural network expansion terms represented as $\emptyset_i$ which are multiplied with the coefficients b_{i}. This paper considers two functional expansions viz. trigonometric expansion as well as Legendre polynomials developed due to their special properties and ease of implementation on hardware setup. The details of both these are discussed below.

The developed FNNAR models also include a second part involving terms based on the feedback of the output ($y(k)$) sampled after a unit delay (1/z). These are multiplied with coefficients a_{i}. It can be noted that coefficients b_{i} and a_{i} are unknown and need to be estimated. Once the models are developed and suitably trained, all the coefficients (b_1, b_2, b_3….) and (a_1, a_2,….)as shown in Fig. 2 get estimated online.

A real time updation of the coefficients (b_1, b_2, b_3….) and (a_1, a_2,….) is carried out using the LMS technique shown as a block in Fig. 2. The estimated coefficients adjust adaptively as per the varied conditions. As soon as the input varies, the model coefficients also get updated online and the new fundamental current component $y(k)$ is estimated.

The developed models are useful to predict the load compensation requirement. It is also important to highlight that only a few initial terms for model estimation are required for both the FNNAR models. The online estimation based on LMS block shown in Fig. 2 works on the principle of evaluating the error $e(k)$ from the obtained peak quantity ($d(k)$) and the quantity $y(k)$ estimated from the model itself. The weights are updated iteratively and the mathematical details are discussed below.

3.1 Development of FNNAR models

Statistical models based on ARMA comprise the benefits of auto-regressive (AR) parameters and moving average (MA) parameters. A combined auto-regressive moving average

(ARMA) model of order p,m is written as

$$y(k) = \sum_{i=1}^{p}(a_i y(k-i)) + \sum_{i=1}^{m}(b_i x(k-i)) \tag{1}$$

where $x(k)$ and $y(k)$ denote the input and output of the model, respectively, a_i, b_i are the coefficients which need to be initialized only once. Thereafter, these coefficients get self-adjusted during real time implementation. Equation (1) represents a general ARMA technique. This paper considers two improved FNNAR models (Legendre as well as trigonometric) developed by the suitable modification of Eq. (1). Further, since the input load current (x) is nonlinear in nature, a suitable activation function (tanh) is added and it is rewritten as Eq. (2). The improved and developed FNNAR analysis involve expansion in the form of functional expansion terms. Two FNNAR functions considered in this paper are highly efficient; and hence, the expansion of the proposed models are represented as

$$y(k) = \sum_{i=1}^{p}(a_i y(k-i)) + \tanh\sum_{i=1}^{m}\left(b_i \emptyset_i(x(k-i))\right) \tag{2}$$

where $y(k)$ denotes the output, a_i, b_i are the coefficients and functional neural network based functional expansion is represented by $\emptyset_i$. This function can be expressed either on the basis of trigonometric expansion model or Legendre polynomial FNNAR model, both these models are now described mathematically in Eq. (3) and Eq. (4).

The first FNNAR model is based and developed using trigonometric expansion terms. For this trigonometric FNNAR model, the functional expansion is written as

$$\emptyset_i = x_i;\ \cos(\pi x_i);\ \sin(\pi x_i);\ \cos(2\pi x_i);\ \sin(2\pi x_i)..p\text{terms} \tag{3}$$

where the p terms are cosine/sine terms of different orders: $\cos(p\pi x_i)$, $\sin(p\pi x_i)$etc.

The second FNNAR series is based on Legendre polynomials and involves the expansion terms as shown in Eq. (4)

$$\emptyset_i = 1;\ x_i;\ 0.5\left(3x_i^2 - 1\right);\ 0.5\left(5x_i^3 - 3x_i\right);\ \ldots\ p\text{terms} \tag{4}$$

The general expansion term (P_{n+1}) using Legendre polynomials of nth order can be estimated using the previous two terms (P_n and P_{n-1}) as shown in recursive relation (5).

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \tag{5}$$

In addition, the first two Legendre polynomials are $P_o = 1$ and $P_1 = x_i$. The other expansion terms shown in Eq. (4) can be suitably derived using Eq. (5). Both the FNNAR models are now made adaptive in nature using the LMS algorithm. The estimated quantity y_{est} is first computed using two series as shown in Eq. (1). These equations also show the delayed output terms obtained by using z transform. Combining this with the output y(k) gives the following

$$y_{est1} = y(k) - \sum_{i=1}^{p}(a_i y(k-i)) = [1 - A(k)]y(k) \tag{6}$$

where $A(k) = \sum_{i=1}^{p}\left(a_i z^{-1}\right)$.

Another contribution to the FNNAR models is from the functional neural network series which is

$$y_{est2} = \sum_{k=1}^{n} w(k)\varphi(k) \tag{7}$$

where weights $w(k)$ are coefficients b_i's and $\emptyset_k$ denote the significant terms of trigonometric model and Legendre polynomial FNNAR model as shown in Eq. (3) and (4).

$$y_{est2} = \tanh\sum_{i=1}^{m}\left(b_i \emptyset_i(x(k-i))\right) \tag{8}$$

$$\text{Thus, } y_{est} = y_{est1} + y_{est2} \tag{9}$$

Equation (9) denotes the output from the developed model. The developed FNNAR models can be used to estimate the output using the simplified expression below.

$$y_{est} = w^T(k)i(k) \tag{10}$$

where the net weight matrix to be updated for both the series can be represented as

$$w(k) = [a_1, a_2, a_3, \ldots . b_1, b_2, b_3 \ldots .]$$

and $i(k) = \left[y_{k-1}, y_{k-2}, y_{k-3}, \ldots . \emptyset_1, \emptyset_2, \emptyset_3 \ldots .\right]$ denotes the net input matrix.

Starting with random initial weights, the weight matrix is updated using the Least Means Square technique as shown in Fig. 2 and updated using Eq. (11)

$$w(k+1) = w(k) + \mu e(k) sin\theta \tag{11}$$

where $e(k) = y(k) - d(k)$ denotes the error between the actual and the estimated quantity, μ denotes a fixed constant for the step-size used for LMS updation and sinθ is the unit template obtained from the voltage waveform. The weight coefficients (a_i, b_i) get converged in real time within a few cycles only. The LMS technique is self-adaptive in nature and updates the weight coefficients iteratively.

3.2 Generation of pulses for control of h bridge converter

If v_p denotes the in-phase supply voltage, V_m the peak value then for single-phase supply, it can be written

$$v_p = V_m \sin\theta \tag{12}$$

The quadrature 90° phase shifted voltage component, v_q is

$$v_q = V_m \cos\theta \tag{13}$$

The value of in-phase template (sinθ) can be computed using

$$\sin\theta = \frac{v_p}{\sqrt{v_p^2 + v_q^2}} \tag{14}$$

The DC link voltage of the compensator needs to be regulated using a Proportional-Integral (PI) controller to estimate the switching loss component, y_{loss}.

$$v_{de}(k) = V_{dc}^*(k) - V_{dc}(k) \tag{15}$$

$$y_{loss}(k) = y_{loss}(k-1) + K_p\,[v_{de}(k) - v_{de}(k-1)] + K_i\,[v_{de}(k)] \tag{16}$$

where K_p and K_i denote the proportional and integral gain constants for the PI controller and V_{dc}*, V_{dc} denote the reference and actual dc link voltage and v_{de} the error. The reference source current (i_s*) is obtained by taking the product of unit template (sinθ) with the total fundamental weight component (y_{tot}) computed using the switching losses (y_{loss}) as

$$y_{tot} = y_{est} + y_{loss} \tag{17}$$

$$i_s^* = y_{tot} * \sin\theta \tag{18}$$

This reference supply current ($i_s{}^*$) is compared with the actual supply current (i_s) over a hysteresis controller so as to generate the necessary gating pulses for the compensator.

3.3 Stability analysis

The stability analysis of the technique is now discussed using the Lyapunov criteria. Taking the Lyapunov function $L(k)$ in Eq. (19) to be square of error 'e' at the kth instant

$$L(k) = e^2(k) \tag{19}$$

$$\dot{L} = \Delta L(k) = e^2(k) - e^2(k-1) \tag{20}$$

where $e(k) = y_{act} - y_{est}$ and $y_{est} = w^T(k)i(k)$ as used in Eq. (10–11) and rate of change of $L(k)$ is determined in Eq. (20).

$$i(k) = \left[y_{k-1},\, y_{k-2},\, y_{k-3},\, \ldots . \emptyset_1,\, \emptyset_2,\, \emptyset_3 \ldots .\right]$$

$$\Delta L(k) = [y_{act} - w^T(k)i(k)]^2 - e^2(k-1) \tag{21}$$

Using the weight updation Eq (11) and rewriting Eq. (21) gives

$$\Delta L(k) = [y_{act} - w(k)\,i(k) - \mu e(k)\sin\theta\, w(k)]^2 - e^2(k-1) \tag{22}$$

$$\Delta L(k) = [e(k) - \mu e(k)\sin\theta\, w(k)]^2 - e^2(k-1) \tag{23}$$

$$\Delta L(k) = e^2(k)[1 - \mu\sin\theta\, w(k)]^2 - e^2(k-1) \tag{24}$$

Equation (24) can be simplified to Eq. (25) assuming μ is quite small and also μ^2(being small constant) can be suitably neglected. Hence, it can be inferred that

$$\Delta L(k) \sim -2e^2(k)\mu\sin\theta \tag{25}$$

This proves the stability of the weight updation technique used for developing FNNAR models.

4 Simulation results

Figures 3 and 4 present the simulation results for shunt compensation using trigonometric-FNNAR and Legendre FNNAR. Figure 3a shows the subplots of the supply voltage (v_s), grid current (i_s), load current (i_l), compensator current (i_c) and the DC link voltage (v_{dc}).The results have been simulated using MATLAB Simulink (2020) for 110 V, 50 Hz system under load variations- two load increments introduced at $t = 1.2$ s and $t = 1.3$ s and two load decrements introduced at $t = 1.4$ s and $t = 1.5$ s. The FNNAR based controller for the H bridge is designed to mitigate PQ problems. The first FNNAR model is based on trigonometric terms and it estimates the fundamental component from the nonlinear load current and then computes the compensator current based on the load profile. This computed current is injected at the common coupling point from the shunt compensator. The result shown in Fig. 3a points out that the grid current is sinusoidal and perfectly in phase with the supply voltage ensuring power factor correction is achieved.

The load current is observed to be quite distorted due to the presence of nonlinear load. The trigonometric FNNAR

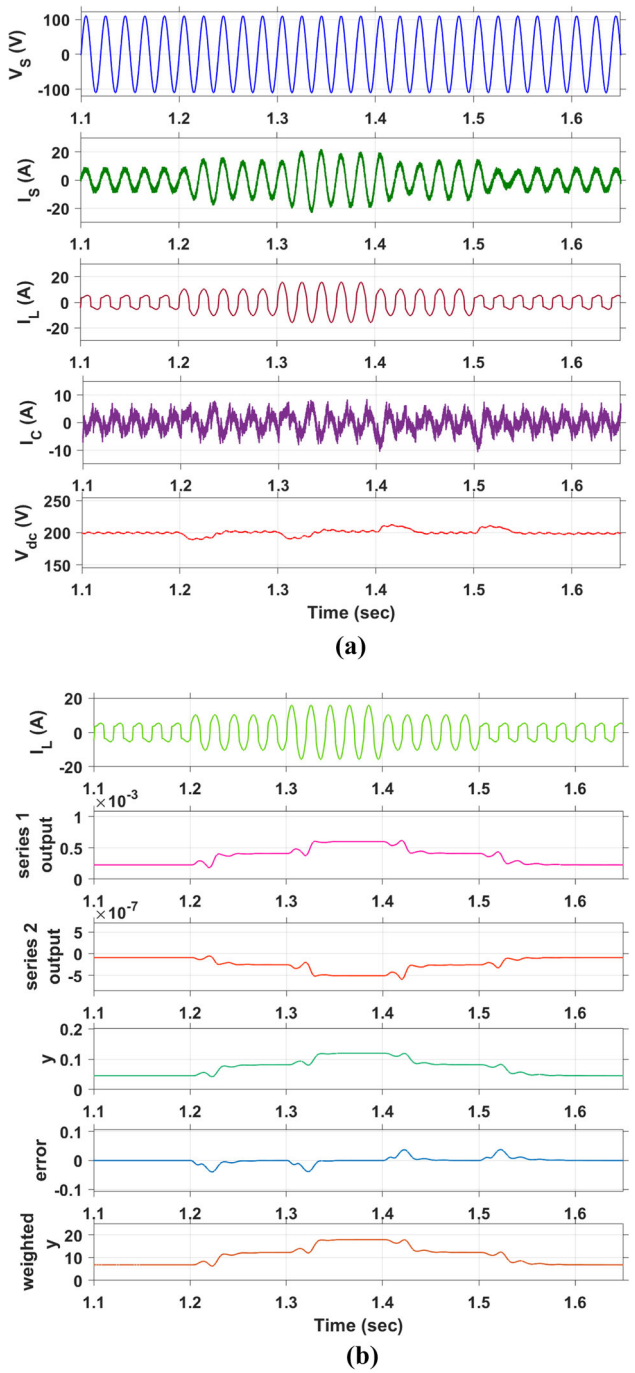


Fig. 3 **a** Simulation results with trigonometric- FNNAR for achieving compensation. **b** Intermediate signals for load compensation using the trigonometric- FNNAR model

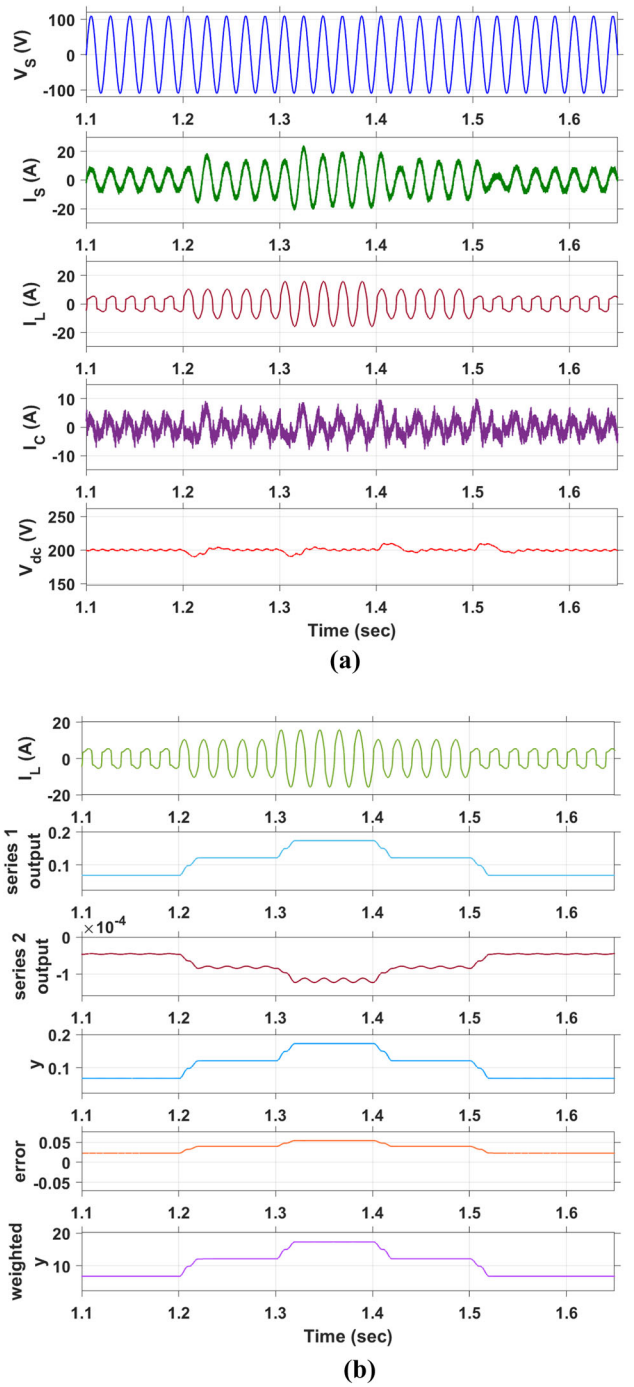


Fig. 4 **a** Simulation results with Legendre- FNNAR for achieving compensation. **b** Intermediate signals using the Legendre- FNNAR model for compensation

model correctly estimates the compensation requirement which is shown in Fig. 3b as weighted 'y' output and ensures improved power quality.

It is also observed from Fig. 3a that the DC link voltage settles down to 200 V during steady state as well as during dynamic load changes introduced at $t = 1.2$ s, 1.3 s,1.4 s and 1.5 s, respectively. Figure 3b shows intermediate results with the developed trigonometric FNNAR where the plots of both the series have been shown along with the estimated weight (y) denoting the fundamental component of load current.

Further, the error between the estimated fundamental quantity and load current is shown which is quite close to

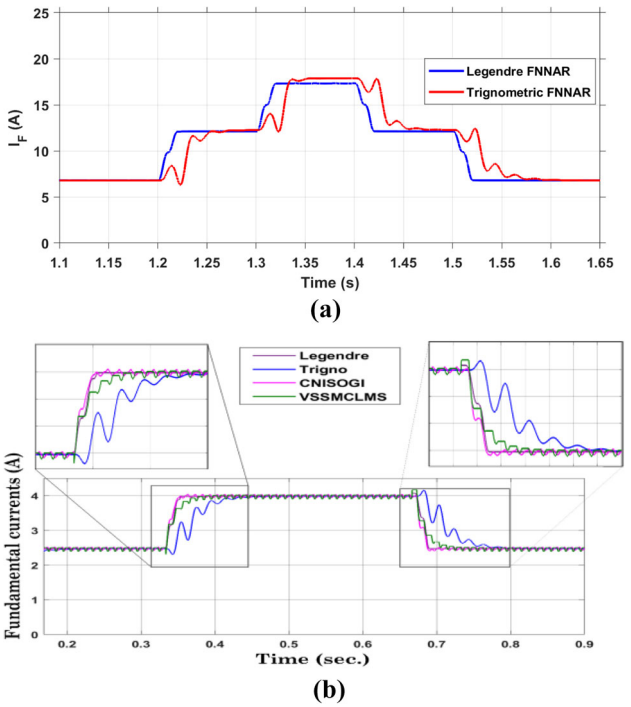


Fig. 5 **a** Results showing the fundamental load current extracted using both the developed FNNAR models. **b** Results of proposed FNNAR models vs. VSSLMS and CNISOGI algorithms

zero. The scaled estimated fundamental quantity is plotted which shows correct value.

Figure 4a shows the results with Legendre functional expansion-based FNNAR. Similar plots have been shown with this developed model viz. supply voltage (v_s), grid current (i_s), load current (i_l), compensator current (i_c) and the DC link voltage (v_{dc}) under the same load variations. The results are recorded under similar load dynamics.

The Legendre FNNAR is also suitably trained using LMS to estimate compensation and improve power quality of the distribution system. The second model is also successful in making the grid current free from harmonics and it is in-phase with the supply voltage. The DC link voltage also settles down to the set reference value at 200 V. Figure 4b shows the intermediate results with developed Legendre FNNAR series. The figure shows subplots showing the variation of output of both the series and the fundamental obtained and scaled down weight component (y), error as well as the properly scaled weighted output. Comparison of simulation results with different algorithms is shown in Fig. 5.

The results shown in Fig. 3 and Fig. 4 clearly show the feasibility of both the proposed FNNAR models in achieving shunt compensation.

The two developed models are now compared under the same set of dynamic load variations as shown in Fig. 5a and Fig. 5b. Figure 5a shows the results when load is incremented at $t = 1.2$ s and $t = 1.3$ s, and decremented at $t = 1.4$ s, and $t = 1.5$ s and its effect is observed on the two developed models. The dynamics of load changes are carried out after time interval of 1 s. A faster convergence within 0.02 s is achieved with Legendre FNNAR model; while, a slower convergence within 0.04 s is obtained with trigonometric FNNAR model. Moreover, the Legendre FNNAR tracking is superior as compared to that of trigonometric FNNAR. It can be noted from Fig. 5a,b that while both the proposed models show tracking in a ripple free manner, the convergence of weight achieved with trigonometric FNNAR shows some undershoot and overshoot and delay before reaching the steady state position. A similar response under load increment as well as load decrement is observed with both the models.

Fast convergence with Legendre expansion-based FNNAR as compared to trigonometric FNNAR is also expected due to more mathematical operations involved in evaluating trigonometric (sine or cosine) functions. Figure 5b shows that load current plotted against the fundamental weight tracked using Legendre FNNAR model with respect to two advanced algorithms VSSCLMS and CNISOGI. It is observed from Fig. 5b that VSSCLMS and CNISOGI algorithms result in producing fundamental current (i_f) with relatively high ripple content. The same i_f tracked under identical load current change at $t = 1.0$ s is almost ripple free as obtained with Legendre FNNAR model. Thus, there is a need of additional filters when VSSCLMS and CNISOGI techniques are used for the extraction of load current and used further for mitigation of power quality issues. Figure 6–7 show the detailed steady state results.

Figure 6 and 7 show the total harmonic distortion (THD) results with both the developed FNNAR models. The difference in the two developed models is predominantly in terms of dynamic response; while, the steady state response of both the models is quite similar and both the models are able to extract the same weight component. The THD values for the supply voltage (v_s), supply current (i_s), load current (i_l) and compensator current (i_c) developed using Legendre model and trigonometric FNNAR model are shown in Fig. 6 and Fig. 7, respectively.

The results shown in Fig. 6 and Fig. 7 point out that for the load current THD of 31.73%, both the developed models are effective in achieving compensation. The THD of the source current reduces to 3.23% with Legendre FNNAR model while it is 3.33% with trigonometric FNNAR model. The supply voltage in both the cases is close to 0.07% indicating an undistorted grid voltage. The compensating current shows high THD of ~ 72–74%. The comparison of Fig. 6 and Fig. 7 also highlights a slight superiority of FNNAR model based on Legendre functions over the trigonometric expansion in terms of steady state response.

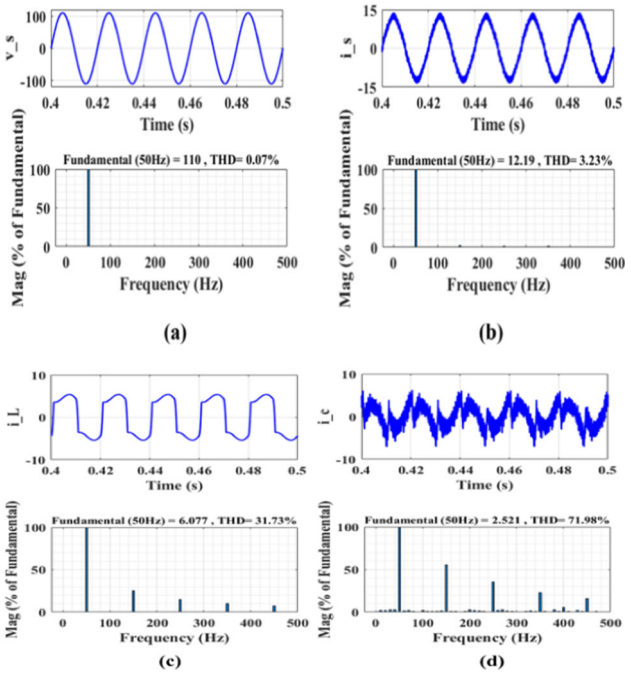


Fig. 6 Waveforms and THD values of (**a**) v_s (**b**) i_s(**c**) i_l and (**d**) i_c using Legendre- FNNAR model

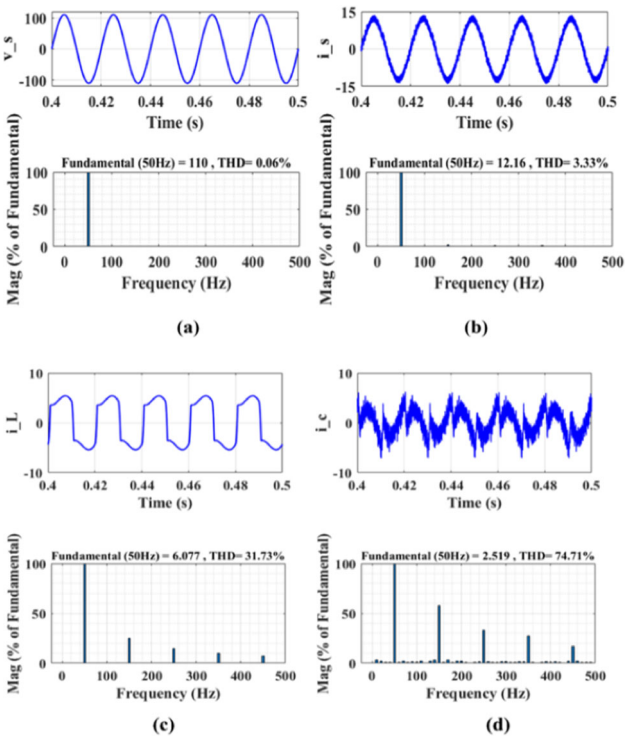


Fig. 7 Waveforms and THD values of (**a**) v_s (**b**) i_s (**c**) i_l and (**d**) i_c using trigonometric- FNNAR model

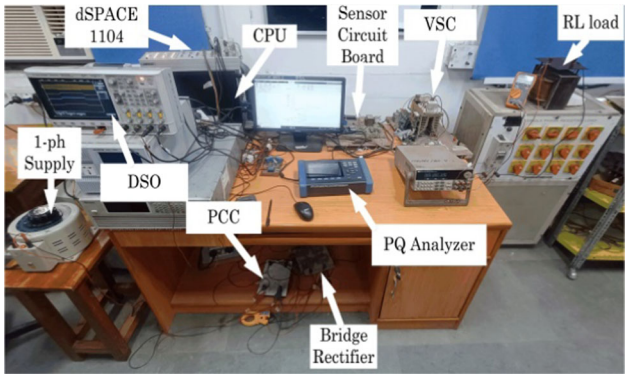


Fig. 8 Photograph of the experimental setup

5 Experimental results

An experimental system of scaled down model is shown in Fig. 8 and it is developed to test the validity of the developed models.

The scaled down system includes a single-phase supply, a diode bridge rectifier-based nonlinear load and a H bridge compensator shown as a voltage source converter (VSC). The parameter values are mentioned in Appendix.

The voltage signals (v_s, v_{dc}) and current signals (i_s, i_l and i_c) are captured using LEM Hall effect sensors (LA-25, LV25). The developed controller is tested on a real time simulator using dSpace1104 and the hardware results are captured on the prototype model using a digital storage oscilloscope (DSO). Figure 9a shows the plot of supply voltage (v_s) and supply current (i_s) which can be observed to be in-phase. The supply current is nearly sinusoidal although the load current (i_l) is quite distorted as shown in Fig. 9b. The compensator current shown in Fig. 9c is extremely distorted as it is required to nullify the harmonics in the supply current. The exact computations are done using the developed FNNAR models. The THD values of the three currents (i_s, i_l and i_c) are 4.2%, 32.28% and 95.85% when the THD of the supply voltage is nearly 1.56%. These observed values are in line with the expected simulation values.

The real and reactive powers for load ($P = 74.7$W, $Q = 33.4$var), supply ($P = 85.4$W, $Q = 22$var) and compensator ($P = 16.4$, $Q = 35$var) are shown in Fig. 9g–i. The active power losses correspond to ~ 16.4W due to the switching of the H bridge compensator and the supply meets a total of 85.4W fulfilling the load demand of 74.7 W and switching loss also. Figure 9h also shows that the load power factor of 0.9128 is improved to 0.9684 on the supply side due to the application of the compensator. Thus, the improvement of power factor as well as sinusoidal waveform of supply current can be easily observed in the hardware results with developed Legendre FNNAR in spite of distorted load current. Figure 10a,b show the dynamic results using simulation

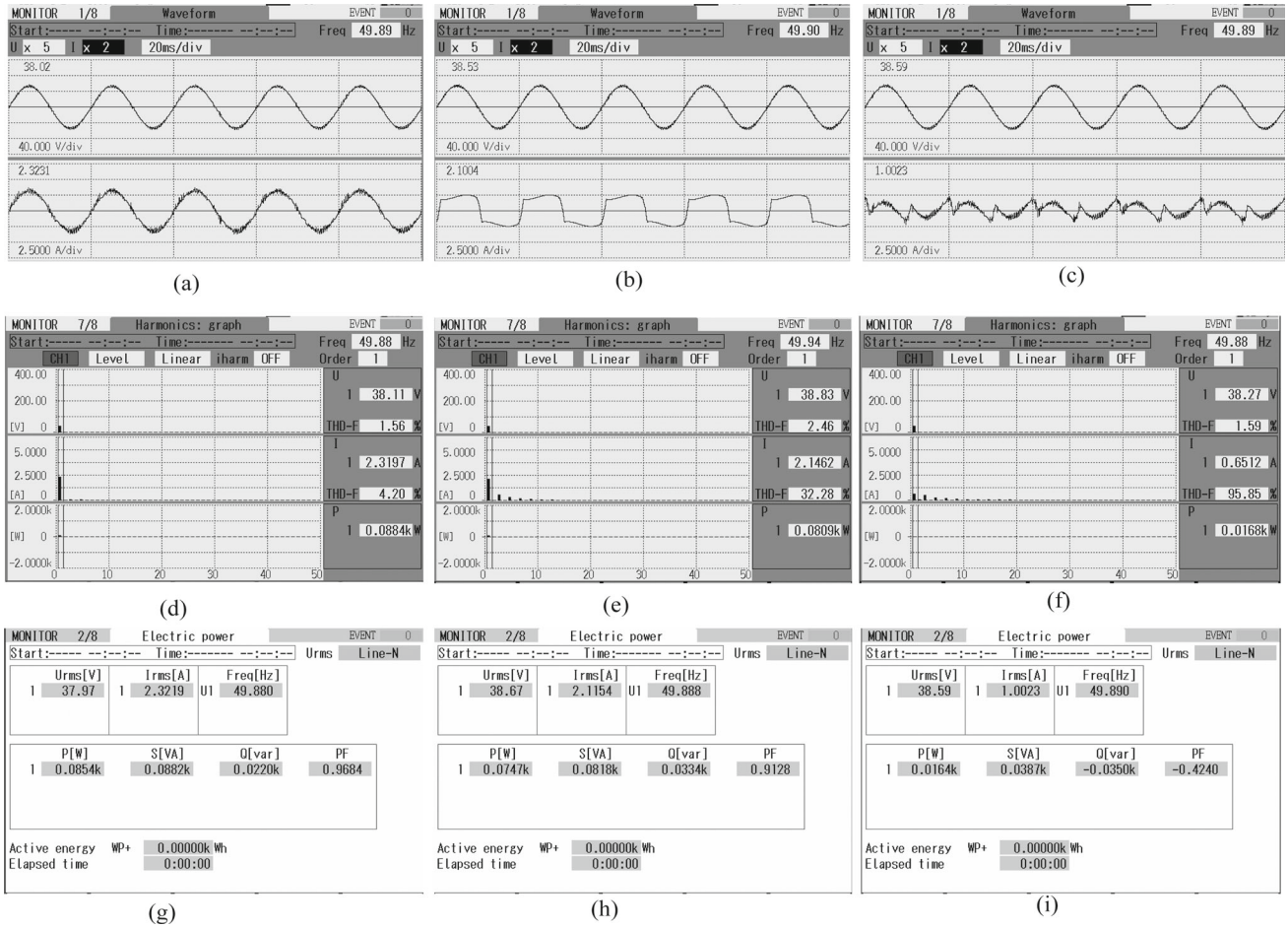


Fig. 9 Hardware results for Legendre FNNAR on the experimental prototype system. **a** Plot of v_s and i_s, **b** plot of v_s and i_l **c** plot of v_s and i_c, **d** THD of v_s (1.56%) and i_s (4.2%), **e** THD of v_s (2.46%) and i_l (32.28%), **f** THD of v_s (1.59%) and i_c 95.85%), **g** real and reactive power values for source, **h** real and reactive power values for load, (**i**) real and reactive power values for compensator

studies as well as captured experimentally which are found to be consistent.

The fundamental current is tracked using Legendre FNNAR technique as well as CNISOGI and VSSCLMS in Fig. 10a, b. It can be clearly noted that proposed Legendre FNNAR results in oscillation free tracking of the fundamental weight component of the load current. Their performance is further compared in Table 1 shown below:

Table 1 clearly shows that time taken for the fundamental weight to converge is least (being less than 2 cycles) for Legendre FNNAR. The current distortion in supply current is the least and since the tracked fundamental weight has negligible ripples; hence, additional filter is not required for Legendre FNNAR in comparison with the advanced techniques such as CNISOGI and VSSCLMS.

The results in Fig. 10 show that the performance comparison of improved techniques such as CNISOGI and VSSCLMS produce satisfactory results. However, slight oscillations are confirmed by simulation and experimental results. Moreover, these advanced techniques require a judicious choice of parameters such as gains, step size, alpha and beta involved in the development of techniques. If the parameters are not appropriate, the working of CNISOGI and VSSCLMS techniques gets affected adversely. The proposed Legendre FNNAR shows improved performance in terms of ripple free weight tracking as evident from the simulation and experimental results in Fig. 10a, b. It can be further noted that a few dominant terms of Legendre FNNAR and trignometric FNNAR are sufficient in accurate modeling and predicting load compensation.

6 Conclusions

This paper discusses the development of two FNNAR models for achieving load compensation in a power distribution system. Both the developed techniques are self-adaptive

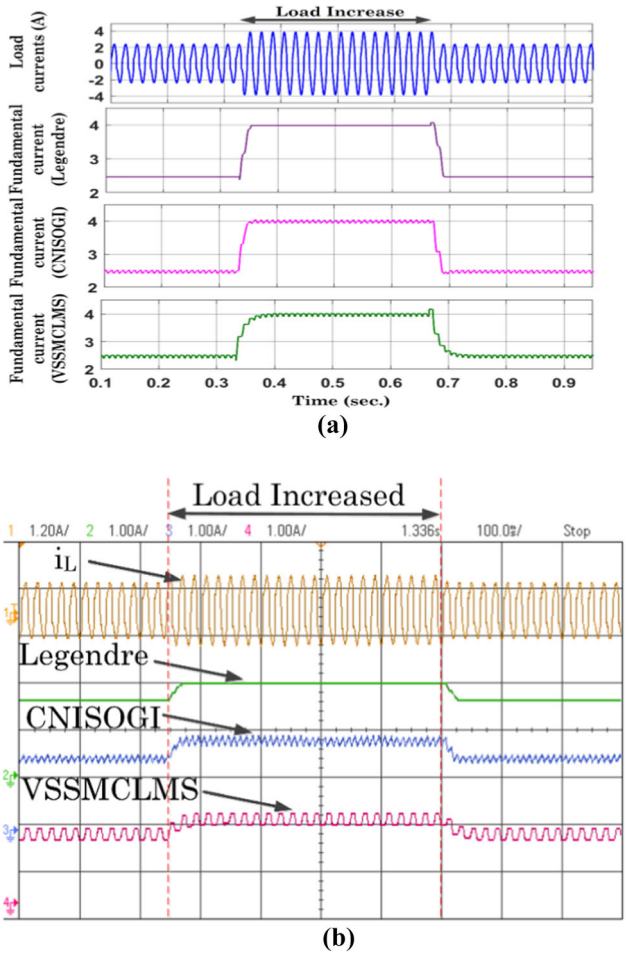


Fig. 10 **a** Simulation results for dynamic load changes for Legendre FNNAR, CNISOGI and VSSCLMS. **b** Experimental results for dynamic load changes for Legendre FNNAR, CNISOGI and VSSCLMS

Table 1 Performance Comparison

Parameter	CNISOGI technique	VSSCLMS technique	Legendre FNNAR
Technique	Non adaptive	Adaptive	Self adaptive
Tracking	Moderate	Moderate	Very good
Weight Convergence	Moderate	Moderate	Fast (< 2 cycles)
Converged weight	High ripples	High ripples	No ripples
% THD of i_s	4.4%	4.1%	3.23%
% THD of i_L	31.73%	31.73%	31.73%
Additional filter requirement	Yes	Yes	No
Dependency on control parameters	Highly Dependent	Highly Dependent	Self adaptive
Complexity	High	High	Moderate

and based on trigonometric and Legendre functional neural networks. Simulation and experimental results have been presented for a single–phase system connected to a nonlinear load at the coupling point. Further, power quality problems such as poor power factor, harmonics, reactive burden, voltage regulation are present in distribution system and can be solved using shunt compensation. The controller is developed using two proposed FNNAR models. The proposed models present advantages of functional neural networks coupled with feedback of delayed output terms. Both the developed models have been trained online using the LMS technique to estimate the fundamental weight of the load current. The two algorithms are able to estimate the correct injection current required for adequate compensation. Simulation results have been shown using both the FNNAR models and suitably compared with advanced techniques such as CNISOGI and VSSCLMS. The Legendre-based FNNAR model outperforms the trigonometric function-based FNNAR model in terms of lower THD value of 3.23% achieved and fast dynamic response (~ 2cycles) as compared to trigonometric FNNAR model. The experimental results are also in agreement with the simulation results. The paper justifies the feasibility of developing FNNAR models for achieving compensation and improving power quality problems in power systems. It also highlights the need of developing accurate models for estimation, prediction and forecasting.

Appendix

Simulation parameters: Supply voltage (v_s) = 110 V, L_s = 3.5 mH, Diode rectifier load with L = 80mH, variable resistance (0-120Ω), dc link voltage (V_{dc}) = 200 V. Models developed in MATLAB Simulink (2020).

Experimental parameters: Supply voltage (v_s) = 40 V, L_s = 3.5mH, Diode rectifier load with L = 80mH, variable resistance (0-100Ω), dc link voltage (V_{dc}) = 80 V.Real time Controller used is dSpace-1104.

Acknowledgement The authors are thankful to SERB, DST.

Author contributions Alka Singh contributed to concept, design and editing. Srishti Singh contributed to basic design layout, simulation and results.

Funding No funding has been received for this research work.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable.

Data availability Not applicable.

References

1. Singh B, Chandra A, Al-Haddad K (2015) Power Quality Problems and Mitigation Techniques. Wiley
2. Razmi D, Lu T, Papari B, Akbari E, Fathi G, Ghadamyari M (2023) An overview on power quality issues and control strategies for distribution networks with the presence of distributed generation resources. IEEE Access 11:10308–10325. https://doi.org/10.1109/ACCESS.2023.3238685
3. Narayanan V, Singh B (2023) "CNISOGI and VSSMCLMS Filters for solving grid and load abnormalities for a microgrid with seamless mode transfer capability. IEEE Trans Ind Appl. https://doi.org/10.1109/TIA.2023.3259387
4. Arya SR, Alam SJ, Ray P (2024) Optimized PI gain in UPQC control based on improved zero attracting normalized LMS. CPSS Trans Power Electronics Appl 9(2):242–251. https://doi.org/10.24295/CPSSTPEA.2024.00007
5. Rai KB, Kumar N, Singh A (2024) Three-phase grid connected shunt active power filter based on adaptive Q-LMF control technique. IEEE Trans Power Electron 39(8):10216–10225. https://doi.org/10.1109/TPEL.2024.3398369
6. Xiaoa H, Huangb D, Panc Y, Liub Y, Song K (2017) Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model. Chemom Intell Lab Syst 161:96–107
7. Chon Ki H, Cohen RJ (1997) Linear and nonlinear ARMA model parameter estimation using an artificial neural network. IEEE Trans Biomed Eng 44(3):168–174
8. Pappas SP, Ekonomou L, Karampelas P, Karamousantas DC, Katsikas SK, Chatzarakis GE, Skafidas PD (2010) Electricity demand load forecasting of the Hellenic power system using an ARMA model. Electric Power Syst Res 80:256–264
9. Hosovský A, Pitel J, Adamek M, Mizakov J, Zidek K (2021) "Comparative study of week-ahead forecasting of daily gas consumption in buildings using regression ARMA/SARMA and genetic-algorithm-optimized regression wavelet neural network models. J Build Eng 34:101955
10. Dash SK, Dash PK (2019) Short-term mixed electricity demand and price forecasting using adaptive autoregressive moving average and functional link neural network. J Mod Power Syst Clean Energy 7(5):1241–1255
11. Cui M, Liu H, Li Z, Tang Y, Guan X (2014) Identification of Hammerstein model using functional link artificial neural network. Neurocomputing 142:419–428
12. Wang Y, Wang D, Tang YI (2020) Clustered hybrid wind power prediction model based on ARMA, PSO-SVM, and clustering methods. IEEE Access 8:17071–17079
13. Jalil M, Samet H, Ghanbar T (2023) Dynamic polynomial models with ARMA coefficients used for modeling the DC series ARC fault in photovoltaic systems. IEEE Trans Industr Inf 19(5):6364–6375
14. Pham HT, Yang BS (2010) Estimation and forecasting of machine health condition using ARMA/GARCH model. Mech Syst Signal Process 24:546–558
15. Chhajer P, Shah M, Kshirsagar A (2022) The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction. Decis Anal J 2(100015):1–12
16. Abassi AR, Baleanu D (2023) "Recent developments of energy management strategies in microgrids: An updated and comprehensive review and classification. Energy Convers Manag 297:117723
17. Abassi AR, Maohammadi M (2023) "Probabilistic load flow in distribution networks: An updated and comprehensive review with a new classification proposal. EPSR 222:109497
18. Agarwal RK, Hussain I, Singh B (2018) Application of LMS-based NN structure for power quality enhancement in a distribution network under abnormal conditions. IEEE Trans Neural Netw Learn Syst 29(5):1598–1607. https://doi.org/10.1109/TNNLS.2017.2677961
19. Chittora P, Singh A, Singh M (2018) Chebyshev functional expansion based artificial neural network controller for shunt compensation. IEEE Trans Industr Inf 14(9):3792–3800. https://doi.org/10.1109/TII.2018.2793347
20. Pandey A, Singh A (2023) A Laguerre polynomial function-based inverter control with low-voltage ride-through capabilities". Int J Circ Theor Appl 51(2):764–786. https://doi.org/10.1002/cta.3463
21. Saxena H, Singh A, Rai JN (2020) Adaptive Spline-based PLL for synchronisation and power quality improvement in distribution system. IET 14(7):1311–1319
22. Rai KB, Singh A, Kumar N (2023) Bernoulli polynomial-based control technique for PV-integrated grid system under distorted sup ply. Int J Circuit Theory Appl 51:3204–3225. https://doi.org/10.1002/cta.3578
23. Sharma U, Singh B (2022) "Functional link neural network for wide load operation of bidirectional onboard charging system of E-Rickshaw. IEEE Trans Ind Appl. https://doi.org/10.1109/TIA.2022.3198626
24. 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
25. Nguyen N. K, Abdeslam D. O, Wira P, Flieller D, Merckle J (2008) Artificial neural networks for harmonic currents identification in active power filtering schemes, In: 2008 34th annual conference of IEEE industrial electronics, Orlando, FL, USA, 2008, pp. 2696-2701, https://doi.org/10.1109/IECON.2008.4758384
26. Singh B, Arya SR (2014) Back-propagation control algorithm for power quality improvement using DSTATCOM. IEEE Trans Industr Electron 61(3):1204–1212. https://doi.org/10.1109/TIE.2013.2258303
27. Singh B, Solanki J (2009) A comparison of control algorithms for DSTATCOM. IEEE Trans Industr Electron 56(7):2738–2745. https://doi.org/10.1109/TIE.2009.2021596

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.

ExLibris RapidILL

Rapid #: -23205873

CROSS REF ID: **6420165**

LENDER: **TKN (University of Tennessee, Knoxville) :: TKN-Library**

BORROWER: **HUL (Harvard University) :: Widener Library**

TYPE:	Article CC:CCG
JOURNAL TITLE:	International journal of organizational analysis
USER JOURNAL TITLE:	The international journal of organizational analysis.
ARTICLE TITLE:	Does spirituality at work promote work-to-family enrichment among Indian female employees?
ARTICLE AUTHOR:	Garg, Naval
VOLUME:	
ISSUE:	
MONTH:	
YEAR:	2024
PAGES:	NA
ISSN:	1934-8835
OCLC #:	

Processed by RapidX: 9/27/2024 7:51:05 AM

This material may be protected by copyright law (Title 17 U.S. Code)

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1934-8835.htm

International Journal of Organizational Analysis

Does spirituality at work promote work-to-family enrichment among Indian female employees?

Naval Garg
Department of University School of Management and Entrepreneurship, Delhi Technological University, Delhi, India, and

Nidhi Sharma
Department of Commerce, Kirori Mal College, University of Delhi, New Delhi, India

Received 9 April 2024
Revised 20 June 2024
16 August 2024
Accepted 18 August 2024

Abstract

Purpose – Based on Indian conceptualisation of workplace spirituality, this study aims to examine the linkage between four dimensions of workplace spirituality (swadharma, authenticity, lokasangraha and sense of community) and work-to-family (WTF) enrichment. It also explored the mediating effect of psychological and social capital and the moderating effect of gratitude.

Design/methodology/approach – A sample of 387 women employees of Indian hospitality industry was collected in three waves. The collected data were analysed in three stages. Firstly, reliability, validity and multicollinearity were assessed using appropriate statistical measures like Cronbach's alpha, composite reliability and average variance explained. Secondly, the relationship between four dimensions of workplace spirituality and WTF enrichment were examined using correlation and hierarchical regression. Several demographic variables like marital status, age, experience and income level were controlled. Thirdly, the moderating effect of gratitude and mediating effects of psychological and social capital were analysed using PROCESS macro.

Findings – The results showed adequate reliability and validity estimates. Also, four dimensions of Indian workplace spirituality were significantly related to WTF enrichment with these dimensions of workplace spirituality collectively explaining 46.8% variations in WTF enrichment. The results also concluded significant meditating effect of psychological and social capital. It also asserted significant moderating effect of gratitude.

Originality/value – The study is based on longitudinal data collected to test seven hypotheses of the study.

Keywords Workplace spirituality, Work-to-family enrichment, Gratitude, Psychological capital, Social capital

Paper type Research paper

Declarations: We hereby declare that the present article has been compiled with Ethical Standards as stated below:

Funding: Not applicable.

Disclosure of potential conflicts of interest: No conflict of interest.

Research involving human participants and/or animals: Informed consent of human respondents taken. Data is not shared owing to the promise of confidentiality with respondents.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Authors' contributions: All authors have contributed equally to the realization of the study and the manuscript. The draft manuscript being submitted for due consideration has been read and approved by all the authors. The authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

International Journal of Organizational Analysis
© Emerald Publishing Limited
1934-8835
DOI 10.1108/IJOA-04-2024-4434

Introduction

Work and family life are two distinct yet overlapping facets of one's life (Philip *et al.*, 2017). With blurring office-home boundaries, individuals often find themselves struggling hard to fulfil professional and personal duties efficiently (Lott and Abendroth, 2023). The challenges are exponentially higher for female professionals (Jafrin *et al.*, 2021), especially in a country like India, which still grapples with the shackles of a patriarchal societal set-up (Costagliola, 2021; Garg, 2022). Indian women are largely responsible for household work, rearing children, caring for elders at home and adhering to customs and rituals. These multiple roles often lead to work overload and stress among women, which results in poor performance on both fronts: home and workplace (Lu, 2011). This phenomenon persists across industries, evermore so in the hospitality sector, which is inherently a demanding industry (Karatepe, 2011; Miao *et al.*, 2021). Within the Indian cultural context, it is characterised by high stress levels and disturbed work–life balance (Mousa *et al.*, 2023) due to job insecurity, inadequate compensation and unsocial long working hours (Pradhan, 2022). Despite these hurdles, the industry is witnessing an increase in female labour participation (Mousa *et al.*, 2023). Thus, to facilitate their positive participation and productivity; it is imperative to explore solutions for the effective management of work–family dynamics.

From a positive organisational scholarship perspective, work–family integration – a concept popularised by Rapaport *et al.* (2002), highlights one such opportunity. It states that experiences gained in one role can lead to better performance in another role and vice versa (Barnett and Hyde, 2001). This phenomenon is referred to as work–family enrichment. It is defined as "the extent to which experiences in one role improve the quality of life in another role" (Greenhaus and Powell, 2006, p. 73). This positive spillover of the work–family interface states that knowledge, skills, competencies and experiences gained in one role make one more confident, empowered, resourceful and competent not only for the same role but for other assignments of the life too (Premchandran and Priyadarshi, 2020). It is a bi-directional concept wherein benefits derived from work could be applied to enrich the family (known as work-to-family [WTF] enrichment) and resources gained from the family could be used to enrich work life (termed as family-to-work [FTW] enrichment). Carlson *et al.* (2006) asserted that developmental resources, positive affect, psycho-social capital and efficiencies derived from involvement in work/family roles can lead to higher levels of performance in family/work settings. It is, thus, expected that investments made by organisations in the development of their female workforce should empower them with resources to better manage their home front as well, and create a virtuous cycle of benefits that trickle back to the workplace in the form of happy, energetic and competent employees.

As India is a land of rich spiritual heritage (Prabhavananda, 2019) with a deep sense of appreciation for gratitude, optimism, hope, faith, brotherhood, collective activity, love and service to mankind (Chidbhavananda, 1983), the present research undertakes to evaluate the role of such factors in promoting WTF enrichment. As a primary objective, this study examines workplace spirituality as an antecedent of WTF enrichment. Taking cue from Pandey *et al.* (2009), Garg (2017), Saxena *et al.* (2020) and Pawar (2019), the study relies on the Indian conceptualisation of workplace spirituality to obtain results. These researchers have highlighted the cultural sensitivity of spirituality with a unanimous suggestion of using Eastern spiritual values in the Indian context. In doing so, this research aims to overcome the limitation of the standalone study conducted in the Eastern context by Hassan *et al.* (2021) on the association between workplace spirituality and WTF enrichment using a Western conceptualisation of workplace spirituality. In addition, the study explores psychological capital (hope, optimism, self-efficacy and resilience) and social capital (friendship, affiliation, bonding, trust and love) as probable mediating factors, and gratitude as a

moderating variable in the relationship between workplace spirituality and WTF enrichment. Based on the resource-based development perspective (Demerouti *et al.*, 2001; Fredrickson, 2001; Hobfoll, 2001), we argue that workplace spirituality initiatives aimed at the women workforce would build personal resources such as their psychological capital and social capital (Hassan *et al.*, 2021; Sarkar *et al.*, 2022). These resources, in turn, should make them more competent and empowered even when pursuing daily household tasks. For instance, workplace spirituality driven psychological resources might make a working mother more efficient and resilient in dealing with challenges associated with child caring and rearing. Similarly, friendship and bonding with others may provide necessary social support, solace and a channel to vent out their frustration and anger. These resources would create energy reservoirs that could be used in any sphere of life (Solat *et al.*, 2020). Furthermore, based on the find-remind-bind theory of gratitude (Algoe, 2012), we propose that gratitude could potentially moderate the relationship between workplace spirituality and WTF enrichment. The theory states that gratitude helps one acknowledge and appreciate goodness in self-other, consistently and regularly reminds one of this goodness, and finally binds one's thoughts and actions to this realised goodness (Algoe *et al.*, 2008). In this background, we propose that female workers equipped with the practice of gratitude would be able to better recognise and use spiritual values learnt over their workplace experience and carry them to other spheres of their lives.

The purpose of the study is to assess the role of workplace spirituality (swadharma, authenticity, lokasangraha and sense of community) in ensuring work–family enrichment among women employees engaged in the Indian hospitality industry. This leads to the following three objectives:

(1) to examine the direct relationship between four dimensions of workplace spirituality and work–family enrichment;

(2) to explore the mediating role of psychological and social capital; and

(3) to investigate the moderating role of gratitude amidst the relationship of workplace spirituality and work–family enrichment.

Literature review and hypothesis development

Workplace spirituality

The Western conceptualisation of workplace spirituality describes it as the employees having "an inner life that nourishes and is nourished by meaningful work that takes place in the context of community" (Ashmos and Duchon, 2000; Krishnakumar and Neck, 2002). It is said to enrich the mind with knowledge and clarity of thoughts, fill the heart with compassion, trust and love and grow the spirit with self-realisation, consciousness and potential optimisation (Karakas, 2010). Workplace spirituality aligns the body, mind, heart and soul by ensuring the search for meaning and purpose of life, authenticity, transcendence, companionship, collective consciousness, mindfulness and contentment at the workplace (Jalan and Garg, 2022; Kumar and Kumar, 2014; Nigri *et al.*, 2024; Petchsawang and Duchon, 2009; Sharma and Garg, 2024). Amid the innumerable definitions (Karakas, 2010), the common underlying theme has been of "meaningful and purposeful job experience, aligned with the authenticity of work and a sense of belongingness and interconnection felt by the worker" (Garg and Saxena, 2020). Although the essence across conceptualisations may be common, researchers have encouraged the use of culture-specific definitions of workplace spirituality that resonate closely with the target respondents (Garg, 2017; Barik and Nayak, 2024). Within the Indian cultural context (Pandey *et al.*, 2016), Indian notion of workplace spirituality has been validated and extensively used (Garg and Saxena, 2020; Sarkar and Garg, 2021; Saxena *et al.*, 2020). It suggests four dimensions of workplace spirituality:

swadharma (*one's righteous action or duties*), authenticity (*being true and honest to oneself and others*), lokasangraha (*other's orientation*) and a sense of community (*a feeling of belongingness among members*). The present study uses this Indian conceptualisation of workplace spirituality for gathering respondent perceptions on the construct.

Western conceptualisation of workplace spirituality revolves around the work of two prominent scholars. Mitroff and Denton (1999) conceived three dimensions of workplace spirituality: ultimate purpose in life, to develop a strong connection to co-workers and other people associated with work and to have consistency (or alignment) between one's core beliefs and the values of their organisation. Besides, Ashmos and Duchon (2000) also indicated three dimensions: sense of inner life, meaningful work and community life. Pandey *et al.* (2016) conceptualisation seems to be more comprehensive and wider than these Western connotations (Garg and Saini, 2024). Swadharma encapsulates meaningful and meditative work aligned with one's aspirations and purpose of life. Thus, it does not only comprise of two Western dimensions but also includes meditative work, which is missed by Western scholars. Authenticity goes beyond mere alignment of one's belief with organisational values and includes one's sincere efforts to remain honest and true to oneself and others. The notion of lokasangraha further widens the scope of Eastern workplace spirituality through aligning service orientation among employees. It seeks an inherent orientation towards serving others through whatever job one does at the workplace. Thus, these Indian dimensions of workplace spirituality provide much wider coverage of workplace spirituality.

Work-to-family enrichment

WTF positive spillover is the utilisation of experiences, skills, values, abilities, affects, attitudes and behaviours learnt at the workplace within a familial set-up (Edwards and Rothbard, 2000). When this utilisation of work resources leads to improved performance while fulfilling household demands like domestic chores, caring and rearing of kids and elder parents, etc., then it is referred to as WTF enrichment (Graves *et al.*, 2007). The WTF enrichment process is driven by the development of an employee's psycho-social capital at the workplace and its efficient application in family interactions (McNall *et al.*, 2009). The psychological resources (such as optimism, self-efficacy, hardiness, etc.) and social capital resources (such as information processing, networking, influence, etc.) gained through work experiences are expected to enhance the quality of family life (Greenhaus and Powell, 2006). For instance, a positive, satisfied, confident and motivated employee is more likely to bring positive emotions, energies, attitudes and behaviours into their interactions with family members, which would lead to deeper associations between them. Similarly, an employee having cordial and enriching relationships with colleagues will have a larger pool of social resources during a familial emergency. The notion of WTF enrichment and its far-reaching implication for professionals and organisations has been supported by several prominent theories like conservation of resource theory, resource-based development theory, social exchange theory and broaden-and-build theory in studies by Chen and Powell (2012), Fung *et al.* (2012), McNall *et al.* (2015), Mishra *et al.* (2019) and Rastogi *et al.* (2016), to name a few.

Workplace spirituality and work-to-family enrichment

Rooted in the Indian school of thought, the four factions namely swadharma, authenticity, lokasangraha and sense of community have been operationalised as the foundation of workplace spirituality for the current study. The following sections propose the relationships between these dimensions and WTF enrichment.

Swadharma

The reference to swadharma is found across Indian scriptures like Shrimad Bhagwad Gita, Kena Upanishad, Ramayana and Shiv Purana (Bhawuk, 2011; Sarkar, 2022). Shrimad Bhagwad Gita, one of the most revered Indian scriptures, elucidates **"**Svadharmenidhanamsreyah para-dharmobhayavahah" (Chapter 3, verse 35), which asserts that one should indulge in work as per his/her dharma, i.e. talents, skills, value system and aspirations. The natural outcome of indulgence in swadharma is expected in the form of meaningful and meditative work (Pandey *et al.*, 2009). Meaningful work can be derived by aligning natural interest and purpose of life with job assignments and job achievements (Pandey *et al.*, 2016). In scenarios where the work assignment of an employee naturally aligns with his/her talent and aspirations, the derived contentment, sense of satisfaction and meaningfulness are poised to get transferred into one's family life too. However, the absence of such synchronisation would call for remedial action. Drawing on the broaden-and-build theory (Fredrickson, 2004a), promoting spiritual awareness among employees is expected to broaden their outlook and encourage "exploratory thoughts and actions" (Garg and Saini, 2024; Rana and Saxena, 2024). This in turn should elicit efforts by the employees to build on necessary skills and talent to better manage their current assignments or get desired assignments. In either case, the newly acquired skills and abilities are bound to enrich one's personal life too. Accordingly, the following hypothesis is proposed:

H1. Swadharma is positively associated with work-to-family enrichment.

Authenticity

The concept of authenticity has been adored in Indian spiritual tradition. One of the most valued teachings from *Śrīmad Bhagavad Gītā*, "Mansa vacaha karmani", that supports a perfect alignment of thoughts, words and actions (Rāmsukhdāsjī, 2007) resonates with the modern conceptualisation of authenticity given by researchers like Garg (2017) and Pandey *et al.* (2016). They describe authenticity in the context of workplaces as harmonising the actions and behaviours of the employees with their true internalised belief systems and values (Garg and Saxena, 2020). It is achieved by practising virtues like honesty, non-violence, trustworthiness, morality and celibacy (Pandey *et al.*, 2009) and can be operationalised in the workplace by employees being open, supportive and sincere to their co-workers (Garg and Saxena, 2020). Organisations could facilitate authenticity in work relationships by promoting a culture based on positive thoughts, clear communication, trust, consistency in actions and absence of bias and misinterpretation (Pareek, 2003; Mittal *et al.*, 2024), where employees adhere to group norms, organisational policies, professional ethics and societal value systems (Dubey *et al.*, 2024). Now, when employees are equipped with such high degree of personal and professional ethics, and they carry the same into the familial settings too, it should lead to better relationships with their spouse, children, parents and the society at large. By bringing their true selves to the family responsibilities and relationships, they should experience higher levels of WTF enrichment. Accordingly, the following hypothesis is proposed:

H2. Authenticity is positively associated with work-to-family enrichment.

Lokasangraha

Lokasangraha means coming together for the betterment and welfare of society (Sebastian, 2003; Sharma, 1999; Tilak, 1915). The modern literature refers to it as other's orientation (Saxena *et al.*, 2020). The concept provides a guiding light to the actions of contemporary business organisations (Pandey *et al.*, 2016). Lokasangraha inspires employees and companies to work for the betterment

of the society (Sarkar and Garg, 2021). The call for social justice, equity, affirmative and mutual relationship demands the propagation of wisdom of other's orientation among the business firms (Jena, 2022). It has been discussed in various Indian scriptures. Shrimad Bhagwad Gita elucidates that every human action should aim towards the welfare of others (Chapter 3, verse 14). Brihadaranyaka Upanishad prays "Om, Sarve bhavantusukhinah, sarvesantuniramayah, Sarve bhadranipasyantu, Ma kashchitduhkabhagbhavet" (verse 1.4.14). The prayer requests the almighty for a prosperous, happy and spiritually fulfilling life for all. Researchers have asserted that the natural outcomes of lokasangraha are compassion, benevolence, sacrifice and harmony (King, 2022; Panda, 2022; Ranjan Chatterjee, 2009), which are essential for every family man/woman. An employee deeply ingrained in compassion and love for others through organisational endeavours is bound to reciprocate the same to his/her family members too. It is further expected that upholding other's orientation would require a mindset of selfless service and concern for the well-being of others (Hilton *et al.*, 2024). This would necessitate balancing out of individual, organisational, and societal interests through judicious planning, implementation and evaluation of efforts. These skills when transferred to family life are poised to make one's family life more worthwhile. Accordingly, the following hypothesis is proposed:

H3. Loksangraha is positively associated with work-to-family enrichment.

Sense of community

Sense of community is defined as:

> […] a feeling of belongingness among members, a feeling that members matter to each other as well as to the group, and a shared belief that members' needs will be sought after through their commitment to being together (McMillan and Chavis, 1986, p. 9).

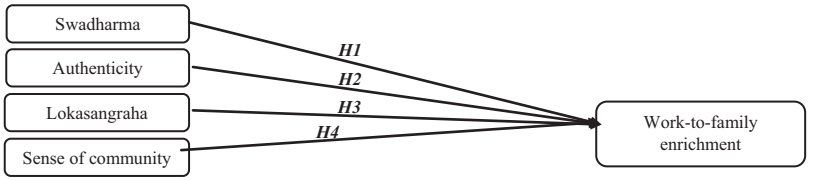
It emphasises interdependence and interconnectedness among employees (Giacalone and Jurkiewicz, 2004; Mahipalan and Sheena, 2018). The feeling of belongingness strengthens through mutual caring and sharing, acknowledging and fulfilling relative obligations and commitments and realising the shared vision, mission and core organisational values (Chawla and Guda, 2013). Interconnection and interdependence are hallmarks of Indian families, wherein women are seen as an indispensable cog in the wheel of the overall family well-being (Garg, 2022; Sode and Chenji, 2024). Sense of community developed at the workplace should make women employees more tolerant and appreciative of the divergent viewpoints of different family members. Moreover, friendships and affiliations acquired at work could provide much-needed social support to the employees and their families even after office hours. Thus, an employee's sense of community could lead to enhanced WTF enrichment. Accordingly, the following hypothesis is proposed:

H4. Sense of community is positively associated with work-to-family enrichment.

Based on the first four hypotheses, the following conceptual model (please refer to Figure 1) is proposed.

Meditating role of psychological and social capital

Positive organisational behaviour, a subject matter of positive psychology, emphasises on psychological resources, which could be measured, developed and managed, with empirical proof of their impact on employees' and organisational performance and well-being (Luthans and Youssef, 2007). These resources are termed as psychological capital. At an individual level, psychological capital promotes learning, growth and performance, and at an



Source: Authors' own

Figure 1. Conceptual model of workplace spirituality and work-family enrichment

organisational level, it offers competitive advantage and leverage through improved employee performance (Luthans *et al.*, 2014). Grounded in evidence-based management, it has been referred to as a multi-dimensional construct that synergistically consists of four resources: hope, efficacy, resilience and optimism (Luthans *et al.*, 2007). Hope is defined as "a positive motivational state that is based on an interactively derived sense of successful (1) agency (goal-directed energy) and (2) pathways (planning to meet goals)" (Snyder *et al.*, 1991, p. 287). Self-efficacy is defined as:

> […] one's conviction (or confidence) about his or her ability to mobilize the motivation, cognitive resources, and courses of action needed to successfully execute a specific task within a given context (Stajkovic and Luthans, 1998, p. 66).

Luthans (2002, p.702) defines resilience as "the capacity to rebound or bounce back from adversity, conflict, and failure". Optimism is defined as a "mood or attitude associated with an expectation about the social or material future - one which the evaluator regards as socially desirable, to his advantage, or for his please" (Tiger, 1979, p.18).

Social capital is defined as:

> […] the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition (Bourdieu and Wacquant, 1992, p. 14).

Social capital consists of resources apropos "whom you know" (Luthans and Youssef, 2007). These resources include social relationships, networks, communities, societies and the flow of information that ensure growth, development and support to employees (Machalek and Martin, 2015). Social capital can manifest in the form of bonding and/or bridging (Woolcock, 1998). Bonding involves linkages and close connections with a homogeneous group that includes family, relatives and close friends. Bridging denotes set of relationships with heterogeneous groups of the community (Melé, 2003).

It has been emphasised that spirituality develops positive thoughts, emotions and behaviour (Aboobaker *et al.*, 2019; Vasconcelos, 2018). The broaden-and-build theory of positive emotion states that positive emotions broaden people's momentary thought-action repertoires and build enduring personal, psychological and social resources (Fredrickson, 2001, p. 147). A similar line of argument is presented by proponents of the resource-based development perspective. Based on the conservation of resource theory (Hobfoll, 2001) and job demand-resource model (Demerouti *et al.*, 2001), it is suggested that workplace spirituality promotes positive emotion and helps in replenishing the resources expended by employees while fulfilling various job demands (Sharma, 2022). It can, therefore be reasoned that workplace spirituality develops psychological and social capital among employees, which facilitates WTF enrichment (Liu *et al.*, 2024). For instance, adherence to the principles of authenticity like honesty, integrity and non-violence would command a lot of courage,

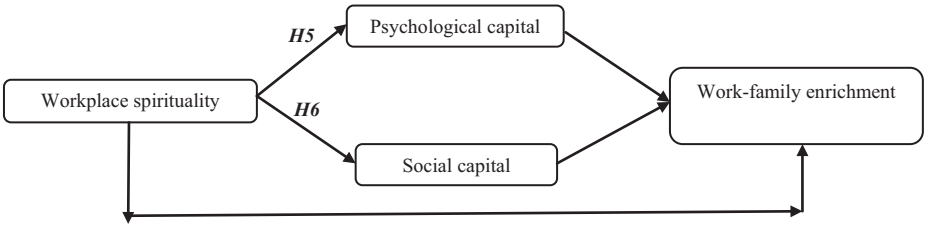
resilience and competence. These psychological resources developed owing to authentic workplace behaviours can be reflected at home where the person brings his true, loyal and dedicated self to the family relationships. Furthermore, an employee true in words and behaviour is poised to win hearts of the colleagues and thereby, develop long-term associations. Similarly, a sense of community developed at the workplace should promote positive affiliations, social relationships and friendships, which are essential elements of social capital that could in turn help to develop strong trust, love and compassion-based personal relationships. It can, therefore, be expected that these resources developed at work would make one effective at the home front too. Based on these theoretical premises, the following hypotheses and conceptual model (please refer to Figure 2) are formulated:

H5. Psychological capital mediates the relationship between workplace spirituality and work-to-family enrichment.

H6. Social capital mediates the relationship between workplace spirituality and work-to-family enrichment.

Moderating role of gratitude

Researchers have theorised gratitude as an emotion, an attitude, an individual trait and a coping strategy (Emmons and McCullough, 2003; Wood *et al.*, 2010). This study conceptualises gratitude as a dispositional trait, which is relatively stable and permanent. As a trait, gratitude is defined as a "general tendency to recognize and respond with grateful emotions to the roles of other people's benevolence in the positive experiences and outcomes that one obtains" (McCullough *et al.*, 2002, p. 112). It represents a state of being grateful (Garg, 2023). There are four facets of trait gratitude: intensity, frequency, span and density (McCullough *et al.*, 2002). A person high on trait gratitude is more likely to experience gratitude more intensely, frequently, deeply and for longer durations. Also, there are three dimensions of trait gratitude: lack of sense of deprivation or sense of abundance, simple appreciation and appreciation for others (Watkins *et al.*, 2014). Drawing on the find-remind-bind theory of gratitude, the present study proposes that gratitude moderates the relationship between workplace spirituality and WTF enrichment. The theory states that gratitude helps one to recognise, acknowledge, consistently remind and appreciate the goodness of others and in situations (Algoe *et al.*, 2008). At the workplace, grateful employees should, therefore, be more appreciative of the psycho-social resources gained through spiritual awareness initiatives. They should be able to better acknowledge the skills and capabilities they are able to learn while fulfilling the work assignments, challenges, responsibilities and



Source: Authors' own

Figure 2. Mediating role of PsyCap and social capital

duties. Inclined to find goodness in every organisational experience and exposure, employees high on gratitude would more likely use these skills and workplace experiences in routine home affairs as well. From these theoretical premises, the following hypothesis and conceptual model (please refer to Figure 3) are formulated:

H7. Gratitude moderates the relationship between workplace spirituality and work-to-family enrichment.

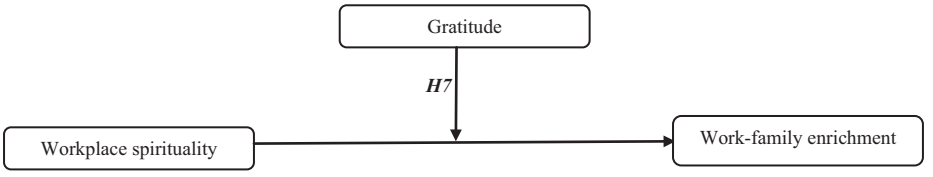
Method

Participants and data collection

The study is based on a longitudinal research design, wherein data were collected in a three-wave survey. Women employees employed in three major players of hospitality industry, i.e. online service aggregators, hotels and restaurants and state and district tourism boards were approached. Twenty entities from each three segments were randomly chosen and approached. Out of these 13, 16 and 10 entities agreed and responded back. Some of these entities were Indian Railways Catering and Tourism Corporation, Ashok Tours and Travels, Balmer Lawrie, MakeMy Trip, Goibibo, Travel Triangle, Agoda, Hotel Westend, Lemon Tree Hotel, Ginger Hotel and Delhi Tourism and Transportation Development Corporation. The women employees were approached through their Human Resource (HR) managers contacted over email or LinkedIn. The purpose and utility of the study were explained to HR managers. To encourage participation in survey, potential respondents were offered an online training on work–family enrichment. Women employees with an experience of minimum one year were allowed to participate, as some time is allowed to translate workplace spirituality into work-family enrichment. Also, employees at the level of supervisor or above were asked to participate. At time T1, information about demographics and workplace spirituality were sought. After one month (T2), gratitude, psychological capital and social capital were measured. And after another one month (T3), WTF enrichment was assessed. The first surveyed yielded 717 responses. Each participant was assigned a unique code. In the second round, 492 responses were received. And lastly, the third round of survey captured data from 399 participants. Twelve responses were discarded for being incomplete. Cochran (1977) has recommended a sample size of 384 at 0.05 significance level, thus, the sample of 387 seems appropriate. Table 1 reports demographic descriptions of the sample.

Instruments

Workplace spirituality. Workplace spirituality was measured using 11 items (Pandey *et al.*, 2016) spiritual climate scale. It captured four dimensions of workplace spirituality – *swadharma*, authenticity, *lokasangraha* (other's orientation) and a sense of community. Items included are *swadharma* (e.g. "working here is a means for realizing my real self",



Source: Authors' own

Figure 3. Moderating role of gratitude

Table1. Demographic profile of samples

Variable	Category	Frequency	%
Marital status	Married	211	54.52
	Unmarried	170	43.93
	Divorced	6	1.55
Age	Below 25 years	54	13.95
	26–35 years	191	49.35
	36–45 years	123	31.78
	Above 45 years	19	4.91
Experience	Below 5 years	51	13.18
	6–15 years	177	45.74
	16–25 years	148	38.24
	Above 25 years	11	2.84
Income level	Below 10 lacs	162	41.86
	11–20 lacs	145	37.47
	Above 20 lacs	80	20.67
Total		*387*	*100*

Source: Primary data

"work itself is enjoyable for me"), authenticity (e.g. "people here own up to mistakes with others in the group"), other's orientation (e.g. "people here try to avoid wastage of any kind", "people here are concerned about the natural environment while working here"), sense of community ("when stuck with a problem, people here feel free to ask for advice from colleagues/advice from a superior/help from their colleagues/help from a superior") on a five-point scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Gratitude. Gratitude was measured using Gratitude Questionnaire-6 developed by McCullough *et al.* (2002). The scale's six items were rated on a seven-point rating scale ranging from 1 (strongly disagree) to 7 (strongly agree). A few items of the scale were "I have so much in life to be thankful for" and "I am grateful for a wide variety of people".

Psychological capital. PsyCap was assessed using 12-item PsyCap scale developed and validated by Luthans *et al.* (2007). The scale captured self-efficacy (three statements), hope (four statements), resilience (three statements) and optimism (two statements) on a six-point rating scale ranging from strongly disagree (1) to strongly agree (6). Items included are "I feel confident contributing to discussions about the company's strategy", "I can get through difficult times at work because I have experienced difficulty before" and "I am optimistic about what will happen to me in the future as it pertains to work".

Social capital. Sixteen-item personal social capital scale developed by Wang *et al.* (2014) was used to measure social capital. It captures two dimensions of social capital-bonding capital and bridging capital- on a five-point rating scale. A few items of the scale were "How do you rate your country fellow" and "Among your relatives, how many you can trust".

Work-to-family enrichment. WTF enrichment was captured with the help of six-item from work–family enrichment scale developed by Carlson *et al.* (2006). The scale measured WTF and FTW enrichment on a five-point rating scale ranging from 1 (completely disagree) to (completely agree). A few items were "My involvement in my family helps me expand my knowledge of new things, and this helps me be a better worker".

Analysis. The collected data were analysed in three stages. Firstly, reliability, validity and multicollinearity were assessed using appropriate statistical measures like Cronbach's alpha (CA), composite reliability (CR) and average variance explained (AVE). Secondly, the

relationship between four dimensions of workplace spirituality and WTF enrichment was examined using correlation and hierarchical regression. Several demographic variables like marital status, age, experience and income level were controlled. Thirdly, the moderating effect of gratitude and mediating effects of psychological and social capital were analysed using PROCESS macro.

Results

Reliability, validity, multicollinearity and common method bias

The CA and CR estimates were used to measure internal consistency reliability. Hair *et al.* (1998) recommended CA and CR values of 0.70 and above to establish reliability of measures. Table 2 suggests CA values of swadharma, authenticity, lokasangraha, sense of community, psychological capital, social capital, gratitude and WTF enrichment are 0.719, 0.746, 0.816, 0.770, 0.852, 0.788, 0.759 and 0.824, respectively. The results also suggested that all CR values are greater than 0.70, which confirms reliability of all measures. The convergent validity was examined using AVE values, whose values should be greater than 0.50 to establish validity of a construct. Table 2 suggests AVE values of swadharma, authenticity, lokasangraha, sense of community, psychological capital, social capital, gratitude and WTF enrichment are 0.548, 0.569, 0.669, 0.618, 0.714, 0.692, 0.622 and 0.738, respectively, which ensured convergent validity of the scales. The divergent validity of four-dimensional workplace spirituality scale was investigated with the help of Fornell and Larcker's (1981) criterion, according to which square root of AVE values should be greater than correlation between four constructs of workplace spirituality. Table 3 established divergent validity of the workplace spirituality scale. As workplace spirituality is a multi-dimensional scale, the issue of multicollinearity was examined using variance inflation factor (VIF), whose recommended values are less than 3 (Kline, 2009). The values of VIF indicated in Table 2 suggested absence of the significant multicollinearity. Although longitudinal research design minimises the chances of common method bias, several measures were used to minimise it. Firstly, pre-validated scales were used in this study. Also, the purpose of the study was discussed with the participants. Secondly, respondents were also assured confidentiality and strict academic use of the collected data. They were also ensured that the collected data would not be shared with any third party without their consent. These minute steps encourage accurate and positive responses (Podsakoff *et al.*, 2012). Finally, Harman's one-factor test was used to examine the significance of common method bias in both current studies. When all items of the study were subjected to factor analysis using varimax rotation,

Table 2. Descriptive statistics, reliability and validity estimates

Variable	Mean	SD	CA	CR	AVE	VIF
Swadharma	3.83	1.145	0.719	0.748	0.548	2.156
Authenticity	3.77	0.849	0.746	0.774	0.569	2.340
Lokasangraha	3.90	1.280	0.816	0.836	0.669	1.563
Sense of community	4.14	0.964	0.770	0.797	0.618	1.440
Psychological capital	4.45	1.748	0.852	0.866	0.714	–
Social capital	3.70	0.795	0.788	0.814	0.692	–
Gratitude	5.28	1.885	0.759	0.784	0.622	–
Work-to-family enrichment	3.90	0.990	0.824	0.848	0.738	

Notes: CA = Cronbach's alpha; CR = Composite reliability; AVE = Average variance explained; VIF = Variance inflation factor
Source: Primary data

Table 3. Correlation matrix

Variable	S	A	L	SC	PsyCap	SoC	G	WTF
S	1 (*0.740*)							
A	0.398*	1 *(0.754)*						
L	0.420*	0.383*	1 *(0.817)*					
SC	0.377*	0.371*	0.319*	1 *(0.786)*				
PsyCap	0.519*	0.538*	0.493*	0.584*	1 *(0.844)*			
SoC	0.533*	0.484*	0.460*	0.376*	0.456*	1 *(0.831)*		
G	0.486*	0.426*	0.385*	0.452*	0.449*	0.521*	1 *(0.788)*	
WTF	0.474*	0.462*	0.402*	0.398*	0.374*	0.436*	0.429*	1 *(0.859)*

Notes: *Sig. at 0.05; S = Swadharma; A = Authenticity; L = Lokasangraha; SC = Sense of community; G = Gratitude; SoC = Social capital; PsyCap = Psychological capital; WTF = Work-to-family enrichment; Square root of AVE mentioned diagonally in italics
Source: Primary data

no single factor was found to explain most of the variations. The single largest factor could explain 29.20% variance. These values recommended insignificant common method variance in this study.

Relationship between workplace spirituality and work-to-family enrichment

Table 2 also reported descriptive statistics including mean and standard deviation of all variables of the study. Table 3 illustrates the correlations between variables of the study. The results reported that four constructs of workplace spirituality are significantly positively correlated with WTF enrichment, i.e. swadharma (r = 0.474, p < 0.05), authenticity (r = 0.462, p < 0.05), lokasangraha (r = 0.402, p < 0.05) and a sense of community (r = 0.398, p < 0.05). Table 4 reported the hierarchical regression results, wherein control variables were entered in Step 1 (Model 1). In Step 2, four dimensions of workplace spirituality were added

Table 4. Results of hierarchical regression

		DV-work-to-family enrichment				
M	Variable	β	SE	Sig.	R^2	F-value
1	Constant	0.730	0.236	2.340	0.086	4.440
	Marital status	0.146	0.011	0.068		
	Age	0.073	0.032	0.064		
	Experience	0.017	0.026	1.859		
	Income level	0.052	0.013	1.534		
2	Constant	1.174	0.642	0.842	0.468	0.004*
	Marital status	0.155	0.045	1.175		
	Age	0.046	0.038	2.323		
	Experience	0.026	0.036	1.248		
	Income level	0.040	0.028	0.740		
	S	0.420	0.168	0.015*		
	A	0.468	0.184	0.037*		
	L	0.382	0.114	0.002*		
	SC	0.452	0.212	0.018*		

Notes: *Sig. at 0.05; S = Swadharma; A = Authenticity; L = Lokasangraha; SC = Sense of community; M = Model; β = Standardised coefficient; SE = Standard error
Source: Primary data

in the regression equation (Model 2). The results revealed that four dimensions of workplace spirituality are significant predictors of WTF enrichment. The regression coefficients of swadharma, authenticity, lokasangraha and a sense of community with WTF enrichment were 0.420, 0.468, 0.382 and 0.452, respectively, at $p < 0.05$ in Model 2, which provided empirical evidences for acceptance of first four hypotheses of the study. These four dimensions of workplace spirituality collectively explain 46.8% of variations in WTF enrichment. These results concluded that workplace spirituality could result in higher levels of WTF enrichment.

Mediating effect of psychological capital and social capital

Table 5 examined the mediating effect of psychological capital and social capital in the relationship of workplace spirituality and work–family enrichment using PROCESS macro in SPSS. The results suggested that indirect effect (via psychological and social capital) (effect size = 0.0980) is significantly greater than direct effect of workplace spirituality on work–family conflict (effect size = 0.4250). It is also observed that zero does not lie between the lower level and upper level of the confidence interval, i.e., LLCI and ULCI, which suggested significance of parallel mediation effect of psychological and social capital. It provided support for *H5* and *H6*. These results concluded that workplace spirituality promotes psychological and social capital among Indian female workers which ultimately results in higher levels of WTF enrichment.

Moderating effect of gratitude

Table 6 explored the moderating effect of gratitude between the relationship of workplace spirituality and WTF enrichment using PROCESS macro in SPSS. As there are different models in PROCESS macro, Model 1 was selected to study the moderation effect. The results highlighted a significant effect of workplace spirituality (t-values = 2.440, $p < 0.05$) and gratitude (t-values = 2.574, $p < 0.05$) on WTF enrichment. Figure 4 reports that gratitude positively moderates the relationship, which suggests that the higher the gratitude of women, the greater is the impact of workplace spirituality on their WTF enrichment. It offered acceptance of the seventh hypothesis of the study.

Discussion

Human beings inevitably carry their whole self to the workplace (Sharma, 2022) and back home. Unrest of the mind, be it family or workplace, is bound to get spilled over to the other sphere. It is, therefore, important to identify the deregulation of the mind caused by stressful

Table 5. Results of parallel multiple mediation

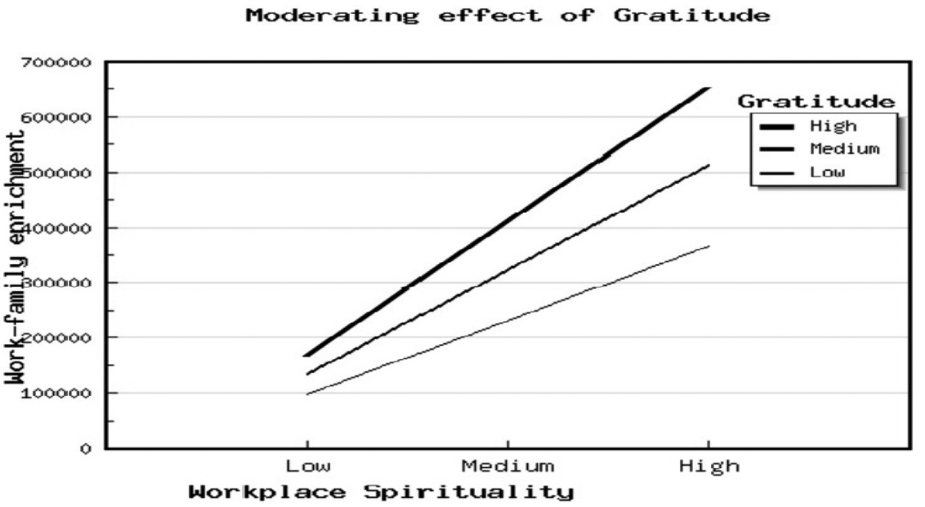
Variable	Effect	SE	LLCI	ULCI
The total effect of workplace spirituality on work-to-family enrichment				
Workplace spirituality	0.5230	0.1003	0.0057	0.1168
Direct effect of workplace spirituality on work-to-family enrichment				
Workplace spirituality	0.0980	0.0054	0.2128	0.2650
Indirect effect of workplace spirituality on work-to-family enrichment				
Total	0.4250	0.1245	0.1780	0.3354
PsyCap	0.2375	0.0340	0.0260	0.1858
Social capital	0.1875	0.0468	0.1642	0.3854

Source: Primary data

Table 6. Results of moderation analysis

Variables	Coefficient	S.E	*t*-value	*p*	LLCI	ULCI
Constant	0.374	0.146	0.856	1.286	0.117	0.693
Workplace spirituality	3.635	2.168	2.440	0.020*	1.775	2.358
Gratitude	1.648	0.840	2.574	0.007*	0.864	1.564
Interaction(WPS*G)	3.440	1.426	2.328	0.037*	0.945	1.640

Notes: WPS = workplace spirituality; G = gratitude * Sig at 0.05
Source: Primary data



Source: Primary data

Figure 4. Moderating role of gratitude

triggers and arrest it in the bud. A realisation that human beings are more than just an anatomical body is what could make the difference. Just like the physical body needs food to survive, the spiritual entity that resides inside the human body too needs nourishment. A nurtured soul is certain to perform well in all domains of life, domestic as well as professional. Taking refuge in spirituality takes the soul on a journey of enlightenment, and brings about "equilibrium of the mind" that turns out maximum work with minimum effort (Chidbhavananda, 1983). In this background, this study proposed a positive relationship between workplace spirituality and work–family enrichment. These results are in accordance with those of Hassan *et al.* (2021), Brough *et al.* (2014) and Garg (2022). These authors also suggested that institutionalisation of spirituality at the workplace helps translation of values, capabilities and experience earned at work into desired skills at home.

Theoretical implications of the study

This study is one of the pioneer studies to investigate the role of workplace spirituality in promoting WTF enrichment in the context of Indian women professionals. Campaigns for

women's emancipation are plenty, nevertheless, one continues to witness social constraints for female professionals in the Indian context who are viewed as primary caregivers and are expected to devote more time to tend to the family needs as compared to the male members of the society (Kumar *et al.*, 2020; Rastogi *et al.*, 2018). Acknowledging such challenges, the present research addresses their need for a balance in the work and family responsibilities. Furthermore, the study uses Pandey *et al.* (2016) spiritual climate scale that captures the notion of workplace spirituality from an Indian perspective. By adopting an Indian conceptualisation of the construct, the data underlying the analysis resonates well with the respondents' cultural orientation. By analysing psychological and social capital as mediating factors, and gratitude as a moderator, the study also delves into the factors that operationalise the relationship between workplace spirituality and WTF enrichment. This comprehensive exploration contributes to numerous theories like the social enhancement theory, resource-based development, broaden-and-build theory of positive emotion and find-remind-bind theory of gratitude.

Practical implications of the results

The study's findings reveal a significant impact of workplace spirituality on WTF enrichment among female employees. The results should catapult the management's initiatives towards institutionalised workplace spirituality interventions. Building a spiritual organisational culture through practices like meditation, yoga, positive visualisation and mindfulness, among others would help build energy reservoirs for a well-balanced life. The Indian hospitality companies might approach spiritual organisations like the Art of Living and Isha Foundation to develop women employee-centric spiritual courses, just like the new recruits as Indian Administrative Officers were administered a course on "Inner engineering", which helps these officers to appreciate the relevance of spirituality at the workplace. Apart from the direct linkage between workplace spirituality and WTF enrichment, the study also reveals the mediating effect of psychological and social capital. HR managers and external Human Resource Development (HRD) consultants may conduct team-building exercises to help women employees realise and cherish the benefits of social affiliation and friendship at work, which will help them gain social capital. The female employees could be trained in gratitude interventions like gratitude journaling, gratitude letters, sharing of positive life events, etc. The companies could take help from gratitude consultants to imbibe the virtue of gratitude among female employees. Gratitude, being contagious (Fredrickson, 2004b), would encourage the energised female staff to pay it forward through their own gratuitous acts. Its repeated practice would reinforce the ethos of kindness and compassion and increase the probability of spiritual values learnt through the workplace experiences being used to balance the family life.

Limitations, scope for future studies and conclusion

Though the study makes significant contributions to both theory and practice, it is not free from limitations. The present study draws inferences from a sample of female professionals in the Indian cultural context. Future studies could, therefore, carry out similar explorations in different countries to validate the findings of the study. In addition, the study's model may be adapted to gather data from male professionals to identify potential differences and/or similarities and enrich the gender-based theories on work–family enrichment. The present research relies on single-source data to collect respondents' perspectives on the study variables. Researchers may consider double-source information in future research to cross-validate the results of this study.

To conclude, the study offers an understanding of the direct and indirect linkages between workplace spirituality and WTF enrichment. It also clarifies the context within which the

relationship gets enhanced. The HR managers could use organization-wide spiritual practices to equip female employees with the necessary psycho-social resources that improve the quality of their work–life and transcend boundaries to enrich their family lives as well. Drawing on positive organisational scholarship literature, promotion of a gratitude-driven organisational culture should heighten the transmission of benefits across role categories and allow women employees to achieve the much-needed balance between their work and family life.

References

Aboobaker, N., Edward, M. and Zakkariya, K.A. (2019), "Workplace spirituality, employee wellbeing and intention to stay", *International Journal of Educational Management*, Vol. 33 No. 1, pp. 28-44, doi: 10.1108/ijem-02-2018-0049.

Algoe, S.B. (2012), "Find, remind, and bind: the functions of gratitude in everyday relationships", *Social and Personality Psychology Compass*, Vol. 6 No. 6, pp. 455-469.

Algoe, S.B., Haidt, J. and Gable, S.L. (2008), "Beyond reciprocity: gratitude and relationships in everyday life", *Emotion*, Vol. 8 No. 3, pp. 425-429, doi: 10.1037/1528-3542.8.3.425.

Ashmos, D.P. and Duchon, D. (2000), "Spirituality at work", *Journal of Management Inquiry*, Vol. 9 No. 2, pp. 134-145, doi: 10.1177/105649260092008.

Barik, S. and Nayak, T. (2024), "The exploration of workplace spirituality research: a scoping review", *Vision*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1177/09722629231225534.

Barnett, R.C. and Hyde, J.S. (2001), "Women, men, work, and family: an expansionist theory", *American Psychologist*, Vol. 56 No. 10, pp. 781-796, doi: 10.1037/0003-066x.56.10.781.

Bhawuk, D. (2011), *Spirituality and Indian Psychology: Lessons from the Bhagavad-Gita*, Springer Science and Business Media, New York; Dordrecht; Heidelberg; London.

Bourdieu, P. and Loïc, W. (1992), *An Invitation to Reflexive Sociology*, The University of Chicago Press, Chicago.

Brough, P., Hassan, Z. and O'Driscoll, M. (2014), "Work-life enrichment", in Dollard, M., Shimazu, A., Bin Nordin, R., Brough, P. and Tuckey, M. (Eds), *Psychosocial Factors at Work in the Asia Pacific*, Springer, Dordrecht, pp. 323-336, doi: 10.1007/978-94-017-8975-2_17.

Carlson, D.S., Kacmar, K.M., Wayne, J.H. and Grzywacz, J.G. (2006), "Measuring the positive side of the work–family interface: development and validation of a work–family enrichment scale", *Journal of Vocational Behavior*, Vol. 68 No. 1, pp. 131-164, doi: 10.1016/j.jvb.2005.02.002.

Chawla, V. and Guda, S. (2013), "Workplace spirituality as a precursor to relationship-oriented selling characteristics", *Journal of Business Ethics*, Vol. 115 No. 1, pp. 63-73, doi: 10.1007/s10551-012-1370-y.

Chen, Z. and Powell, G.N. (2012), "No pain, no gain? A resource-based model of work-to-family enrichment and conflict", *Journal of Vocational Behavior*, Vol. 81 No. 1, pp. 89-98.

Chidbhavananda, S. (1983), *Daily Divine Digest*, Sri Ramakrishna Tapovanam Tirupparaitturai, India.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., John Wiley & Sons, New York.

Costagliola, A. (2021), "Labor participation and gender inequalities in India: traditional gender norms in India and the decline in the labor force participation rate (LFPR)", *The Indian Journal of Labour Economics*, Vol. 64 No. 3, pp. 531-542.

Demerouti, E., Bakker, A.B., Nachreiner, F. and Schaufeli, W.B. (2001), "The job demands-resources model of burnout", *Journal of Applied Psychology*, Vol. 86 No. 3, pp. 499-512, doi: 10.1037/0021-9010.86.3.499.

Dubey, M.J., Ghosh, R., Das, G., Roy, D., Das, S., Chakraborty, A.P., Chatterjee, S., Dubey, S. and Benito-León, J. (2024), "Beyond belief and practice: an exploratory literature review and discussion of the differential impact of spirituality and religiosity on mental health disorders", *Journal of Religion and Health*, Vol. ahead-of-print No. ahead-of-print, pp. 1-14, doi: 10.1007/s10943-024-02090-9.

Edwards, J.R. and Rothbard, N.P. (2000), "Mechanisms linking work and family: clarifying the relationship between work and family constructs", *The Academy of Management Review*, Vol. 25 No. 1, pp. 178-199.

Emmons, R.A. and McCullough, M.E. (2003), "Counting blessings versus burdens: an experimental investigation of gratitude and subjective well-being in daily life", *Journal of Personality and Social Psychology*, Vol. 84 No. 2, pp. 377-389, doi: 10.1037/0022-3514.84.2.377.

Fornell, C.G. and Larcker, D.F. (1981), "Evaluating structural equation models with unobservable variables and measurement error", *Journal of Marketing Research*, Vol. 18 No. 1, pp. 39-50, doi: 10.1177/002224378101800104.

Fredrickson, B.L. (2001), "The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions", *American Psychologist*, Vol. 56 No. 3, pp. 218-226, doi: 10.1037/0003-066X.56.3.218.

Fredrickson, B.L. (2004a), "The broaden–and–build theory of positive emotions", *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, Vol. 359 No. 1449, pp. 1367-1377.

Fredrickson, B.L. (2004b), "'Gratitude, like other positive emotions, broadens and Builds'", in Emmons, R.A. and McCullough, M.E. (Eds), *The Psychology of Gratitude*, pp. 144-166. Oxford University Press, New York, NY.

Fung, N.S., Ahmad, A. and Omar, Z. (2012), "Work-family enrichment: it's mediating role in the relationships between dispositional factors and job satisfaction", *International Journal of Academic Research in Business and Social Sciences*, Vol. 2 No. 11, pp. 73-88.

Garg, N. (2017), "Workplace spirituality and employee well-being: an empirical exploration", *Journal of Human Values*, Vol. 23 No. 2, pp. 129-147, doi: 10.1177/0971685816689741.

Garg, N. (2022), "Gratitude and work-family enrichment among Indian female workforce: exploring the mediating role of psychological capital", *International Journal of Work Organisation and Emotion*, Vol. 13 No. 1, pp. 1-17.

Garg, N. (2023), "Gratitude and work-family enrichment among Indian female workforce: exploring the mediating role of psychological capital", *International Journal of Work Organisation and Emotion*, Vol. 13 No. 1, pp. 1-17, doi: 10.1504/ijwoe.2022.10043234.

Garg, N. and Saxena, A. (2020), "Analyzing the inter-relation between workplace spirituality and constructive deviance", *Asian Journal of Business Ethics*, Vol. 9 No. 1, pp. 121-141, doi: 10.1007/s13520-020-00102-x.

Garg, N. and Saini, D. (2024), "Workplace spirituality to cultivate innovative work behaviour: the moderating role of perceived working conditions", *International Journal of Sociology and Social Policy*, Vol. 44 Nos 1/2, pp. 22-42, doi: 10.1108/IJSSP-03-2023-0050.

Giacalone, R.A. and Jurkiewicz, C.L. (2004), "Toward a science of workplace spirituality", In Giacalone, R.A. and Jurkeiwicz, C.L. (Eds), *Handbook of Workplace Spirituality and Organizational Performance*, New York, NY.

Graves, L.M., Ohlott, P.J. and Ruderman, M.N. (2007), "Commitment to family roles: effects on managers' attitudes and performance", *Journal of Applied Psychology*, Vol. 92 No. 1, pp. 44-56, doi: 10.1037/0021-9010.92.1.44.

Greenhaus, J.H. and Powell, G.N. (2006), "When work and family are allies: a theory of work-family enrichment", *Academy of Management Review*, Vol. 31 No. 1, pp. 72-92, doi: 10.5465/amr.2006.19379625.

Hair, J.E. Jr., Anderson, R.E., Tatham, R.L. and Black, W.C. (1998), *Multivariate Data Analysis*, Prentice-Hall, Upper Saddle River, NJ.

Hassan, Z., Tnay, J.S., Sukardi Yososudarmo, S.M. and Sabil, S. (2021), "The relationship between workplace spirituality and work-to-family enrichment in selected public sector organizations in Malaysia", *Journal of Religion and Health*, Vol. 60 No. 6, pp. 4132-4150, doi: 10.1007/s10943-019-00971-y.

Hilton, S.K., Puni, A. and Yeboah, E. (2024), "Leadership practices and job involvement: does workplace spirituality moderate the relationship?", *Cogent Business and Management*, Vol. 11 No. 1, p. 2316582, doi: 10.1080/23311975.2024.2316582.

Hobfoll, S.E. (2001), "The influence of culture, community, and the nested-self in the stress process: advancing conservation of resources theory", *Applied Psychology*, Vol. 50 No. 3, pp. 337-421, doi: 10.1111/1464-0597.00062.

Jafrin, N., Mahi, M., Masud, M.M. and Ghosh, D. (2021), "Demographic dividend and economic growth in emerging economies: fresh evidence from the SAARC countries", *International Journal of Social Economics*, Vol. 48 No. 8, pp. 1159-1174, doi: 10.1108/IJSE-08-2020-0588.

Jalan, S. and Garg, N. (2022), "Exploring linkages between workplace spirituality and occupational stress among Indian police personnel", *South Asian Journal of Human Resources Management*, Vol. 9 No. 1, pp. 53-78, doi: 10.1177/23220937211059626.

Jena, L.K. (2022), "Does workplace spirituality lead to raising employee performance? The role of citizenship behavior and emotional intelligence", *International Journal of Organizational Analysis*, Vol. 30 No. 6, pp. 1309-1334, doi: 10.1108/IJOA-06-2020-2279.

Karakas, F. (2010), "Spirituality and performance in organizations: a literature review", *Journal of Business Ethics*, Vol. 94 No. 1, pp. 89-106, doi: 10.1007/s10551-009-0251-5.

Karatepe, O.M. (2011), "Do job resources moderate the effect of emotional dissonance on burnout? A study in the city of Ankara, Turkey", *International Journal of Contemporary Hospitality Management*, Vol. 23 No. 1, pp. 44-65.

King, A.S. (2022), "Hinduism: the culture of peace and the ethics of war", in Mitchell, J., Millar, S.R., Po, F. and Percy, M. (Eds), *The Wiley Blackwell Companion to Religion and Peace*, Wiley Publications, New York, NY, pp. 195-215, doi: 10.1002/9781119424420.ch17.

Kline, R.B. (2009), "Chapter 7, Measurement", *Becoming a Behavioural Science Researcher: A Guide to Producing Research that Matters*, Guildford Press, NY.

Krishnakumar, S. and Neck, C.P. (2002), "The 'what', 'why' and 'how' of spirituality in the workplace", *Journal of Managerial Psychology*, Vol. 17 No. 3, pp. 153-164, doi: 10.1108/02683940210423060.

Kumar, V. and Kumar, S. (2014), "Workplace spirituality as a moderator in relation between stress and health: an exploratory empirical assessment", *International Review of Psychiatry*, Vol. 26 No. 3, pp. 344-351, doi: 10.3109/09540261.2014.924909.

Kumar, P., Chakraborty, S. and Kumar, A. (2020), "Support system's impact on work-life interface: a study of part-time adult students in India", *Indian Journal of Industrial Relations*, Vol. 55 No. 3, pp. 491-505.

Liu, Y., Si, Z., Shi, Y., Li, B., Liu, P., Liu, S. and Sun, Q. (2024), "'Win-win': dual-path influence of workplace spirituality on work-family enrichment", *Journal of Occupational and Organizational Psychology*, Vol. 97 No. 3, pp. 841-863, doi: 10.1111/joop.12495.

Lott, Y. and Abendroth, A.K. (2023), "Affective commitment, home-based working and the blurring of work–home boundaries: evidence from Germany", *New Technology, Work and Employment*, Vol. 38 No. 1, pp. 82-102.

Lu, L. (2011), "A Chinese longitudinal study on work/family enrichment", *Career Development International*, Vol. 16 No. 4, pp. 385-400, doi: 10.1108/13620431111158797.

Luthans, F. (2002), "The need for and meaning of positive organizational behaviour", *Journal of Organizational Behaviour*, Vol. 23 No. 6, pp. 695-706, doi: 10.1002/job.165.

Luthans, F. and Youssef, C.M. (2007), "Emerging positive organizational behaviour", *Journal of Management*, Vol. 33 No. 3, pp. 321-349.

Luthans, B.C., Luthans, K.W. and Avey, J.B. (2014), "Building the leaders of tomorrow", *Journal of Leadership and Organizational Studies*, Vol. 21 No. 2, pp. 191-199, doi: 10.1177/1548051813517003.

Luthans, F., Avolio, B.J., Avey, J.B. and Norman, S.M. (2007), "positive psychological capital: measurement and relationship with performance and satisfaction", *Personnel Psychology*, Vol. 60 No. 3, pp. 541-572, doi: 10.1111/j.1744-6570.2007.00083.x.

McCullough, M.E., Emmons, R.A. and Tsang, J.A. (2002), "The grateful disposition: a conceptual and empirical topography", *Journal of Personality and Social Psychology*, Vol. 82 No. 1, pp. 112-127, doi: 10.1037/0022-3514.82.1.112.

McMillan, D.W. and Chavis, D.M. (1986), "Sense of community: a definition and theory", *Journal of Community Psychology*, Vol. 14 No. 1, pp. 6-23, doi: 10.1002/1520-6629(198601)14:1<6::AID-JCOP2290140103>3.0.CO;2-I.

McNall, L.A., Masuda, A.D. and Nicklin, J.M. (2009), "Flexible work arrangements, job satisfaction, and turnover intentions: the mediating role of work-to-family enrichment", *The Journal of Psychology*, Vol. 144 No. 1, pp. 61-81.

McNall, L.A., Scott, L.D. and Nicklin, J.M. (2015), "Do positive affectivity and boundary preferences matter for work–family enrichment? A study of human service workers", *Journal of Occupational Health Psychology*, Vol. 20 No. 1, pp. 93-104, doi: 10.1037/a0038165.

Machalek, R. and Martin, M.W. (2015), "Sociobiology and sociology: a new synthesis", in Wright, J.D. (Ed), *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed. Elsevier, Oxford.

Mahipalan, M. and Sheena, S. (2018), "Mediating effect of engagement on workplace spirituality–job involvement relationship: a study among generation Y professionals", *Asia-Pacific Journal of Management Research and Innovation*, Vol. 14 Nos 1/2, pp. 1-9, doi: 10.1177/2319510x18810995.

Melé, D. (2003), "Organizational humanizing cultures: do they generate social capital?", *Journal of Business Ethics*, Vol. 45, pp. 3-14, doi: 10.1023/A:1024112226673.

Miao, C., Humphrey, R.H. and Qian, S. (2021), "Emotional intelligence and job performance in the hospitality industry: a meta-analytic review", *International Journal of Contemporary Hospitality Management*, Vol. 33 No. 8, pp. 2632-2652.

Mishra, P., Bhatnagar, J., Gupta, R. and Wadsworth, S.M. (2019), "How work–family enrichment influence innovative work behaviour: role of psychological capital and supervisory support", *Journal of Management and Organization*, Vol. 25 No. 1, pp. 58-80.

Mitroff, I.I. and Denton, E.A. (1999), "A study of spirituality in the workplace", *Sloan Management Review*, Vol. 40 No. 4, pp. 83-92, available at: https://dialnet.unirioja.es/servlet/articulo?codigo=2486721

Mittal, A., Mishra, G. and Bishnoi, S. (2024), "Workplace spirituality and job satisfaction of higher education teachers in India: examining occupational stress as a mediator", *Business Perspectives and Research*, Vol. 12 No. 2, pp. 296-314, doi: 10.1177/22785337221148822.

Mousa, M., Abdelgaffar, H., Salem, I.E., Elbaz, A.M. and Chaouali, W. (2023), "Religious, contextual and media influence: determinants of the representation of female tour guides in travel agencies", *International Journal of Contemporary Hospitality Management*, Vol. 35 No. 9, pp. 3172-3192, doi: 10.1108/IJCHM-05-2022-0650.

Nigri, G., Ramirez, L.F. and Sferrazzo, R. (2024), "The civil company: place-based corporate social responsibility in dialogue with business ethics and workplace spirituality", in Del Baldo, M., Baldarelli, M.G. and Righini, E. (Eds), *Place Based Approaches to Sustainability Volume I. Palgrave Studies in Sustainable Business in Association with Future Earth*, Palgrave Macmillan, Cham, pp. 177-199, doi: 10.1007/978-3-031-41606-4_9.

Panda, A. (2022), "Effective organizational leadership in Indian context", in: Pandey, A., Budhwar, P. and Bhawuk, D.P.S. (Eds) *Indigenous Indian Management. Palgrave Studies in Indian Management* Palgrave Macmillan, Cham, pp. 231-269, doi: 10.1007/978-3-030-87906-8_8.

Pandey, A., Gupta, R.K. and Arora, A.K. (2009), "Spiritual climate of business organizations and its impact on customers' experience", *Journal of Business Ethics*, Vol. 88 No. 2, pp. 313-332, doi: 10.1007/s10551-008-9965-z.

Pandey, A., Gupta, R.K. and Kumar, P. (2016), "Spiritual climate and its impact on learning in teams in business organizations", *Global Business Review*, Vol. 17 No. 3_suppl, pp. 1S-14S.

Pareek (2003), *Training Instruments for Human Resource Development*, McGraw Hill, New Delhi.

Pawar, B.S. (2019), *Employee Performance and Well-being: Leadership, Justice, Support, and Workplace Spirituality*, 1st ed., Routledge India, London, doi: 10.4324/9780429244193.

Petchsawang, P. and Duchon, D. (2009), "Measuring workplace spirituality in an Asian context", *Human Resource Development International*, Vol. 12 No. 4, pp. 459-468.

Philip, J., Najmi, H., Orudzheva, L. and Struckell, E. (2017), "Work life overlap in the millennial generation: the role of ubiquitous technology", *Journal of Strategic Innovation and Sustainability*, Vol. 12 No. 1, pp. 80-98.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y. and Podsakoff, N.P. (2003), "Common method biases in behavioural research: a critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88 No. 5, pp. 879-903, doi: 10.1037/0021-9010.88.5.879.

Prabhavananda, S. (2019), *The Spiritual Heritage of India*, Routledge, London.

Pradhan, S. (2022), "Work stressors and job outcomes: an empirical investigation of frontline service employees in the Indian hospitality industry", *Evidence-Based HRM: a Global Forum for Empirical Scholarship*, Vol. 10 No. 1, pp. 35-52, doi: 10.1108/EBHRM-11-2020-0143.

Premchandran, R. and Priyadarshi, P. (2020), "Work-family enrichment: evidence from India on source attribution", *International Journal of Manpower*, Vol. 41 No. 8, pp. 1269-1285, doi: 10.1108/ijm-07-2018-0235.

Rāmsukhdāsjī, S. (2007), *Śrīmad Bhagavad Gītā: Sādhaka-Sañjivanī*, Gītā Press, Gorakhpur.

Rana, M. and Saxena, P. (2024), "Encouraging workplace spirituality through its evolution, features and components", *Educational Administration: Theory and Practice*, Vol. 30 No. 5, pp. 5225-5233, doi: 10.53555/kuey.v30i5.3769.

Ranjan Chatterjee, S. (2009), "Managerial ethos of the Indian tradition: relevance of a wisdom model", *Journal of Indian Business Research*, Vol. 1 Nos 2/3, pp. 136-162.

Rapaport, R., Bailyn, L., Fletcher, J.K. and Pruitt, B.H. (2002), *Beyond Work –Family Balance*, Jossey-Bass, San Francisco.

Rastogi, M., Rangnekar, S. and Rastogi, R. (2016), "Flexibility as a predictor of work–family enrichment", *Global Journal of Flexible Systems Management*, Vol. 17 No. 1, pp. 5-14.

Rastogi, M., Rangnekar, S. and Rastogi, R. (2018), "Enhancing quality of work life in India: the role of workplace flexibility", *Industrial and Commercial Training*, Vol. 50 No. 5, pp. 234-249.

Sarkar, A. (2022), "Deconstructing ahimsa at the workplace: its relevance and significance", *South Asian Journal of Business Studies*, Vol. 11 No. 3, pp. 260-275, doi: 10.1108/SAJBS-05-2020-0144.

Sarkar, A. and Garg, N. (2021), "A multilevel model exploring the relationships between individual spirituality, spiritual climate, and nonviolent work behaviour", *Conflict Resolution Quarterly*, Vol. 38 No. 3, pp. 209-229, doi: 10.1002/crq.21298.

Sarkar, A., Garg, N. and Palo, S. (2022), "Workplace spirituality and psychological capital (PsyCap) of the employees: a cross-sectional study", *Global Business Review*, doi: 10.1177/09721509221111445.

Saxena, A., Garg, N., Punia, B.K. and Prasad, A. (2020), "Exploring role of Indian workplace spirituality in stress management: a study of oil and gas industry", *Journal of Organizational Change Management*, Vol. 33 No. 5, pp. 779-803, doi: 10.1108/jocm-11-2019-0327.

Sebastian, C.D. (2003), "The essential teachings of Bhagavad Gita", *Jnanatirtha: International Journal of Sacred Scriptures*, Vol. 3 No. 2, pp. 193-203.

Sharma, N. (2022), "Impact of workplace spirituality in reducing emotional labour among academics", in Garg, N. and Punia, B. (Eds), *Handbook of Research on Integrating Spirituality in Modern Workplaces*, IGI Global, Hershey, PA, USA, pp. 307-339.

Sharma, S. (1999), "Corporate Gita: lessons for management, administration and leadership", *Journal of Human Values*, Vol. 5 No. 2, pp. 103-123.

Sharma, N. and Garg, N. (2024), "Workplace spirituality and organisational effectiveness in the Indian higher education sector: exploring the relationship and mediating role of positive organisational scholarship factors", *International Journal of Indian Culture and Business Management*, Vol. 32 No. 3, pp. 413-439, doi: 10.1504/IJICBM.2024.140098.

Snyder, C.R., Irving, L. and Anderson, J. (1991), "Hope and health: Measuring the will and the ways", In Snyder C.R. and Forsyth D.R. (Eds), *Handbook of Social and Clinical Psychology: The Health Perspective*, Pergamon Press, UK, pp. 285-305.

Sode, R. and Chenji, K. (2024), "The mediating role of workplace spirituality: exploring the relationship between, self-transcendence, spiritual transcendence, and innovative work behavior", *Acta Psychologica*, Vol. 245, p. 104228, doi: 10.1016/j.actpsy.2024.104228.

Solat, S., Abrar, M., Shabbir, R., Bashir, M., Saleem, S. and Saqib, S. (2020), "Workplace interactional demands and Work-Family enrichment: an investigation from the service sector", *Frontiers in Psychology*, Vol. 11, p. 1476, doi: 10.3389/fpsyg.2020.01476.

Stajkovic, A.D. and Luthans, F. (1998), "Self-efficacy and work-related performance: a metaanalysis", *Psychological Bulletin*, Vol. 124 No. 2, pp. 240-261, doi: 10.1037/0033-2909.124.2.240.

Tiger, L. (1979), *Optimism: The Biology of Hope*, Simon and Schuster, New York.

Tilak, B.G. (1915), *Shrimad Bhagavad Gita Rahasya*, (26th ed.). Tilak Brothers, Pune.

Vasconcelos, A.F. (2018), "Workplace spirituality: empirical evidence revisited", *Management Research Review*, Vol. 41 No. 7, pp. 789-821, doi: 10.1108/mrr-07-2017-0232.

Wang, P., Chen, X., Gong, J. and Jacques-Tiura, A.J. (2014), "Reliability and validity of the personal social capital scale 16 and personal social capital scale 8: two short instruments for survey studies", *Social Indicators Research*, Vol. 119, pp. 1133-1148, doi: 10.1007/s11205-013-0540-3.

Watkins, P.C., Uhder, J. and Pichinevskiy, S. (2014), "Grateful recounting enhances subjective well-being: the importance of grateful processing", *The Journal of Positive Psychology*, Vol. 10 No. 2, pp. 91-98, doi: 10.1080/17439760.2014.927909.

Wood, A.M., Froh, J.J. and Geraghty, A.W. (2010), "Gratitude and well-being: a review and theoretical integration", *Clinical Psychology Review*, Vol. 30 No. 7, pp. 890-905, doi: 10.1016/j.cpr.2010.03.005.

Woolcock, M. (1998), "Social capital and economic development: toward a theoretical synthesis and policy framework", *Theory and Society*, Vol. 27 No. 2, pp. 151-208, available at: www.jstor.org/stable/657866

Further reading

Afsar, B. and Rehman, M. (2015), "The relationship between workplace spirituality and innovative work behaviour: the mediating role of perceived person–organization fit", *Journal of Management, Spirituality and Religion*, Vol. 12 No. 4, pp. 329-353, doi: 10.1080/14766086.2015.1060515.

Agrawal, M. and Mahajan, R. (2021), "Work–family enrichment: an integrative review", *International Journal of Workplace Health Management*, Vol. 14 No. 2, pp. 217-241, doi: 10.1108/ijwhm-04-2020-0056.

Bochner, S. and Brockwood, K.J. (2003), "Organisational culture and climate", in O'Driscoll, M., Taylor, P. and Kalliath, T. (Eds), *Organisational Psychology in Australia and New Zealand*, Oxford University Press, Melbourne, (PDF) The Relationship Between Workplace Spirituality and Work-to-Family Enrichment in Selected Public Sector Organizations in Malaysia, pp. 302-318.

Fischer, A. and LaFrance, M. (2015), "What drives the smile and the tear: why women are more emotionally expressive than men", *Emotion Review*, Vol. 7 No. 1, pp. 22-29.

Lingard, H., Francis, V. and Turner, M.C. (2010), "Work–family enrichment in the Australian construction industry: implications for job design", *Construction Management and Economics*, Vol. 28 No. 5, pp. 467-480, doi: 10.1080/01446190903511268.

Locklear, L.R., Sheridan, S. and Kong, D.T. (2022), "Appreciating social science research on gratitude: an integrative review for organizational scholarship on gratitude in the workplace", *Journal of Organizational Behavior*, Vol. 44 No. 2, pp. 225-260, doi: 10.1002/job.2624.

Lyness, K.S. and Judiesch, M.K. (2014), "Gender Egalitarianism and work-life balance for managers. Multisource perspectives in 36 countries", *Applied Psychology*, Vol. 63 No. 1, pp. 96-129.

Möller-Leimkühler, A.M. (2002), "Barriers to help-seeking by men: a review of sociocultural and clinical literature with particular reference to depression", *Journal of Affective Disorders*, Vol. 71 Nos 1/3, pp. 1-9.

Offermann, L.R. and Gowing, M.K. (1990), "Organizations of the future: changes and challenges", *American Psychologist*, Vol. 45 No. 2, pp. 95-108, doi: 10.1037/0003-066x.45.2.95.

Siu, O.L., Lu, J.F., Brough, P., Lu, C.Q., Bakker, A.B., Kalliath, T., O'Driscoll, M., Phillips, D.R., Chen, W.Q., Lo, D. and Sit, C. (2010), "Role resources and work–family enrichment: the role of work engagement", *Journal of Vocational Behavior*, Vol. 77 No. 3, pp. 470-480, doi: 10.1016/j.jvb.2010.06.007.

Stajkovic, A.D. (2006), "Development of a core confidence-higher order construct", *Journal of Applied Psychology*, Vol. 91 No. 6, pp. 1208-1224, doi: 10.1037/0021-9010.91.6.1208.

About the authors

Naval Garg is Assistant Professor (HR and OB) at Delhi Technological University, New Delhi, India. He has published more than 50 research papers in various national and international journals of repute. His areas of interest are workplace spirituality, gratitude and compassion.

Nidhi Sharma is Assistant Professor at Kirori Mal College, University of Delhi, India. She is an alumnus of Shri Ram College of Commerce and Delhi School of Economics. As a researcher, her interest lies in the area of HRM and OB. She has published quantitative and qualitative research papers on organisational effectiveness, workplace spirituality and positive psychology in leading national and international journals. Nidhi Sharma is the corresponding author and can be contacted at: nidhisharma810@gmail.com

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**

Physica Scripta

PAPER

Effect of Dy^{3+} ions on physical, structural and thermal properties of ZnNaPbB glasses

RECEIVED
19 February 2024

REVISED
22 May 2024

ACCEPTED FOR PUBLICATION
11 September 2024

PUBLISHED
23 September 2024

Manjeet[1], **Anu**[2], **Nisha Deopa**[3], **Anand Kumar**[3] and **Rishi Pal Chahal**[1]

[1] Department of Physics, Chaudhary Bansi Lal University, Bhiwani 127021, Haryana, India
[2] Department of Applied Physics, Delhi Technological University, New Delhi 110042, India
[3] Department of Physics, Chaudhary Ranbir Singh University, Jind 126102, Haryana, India

E-mail: rishiphy@gmail.com

Keywords: FT-IR, borate glasses, dysprosium, decay spectroscopy

Abstract

In this report, inspired by the excellent spectroscopic characteristics of dysprosium (Dy^{3+}) ions doped borate glasses, a series with the composition (70-x)B_2O_3-10ZnO-10PbO-10Na_2CO_3-xDy_2O_3 (here 'x' have the values as 0.1, 0.5, 1.0, 1.5 and 2.0 in mol %) was made using the melt quenching process and named as ZnNaPbBDy0.1 to ZnNaPbBDy2.0. Here we study the physical, structural and thermal properties of Dy^{3+} ions doped Zinc Sodium Lead Borate (ZnNaPbB) glasses through different characterization techniques like Fourier transform infrared (FT-IR), Raman spectroscopy, Differential scanning calorimetry (DSC) and decay spectroscopy. Several physical parameters were evaluated to assess the effect of Dy^{3+} inclusion. FT-IR and Raman spectra confirm the presence of BO_3 and BO_4 structural units. The data recorded by DSC spectra have been used to predict the glass transition temperature (T_g). The nature of decay spectra was estimated which diverges from exponential to non-exponential at higher concentrations. The decay spectral data is used to calculate the experimental lifetimes for all the glass samples. The Inoakutti Hirayama (I-H) model and Dexter's plot are used to confirm the type of interaction between the donor and acceptor Dy^{3+}- Dy^{3+} ions is dipole–dipole in nature that is responsible for the emission intensity quenching with concentration. All the above results might be advantageous for variety of applications in photonic devices.

1. Introduction

Currently, glasses are found to be a perfect substitute for white LED (light emitting diodes) because they are favorable hosts for Rare Earth (RE) elements. Glasses doped with RE ions are of great significance due to their unique spectral and luminescence characteristics in the visible region through 4f-4f transitions and are more suitable for several high tech and environmental applications. RE-doped glasses offer superior properties like high transparency, low power consumption and low thermal expansion, making them popular in optoelectrical applications like solid-state lasers, sensors, and display devices [1–4]. The oxide glasses, such as phosphate, borosilicate, telluride, borate and silicate are stable matrices so if they are doped with RE ions (trivalent) emit strong luminescence and utilized in a variety of optoelectronic applications [5–7]. It has been found that of the glasses mentioned above, the glasses that contains boric oxide are more suitable as host glass matrices for the addition of RE ions as boric oxide is a significant glass fluxing and forming ingredient due to their better thermal soluble nature, low melting point, high viscose agent and sufficient solubility of RE ions. Borate glasses are more interesting than other oxide glasses because they have two different forms of B atom co-ordinations. The three-dimensional network of borate glasses is interconnected via the B-O-B links that connect with BO_3 triangles and transform into di, tri, tetra, and penta borate units when a modifier is added. This structural versatility offers unique advantages, particularly when incorporating RE ions for luminescent materials. The diverse borate network allows for precise control over the local environment of the RE ions, influencing their luminescent properties. Borate glasses can provide a suitable environment for RE ions to exhibit intense luminescence with high quantum efficiency [8, 9]. Over the years, multiple researchers have investigated the

© 2024 IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

luminescence and other properties of RE ions in several host matrixes. For white light emission applications, many borate glasses with RE ions doping have been investigated up to this point such as zinc-sodium-bismuth-borate (Hegde *et al* 2019) [10], lithium-strontium-zinc-borate (Ahmad *et al* 2020) [11], lead-bismuth-borate (Divina *et al* 2021) [12], sodium-zinc-borate (Hemalatha *et al* 2022) [13], fluoride bismuth borate (Dahiya *et al* 2022) [14], Bi-B (Mukamil *et al* 2022) [15], barium-borate and lithium-alumino-borate (Grune *et al* 2022) [16] and ZnKAlB (Manjeet *et al* 2022) [17] glasses. However, the borate glasses have a high phonon energy range (1300–1500 cm^{-1}) therefore to reduce this energy we may add some heavy metal oxides like lead oxide (PbO), zinc oxide (ZnO) and sodium oxide (Na_2O). Such low-phonon-energy glasses can enhance radiative emissions by reducing non-radiative relaxation and thus can improve quantum efficiency. The lower phonon energy glass materials are more suitable for a range of luminescence applications. Indeed, lead oxide is widely recognized as one of the most effective network modifiers for borate glasses. Its addition to glass compositions can significantly alter their properties, making it a popular choice in the production of commercial low-temperature glasses. Lead oxide plays a multifaceted role in glass formation. In early studies, it was observed that PbO can act both as a network former and a network modifier, depending on the glass composition and processing conditions. The combination of lead oxide with RE oxides in borate glasses can lead to synergistic effects, influencing the optical and luminescent properties of the glasses [18]. Zinc oxide indeed plays a unique structural role in many oxide glasses, offering a dual behavior as it can be glass-forming or glass-modifying. As zinc oxide has excellent thermal conductivity, high refractive index and UV protection, it has several uses in material science. Glasses become observably sublime, covalent and thermally stable when ZnO is added, which may reduce the time required to solidify the glasses [19, 20]. Additionally, the tendency of the borate glasses to form glass can be improved by adding alkali ions at the right concentration. Any alkali oxide, like Na_2O, can cause the structure to tighten or compact by converting BO3 triangular units into BO4 tetrahedral ones. The combination of alkali Na_2O, PbO and ZnO with B_2O_3 as a network former can be used as a proficient optical host material for Dy^{3+} ions. Thus, we have synthesized a Zinc Sodium Lead Borate (ZnNaPbB) glass as a noble glass host, motivated by the research findings mentioned above. Furthermore, the RE elements having +3 valence state exhibit luminescence properties and the valence states with +4 and +2 oxidation numbers are stable states mainly because of fully empty, half-full and fully-filled shells such as $4f^{\circ}$, $4f^{7}$, and $4f^{14}$ for some RE elements. RE elements possess a unique property that makes them different from other optically active ions like transition elements. Doping of trivalent RE ions in borate glasses has a variety of advantages like good transparency, low cost easy manufacturing and isotopic refractive index. The Dy^{3+} ions give PL spectra consisting of two intense bands of blue emission around (470–500 nm) and yellow emission around (570–600 nm) parallel to transitions ($^4F_{9/2} \rightarrow {}^6H_{15/2}$) and ($^4F_{9/2} \rightarrow {}^6H_{13/2}$), respectively [13, 21]. To generate white light emission with an acceptable yellow-to-blue (Y/B) intensity ratio, researchers must tune the local field environment of Dy^{3+} ions doped glasses. The suitable luminescence intensity ratio (Y/B) is obtained due to the optimization of the local field environment that is glass chemical composition and by altering the amount of Dy^{3+} ions concentration in the glass samples. Amorphous and crystalline materials that have been doped with Dy^{3+} ions exhibit fluorescence in the yellow to orange (570–590 nm) region which is advantageous for white LEDs, lasers and telecommunications [1, 2, 16]. Hegde *et al* in their research [10] reported the luminescent and photoemission features of Dy^{3+} ions doped ZNBB glasses. They observed luminescence property at 387 nm excitation wavelength for yellow and blue intense emission around 480 nm and 583 nm, respectively. Recently, Ahmad *et al* [11] reported spectroscopic, radiative and colorimetric characteristics of borate glasses with doping of Dy^{3+} ions. The absorption spectra of their glasses consist of 12 bands in all and the hypersensitive band was observed at 1270 nm. Zaman *et al* [22] study the luminescence comparison of borate glasses (i.e., NMgBiB and NBaBiB) with Dy^{3+} ions for white light emission. Divina *et al* [12] reported Dy^{3+} ions as an optical probe to examine the luminescent characteristics of alkali Pb-Bi borate glasses for white light emission. They make efforts to show the difference that occurs in the structure of the host network with the addition of different modifiers. Manjeet *et al* [17] reported various structural, optical and photoluminescence studies of Dy^{3+} ions doped ZnKAlB glass samples. They reported that the inclusion of network modifiers and rare earth ions improved the physical properties and thermal stability of the prepared glasses. Hemalatha *et al* from the literature [13] reported the insertion of RE ions in sodium-zinc-borate glasses to study spectroscopic and white light-emitting characteristics. These studies show that the borate glass with the proper modifier has a key influence in enhancing the luminescence characteristics of Dy^{3+} ions, which leads to enhanced white light emission for w-LED applications [2, 12, 22]. We are motivated to thoroughly examine the physical, structural, and thermal characteristics of the Dy^{3+} ion doped ZnNaPbB glasses because the majority of the research work in this area has concentrated on the analysis of certain finite features of glasses. Before using any glass system for optical or other technical applications, it is essential to understand its basic physical, structural, and thermal features. From this angle, our main objective is to investigate the effects of doping of Dy^{3+} ions in the as-prepared glasses on their thermal, structural and physical traits. Thus, the present research is to examine the structural and physical modifications along with thermal stability of as-prepared glasses by varying the quantity of dysprosium oxide and the glass network environment.

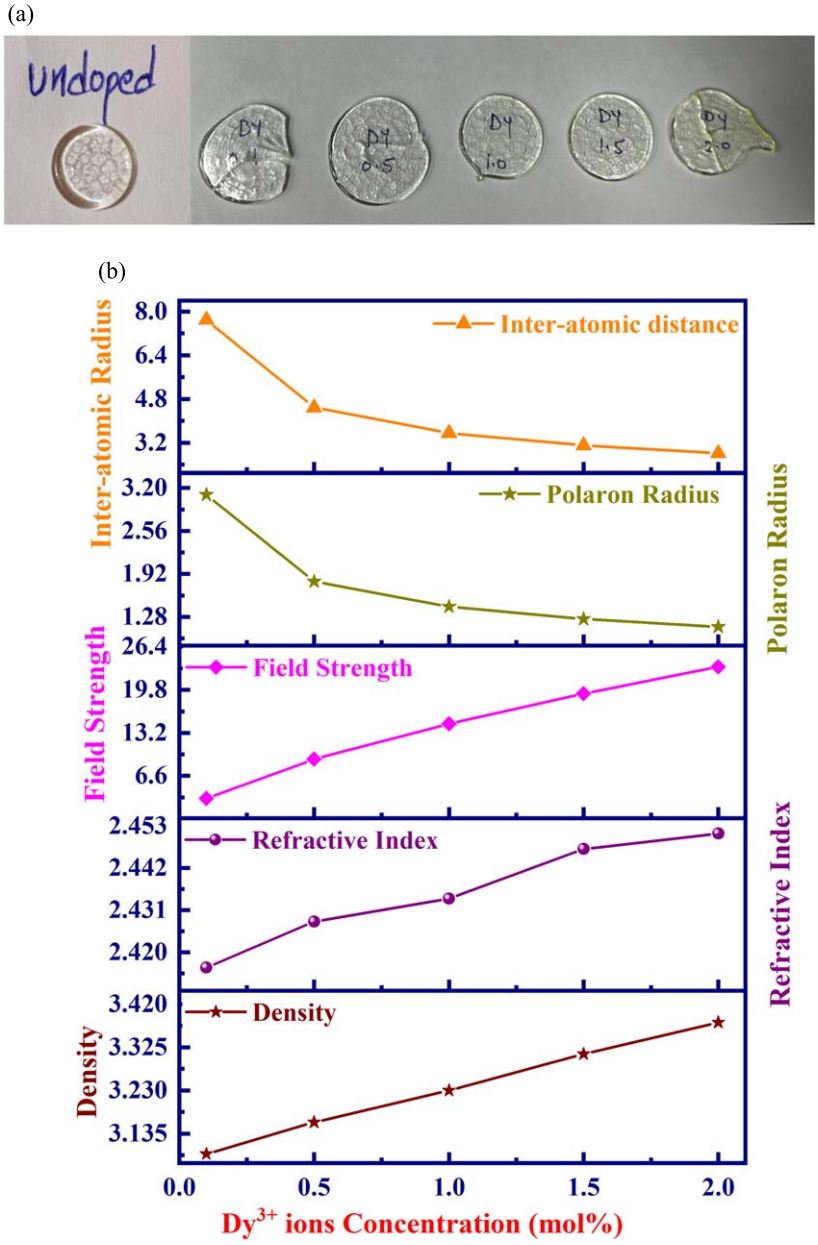


Figure 1. (a): Pictures of all Dy^{3+} ions doped ZnNaPbB glasses. (b): Variation of density, refractive index (n), field strength, inter-ionic distance (Å) and polaron radius with increasing Dy^{3+} ions concentrations in ZnNaPbB glass.

2. Material and characterizations

In this work, a series of six glass samples were produced with the composition: (70-x) B_2O_3-10ZnO-10PbO-10Na_2CO_3-xDy_2O_3, here x has values from 0.0, 0.1, 0.5, 1.0, 1.5 & 2.0 (in mol %) using the melt quenching method. For our ease, the glasses are named ZnNaPbB, ZnNaPbBDy0.1 to ZnNaPbBDy2.0, respectively. The glass sample is produced by using pure chemical reagents including boric acid H_3BO_3, PbO, ZnO and Na_2CO_3, Dy_2O_3. The process involves grinding chemicals in an agate mortar to create a homogenous powder of samples. The homogenous powder is heated in an electric furnace at 1250 °C for 2 h and 30 min, forming a bubble-free liquid. The melted form of the sample is quenched between the brass plates to create transparent glasses of nearly 2 mm thickness. After annealing at 350 °C for 3 h, the transparent colorless ZnNaPbBDy glass samples were obtained in solid form, and are used for their physical, structural and thermal studies. All of the ZnNaPbBDy glasses are shown in figure 1(a) are clear and round in form.

The refractive indices of the prepared glass samples have been calculated using Brewster's angle method. FT-IR spectra were observed from the Perkin Elmer FT-IR spectrometer. The thermo-grams of the synthesized ZnNaPbBDy glasses have been recorded from the DSC of model number Q600 SDT with a heating rate of

Table 1. Physical parameters for Dy^{3+} ions doped ZnNaPbB glasses.

Physical Parameters	Concentration of Dy^{3+} ions in mol%				
	ZnNaPbB Dy0.1	ZnNaPbB Dy 0.5	ZnNaPbB Dy 1.0	ZnNaPbB Dy 1.5	ZnNaPbB Dy2.0
Thickness (t) (in cm)	0.17	0.17	0.16	0.17	0.16
Density (d) (in g/cm^3)	3.09	3.16	3.23	3.31	3.38
Average molecular weight (M_{avg}) (in g/mol)	84.65	85.89	87.45	89.01	90.56
Refractive Index (n)	2.416	2.428	2.434	2.447	2.451
Concentration of Dy^{3+}(N) ($\times 10^{22}$ ions cm^{-3})	0.220	1.108	2.224	3.359	4.495
Dielectric Constant (ε)	5.838	5.897	5.925	5.989	6.006
Optical dielectric constant ($\varepsilon-1$)	4.838	4.897	4.925	4.989	5.006
Molar Volume (V_m) (in cm^3/mol)	27.394	27.180	27.074	26.890	26.793
Field Strength (F) ($\times 10^{15}$cm^{-2})	3.123	9.179	14.610	19.231	23.353
Inter-atomic distance (r_i) (in Å)	7.691	4.486	3.556	3.099	2.812
Polaron radius (r_p) (in Å)	3.099	1.808	1.433	1.249	1.133
Reflection Loss (R_L)	0.1719	0.1736	0.1744	0.1763	0.1768
Molar refraction (R_m) (in cm^{-3})	16.909	16.855	16.825	16.793	16.754
(R_m/V_m)	0.6172	0.6201	0.6215	0.6245	0.6253
Metallization (M)	0.3828	0.3799	0.3785	0.3755	0.3747
Oxygen packing density (OPD)	401.54	404.70	406.29	409.08	410.55
Molecular electronic polarizability ($\times 10^{-23}$)	6.703	1.336	0.667	0.444	0.332
Optical Electronegativity (χ_{opt})	0.875	0.864	0.859	0.848	0.845

10 °C min^{-1}. Raman spectra of glass samples were recorded by the Raman spectrometer (model Alpha300 Make-WITec). Further, the decay spectra of prepared glass samples were recorded via Horiba quanta master.

3. Results and discussion

3.1. Physical parameters

The various physical parameters such as molar volume, dielectric constant, field strength, inter-atomic distance, metallization, polaron radius and reflection losses were evaluated for the synthesized glass samples using experimentally calculated density and refractive index from the expressions used in the literature [9, 11, 23, 24] and tabulated in table 1. Then the thickness of synthesized ZnNaPbBDy glasses was recorded using a vernier caliper and found to be in the range of 0.16–0.17 cm. The changes in the glass structure brought about by compositional differences may be investigated through density measurement since they are impacted by the change in structural compactness and coordination number with the addition of various modifiers. The density of the synthesized glass samples was determined by the Archimedes principle at room temperature and considering distilled water as a buoyant liquid. The calculated value of density rises gradually with the increase of Dy^{3+} ions concentration as presented in figure 1(b). The reason for the increase in density is the replacement of borate ions having lower density and molecular masses with dysprosium ions of higher molecular masses. The Average molecular weight (M_{avg}, in g/mol) of the synthesized samples rises with the inclusion of heavy Dy^{3+} ions. The calculated values of M_{avg} for all the ZnNaPbBDy glasses are listed in table 1. The refractive index assesses a material's light-transmitting capacity. The refractive index of a material depends upon its density and observed values increase slightly with the increase in density and Dy^{3+} ions in the as-prepared glass samples. Figure 1(b) depicts the change in refractive index by adding Dy^{3+} ions concentration in the ZnNaPbB glass sample. The parameters directly perceived from refractive indices such as dielectric constant (ε), reflection losses (R_L) and optical dielectric constant ($\varepsilon-1$) increase with the increase of Dy^{3+} ions content, as presented in table 1. Further, the inclusion of Dy_2O_3 consequences increases the Dy^{3+} ion concentration (N) of ZnNaPbB glasses. The molar volume (V_m) decreases with Dy^{3+} ions concentration, which is consistent with a decrease in bond length or interatomic separation between the atoms, as seen in table 1. The calculated values of field strength continuously rise with the inclusion of Dy^{3+} ions concentration whereas the values of inter-ionic distance and polaron radius decrease, as shown in table 1. Figure 1(b) depicts the variation of the above-said factors with rising Dy^{3+} ion concentrations in ZnNaPbB glass. The average inter-ionic distance between Dy^{3+} ions and oxygen is smaller, which raises the relative bond strengths and, as a result, enhances the field strength near the Dy^{3+} ions in the glass network [10]. The calculated value of molar refraction for the synthesized glasses reduces with the increases in Dy^{3+} ions is an indication of the formation of the NBOs with Dy_2O_3 content in the host network [25]. The statement provided by Dimitrov and Komatsu [26] in their research is used to identify the nature of glasses that is if the R_m to V_m ratio is smaller than one reflects a non-metallic nature but if the ratio is

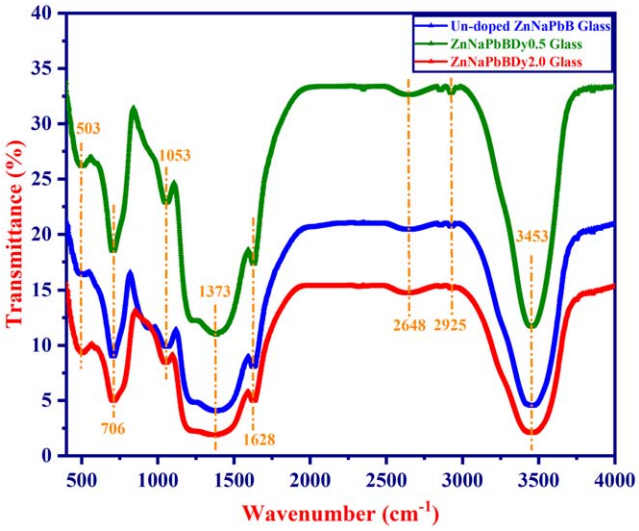


Figure 2. FT-IR spectra of an un-doped ZnNaPbB and doped (ZnNaPbBDy0.5 & ZnNaPbBDy2.0) glasses.

more than one reflects the metallic nature. Here, the estimated values of the R_m to V_m ratio were found to be smaller than one which shows the non-metallic nature of present glass samples. Further, metallization (M) is another parameter to identify the insulating characteristics of the glass samples and it is obtained by deducting the ratio R_m to V_m from one [27]. The observed values of 'M' for the synthesized ZnNaPbBDy glasses are less than one and in decreasing order consequences the extent of the conduction band is increased hence the optical band gap is reduced [28]. The oxygen packing density (OPD) for as-prepared glasses lies in between 400 to 410 and continuously increases with the inclusion of Dy^{3+} ions content, as listed in table 1. The molecular electronic polarizability and optical electronegativity (χ_{opt}) depend upon the refractive index of glasses and both decrease with the increase of doping ions in the host network. All the above-said parametric orders are in concurrence with published research in this field with Dy^{3+} ions doped borate glasses [24, 29, 30].

3.2. FT-IR and Raman spectroscopy

FT-IR and Raman spectroscopies are generally used to find the pieces of evidence of the specific structural units and different functional groups presented in the as-prepared glass samples. FT-IR spectra have been recorded for the un-doped ZnNaPbB, ZnNaPbBDy0.5 and ZnNaPbBDy2.0 glass samples at room temperature for the range of 4000 to 400 cm^{-1} wavenumber and displayed in figure 2. All three spectra are relatively similar in band positions and each spectrum contains eight intense bands centered about 503 cm^{-1}, 706 cm^{-1}, 1053 cm^{-1}, 1373 cm^{-1}, 1628 cm^{-1}, 2648 cm^{-1}, 2925 cm^{-1} and 3453 cm^{-1} which confirms the presence of Pb-O & ZnO_4 bond vibrations, different vibrations of borate unit and hydrogen bonding. The present bands are recognized and assigned from the previously published research [5, 8, 17, 31] and are recorded in table 2. The origin of molecular vibrational bands leads to out-of-plane molecule deformation, including stretching, bending, twisting, rocking, and wagging, causing a shift in the wavenumber of the infrared spectrum. The variations of the molecular dipole moments are the main reason for the bands observed in the FT-IR spectra [14]. Approximation of OH content in a glass sample is another valuable use of FT-IR spectroscopy. It is important to find the existence of OH content in glasses since it affects the glass viscosity, thermal expansion as well the reduction of radiative decay, which lowers quantum efficiency [32]. Mostly, a glass with a high concentration of OH groups will have more traps, cracks and impurities which can reduce the luminous intensity and overall strength of as-prepared glasses. Thus, it is important to confirm their presence in the synthesized glasses which can be estimated from the equation given below [32]:

$$\alpha_{OH} = \frac{\ln\left(\frac{T_0}{T_D}\right)}{l} \quad (1)$$

T_o use here is the value of the highest transmission, T_D denotes the transmission at 3000 cm^{-1} and l is the width of glass samples. For a glass to have high quantum efficiency, its OH content needs to be as low as possible. The calculated values of OH content (α_{OH}) from the above expression of un-doped ZnNaPbB, ZnNaPbBDy0.5 and ZnNaPbBDy2.0 glass samples are 0.908, 0.214 and 2.371 ppm, respectively. Among all these, the glass with 0.5 mol% Dy^{3+} ions possesses the lowest α_{OH} value resulting in high efficiency for optoelectronic applications [30, 32]. Further, to understand the surroundings, structure and dynamics of prepared glass samples, Raman

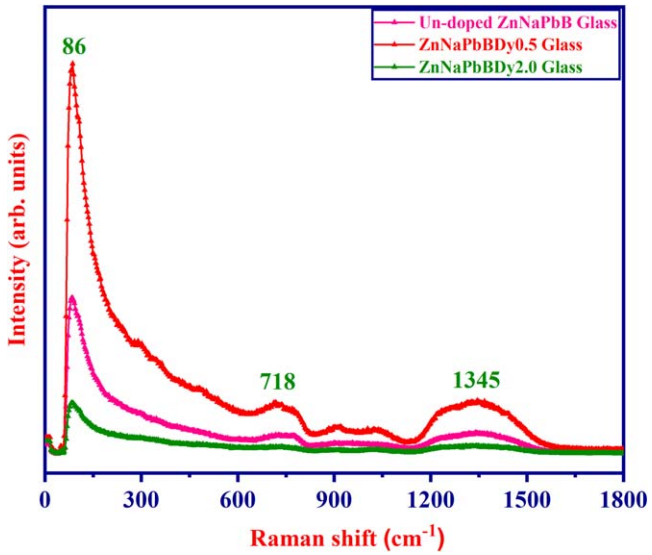


Figure 3. Raman spectra of un-doped ZnNaPbB and doped (ZnNaPbBDy0.5 & ZnNaPbBDy2.0) glasses.

Table 2 Assignment of identified FT-IR band positions of un-doped ZnNaPbB and Dy^{3+} ions doped ZnNaPbBDy0.5 and ZnNaPbBDy2.0 glasses.

Band Positions (cm^{-1})	Band assignments	References
503	Pb-O bond vibrations and vibrations of metal cation in ZnO_4 units.	[8]
706	The bending of B-O-B linkage in the borate networks and vibration of bridging oxygen atoms.	[8, 17]
1053	The B-O stretching vibration of tetrahedral [BO_4] structural units in di-borate group.	[8]
1373	The existence of asymmetric stretching of NBOs (non-bridging oxygen) of BO_3 units in Metaborate, Pyroborate, Orthoborate group.	[8, 17]
1628	The B-O bond stretching vibrations of borate units in which boron atoms are co-ordinated with three oxygen atoms.	[8, 17]
2648	Due the presence of hydrogen bonding	[8]
2930	Existence of Hydrogen Bond, organic contaminants.	[17]
3453	Symmetric stretching of hydrogen bond in O-H group.	[8, 17]

spectra were recorded. The Raman spectra for un-doped ZnNaPbB, ZnNaPbBDy0.5 and ZnNaPbBDy2.0 glass samples were recorded up to 1800 cm^{-1} and are displayed in figure 3. All three spectra are identical in band position but varying in band intensities, maximum intense peaks were observed for the ZnNaPbBDy0.5 glass. Each spectrum is composed of three peaks all of which one is highly intense at 86 cm^{-1} and two board bands near 718 cm^{-1} and 1345 cm^{-1}. The band at 86 cm^{-1} is attributed to the vibrations of Zn^{3+} metal ions, the broad band in the region 650 to 850 cm^{-1} is ascribed to the Zn–O bond stretching vibrations of ZnO_6 unit and the wide band in the region 1150 to 1550 cm^{-1} is due to the vibrations of the pyroborate group and B-O bond stretching of trigonal BO_3 units [32–34].

3.3. Thermal analysis

The DSC spectrum can be used to determine the transition temperature (T_g) and thermal stability of glass samples. Since it is well known that the variations in the coordination number of network-forming atoms or production of non-bridging oxygens (NBOs) will affect the T_g, thermal analysis of the glasses is more important. The temperature of a sample and a reference are raised using different quantities of heat in thermo-analytical techniques such as DSC, and the temperatures are then observed as a function of temperature. DSC thermographs of synthesized glasses named un-doped ZnNaPbB, ZnNaPbBDy0.5 and ZnNaPbBDy2.0 glasses were recorded for the temperature 200 °C to 900 °C and represented by figure 4. The figure also included the plot between the derivatives of heat flow versus temperature to determine the accurate values of T_g. It is noticeable in figure 4 that thermo-graphs of all three glass samples comprise a minima between 450 °C and 500 °C and from references [35] it has been discussed by the authors that the glass transition temperature (T_g) values of glass sample may be predicted by the existence of minima in the dH/dT versus temperature. The assessment of the present investigation is the T_g values of un-doped ZnNaPbB, ZnNaPbBDy0.5 and ZnNaPbBDy2.0 glass samples

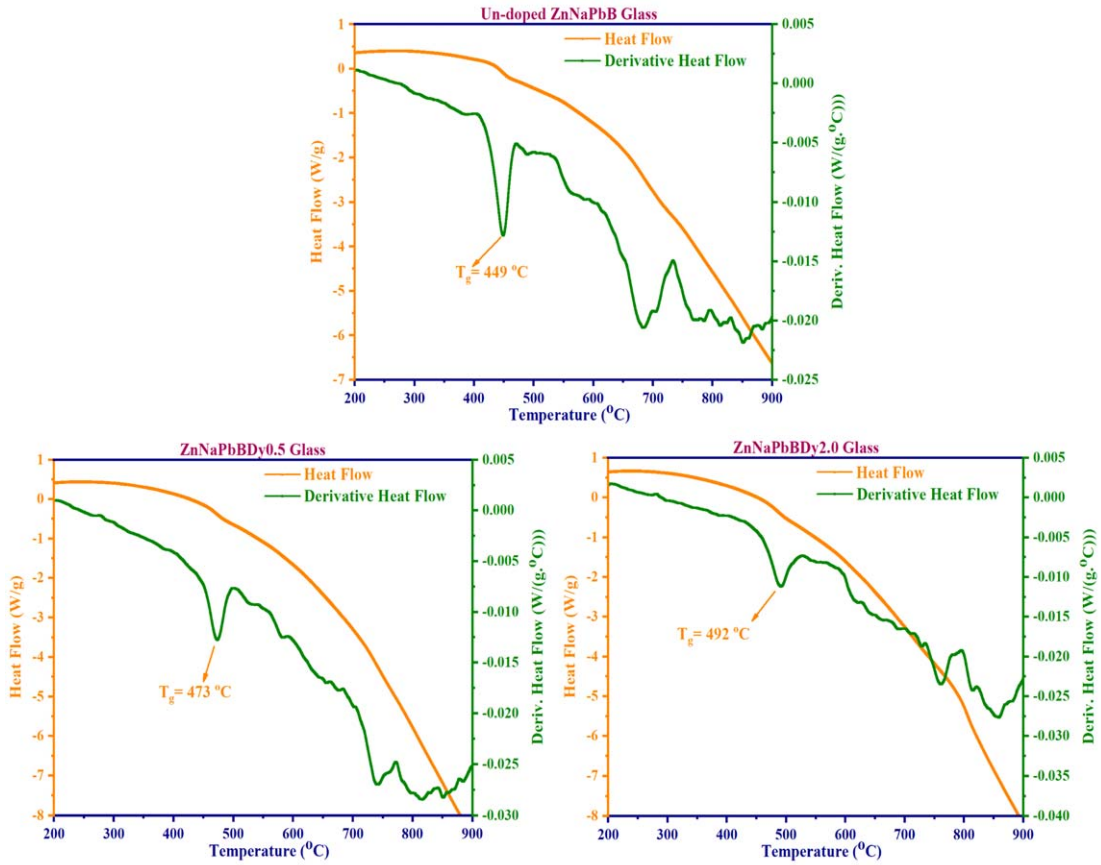


Figure 4. DSC thermo-grams for an un-doped ZnNaPbB and doped (ZnNaPbBDy0.5 & ZnNaPbBDy2.0) glasses.

are 449 °C, 473 °C and 492 °C, correspondingly. The variation in the T_g values is dependable on the structural changes in the network of glasses and the elevation in T_g can be ascribed to the partial replacement of B_2O_3 by Dy_2O_3 with a great number of cations [5, 13, 35].

3.4. Decay spectral studies

To examine the compatibility of as-prepared ZnNaPbBDy glasses for lasing medium and to measure the lifetime of fluorescence energy level $^4F_{9/2}$, the decay spectra were recorded at room temperature for the transition $^4F_{9/2} \rightarrow {}^6H_{13/2}$ by monitoring excitation and emission wavelength as 350 nm and 575 nm, respectively and presented in figure 5. Then, the decay data is fitted to different exponential decay equations and it has been found that a bi-exponential equation is the most suitable fitting for all doping concentrations. The decay data is fitted bi-exponentially using the following expression [36]:

$$I = I_0 + A_1 \quad \exp\left(-\frac{t}{\tau_1}\right) + A_2 \exp\left(-\frac{t}{\tau_2}\right) \tag{2}$$

where, I_0 and I is used for the PL intensities at the time 0 and t, respectively, A_1 and A_2 are fitting constants, τ_1 and τ_2 is used for the long and short lifetimes and t is time in ms. It was perceived that the profiles of all the synthesized glasses exhibit exponential behavior for lower concentration (up to 0.5 mol %) and it became non-exponential for higher concentration, as shown in figure 5. This behavioral change can be explained by energy transfer between excited-unexcited donor–acceptor Dy^{3+}- Dy^{3+} ions. The experimental decay lifetime (τ_{exp}) is calculated from the exponentially fitted curve by taking the e-folding time of the decay spectral intensity and the values are 678, 443, 239, 86 and 50 μs [37]. To determine the experimental lifetime for the as-prepared glass samples the following expression can be used [36]:

$$\tau_{exp} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \tag{3}$$

The values of τ_{exp} for synthesized glass samples decrease continuously with the increment of Dy^{3+} ions as shown in figure 6, it may arise from the energy transfer among Dy^{3+}-Dy^{3+} ions through the CR mechanism and

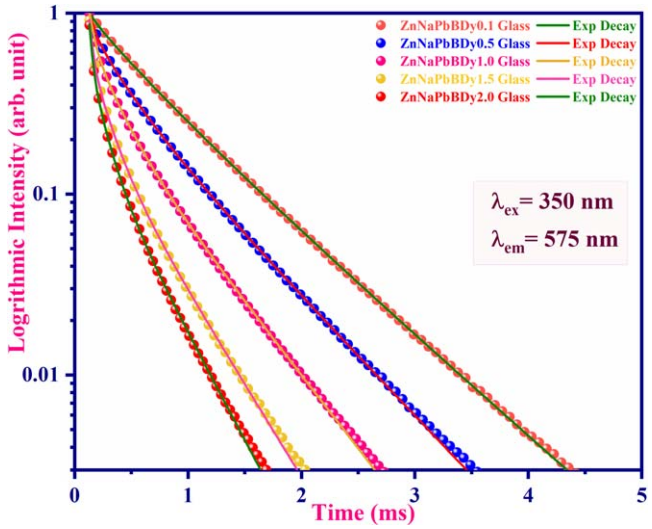


Figure 5. The decay spectra for $^4F_{9/2} \rightarrow {}^6H_{13/2}$ emission transition of Dy^{3+} ions doped ZnNaPbB glasses under 350 nm excitation wavelengths.

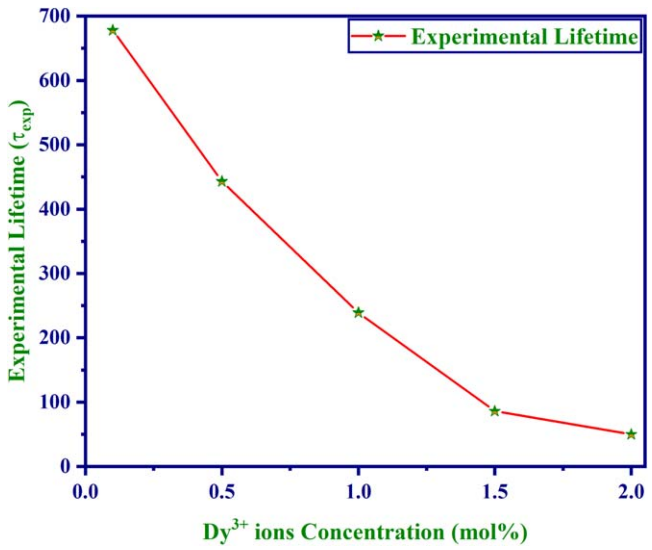


Figure 6. Variation of experimental lifetime (τ_{exp}) with Dy^{3+} ions concentration in ZnNaPbB glass.

consequential quenching in concentration. Our understanding of the non-exponential decay curves, which lower lifetime values and quench concentrations, might be understood by the CR process.

3.5. Energy transfer analysis and Dexter theory

The decrease in the intensity of PL peaks of the as-prepared samples is shown in [38] after a particular concentration specifically known as the critical concentration of doping ions resulting in the concentration quenching via resonant energy transfer (RET) and cross-relaxation channels (CRCs) amid the Dy^{3+}-Dy^{3+} ions exhibiting the self-quenching process. The partial energy level diagram of the ZnNaPbBDy glasses is displayed in figure 7, which explains the excitation, emission, non-radiative transition, RET and C-R mechanism. From figure 7, it is observed that the Dy^{3+} ions are excited to different exciting levels from the ground state ($^6H_{15/2}$) and then drop down to the meta-stable state ($^4F_{9/2}$) via a non-radiative process which results in the population inversion of this state. Usually, the RET mechanism occurs between the ground state ($^6H_{15/2}$) and the meta-stable excited state ($^4F_{9/2}$). The pertinent CRCs among the acceptor and donor Dy^{3+} ions of the present work are:

CRC1: $^4F_{9/2} + {}^6H_{15/2} \rightarrow ({}^6F_{11/2} + {}^6H_{9/2}) + {}^6F_{3/2}$

CRC2: $^4F_{9/2} + {}^6H_{15/2} \rightarrow ({}^6F_{9/2} + {}^6H_{7/2}) + {}^6F_{7/2}$

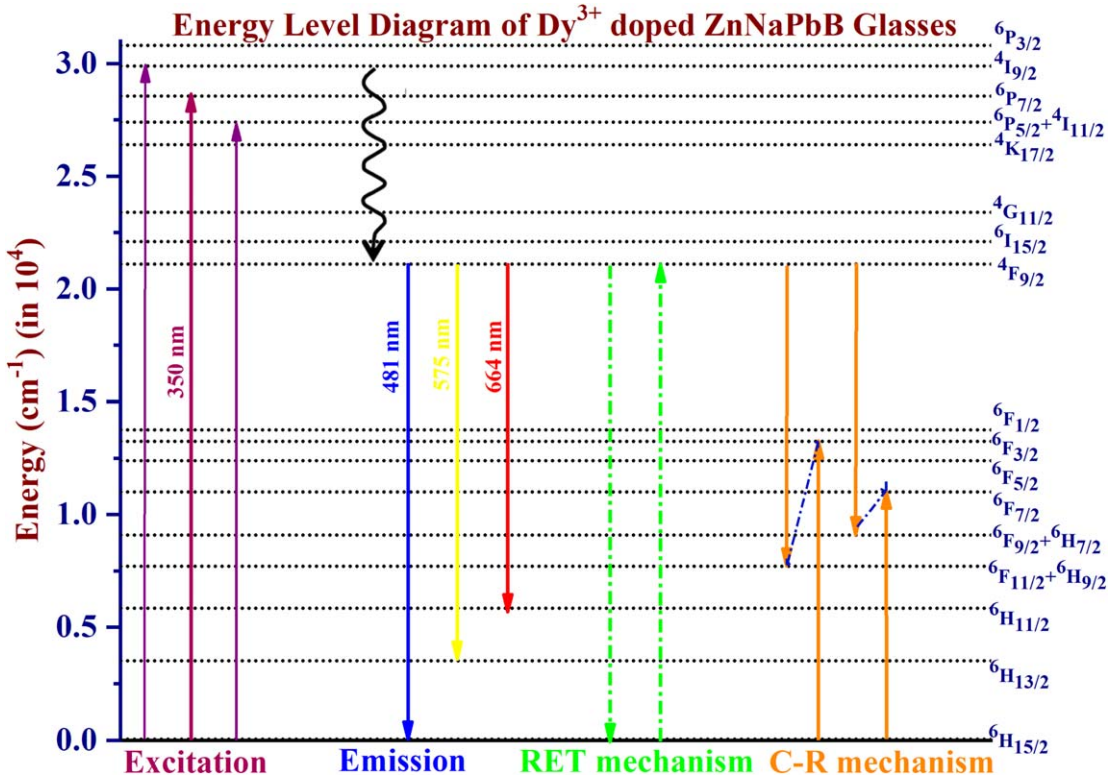


Figure 7. Energy level diagram of Dy^{3+} ions doped ZnNaPbB glasses representing radiative and non-radiative transitions.

RET: $^4F_{9/2} \rightarrow {}^6H_{15/2}$

Further, the non-radiative energy transfer processes and concentration quenching can be explained by multipole-multipole interaction between Dy^{3+}-Dy^{3+} ions. There are three types of multipole-multipole interactions Dipole–Dipole (D-D), Dipole-Quadrupole (D-Q) and Quadrupole-Quadrupole (Q-Q). The type of this multipole-multipole interaction can be identified by using Dexter's theory. Dexter established a correlation between doping concentration and emission intensity to examine the type of interaction. The mathematical expression developed by Dexter is as follows [39]:

$$\log\left(\frac{I}{x}\right) = \log C - \frac{s^{*}\log(x)}{d} \tag{4}$$

where d (=3) denotes the dimension of the compound, I is emission intensity, x is dopant ions concentration and C represents a constant that depends on dopant ions concentration. The nature of interaction among Dy^{3+} ions is identified by the 's' value, which can be 6, 8 and 10 for D-D, D-Q and Q-Q interactions, respectively [5, 30]. The PL emission data of the glasses under investigation are used from the [38] and hence a linear fitted graph between log *(I/x)* and log*(x)* is shown in figure 8. The slope of the fitted curve reflects the value of the (s/d) ratio which is −1.917 for the present study; therefore, the index (s) of electric multipole-multipole interaction is 5.75 which are close to 6. Consequently, the dominant interaction amid the Dy^{3+}-Dy^{3+} ions is dipole–dipole in nature for the synthesized ZnNaPbBDy glasses.

3.6. Inoakutti-Hirayama (I-H) Model

There are two ways for an exciting ion or molecule might return to its ground state. Firstly, direct energy transfer from donor ions to the acceptor ions. Secondly, before it reaches the acceptor ions, energy is transferred from the donor to the donor ion. The I-H model efficiently explains the mechanism of energy transfer and the non-exponential behavior of decay spectra at high concentrations. The PL decay intensity can be expressed as [32]:

$$I_t = I_0 \exp\left\{-\frac{t}{\tau_0} - Q\left(\frac{t}{\tau_0}\right)^{3/S}\right\} \tag{5}$$

Where, I_0 represents the intensity at t is zero, τ_0 represents the intrinsic decay time of the donor ions and Q is the energy transfer parameter given by the expression [32]:

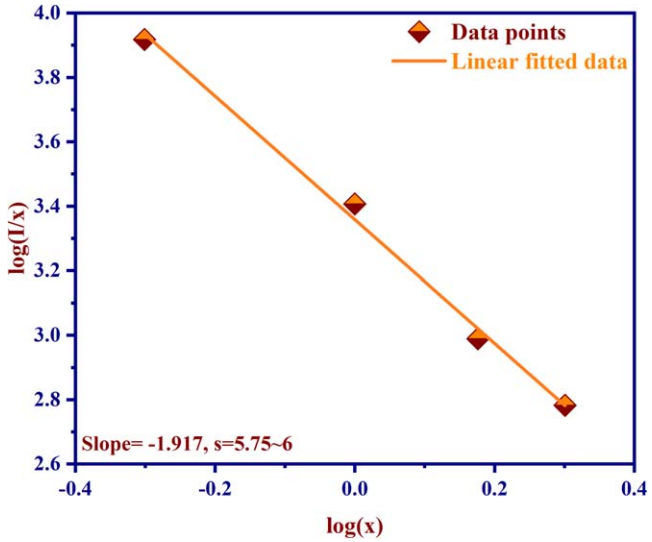


Figure 8. Dexter plot of log (I/x) versus log(x).

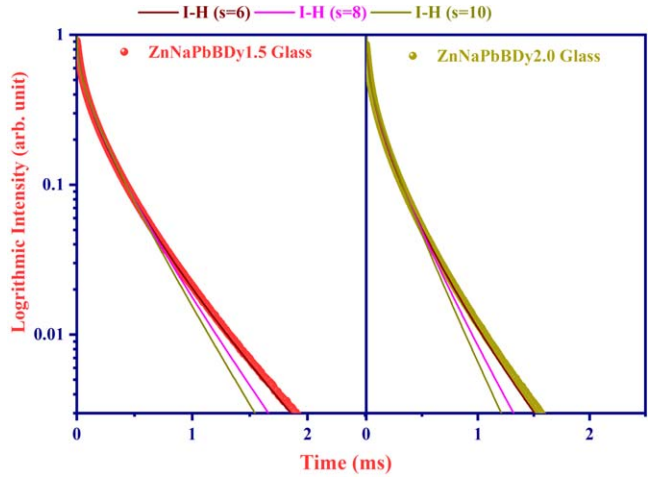


Figure 9. I-H fitted decay profiles of ZnNaPbBDy1.5 and ZnNaPbBDy2.0 glasses.

$$Q = \frac{4}{3}\pi\Gamma\left(1 - \frac{3}{S}\right)N_a R_0^3 \tag{6}$$

The type of energy transmission interaction is estimated from the 's' value which is D-D interaction for s = 6, D-Q for s = 8 and Q-Q for s = 10, N_a is used for the concentration of acceptor ions and R_0 is a distance of critical transfer. The factor Γ has unlike values depending on the type of interaction i.e., for s = 6 Γ(X) is 1.77, for s = 8 Γ (X) is 1.43 and for s = 10 Γ(X) is 1.3. The I-H curve of ZnNaPbBDy1.5 and ZnNaPbBDy2.0 glasses was fitted for the s = 6, 8 and 10, it was found that the fitting with s = 6 is best fitting for decay spectra reflects that the D-D interaction controls Dy^{3+}-Dy^{3+} ions energy transfer [40]. The best-fitted curve of ZnNaPbBDy1.5 and ZnNaPbBDy2.0 glasses is shown in figure 9. This demonstrates that the principal mechanism of lifespan quenching in the as-prepared glasses is energy transfer via D-D interactions between the Dy^{3+} ions. A similar result is observed in Dexter's theory analysis.

4. Conclusions

In brief, Dy^{3+} ions doped zinc-sodium-lead-borate glass samples were prepared by melt-quenching technique and are used to analysis their fundamental physical, structural and thermal properties. The several physical properties calculated from the previous work reflect that the addition of Dy^{3+} ions enhanced the physical features of the host glass matrix. The physical properties of the glasses were found to be sensitive to Dy^{3+} doping

concentration. The observed values of 'M' for the synthesized glasses are less than one and in decreasing order consequences the extent of the conduction band is increased hence the optical band gap is reduced. The FT-IR spectrum provides identification information about different functional groups and OH content. A small OH concentration enhances the host glass's radiative emission capability and validates its appropriateness for photonic applications. Raman spectra confirmed the presence of borate functional groups (i.e., BO_3, BO_4) and various types of bonding present in the prepared glass matrices. DSC thermograms were used to compute the glass transition temperature for the current glasses and found to be in the range of 450 °C–500 °C. The decay curves show non-exponential behavior owing to the energy transfer taking place through cross-relaxation processes. The I-H fitting of decay curves and Dexter plot using photoluminescence emission spectra of the glasses indicates that the energy transfer between Dy^{3+}-Dy^{3+} ions is mainly dipole–dipole in nature and is the cause for observed luminescence quenching for higher concentration. The optical and luminescence properties of the as-prepared glasses were already reported in the previous article which confirms their suitability for while light emission applications and the findings of the present work is in agreement with the previously published results. Thus, this newly proposed glass composition may contribute to the development of photonic devices.

Acknowledgments

Two of the authors (Dr. Anand Kumar & Dr. Nisha Deopa) would like to thank Chaudhary Ranbir Singh University (CRSU), Jind (Haryana) India for providing financial support (seed money) through minor research project (MRP) (letter No. CRSU/IQAC/2023/230 dated 06/04/2023 & CRSU/IQAC/2023/223 dated 06/04/2023, respectively).

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Manjeet https://orcid.org/0009-0000-4234-6542
Anu https://orcid.org/0000-0002-0753-1141
Rishi Pal Chahal https://orcid.org/0000-0001-9974-6520

References

[1] Sai Dinesh P, Kumar M and Ratnakaram Y C 2024 Impact of modifiers on structural and optical properties of Dy^{3+} doped different lithium tetraborate glasses for w-LED applications *Infrared Phys.* **136** 105029
[2] Yaduvanshi N, Basavapoornima C, Alqarni A S, Kumar Kagola U, Uthayakumar T, Madhu A and Srinatha N 2024 Enhanced white light emission and quantum efficacy of borate-zinc-lithium-aluminium glasses doped with Dy_2O_3 for potential white light emission applications *Opt. Mater.* **151** 115359
[3] Ravita, Prasad A, Rohilla P, Shandilya A and Rao A S 2023 Photoluminescence and energy transfer studies on Tm^{3+}/Dy^{3+}/Eu^{3+} doped borosilicate glasses for color tunability and warm white light generation *J. Non. Cryst. Solids* **606** 122192
[4] Pawaria S, Ahlawat J, Bala M, Dahiya S, Ohlan A, Punia R and Maan A S 2022 Structural and optical characterization of semiconducting lithium modified zinc borate glassy system for UV band reject filter *J. Mol. Struct.* **1270** 133836
[5] Manjeet, Kumar A, Anu, Lohan R, Deopa N, Kumar A, Chahal R P, Dahiya S, Punia R and Rao A S 2023 Impact of Sm^{3+} ions on structural, thermal, optical and photoluminescence properties of ZnO-Na_2O-PbO-B_2O_3 glasses for optoelectronics device applications *Opt. Mater. (Amst).* **139** 113778
[6] Monisha M, Murari M S, Sayyed M I, Al-Ghamdi H, Almuqrin A H, Lakshminarayana G and Kamath S D 2021 Thermal, structural and optical behaviour of Eu^{3+} ions in Zinc Alumino Boro-Silicate glasses for bright red emissions *Mater. Chem. Phys.* **270** 124787
[7] Shoaib M, Rooh G, Rajramakrishna R, Chanthima N, Kiwsakunkran N and Kim H J 2019 Comparative study of Sm^{3+} ions doped phosphate based oxide and oxy-fluoride glasses for solid state lighting applications *J. Rare Earths* **37** 374–82
[8] Rani P R, Venkateswarlu M, Mahamuda S, Swapna K, Rao A S and Vijaya Prakash G 2019 Structural, absorption and photoluminescence studies of Sm^{3+} ions doped barium lead alumino fluoro borate glasses for optoelectronic device applications *Mater. Res. Bull.* **110** 159–68
[9] Azizan S A, Hashim S, Razak N A, Mhareb M H A, Alajerami Y S M and Tamchek N 2014 Physical and optical properties of Dy^{3+}: Li_2O-K_2O-B_2O_3 glasses *J. Mol. Struct.* **1076** 20–5
[10] Hegde V, Chauhan N, Viswanath C S D, Kumar V, Mahato K K and Kamath S D 2019 Photoemission and thermoluminescence characteristics of Dy^{3+}-doped zinc sodium bismuth borate glasses *Solid State Sci.* **89** 130–8
[11] Ahmad A U, Hashim S and Ghoshal S K 2020 Physical, thermal and absorption traits of lithium strontium zinc borate glasses: Sensitiveness on Dy^{3+} doping *J. Alloys Compd.* **844** 156176
[12] Divina R, Teresa P E and Marimuthu K 2021 Dy^{3+} ion as optical probe to study the luminescence behavior of Alkali lead bismuth borate glasses for w-LED application *J. Alloys Compd.* **883** 160845
[13] Hemalatha S, Nagaraja M, Madhu A and Srinatha N 2022 Rare earth (Dy^{3+}-ion) induced near white light emission in sodium-zinc-borate glasses *Results Opt.* **9** 100275

[14] Dahiya J, Hooda A, Agarwal A and Khasa S 2022 Effect of dysprosium and samarium RE ion Co-doping on photoluminescence behaviour of novel alkali fluoride bismuth borate glasses: a white LED material *Opt. Mater. (Amst).* **134** 113162
[15] Mukamil S *et al* 2022 Photoluminescence and optical transition properties of Sm^{3+} activated lead-borate glasses *Opt. Mater. (Amst).* **129** 112486
[16] Grüne M and Schweizer S 2022 Comparison of Dy^{3+}-doped barium borate and lithium aluminoborate glass *Opt. Mater. (Amst).* **128** 112339
[17] Manjeet, Kumar A, Anu, Ravina, Deopa N, Kumar A, Chahal R P, Dahiya S, Punia R and Rao A S 2022 Structural, thermal, optical and luminescence properties of Dy^{3+} ions doped zinc potassium alumino borate glasses for optoelectronics applications *J. Non. Cryst. Solids* **588** 121613
[18] Rada S, Ristoiu T, Rada M, Coroiu I, Maties V and Culea E 2010 Towards modeling gadolinium-lead-borate glasses *Mater. Res. Bull.* **45** 69–73
[19] Pascuta P, Borodi G, Jumate N, Vida-Simiti I, Viorel D and Culea E 2010 The structural role of manganese ions in some zinc phosphate glasses and glass ceramics *J. Alloys Compd.* **504** 479–83
[20] Szumera M, Wacławska I and Sułowska J 2016 Influence of CuO and ZnO addition on the multicomponent phosphate glasses: spectroscopic studies *J. Mol. Struct.* **1114** 78–83
[21] Aljewaw O B, Karim M K A, Kamari H M, Zaid M H M, Noor N M, Isa I N C and Mhareb M H A 2020 Impact of Dy_2O_3 substitution on the physical, structural and optical properties of lithium-aluminium-borate glass system *Appl. Sci.* **10** 8183
[22] Zaman F, Srisittipokakun N, Rooh G, Khattak S A, Kaewkhao J, Rani M and Kim H J 2021 Comparative study of Dy^{3+} doped borate glasses on the basis of luminescence and lasing properties for white-light generation *Opt. Mater. (Amst).* **119** 111308
[23] Ichoja A, Hashim S, Ghoshal S K, Hashim I H and Omar R S 2018 Physical, structural and optical studies on magnesium borate glasses doped with dysprosium ion *J. Rare Earths* **36** 1264–71
[24] Mohan S, Kaur S, Singh D P and Kaur P 2017 Structural and luminescence properties of samarium doped lead alumino borate glasses *Opt. Mater. (Amst).* **73** 223–33
[25] Rekha Rani P, Venkateswarlu M, Mahamuda S, Swapna K, Deopa N and Rao A S 2019 Spectroscopic studies of Dy^{3+} ions doped barium lead alumino fluoro borate glasses *J. Alloys Compd.* **787** 503–18
[26] Dimitrov V and Komatsu T 1999 Electronic polarizability, optical basicity and non-linear optical properties of oxide glasses *J. Non. Cryst. Solids* **249** 160–79
[27] Herzfeld K F 1927 On atomic properties which make an element a metal *Phys. Rev.* **29** 701–5
[28] Duffy J A and Ingram M D 1976 An interpretation of glass chemistry in terms of the optical basicity concept *J. Non. Cryst. Solids* **21** 373–410
[29] Kaur R, Bhatia V, Kumar D, Rao S M D, Singh S P and Kumar A 2019 Physical, structural, optical and thermoluminescence behavior of Dy2O3 doped sodium magnesium borosilicate glasses *Results Phys.* **12** 827–39
[30] Anu, Deopa N and Rao A S 2022 Structural and luminescence characteristics of thermally stable Dy^{3+} doped oxyfluoride strontium zinc borosilicate glasses for photonic device applications *Opt. Laser Technol.* **154** 108328
[31] Sangeetha G, Sekhar K C, Hameed A, Ramadevudu G, Chary M N and Shareefuddin M 2021 Influence of CaO on the structure of zinc sodium tetra borate glasses containing Cu^{2+} ions *J. Non. Cryst. Solids* **563** 120784
[32] George H, Deopa N, Kaur S, Prasad A, Sreenivasulu M, Jayasimhadri M and Rao A S 2019 Judd-Ofelt parametrization and radiative analysis of Dy^{3+} ions doped sodium bismuth strontium phosphate glasses *J. Lumin.* **215** 116693
[33] Raman C V and Krishnan K S 1928 A new type of secondary radiation *Nature* **121** 501–2
[34] Reddy Y L P, Waaiz M, Ahmed S N and Reddy C V K 2017 Optical properties of dysprosium (Dy^{3+}) doped fluoroborate glasses *Int. J. Innov. Sci. Eng. Technol.* **4** 4–8
[35] Bala M, Agrohiya S, Dahiya S, Ohlan A, Punia R and Maan A S 2020 Effect of replacement of Bi_2O_3 by Li_2O on structural, thermal, optical and other physical properties of zinc borate glasses *J. Mol. Struct.* **1219** 128589
[36] Vijaya Babu K and Cole S 2018 Luminescence properties of Dy^{3+}-doped alkali lead alumino borosilicate glasses *Ceram. Int.* **44** 9080–90
[37] Suthanthirakumar P and Marimuthu K 2016 Investigations on spectroscopic properties of Dy^{3+} doped zinc telluro-fluoroborate glasses for laser and white LED applications *J. Mol. Struct.* **1125** 443–52
[38] Manjeet, Ravina, Amit, Poria K, Deopa N, Kumar A and Chahal R P 2023 Optimization of dysprosium ions doped borate glasses for photoluminescence applications *Mater. Lett.* X **19** 1–3
[39] Dexter D L 1953 A theory of sensitized luminescence in solids *J. Chem. Phys.* **21** 836–50
[40] Arunkumar S, Venkataiah G and Marimuthu K 2015 Spectroscopic and energy transfer behavior of Dy^{3+} ions in $B_2O_3TeO_2PbOPbF_2Bi_2O_3CdO$ glasses for laser and w-LED applications *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **136** 1684–97

Bulletin of Engineering Geology and the Environment (2024) 83:389
https://doi.org/10.1007/s10064-024-03864-x

ORIGINAL PAPER

Effect of load frequency and amplitude of displacement on soft synthetic rock joints under cyclic shear loads

S. M. Mahdi Niktabar[1] · K. Seshagiri Rao[2] · Amit Kumar Shrivastava[3] · Jiří Ščučka[1]

Received: 6 March 2024 / Accepted: 13 August 2024
© The Author(s) 2024

Abstract

If dynamic loads propagate in the rock mass, they can generate vibrations or cyclic movements along the joints with different frequency and displacement. Hence, evaluation of rock joints behaviour under different cyclic movements significantly contributes to increasing the safety and stability of rock mass. In this study a series of cyclic shear tests was conducted on soft synthetic rock joints (with asperity angle of 30°- 30°) by shear apparatus to investigate the effects of shear load frequency and amplitude of displacement on the joint. Varying load frequencies (0.01, 0.05 and 0.1 Hz,) and displacement amplitudes ($\pm 4, \pm 6$, and ± 8 mm) for the three different normal stresses (0.1, 0.5 and 1 MPa) were applied. The results indicated that the first peak shear stress on the joint was significantly affected by the frequency of cyclic shear load and the subsequent peaks of shear stress were largely influenced by the amplitude of displacement. Reduction of post peak shear stress was significant for the first cycle as the load frequency increased. For low normal stress, shear load frequency of 0.05 Hz and one-third of asperity base length as displacement were the certain values that shear mechanism changed on slopes of the joint asperities. Shear mechanism also demonstrated that asperity diminishing did not depend on load direction at high frequency. Whereas, at low frequency of loads, it depended on the direction of starting cyclic load. Comparison of shear strength envelopes indicated an opposite behaviour of the joint between load frequencies and displacement amplitudes.

Keywords Shear resistance · Soft rock joints · Cyclic shear load · Frequency of load · Displacement amplitude · Shear mechanism

Introduction

Earthquakes and other seismic activities (due to excavation in mining or underground structures) propagate dynamic loads on rock masses (Li et al. 2014, 2022). These dynamic loads can lead to a number of cyclic shear movements with different frequencies and amplitudes of displacement along the joints in the rock mass. Hence, shear resistance of the joints can change under cyclic loads and affect the rock mass stability. Study of cyclic shear behaviour of jointed rocks under different dynamic parameters plays an important role in analysis, design and safety of rock structures.

Shear behaviour of rock joints with focusing on shear velocity was investigated in the last decades. Results indicated that shear resistance was increased with increasing shear velocity (Schneider 1977; Hencher 1981; Gillette et al. 1983; Barla et al. 1990; Barbero et al. 1996; Qi et al. 2020; Dang et al. 2021). However, other studies showed that shear resistance was decreased with increasing shear velocity (Scholz and Engelder 1976; Zheng et al. 2015; Tang and Wong 2016; Meng et al. 2019; Tiwari and Latha 2019; He et al. 2023). No velocity dependency was reported by Hassani and Scoble (1985) and Li et al. (2021) for shear resistance in sandstone (0.012–0.180 mm/min) and planar joint respectively.

Crawford and Curran (1981) and Atapour and Moosavi (2014) observed different shear behaviour in soft and hard jointed rocks. The results indicated that the shear resistance

✉ S. M. Mahdi Niktabar
mahdi.niktabar@ugn.cas.cz

[1] Department of Laboratory Research On Geomaterials, Institute of Geonics of the Czech Academy of Sciences, Studentska 1768/9, 708 00 Ostrava-Poruba, Czech Republic

[2] Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, Delhi 110016, India

[3] Department of Civil Engineering, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India

Published online: 09 September 2024

was increased with increasing shear velocity on soft rocks, but it declined for hard rocks. Thus, the shear resistance under different shear velocity is probably influenced by the type of rocks.

In recent years, focusing on cyclic shear loads has increased due to investigation of joint behaviour under dynamic loads. However, most studies on cyclic shear loads were conducted under constant frequency and displacement amplitude (Hutson and Dowding 1990; Qiu et al. 1993; Huang et al. 1993; Jing et al. 1993; Homand-Etienne et al. 1999; Lee et al. 2001; Desai et al. 2005; Ferrero et al. 2010; Nemcik et al. 2014; Niktabar et al. 2017; Dong et al. 2020; Wu et al. 2023). As the seismic events can occur with different intensity and distance from the rock mass, it is also necessary to study the behaviour of rock joints at different shear load frequencies and displacement amplitudes (Ramulu et al. 2009; Yang et al. 2019; Gómez et al. 2020).

Shear displacement and its effect on rock joints under cyclic shear loads were investigated by Jafari et al. (2003). They reported that minor degradation occurred on rock joints with low amplitude of displacement. But asperities significantly demolished with large displacement amplitude. Similar joint degradation was observed by Ferrero et al. (2010), Wu et al. (2019) and Thongprapha et al. (2020). Damage of joint asperities at lower shear amplitudes was smaller compared to the larger amplitudes. Nguyen (2013) studied the behaviour of jointed rocks under cyclic shear loading with different frequencies (dynamic condition). The results showed that peak shear stress under dynamic conditions was 30% greater than that under static conditions. Mirzaghorbanali et al. (2014) and Guansheng et al. (2020) investigated the influence of shear rate on rock joints for cyclic loads under constant normal stiffness. The results indicated that shear resistance decreased with increasing shear rate on (unfilled) joint. However, opposite behaviour was observed in shear resistance of (infilled) joints i.e. shear resistance was increased in the joint with increase in shear rate. Wu et al. (2024) conducted cyclic shear tests on rock joints with same and different joint wall material. They reported shear strength of jointed rock with different joint wall material (siltstone/mudstone) was between shear strength of the same joint wall materials (siltstone and mudstone). But it was closer to the joint with lower shear strength of material (mudstone). The results showed that greater normal stress, rougher surface morphology, larger shear displacement and faster shear rate have significant effects on deterioration of rock joints with different wall material.

Study shear behaviour of soft rock joints (or weathered joints) is important because of its low shear resistance compared to hard rock joints (Dearman et al. 1978; Woo et al. 2010; Özvan et al. 2014; Ram and Basu 2019; Nian et al. 2024). Shear behaviour of these jointed rocks in seismic active regions where the joints are subjected to cyclic loads is significant. However, few studies have been conducted on shear behaviour of soft rock joints (Indraratna et al. 1998; Lu and Zhang 2021; Ban et al. 2023; Niktabar et al. 2023) and influence of load frequency and displacement amplitude on these jointed rocks are ignored. In addition there is a lack of study for the effect of these dynamic parameters on shear mechanism of the joint asperities. Hence, in this study, a series of cyclic shear tests was conducted on simulated soft rock joints with different load frequency and displacement amplitude to investigate the effect of these dynamic parameters on shear behaviour of the joints. Moreover, shear mechanisms on soft rock joints under cyclic loads with changing frequency and displacement amplitude are described in this research. Hence, the results could provide useful evaluation and analysis for slope stability and foundation on soft rock joints in seismic regions.

Preparation of specimens

Preparation of soft (or weathered) jointed rocks as natural specimens with the same joint roughness is very difficult. Hence, plaster of Paris was selected as model material. Casting of this material is easy and its long-term strength is independent of time once the chemical hydration is completed. Specimens were prepared by casting the plaster into special steel moulds. Due to easy removal of the specimen without breaking the asperities, the inside of the moulds and the surface of the asperity plates were lubricated with oil. The plaster paste with water-powder ratio of 0.6 was mixed by a mixer machine for 2 min and then poured into the moulds placed on a vibrating table. The moulds with the paste were vibrated (2 Hz frequency) for 30 s to remove the air bubbles from the paste (Fig. 1a). Final setting time of plaster of Paris with 0.6 water powder ratio was 50 min. Two parts of the sample is removed from the moulds ultimately and placed separately in proper condition in such a way the asperities were in upward position. The specimens (hardened) were removed from the moulds after 60 min and then left for 20 days at laboratory temperature or 25 °C for the air curing (Fig. 1b).

The specimens or soft synthetic rock joints with asperity angles of 30°- 30° and size of 299 mm 299 mm 125 mm were formed. Height and base length of asperities are 5 mm and 17 mm respectively based on 5 mm gape between the shear boxes and 30^0 angle of asperities. The hardened plaster has a low uniaxial compressive strength and it is suitable for simulating the mechanical behaviour of weathered joints and soft rocks (covering a series of sedimentary rocks) like siltstone, claystone, shale, coal and rock salt. Mechanical properties of the specimens are presented in Table 1.

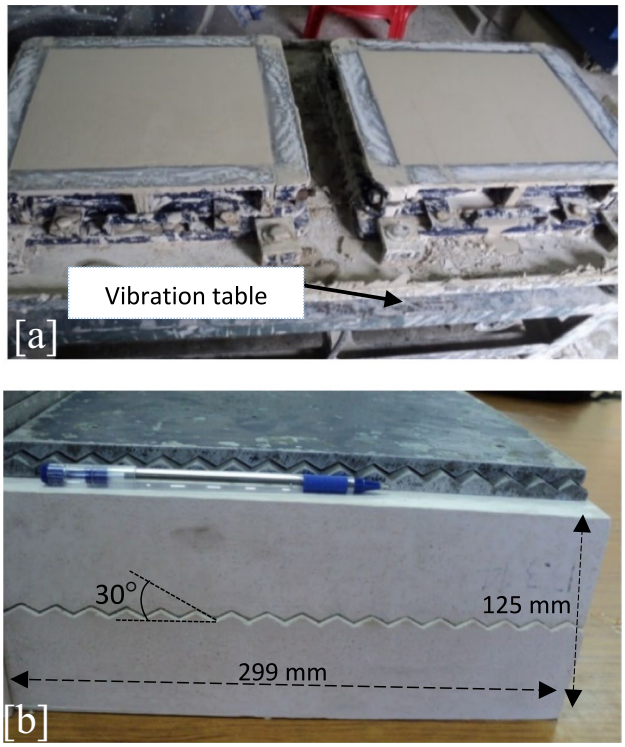


Fig. 1 **a** Moulds fixed on the vibration table during specimen preparation and (**b**) prepared joint with asperity angle of 30°- 30°

Table 1 Mechanical properties of the specimens

Parameter	Value
Uniaxial compressive strength	6 MPa
Tangent modulus	2.45 GPa
Poisson's ratio	0.27
Cohesion	0.72 MPa
Intact friction angle	51.3°

Shear machine and testing procedure

Automatic static and cyclic shear testing machine (Niktabar et al. 2018) was utilized in this experiments (Fig. 2). The shear machine is an automatic apparatus and it is a combination of the four main units, including frame and loading unit (1), hydraulic power pack (2), data acquisition (3), and cooling unit (4). The size of each shear box is 300 mm × 300 mm × 448 mm and 5 mm gap is kept between the boxes by needle linear bearing (Shrivastava and Rao 2013). Maximum capacities of normal and horizontal load cells are 500 kN and 1000 kN, respectively. Calibration of the load cells and linear variation displacement transducers (LVTD) was conducted before the experiment. Software of the shear machine automatically collects data during the shear loading. This apparatus has capability to carry out cyclic shear tests (Fig. 3) on jointed rocks for different frequencies (f_s) and displacement amplitudes (a_s).

A series of cyclic shear tests were performed on the jointed specimens for different load frequencies (0.01, 0.05 and 0.1 Hz) by keeping amplitude of displacement constant (± 8 mm). Another series of tests were conducted for different displacement amplitudes (± 4, ± 6 and ± 8 mm) while keeping frequency of shear loads constant (0.01 Hz). This range of displacement amplitudes were selected on the joint because rock joints can move several millimeters or few centimeters during vibrations or seismic events. In addition, frequencies were selected based on dynamic loads from earthquakes which usually have low frequencies or less than 1 Hz. In this study during each load frequency, shear velocity is constant, only the direction of movement has changed repeatedly. However for different frequencies (different cyclic shear test), the shear velocity has changed, i.e., with increasing frequency, shear velocity has increased. Three different normal stresses 0.1 MPa (low), 0.5 MPa (medium),

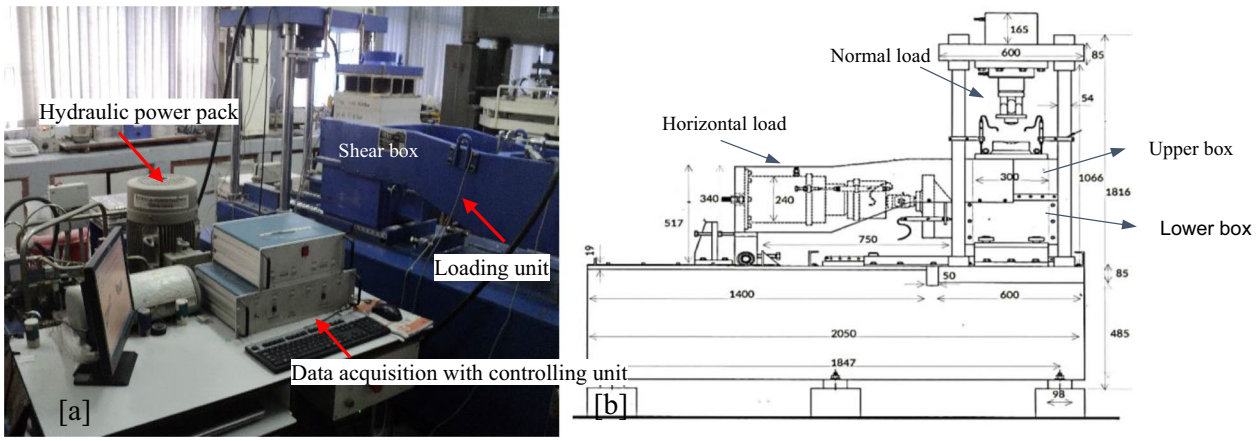


Fig. 2 **a** Photograph of automatic static and cyclic shear testing machine, and (**b**) schematic diagram of loading unit (all dimensions are mm)

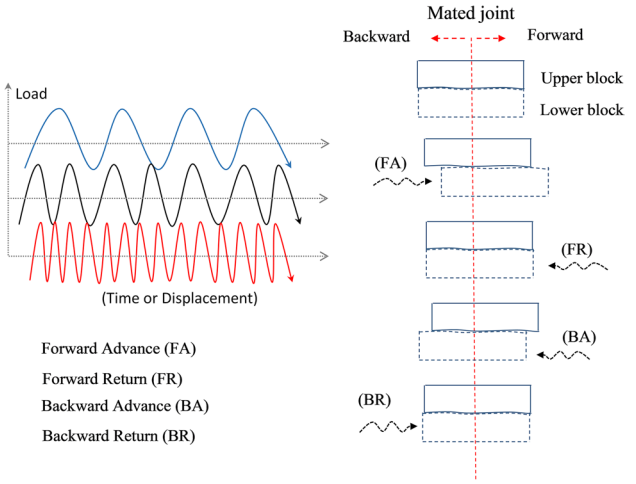


Fig. 3 Cyclic shear load on the joint with different load frequency and amplitude of displacement

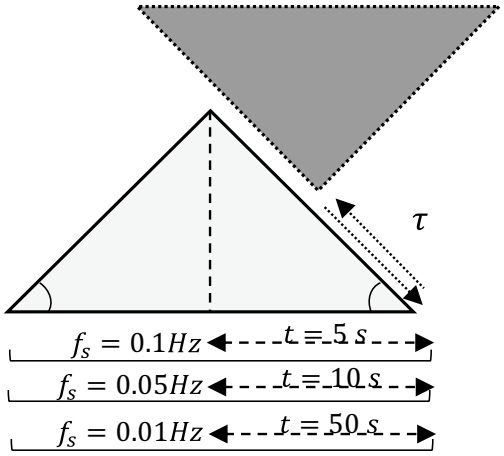


Fig. 4 Three different frequencies and shearing times on each slope of asperity for every cyclic load

1 MPa (high) and thirty shear loading cycles were applied for each shear test. The tests were conducted under constant normal load condition (normal stiffness was kept zero). Shear test parameters are presented in Table 2.

Results and discussion

Effect of cyclic shear load frequency on the joint

Cyclic shear tests with different load frequencies (0.01, 0.05, 0.1 Hz) conducted on the joint to investigate the influence of load frequency on shear resistance of the joint asperities (Fig. 4). The results as shear stress and normal displacement (dilation) vs. horizontal displacement for low normal stress (0.1 MPa) are presented in Fig. 5. With increasing the frequency of cyclic shear load from 0.01 Hz to 0.05 Hz (Fig. 5a and b), the first peak shear stress of the joint was increased from 0.34 MPa to 0.42 MPa (23%). When frequency increased from 0.01 Hz to 0.1 Hz (Fig. 5a and c), peak shear stress also increased from 0.34 MPa to 0.48 MPa (41%), i.e. shear resistance on soft rock joints was increased with rising shear load frequency (Table 3). Similar behaviour was observed for the joint shear resistance at normal stress of 0.5 and 1 MPa (Fig. 13 and Fig. 14 in appendix). Reduction of post peak shear stress was significant with increasing shear load frequency i.e., angle of slope for the post peak shear stress increased (negatively) as the load frequency increased. It is clear for the first cycle (Fig. 5a-c; Fig. 14a-c in appendix), but it was not observed for the next cycles.

Joint with low frequency of load (0.01 Hz) was under compression (positive normal displacement) after several cycles (Fig. 5d). But the joint with high frequency of load (0.1 Hz) could dilate (negative normal displacement) even at the last cycle (Fig. 5f), i.e. the subsidence of joint asperities was more significant at lower shear frequency under cyclic loads.

Peak shear stress for different frequencies and number of cycles are presented in Fig. 6a and Table 3. The results show that frequency of load has a significant effect on the first peak shear stress. With increasing the number of cycles ($N > 2$), the influence of frequency was insignificant on the joint. The maximum difference of peak shear stress was observed between the first and the second cycles. With increasing frequency of shear load, the difference of peak shear stress between the first and the last cycle ($N = 30$) was increased. It was 59%, 64% and 67% for shear frequencies of 0.01, 0.05 and 0.1 Hz respectively. These results demonstrated that shear resistance reduction at higher frequency (0.1 Hz) was greater than that at lower frequency (0.01 Hz) under cyclic movements.

Joint dilation and peak joint dilation vs. number of cycles for different shear load frequencies are illustrated in Fig. 5d-f and Fig. 6b respectively. For low shear frequency (0.01 Hz), the peak joint dilation at forward movement (+8 mm) was higher than backward movement (-8 mm) for the first cycle. This difference was 33.4% (Fig. 5d). However, the difference of peak joint dilation between the both directions gradually decreased and it was zero after several cycles ($N = 22$).

Table 2 Shear test parameters under cyclic shear loads

Test parameter	Value
Frequency of shear load, f_s [Hz]	0.01, 0.05, 0.1
Amplitude of displacement, a_s [mm]	$\pm 4, \pm 6, \pm 8$
Normal load, P [MPa]	0.1, 0.5, 1
Number of cycle, N	30

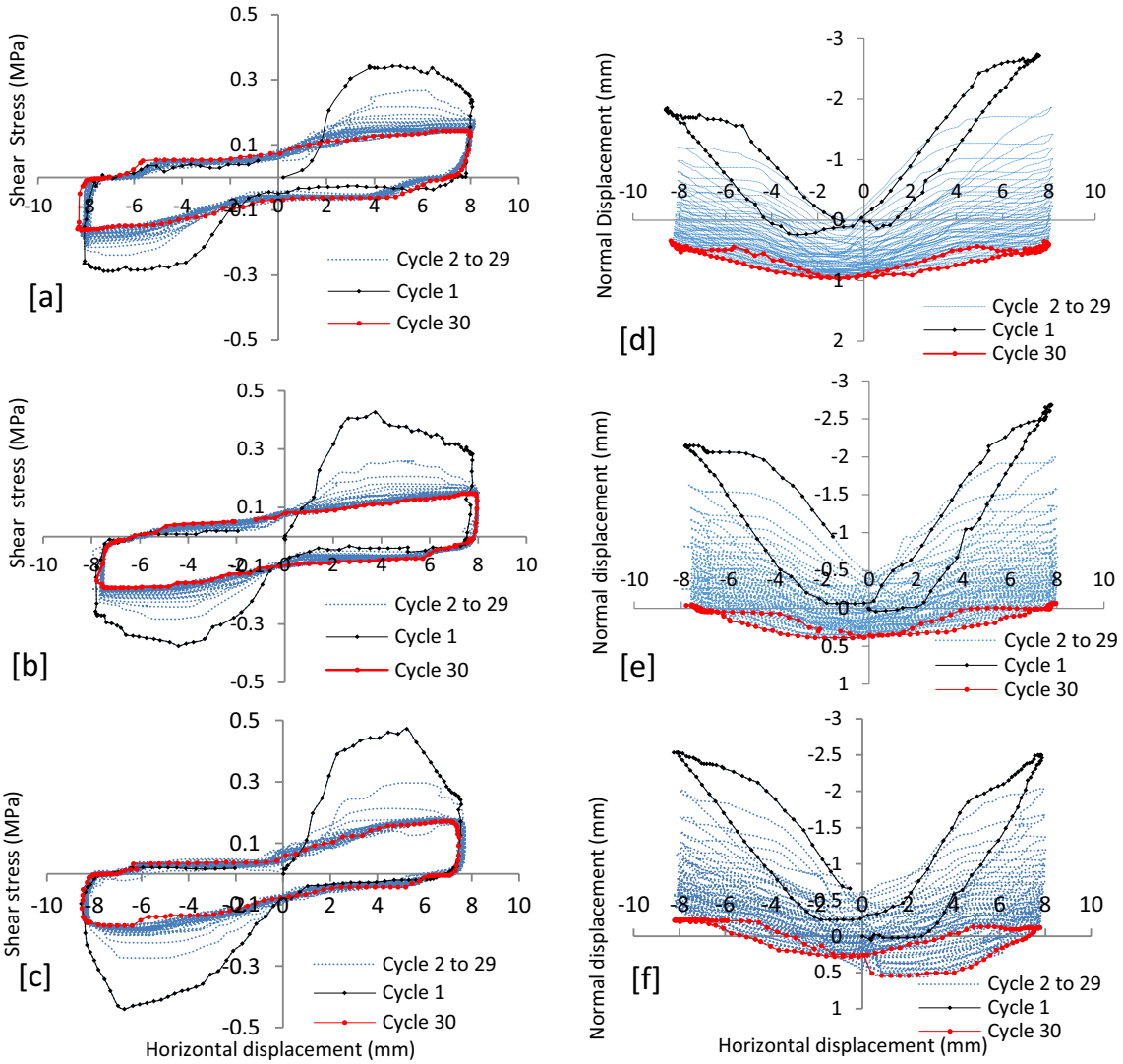


Fig. 5 **a**, **b**, **c** Shear stress and (**d**, **e**, **f**) normal displacement vs. horizontal displacement of the joint with different frequencies, (0.01, 0.05 and 0.1 Hz respectively) under cyclic shear load ($P=0.1$ MPa)

Table 3 Peak shear stress on the joint for different frequencies (f_s) and number (N) of shear cycle ($P=0.1$ MPa)

N	$f_s=0.01$ Hz	$f_s=0.05$ Hz	$f_s=0.1$ Hz
1	0.34	0.42	0.48
2	0.26	0.26	0.29
3	0.21	0.21	0.21
15	0.15	0.15	0.16
30	0.14	0.15	0.16

With increasing the frequency of cyclic shear load to 0.05 and 0.1 Hz, the difference of peak joint dilation between forward and backward movements was decreased to 20.1% and 1.2% (negligible) for the first cycle (Fig. 5e and Fig. 5f) i.e. as the frequency of cyclic shear load increased, peak joint dilation between the both directions became equal under cyclic shear loads. The peak joint dilation decreased with the same rate for the both directions as the number of cycles increased.

Effect of cyclic shear displacement on the joint

Three different displacement amplitude for cyclic shear tests was selected based on fraction of the base length of asperity ($\lambda/2\simeq8$ mm, $\lambda/3\simeq6$ mm and $\lambda/4\simeq4$ mm) as shown in Fig. 7. Shear test results for different displacement amplitudes at low normal stress (0.1 MPa) are presented in Fig. 8 (for 0.5 and 1 MPa normal stresses are presented in Fig. 15 and Fig. 16 in appendix). The graphs show shear stress and normal displacement (dilation) vs. horizontal displacement. Peak shear stress on the joint was constant

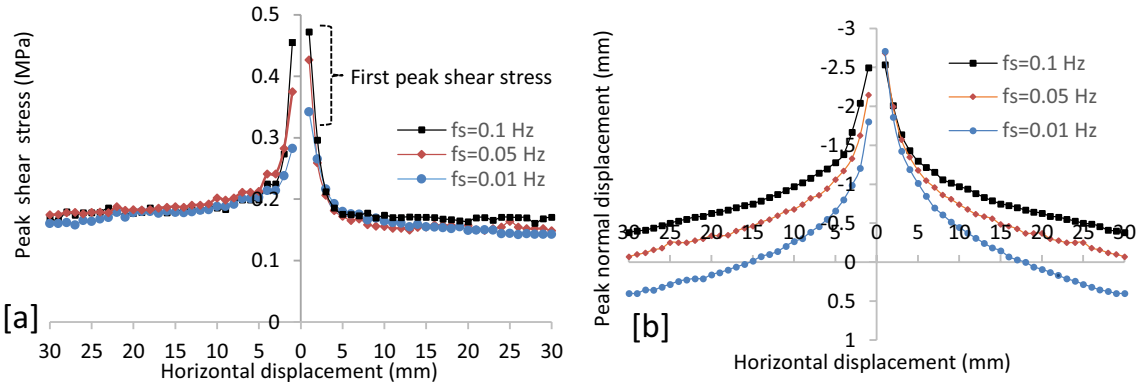


Fig. 6 (**a**) Peak shear stress and (**b**) peak normal displacement on the joint for different load frequencies, (0.01, 0.05 and 0.1 Hz) under 30 cyclic shear load ($P = 0.1$ MPa)

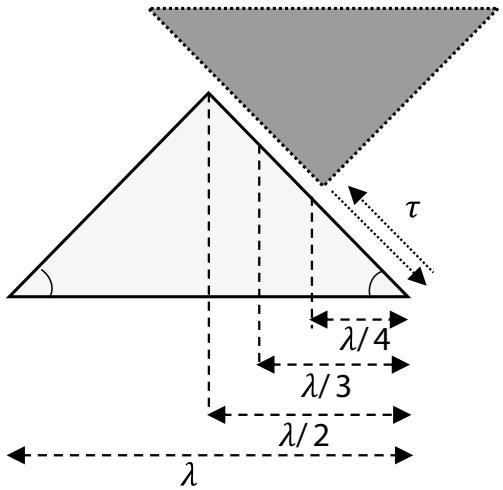


Fig. 7 Movement of asperities for different displacement amplitude under cyclic shear loads

as the number of shear cycles increased for displacement of ± 4 mm (Fig. 8a) i.e. the number of cycles was not effective on peak shear stress at low level of displacement amplitude until $\lambda/4$ (Table 4). With increasing displacement amplitude to ± 6 mm or $\lambda/3$ (Fig. 8b), peak shear stress started to decrease as the number of shear cycles increased (Similar behaviour was observed on the joint for 0.5 and 1 MPa normal stresses).

This reduction for peak shear stress was more significant when displacement amplitude was ± 8 mm or $\lambda/2$ (Fig. 8c). The peak shear stress was decreased by 33% and 57% for ± 6 mm and ± 8 mm amplitudes during thirty cycles.

Joint was under compression at low shear amplitude (± 4 mm) during all cycles (Fig. 8d). When the shear amplitude increased (± 6 and ± 8 mm), dilation was observed on the joint (Fig. 8e and f). Due to diminishing asperities, this dilation was declined as the number of cycles increased. Peak shear stress vs. number of cycles is demonstrated for different displacement amplitudes in Fig. 9a. The first peak shear stress was the same for different displacement amplitudes, but the other peaks of shear stress in the same range of displacements were different for the next cycles.

Comparison of shear tests between different load frequencies and displacements amplitudes (Fig. 6a and Fig. 9a) shows that frequency of shear load has a significant effect on the first peak shear stress. This is due to changing frequency leads to changing shear velocity and it has a significant effect on the joint for the first peak shear stress. After the first cycle, the effect of shear frequency is marginal for the next peak shear stresses. It might be because initial deformation was conducted on the joint. However, for different displacement, the first peak shear stress was the same. It is because the first peak shear stress is independent of displacement (after the peak) for the first cycle. A significant difference of peak shear stress was observed for the next cycles. The reason is that the next peak shear stress is dependent on previous shear displacement; hence, when displacement is changing then peak shear stress is different for the next cycles.

Joint dilation and peak joint dilation vs. number of cycles for different displacement amplitudes are demonstrated in Fig. 8d-f and Fig. 9b respectively. According to the graphs, peak joint dilation values at forward and backward movements are equal at low displacement amplitude (± 4 mm) for thirty cycles. With increasing amplitude of displacement (± 6 mm) the difference of peak joint dilation between the both movements appeared and it was 14.9% for the first cycle. For the amplitude of ± 8 mm, the difference of peak dilation was increased by 33.4% i.e. with increasing the amplitude of displacement, the difference of diminishing between the first and the second slopes of asperities became greater under cyclic loads. This difference gradually decreased with increasing the number of cycles.

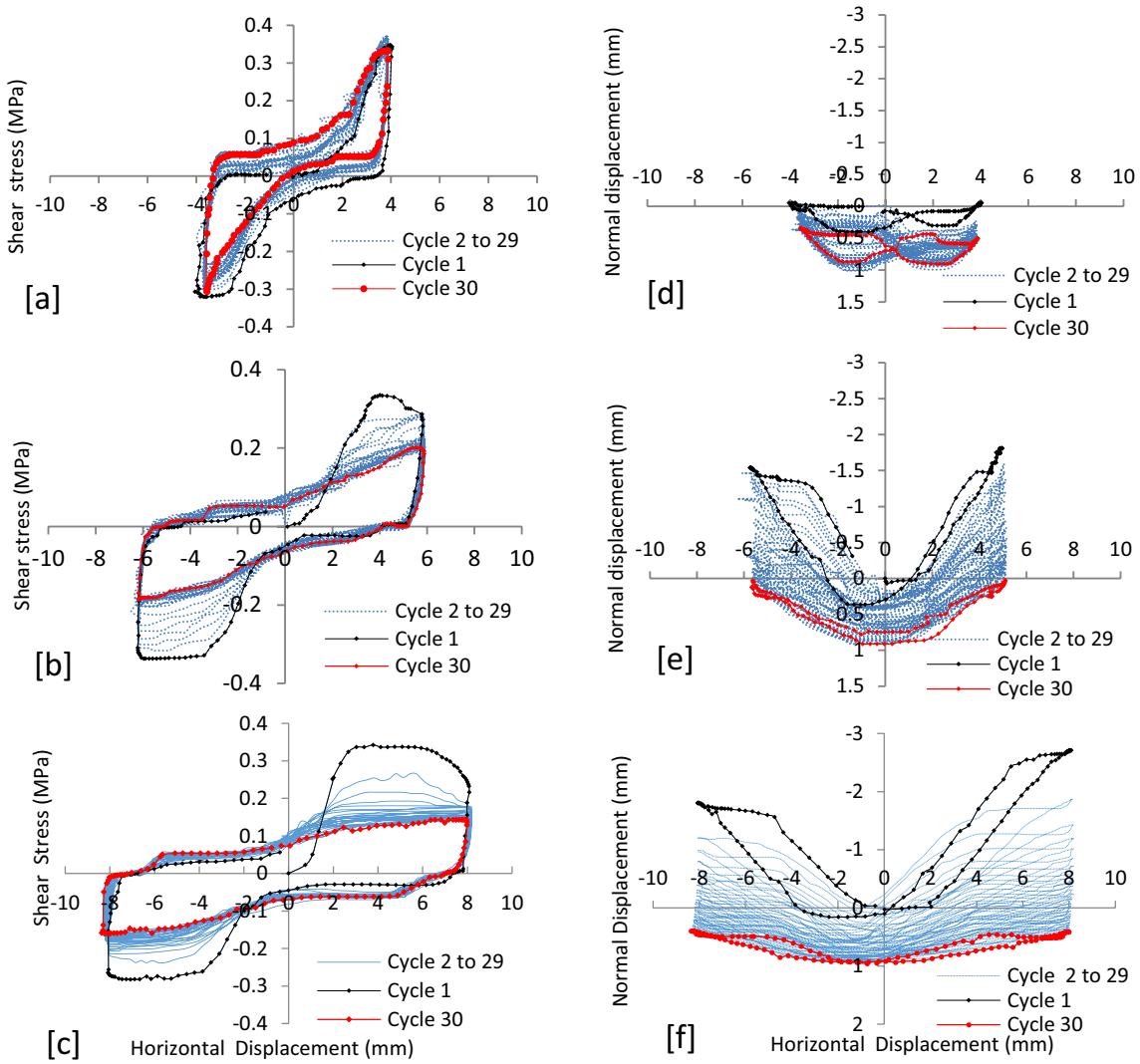


Fig. 8 **a**, **b**, **c** Shear stress and (**d**, **e**, **f**) normal displacement vs. horizontal displacement of the joint with different displacement amplitudes (4 mm, 6 mm and 8 mm respectively) under cyclic shear load ($P = 0.1$ MPa)

Table 4 Peak shear stress on the joint for different displacement amplitudes (a_s) and number (N) of shear cycle ($P = 0.1$ MPa)

N	$a_s = \pm 4$ mm ($\simeq \lambda/4$)	$a_s = \pm 6$ mm ($\simeq \lambda/3$)	$a_s = \pm 8$ mm ($\simeq \lambda/2$)
1	0.34	0.33	0.33
2	0.32	0.28	0.26
3	0.32	0.28	0.22
15	0.34	0.22	0.15
30	0.34	0.20	0.14

Shear strength envelope for cyclic shear load

Joints under each cycle of shear load exhibited double peak shear stress as expected, i.e. one peak at the forward movement and another peak at the backward movement. Shear strength envelopes for different shear load frequencies and displacement amplitudes are presented in Fig. 10.

The envelopes were plotted for several cycles based on peak shear stress under three different normal stresses (0.1 MPa as shown in Figs. 5 and 8; also 0.5 and 1 MPa as shown in Figs. 13, 14, 15, 16 in the appendix). At the first cycle, the shear strength envelope was curvilinear (Fig. 10a) and it rose with increasing load frequency from 0.01 to 0.1 Hz. The envelope was greater for the first forward movement compared to the first backward movement; therefore, frequency has a significant effect on shear strength envelope of the joint at initial joint movement. Shear strength envelope was linear for the next cycles (Fig. 10b and Fig. 10c; Fig. 17a-c in appendix), and no significant change was observed as the load frequency increased i.e., after the first

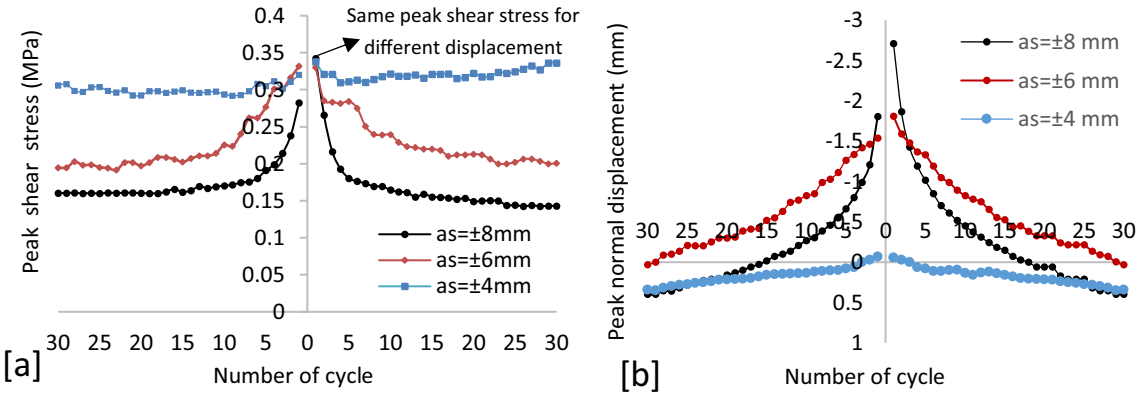


Fig. 9 (**a**) Peak shear stress and (**b**) peak normal displacement on the joint with different amplitudes of displacement (± 4, ± 6, ± 8 mm) under 30 cyclic shear load ($P = 0.1$ MPa)

cycle, no effect of frequency was observed on the shear strength envelope of the joint.

However, shear strength envelope was the same as displacement amplitude increases ($\pm 4, \pm 6$, and ± 8 mm) for the first cycle (first forward movement; Fig. 10d), i.e., displacement amplitude has no significant effect on shear strength envelope on the joints at initial joint movement. Shear strength envelope decreases as the displacement amplitude increases for the next cycles (Fig. 10e and f; Fig. 17d-f in appendix). Therefore, influence of displacement amplitude was observed after the first cycle. Shear strength envelope for 6 mm ($\lambda/3$) displacement amplitude was between the envelopes of 4 mm ($\lambda/2$) and 8 mm ($\lambda/4$) displacements amplitudes. It was close to the envelope of 4 mm at initial cycles, then after several cycles (15 cycles) it was close to the envelope of 8 mm displacement amplitude.

The opposite effect of load frequency and displacement amplitude was observed on the shear strength envelope of the joint (with considering one peak shear stress for each cycle or peak shear stress in forward movement). Shear strength envelope for varying frequencies was different at the first cycle and it was the same from the next cycle (Fig. 10a-c). However, the shear strength envelope for varying displacement amplitude was the same at the first cycle and it was different from the next cycles (Fig. 10d-f).

Shear mechanism on the joint by changing load frequency and displacement amplitude

The joints behaviour indicated that shear mechanism is not the same at low (0.01 Hz) and high (0.1 Hz) frequency of cyclic loads. At low frequency of cyclic loads, peak joint dilation in forward and backward movements was different (Fig. 11a) i.e. diminishing the both slopes of asperities was unequal during cyclic loads and it shows a significant deformation on asperities during the backward movements. It is due to a sufficient time to deform second slopes (left slopes) of the asperities when diminishing is conducting on the first slopes (right slopes) of asperities. Second slopes of asperities are influenced by diminishing the first slopes at low frequency of cyclic loads i.e. diminishing on each slope of asperity is dependent on deformation or diminishing on another slope of asperity for low frequency of cyclic loads. Based on the dilation behaviour (Fig. 11a) the asperity damage was more severe in the backward movement. This behaviour may likely be due to the weakening of asperities caused by the formation of micro-cracks during the forward movement. Hence, during backward movement diminishing is more significant on the second slopes of asperities, with respect to the first slopes of asperities for each cyclic load. In consequence, peak joint dilation was smaller in the backward movements with respect to the forward movements.

Similar joint dilation in forward and backward movements for each cycle at high frequency of loads demonstrated that diminishing the both slopes of asperities were equal (Fig. 11c). It is due to no sufficient time to deform the second slopes of the asperities when diminishing is conducting on the first slopes (right slopes) of asperities i.e. there is no deformation effect on the second slopes of asperities when diminishing is carrying out on the first slopes of asperities. Hence, diminishing on each slope of asperity is independent for high frequency of cyclic loads. In consequence the peak joint dilation was the same in the both forward and backward movements.

Based on Patton's model (1966) on rock joint, the forward and backward shear resistance at low normal stress are presented in the below:

$$\tau_f = \sigma_n \tan\left(\phi_b + i_f\right) \tag{1}$$

$$\tau_b = \sigma_n \tan\left(\phi_b + i_b\right) \tag{2}$$

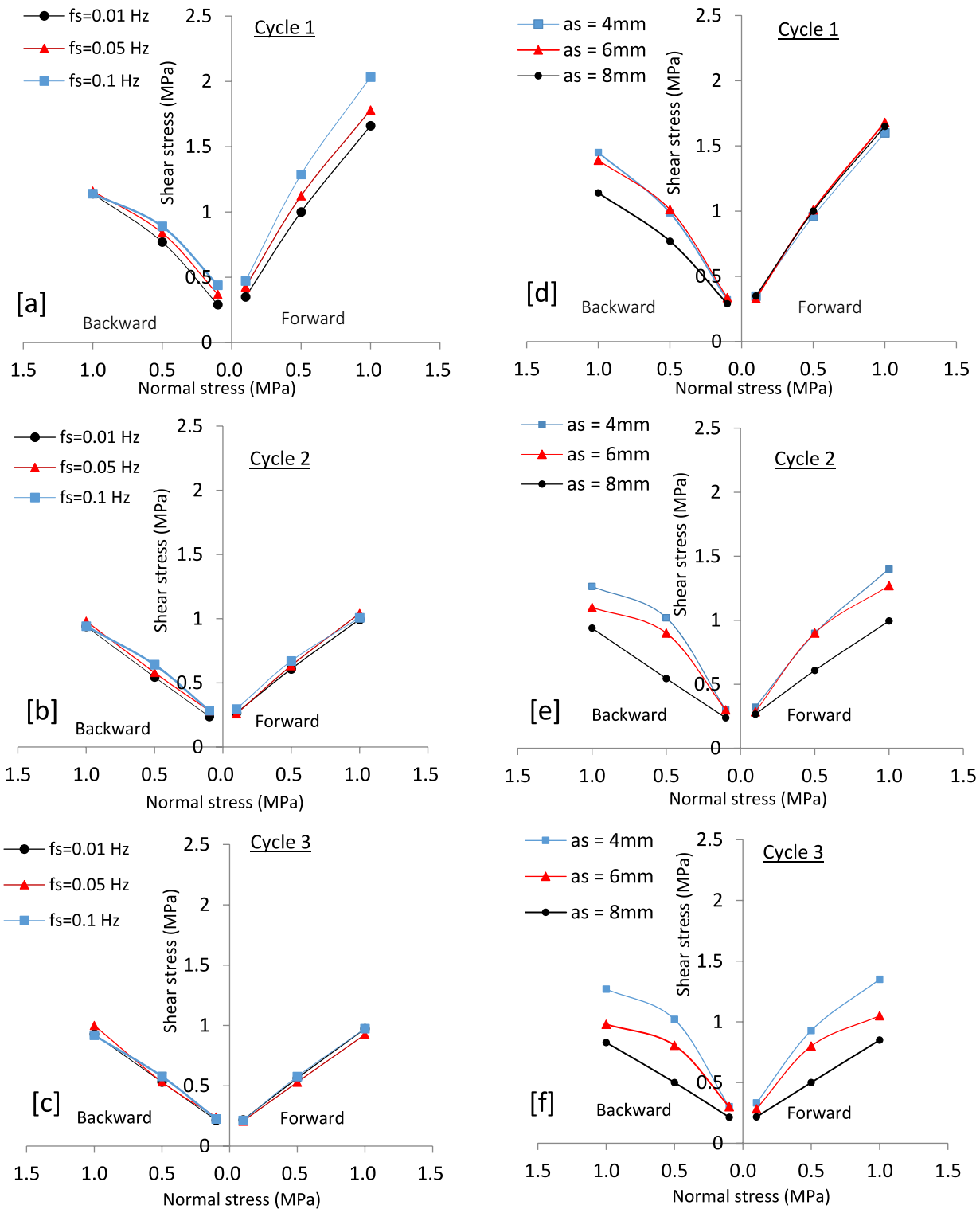


Fig. 10 Shear strength envelope of the joint under (**a**, **b**, **c**) varying shear load frequencies (fS = 0.01, 0.05, 0.1 Hz) and (**d**, **e**, **f**) amplitudes of displacement ($aS = \pm 4, \pm 6, \pm 8$ mm) for three different cycles respectively

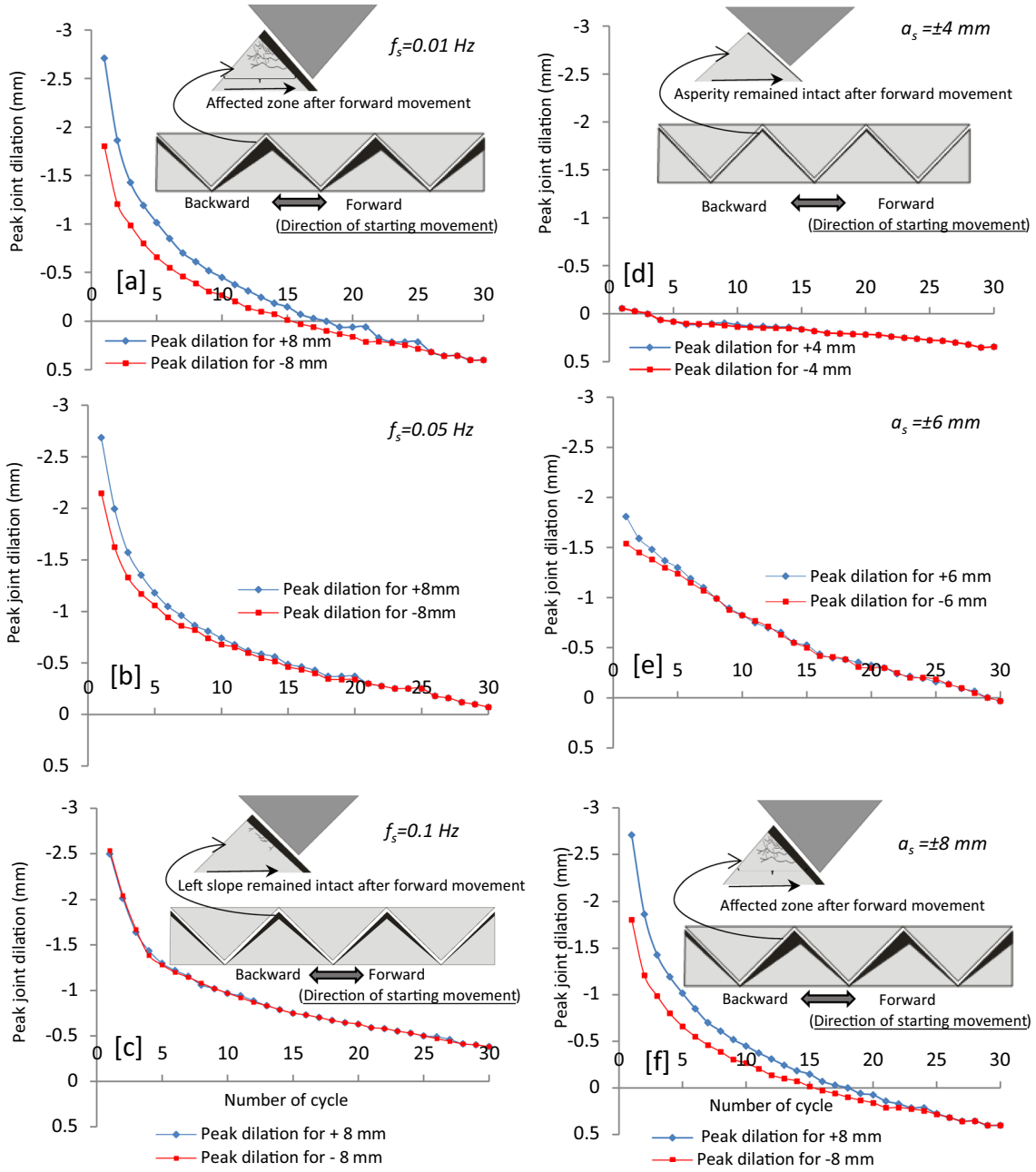


Fig. 11 Peak joint dilation vs. number of cycles for (**a**, **b**, **c**) varying load frequencies and (**d**, **e**, **f**) different displacement amplitudes under cyclic loads (illustrated with diminishing on slopes of asperities, $P = 0.1$ MPa)

where,

$\tau_f and \tau_b$ = peak shear stresses in forward and backward movements
$i_f and i_b$ = slopes of asperities in forward and backward movements
σ_n = normal stress
ϕ_b = basic friction angle

a) If $f_s \geq 0.1$ Hz, then $i_f = i_B$ and $\tau_f = \tau_b$
b) If $f_s = 0.05$ Hz, then $i_f \geq i_b$ and $\tau_f \geq \tau_b$
c) If $f_s \leq 0.01$ Hz then $i_f > i_b$ and $\tau_f > \tau_b$

When frequency is high or $f_s \geq 0.1$ Hz, then $\tau_f = \tau_b$ for each cycle, because deformation on each slope is small and it has not any effect on another slope of asperity, i.e. the time is very short for high frequencies that have no possibility to deform the second slopes of asperities when shear is running on the first slopes of asperities.; therefore, $i_f = i_B$ (Fig. 11c). With decreasing frequency or $f_s = 0.05$ Hz, then $\tau_f \geq \tau_b$ i.e. influence of forward movement on the second slopes of asperities starts and it can effect

on shear deformation in backward movement; hence, there is a time to deform slightly second slopes of asperities when shear is conducting on the first slopes of asperities, in consequence $i_f \geq i_b$ (Fig. 11b) and shear resistance starts to decrease in backward movement with respect to forward movement.

When frequency is low or $f_s \leq 0.01$ then $\tau_f > \tau_b$ and there is a significant deformation on second slopes of asperities i.e. there is sufficient time to deform second slopes of asperities when shear is running on the first slopes of asperities; hence, $i_f > i_b$ after each cycle (Fig. 11a). In this study, frequency of 0.05 Hz is a certain load frequency that shear mechanism changes on slopes of asperities.

Based on peak shear stress and joint dilation, it can be stated that the shear mechanism on the joint at high frequency of shear load does not depend on load direction (forward or backward movement) because the same diminishing was observed on the both slopes of asperities. However, at low frequency, the shear mechanism depends on the direction of starting cyclic loads (forward movement) because different diminishing has occurred on the slopes of asperities in forward and backward movements. These phenomena were observed at low normal stress. As the normal stress increased on the joint (0.5 and 1 MPa), most of the joint asperities were demolished at initial cycles (Figs. 13 and 14 in appendixes). After a few cycles, the joint behaves the same as a planar joint due to failure of all asperities.

Similar joint dilation in forward and backward movements at low displacement amplitude or $a_s \leq \lambda/4$, demonstrated that the shear mechanism was sliding only without significant degradation of the asperities (Fig. 11d) i.e. the joint asperities mostly have elastic behaviour during this displacement amplitude under cyclic loads; therefore, $i_f = i_B$ and direction of the forward or the backward movements has no any effect on asperities. With increasing displacement amplitude to $\lambda/3$ then influence of forward movement on the second slope of asperity starts (Fig. 11e) and it can affect shear deformation (starting elastoplastic behaviour on the joint asperities) in backward movement, hence, $i_f \geq i_b$ and shear resistance starts to decrease in backward movement with respect to forward movement. In case of higher displacement amplitude (± 8 mm) or $\lambda/3 < a_s \leq \lambda/2$, joint dilation in forward and backward movements was different significantly and it shows greater deformation (due to elastoplastic behaviour) on the second slope of asperities during the backward movements; therefore, shear mechanism is depend on the direction of starting cyclic load (Fig. 11f). Based on the description above, $\lambda/3$ is a certain displacement amplitude that shear mechanism can change on slopes of asperity before and after this value for shear cycles. Changing shear resistance based on different displacement amplitudes and Eq. 1 and Eq. 2 is presented in the below:

a) If $a_s \leq \lambda/4$, then $i_f = i_B$ and $\tau_f = \tau_b$
b) If $a_s = \lambda/3$, then $i_f \geq i_b$ and $\tau_f \geq \tau_b$
c) If $\lambda/3 < a_s \leq \lambda/2$, then $i_f > i_b$ and $\tau_f > \tau_b$

With increasing displacement amplitude or decreasing load frequency on the joint, peak joint dilation was difference between forward and backward movements (Fig. 11). This is because of unequal asperity diminishing or grater diminishing was observed on the second slopes of asperities (backward movements) during cyclic loads. Hence, failure on rock joints is more expected to happen in the opposite direction of starting cyclic loads when frequency of load decreases or displacement amplitude increases.

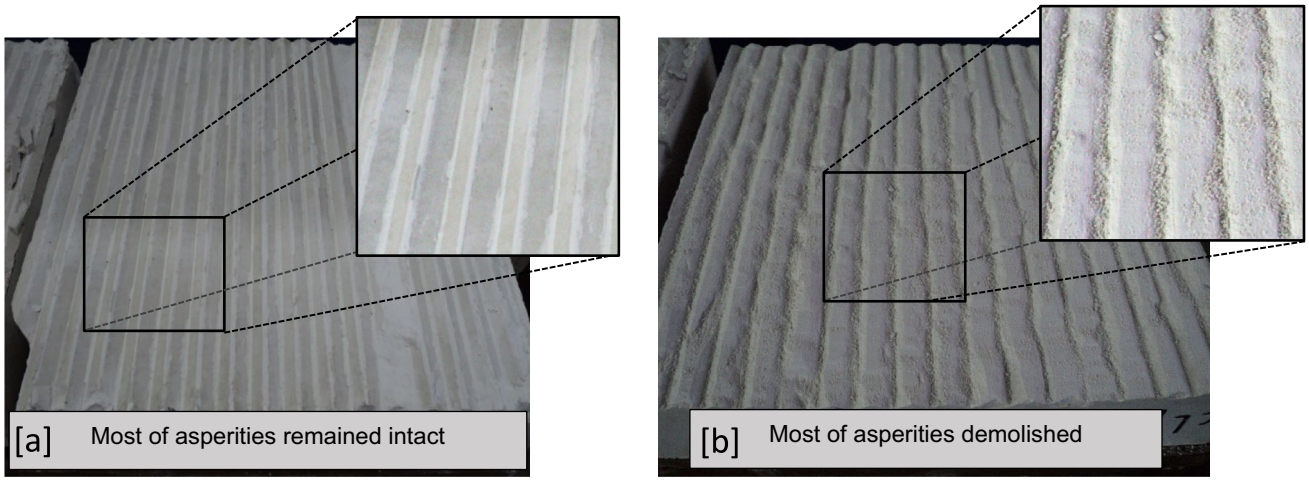


Fig. 12 Joint surface after 30 shear cycles for (**a**) λ/4 and (**b**) λ/2 displacement amplitudes, $P = 0.1$ MPa

Specimens were removed from the shear box to observe the effect of displacement amplitude on the joint surface. In the case of joint with low displacement amplitude (λ/4), e.g. low magnitude earthquakes most of the asperities remained intact (Fig. 12a) due to joint sliding during cyclic shear loads. With increasing displacement amplitude (λ/3), the degradation of asperities started. For high displacement amplitude (λ/2), e.g. high magnitude earthquakes, the degradation of asperities was significant (Fig. 12b).

This study conducted on regular joint to simplified behaviour of rock joints under different load frequency and displacement amplitude; however, rock joints are irregular and behaviour of irregular joints may have some deviation. Hence, extending this study on irregular joints (or natural rock joints) with different roughness is suggested to obtain shear behaviour of real rock joints in seismic regions.

Conclusions

A series of cyclic shear tests with different load frequencies and displacement amplitudes was conducted on soft synthetic rock joints under three levels of normal stress. The following conclusions can be obtained:

a) The first peak shear stress on the joint is significantly influenced by the frequency of cyclic loads and subsequent peaks of shear stress are largely influenced by the amplitude of displacement. Moreover, reduction of post peak shear stress is significant for the first cycle as the load frequency increases.
b) Shear load frequency of 0.05 Hz is a certain load frequency that shear mechanism changes on slopes of asperities under low normal stress. When frequency is 0.1 Hz (or $f_s \geq 0.1$ Hz), deformation on each slope of asperity is independent and it has not any effect on another slope of asperity; therefore, shear resistance in both directions (forward and backward) are equal. As the load frequency decreases to 0.05 Hz, the influence of diminishing on the second slopes of asperities starts. When frequency is 0.01 Hz (or $f_s \leq 0.01$) effect of the first diminishing (from the first slopes) on the second slopes of asperities is significant and the difference of shear resistance between the both directions increases.
c) One-third of asperity base length ($\lambda/3$) is a certain displacement amplitude under low normal stress for cyclic shear loads because shear mechanism changes from this displacement. Equal or less than one quarter of asperity base length ($\lambda/4$) has no effect on shear resistance on the joints due to sliding dominating along the shear plan. As displacement amplitude rises to one-third of asperity base length ($\lambda/3$), shear mechanism changes from sliding to diminishing of asperities and shear resistance starts to decrease. In case of higher displacement amplitude until half of asperity base length ($\lambda/2$) shear resistance decreases significantly and difference of shear resistance between the both directions increases
d) Shear mechanism on the joint asperities under low normal stress for high frequency of loads ($f_s \geq 0.1$ Hz) does not depend on load direction because diminishing on the both slopes of asperities is equal on the joint. However, shear mechanism depends on the direction of starting the loads at low frequency of loads ($f_s \leq 0.05$) or high displacement amplitude ($a_s \geq \lambda/3$). It is due to greater diminishing on the (second) slopes of asperities during backward movements on the joint. In consequence, failure on rock joints or rock mass is more expected to happen in the opposite direction of starting the loads. These results provide useful insight in practical geotechnical studies to expect orientation of rock mass failure with respect to direction of dynamic loads (e.g. earthquake or blasting).
e) The rock joint exhibits an opposite shear strength envelope for load frequencies and displacement amplitudes under cyclic loads. Frequency has a significant effect on shear strength envelope of the joint at initial joint movement (first cycle), but it is marginal for the next movements (or next cycles). However, displacement amplitude has no influence on shear strength envelope on the joints at initial joint movement, but it has a great influence on displacement amplitude after the next movements.

Appendix

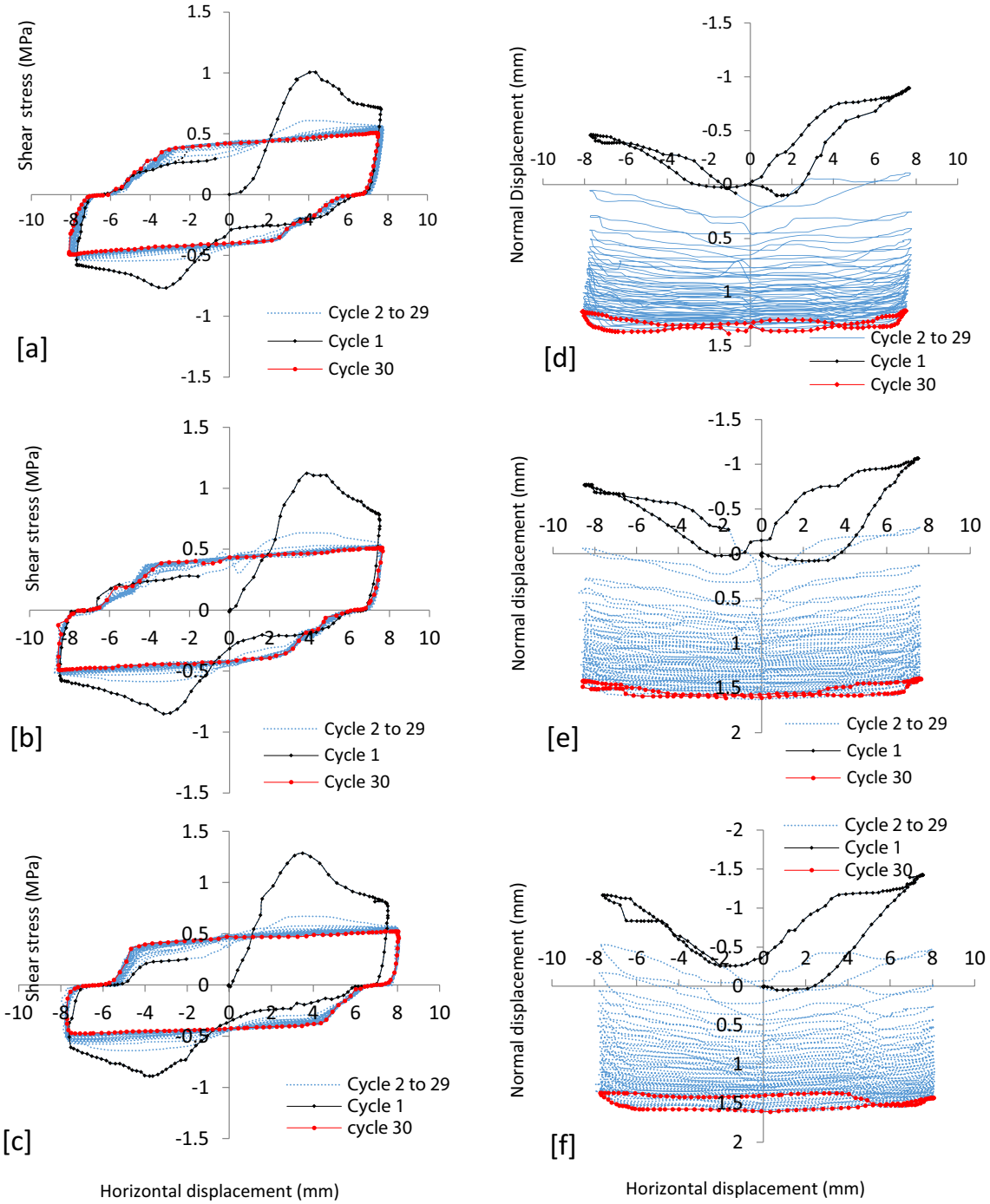


Fig. 13 (**a**, **b**, **c**) Shear stress and (**d**, **e**, **f**) normal displacement vs. horizontal displacement of the joint for cyclic shear load with different frequencies, 0.01, 0.050 and 0.1 Hz respectively ($P = 0.5$ MPa)

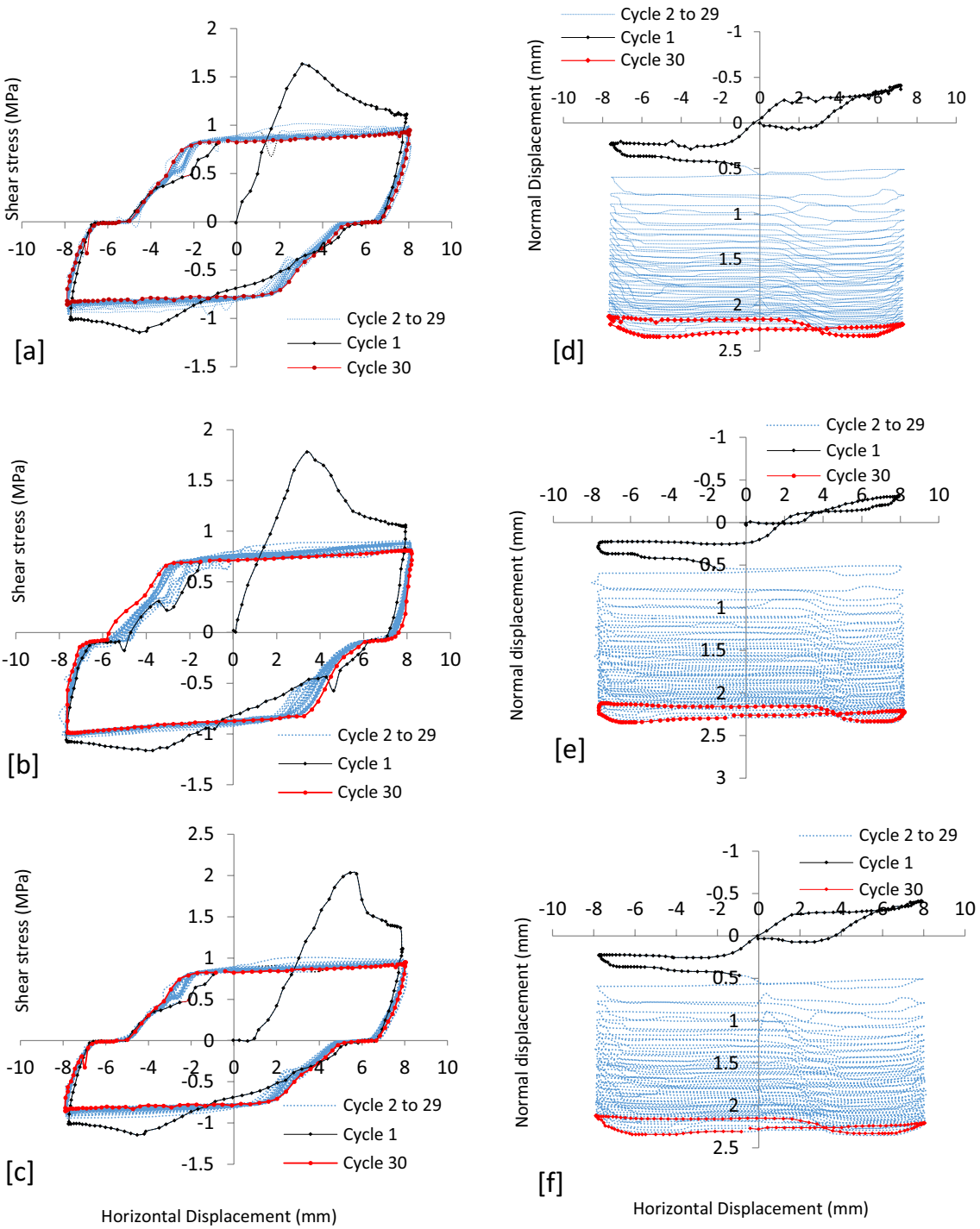


Fig. 14 (**a**, **b**, **c**) Shear stress and (**d**, **e**, **f**) normal displacement vs. horizontal displacement of the joint for cyclic shear load with different frequencies, 0.01, 0.05 and 0.1 Hz respectively ($P = 1$ MPa)

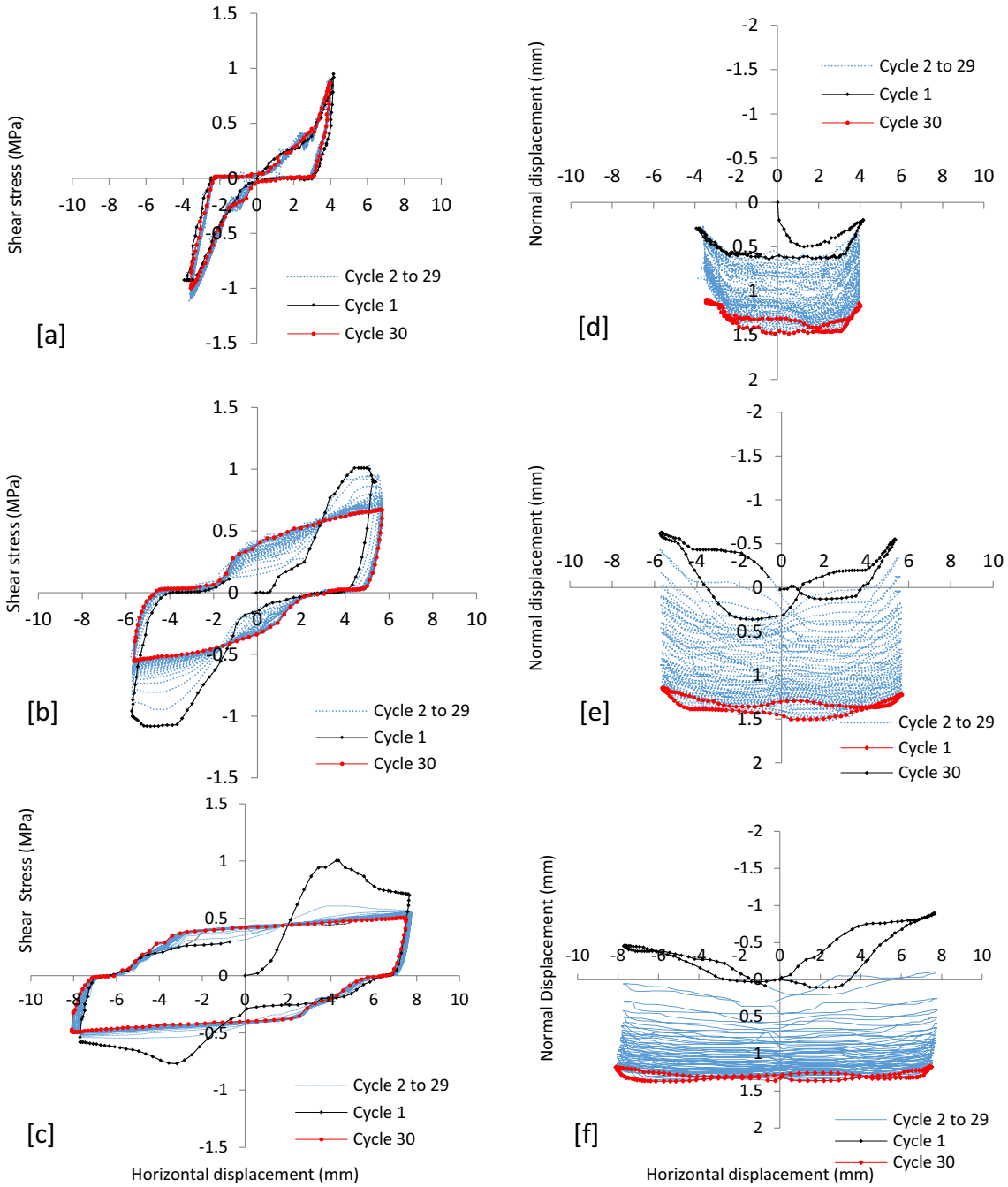


Fig. 15 (**a**, **b**, **c**) Shear stress and (**d**, **e**, **f**) normal displacement vs. horizontal displacement of the joint for cyclic shear load with different amplitudes, 4, 6 and 8 mm respectively ($P = 0.5$ MPa)

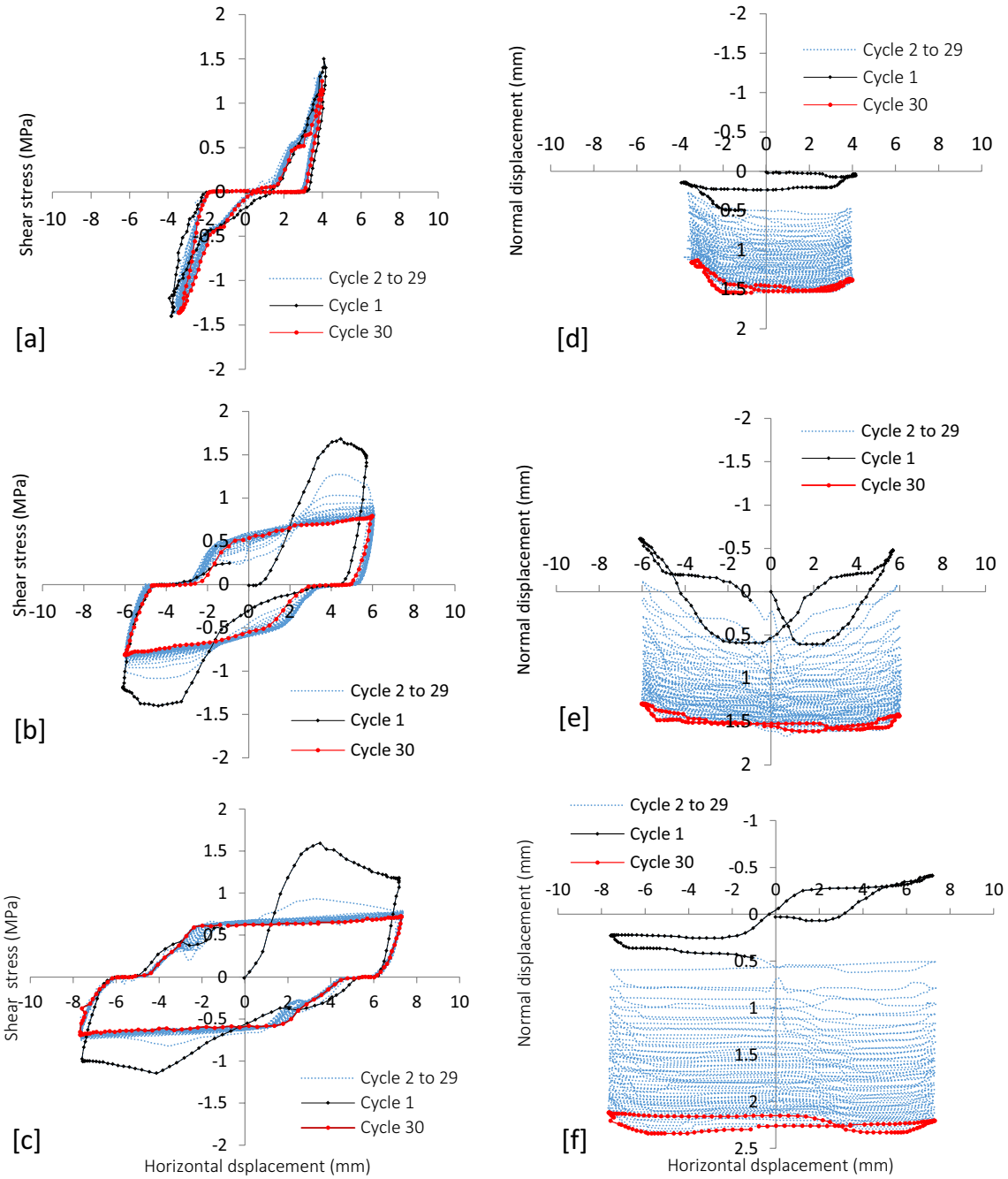


Fig. 16 (**a**, **b**, **c**) Shear stress and (**d**, **e**, **f**) normal displacement vs. horizontal displacement of the joint for cyclic shear load with different amplitudes, 4, 6 and 8 mm respectively ($P = 1$ MPa)

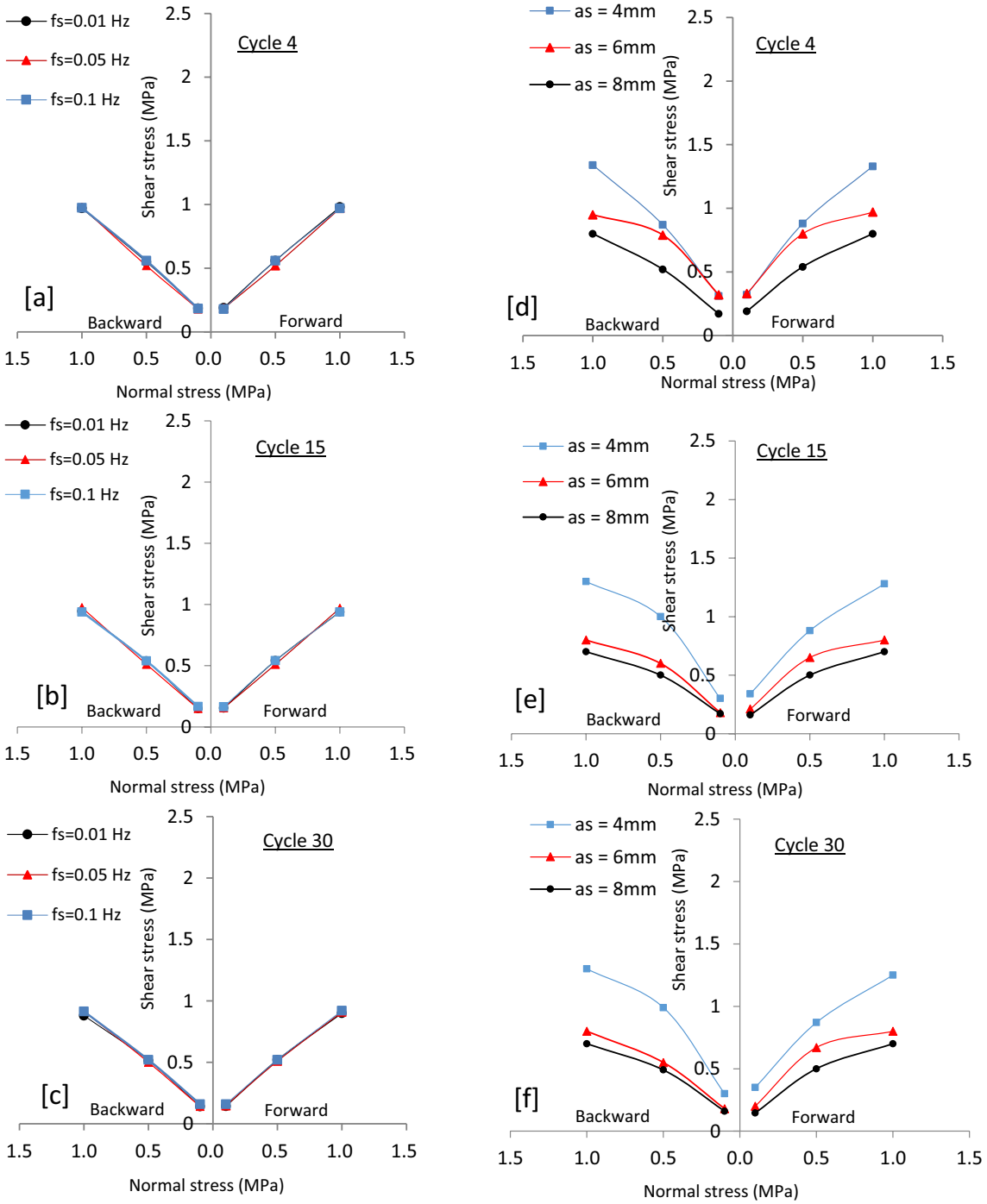


Fig. 17 Shear strength envelope of the joint under (**a**, **b**, **c**) varying shear load frequencies (fS = 0.01, 0.05, 0.1 Hz) and (**d**, **e**, **f**) amplitudes of displacement ($aS = \pm 4, \pm 6, \pm 8$ mm) for different cycles (4, 15, 30) respectively

Data availability statement Some or all data, used in this study are available from the corresponding author by request. There is no generated code in the present study.

Declarations

Conflicts of interest There are no conflict of interest to declare, and we believe that the manuscript and the study do not violate the journal's policies.

References

Atapour H, Moosavi M (2014) The influence of shearing velocity on shear behavior of artificial joints. Rock Mech Rock Eng 47(1):1745–1761. https://doi.org/10.1007/s00603-013-0481-9

Ban L, Tao Z, Du W, Hou Y (2023) A consecutive joint shear strength model considering the 3D roughness of real contact joint surface. Int J Min Sci Technol 33(5):617–624. https://doi.org/10.1016/j.ijmst.2022.12.009

Barbero M, Barla G, Zaninetti A (1996) Dynamic shear strength of rock joints subjected to impulse loading. Int J Rock Mech Min Sci 33(2):141–151. https://doi.org/10.1016/0148-9062(95)00049-6

Barla G, Barbero M, Scavia C, Zaninetti A (1990) Direct shear testing of single joints under dynamic loading. In: Barton N, Stephansson O (eds) Rock Joints. A. A. Balkema, Rotterdam, pp 447–454

Crawford AM, Curran JH (1981) The influence of shearing velocity on the frictional resistance of rock discontinuities. Int J Rock Mech Min Sci 18(1):505–515. https://doi.org/10.1016/0148-9062(81)90514-3

Dang W, Chen J, Huang L (2021) Experimental study on the velocity-dependent frictional resistance of a rough rock fracture exposed to normal load vibrations. Acta Geotech 16:2189–2202. https://doi.org/10.1007/s11440-021-01168-y

Dearman WR, Baynes FJ, Irfan TY (1978) Engineering grading of weathered granite. Eng Geol 12(1):345–374. https://doi.org/10.1016/0013-7952(78)90018-2

Desai CS, Pradhan SK, Cohen D (2005) Cyclic testing and constitutive modeling of saturated sand–concrete interfaces using the disturbed state concept. Int J Geomech 5(4):286–294. https://doi.org/10.1061/(ASCE)1532-3641(2005)5:4(286)

Dong S, Feng W, Yin Y, Hu R, Dai H, Zhang G (2020) Calculating the permanent displacement of a rock slope based on the shear characteristics of a structural plane under cyclic loading. Rock Mech Rock Eng 53(1):4583–4598. https://doi.org/10.1007/s00603-020-02188-y

Ferrero AM, Migliazza MR, Tebaldi G (2010) Development of a new experimental apparatus for the study of the mechanical behaviour of a rock discontinuity under monotonic and cyclic loads. Rock Mech Rock Eng 43(6):685–695. https://doi.org/10.1007/s00603-010-0111-8

Gillette DR, Sture S, Ko HK, Gould M, Scott G (1983) Dynamic behavior of rock joints. In: Proceedings of the 24th US symposium on rock mechanics, pp 163–79

Gómez S, Sanchidrián JA, Segarra P (2020) Near-field vibration from blasting and rock damage prediction with a full-field solution. Int J Rock Mech Min Sci 134 (1). https://doi.org/10.1016/j.ijrmms.2020.104357

Guansheng H, Hongwen J, Yujing J, Richeng L, Jiangyu W (2020) Effect of cyclic loading on the shear behaviours of both unfilled and infilled rough rock joints under constant normal stiffness conditions. Rock Mech Rock Eng 53(1):31–57. https://doi.org/10.1007/s00603-019-01866-w

Hassani FP, Scoble MJ (1985) Frictional mechanism and properties of rock discontinuities. In: Paper on fundamental of rock joints proceeding of the international symposium on fundamental of rock joints, Lulea Sweden Centek Pub, Bjorkliden, pp 15–20

He L, Lu X, Copeland T, Chen J, Yan J, Zhao W (2023) The Impact of Shear Rate on the Mechanical Behavior of Rock Joints Under Multiple-Influencing Factors. Rock Mech Rock Eng 56(10):7397–7414. https://doi.org/10.1007/s00603-023-03432-x

Hencher SR (1981) Friction parameters for the design of rock slopes to withstand earthquake loading. Proc of the Conference on Dams and Earthquake. https://doi.org/10.1680/dae.01237.0012

Homand-Etienne F, Lefeˆvre F, Belem T, Souley M (1999) Rock joints behaviour under cyclic direct shear tests. In: Amadei, Kranz, Scott, Smealie (eds) Rock mechanics for industry, vol 2. Balkema, Rotterdam, pp 399–406

Huang X, Haimson BC, Plesha ME, Qiu X (1993) An Investigation of the Mechanics of Rock Joints. Int J Rock Mech Min Sci 30(3):257–269. https://doi.org/10.1016/0148-9062(93)92729-A

Hutson RW, Dowding CH (1990) Joint asperity degradation during cyclic shear. Int J Rock Mech Min Sci 27(2):109–119. https://doi.org/10.1016/0148-9062(90)94859-R

Indraratna B, Haque A, Aziz N (1998) Laboratory modelling of shear behaviour of soft joints under constant normal stiffness conditions. Geotech Geol Eng 16:17–44. https://doi.org/10.1023/A:1008880112926

Jafari MK, Amini KH, Pellet F, Boulon M, Buzzi O (2003) Evaluation of shear strength of rock joints subjected to cyclic loading. Int J Soil Dyn Earth Eng 23(7):619–630. https://doi.org/10.1016/S0267-7261(03)00063-0

Jing L, Stephansson O, Nordlund E (1993) Study of rock joints under cyclic loading conditions. Rock Mech Rock Eng 26(3):215–232. https://doi.org/10.1007/BF01040116

Lee HS, Park YJ, Cho TF, You KH (2001) Influence of asperity degradation on the mechanical behavior of rough rock joints under cyclic shear loading. Int J Rock Mech Min Sci 38(7):967–980. https://doi.org/10.1016/S1365-1609(01)00060-0

Li JC, Li HB, Jiao YY, Liu YQ, Xia X, Yu C (2014) Analysis for oblique wave propagation across filled joints based on thin-layer interface model. J Appl Geophys 102:39–46. https://doi.org/10.1016/j.jappgeo.2013.11.014

Li H, Deng J, Yin J, Qi S, Zheng B, Zhu J (2021) An experimental and analytical study of rate-dependent shear behaviour of rough joints. Int J Rock Mech Min Sci 142(1):104702. https://doi.org/10.1016/j.ijrmms.2021.104702

Li Z, Wu W, Li J, Zhao J (2022) Dynamic tensile failure of a V-shaped canyon induced by vertically travelling SV waves. Soil Dyn Earthq Eng 162:107458. https://doi.org/10.1016/j.soildyn.2022.107458

Lu HF, Zhang QZ (2021) Investigations on shear properties of soft rock joints under grouting. Rock Mech Rock Eng 54(1):1875–1883. https://doi.org/10.1007/s00603-021-02366-6

Meng F, Wong LNY, Zhou H, Guangtan C (2019) shear rate effects on the post-peak shear behaviour and acoustic emission characteristics of artificially split granite joints. Rock Mech Rock Eng 52(1):2155–2174. https://doi.org/10.1007/s00603-018-1722-8

Mirzaghorbanali A, Nemcik J, Aziz N (2014) Effects of shear rate on cyclic loading shear behaviour of rock joints under constant normal stiffness conditions. Rock Mech Rock Eng 47(5):1931–1938. https://doi.org/10.1007/s00603-013-0453-0

Nemcik J, Mirzaghorbanali A, Aziz N (2014) An elasto-plastic constitutive model for rock joints under cyclic loading and constant normal stiffness. Geotech Geol Eng 32(2):321–335. https://doi.org/10.1007/s10706-013-9716-5

Nguyen VM (2013) Static and dynamic behaviour of joints in schistose rock: lab testing and numerical simulation. Ph.D. dissertation, Geotechnical Institute, TU Bergakademie Freiberg

Nian G, Chen Z, Zhu T, Zhang L, Zhou Z (2024) Experimental study on the failure of fractured rock slopes with anti-dip and strong weathering characteristics under rainfall conditions. Landslides 21:165–182. https://doi.org/10.1007/s10346-023-02142-3

Niktabar SMM, Rao KS, Shrivastava AK (2017) Effect of rock joint roughness on its cyclic shear behavior. Rock Mech Geotech Eng 9(6):1071–1084. https://doi.org/10.1016/j.jrmge.2017.09.001

Niktabar SMM, Rao KS, Shrivastava AK (2018) Automatic static and cyclic shear testing machine under constant normal stiffness boundary conditions. Geotech Testing J 41(3):508–525. https://doi.org/10.1520/GTJ20170083

Niktabar SMM, Rao KS, Shrivastava AK, Ščučka J (2023) Effect of Varying Normal Stiffness on Soft Rock Joints under Cyclic Shear Loads. Materials 16(12):4272. https://doi.org/10.3390/ma16124272

Özvan A, Dinçer I, Acar A, Özvan B (2014) The effects of discontinuity surface roughness on the shear strength of weathered granite joints. Bull Eng Geol Environ 73:801–813. https://doi.org/10.1007/s10064-013-0560-x

Patton FD (1966) Multiple modes of shear failure in rock and related materials. Ph.D.dissertation, University of Illinois, Urbana–Champaign

Qi SW, Zheng BW, Wu FQ, Huang XL, Guo SF, Zhan ZF, Zou Y, Barla G (2020) A new dynamic direct shear testing device on rock joints. Rock Mech Rock Eng 53(10):4787–4798. https://doi.org/10.1007/s00603-020-02175-3

Qiu X, Plesha ME, Huang X, Haimson BC (1993) An investigation of the mechanics of rock joints. Part II—analytical investigation. Int J Rock Mech Min Sci 30(1):271–287. https://doi.org/10.1016/0148-9062(93)92730-E

Ram BK, Basu A (2019) Shear behavior of unfilled-planar quartzitic rock joints with reference to weathering grade of joint surfaces. Rock Mech Rock Eng 52(1):4113–4121. https://doi.org/10.1007/s00603-019-01815-7

Ramulu M, Chakraborty A, Sitharam T (2009) Damage assessment of basaltic rock mass due to repeated blasting in a railway tunnelling project—a case study. Tunn Undergr Space Technol 24(2):208–221. https://doi.org/10.1016/j.tust.2008.08.002

Schneider HJ (1977) The time dependence of friction of rock joints. Bull Eng Geol Environ 16(1):235–239. https://doi.org/10.1007/BF02591495

Scholz CH, Engelder JT (1976) The role of asperity indentation and ploughing in rock friction — I: asperity creep and stick-slip. Int J Rock Mech Min Geomech Abstr 13(5):149–154. https://doi.org/10.1016/0148-9062(76)90819-6

Shrivastava AK (2013) Rao KS (2013) Development of a large-scale direct shear testing machine for unfilled and infilled rock joints under constant normal stiffness conditions. Geotech Testing J 36(5):670e9. https://doi.org/10.1520/GTJ20120155

Tang ZC, Wong LNY (2016) Influences of normal loading rate and shear velocity on the shear behavior of artificial rock joints. Rock Mech Rock Eng 49(6):2165–2172. https://doi.org/10.1007/s00603-015-0822-y

Thongprapha T, Liapkrathok P, Chanpen S, Fuenkajorn K (2020) Frictional behavior of sandstone fractures under forward-backward pre-peak cyclic loading. J Struct Geol 138(1):104106. https://doi.org/10.1016/j.jsg.2020.104106

Tiwari G, Latha GM (2019) Shear velocity-based uncertainty quantification for rock joint shear strength. Bull Eng Geol Environ 78:5937–5949. https://doi.org/10.1007/s10064-019-01496-0

Woo I, Fleurisson J-A, Park H-J (2010) Influence of weathering on shear strength of joints in a porphyritic granite rock mass in Jechon area. South Korea Geosci J 14(3):289–299. https://doi.org/10.1007/s12303-010-0026-0

Wu XZ, Jiang YJ, Gong B, Guan ZC, Deng T (2019) Shear performance of rock joint reinforced by fully encapsulated rock bolt under cyclic loading condition. Rock Mech Rock Eng 52(1):2681–2690. https://doi.org/10.1007/s00603-018-1698-4

Wu X, Zheng H, Jiang Y, Deng T, Xiao G, Wang Y (2023) Effect of Cyclic Shear Loading on Shear Performance of Rock Bolt Under Different Joint Roughness. Rock Mech Rock Eng 56(3):1969–1980. https://doi.org/10.1007/s00603-022-03174-2

Wu Q, Liu Z, Tang H, Wang L, Huo X, Cui Z, Li S, Zhang B, Lin W (2024) Experimental investigation on shear strength deterioration at the interface between different rock types under cyclic loading. J Rock MechGeotech Eng. https://doi.org/10.1016/j.jrmge.2023.11.019

Yang J, Cai J, Yao C, Li P, Jiang Q, Zhou C (2019) Comparative study of tunnel blast-induced vibration on tunnel surfaces and inside surrounding rock. Rock Mech Rock Eng 52(11):4747–4761. https://doi.org/10.1007/s00603-019-01875-9

Zheng BW, Qi SW, Guo SF, Huang XL (2015) The influence of shear deformation velocity on the strength characteristics of rock joints. In: 13th International congress of rock mechanics, 10-143 May, Montreal

131 (2024) 205459

Contents lists available at ScienceDirect

Gas Science and Engineering

journal homepage: www.journals.elsevier.com/gas-science-and-engineering

Effect of promoters in hydrates based carbon dioxide capture: A review

Aseem Dubey, Akhilesh Arora *

Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi, 110042, India

ARTICLE INFO

Keywords:
Gas hydrate capture
Thermodynamic promoters
Kinetic promoters
Synergist
Hybrid integration
Challenges

ABSTRACT

Climate change necessitates a paradigm shift toward using clean energy technologies and carbon dioxide (CO_2) capture from emitting sources to alleviate the greenhouse effect and meet sustainable development objectives. Among the competing carbon capture technologies, high CO_2 storage capacity, use of water, and low capturing cost are critical advantages for applying gas hydrate based CO_2 capture. However, high equilibrium pressure and slow kinetics are issues in gas hydrate formation. The review contributes an exhaustive analysis of hydrate based technology and the additives used to stimulate hydrate formation, focusing on thermodynamic and kinetic promoters. Furthermore, the recent emerging research progress of hydrate-based CO_2 capture encompassing nanoparticles, porous materials, synergists, and hybrid hydrate-membrane methods is also presented. Heat integration in the process of novel hydrate-membrane with promoters has drawn attention appreciably to tackle energy, cost, and ecological concerns, including improvement in technology readiness level. Furthermore, innovative approaches, challenges, and aligned research and development opportunities are presented, which will assist the researchers in exploring advanced promoters and hybrid integrated technologies to compete with existing CO_2 capture technologies.

1. Introduction

Carbon dioxide (CO_2) and methane are the two primary forms of carbon in the atmosphere and play a significant role in global warming. However, methane has a shorter life and lower atmospheric concentration than carbon dioxide. The impact of carbon dioxide as a greenhouse gas has been more significant, and its anthropogenic emissions contribute to more than 60.0% of global warming (Rodhe, 1990). The CO_2 emission from fossil fuel electric power plants has been accounted for about 33% globally (Song et al., 2013). In the year 2018, the leading CO_2 emitters were China (30.0% of the global), the United States (15.0%), India (7.0%), European countries (13%), and 35.0% from the rest of the world (IEA, 2019). In 2021, the global carbon dioxide emissions amounted to 3.63 billion tons, nearly 89% of which was contributed by fossil fuel utilization (IEA, 2022). The atmospheric concentration of CO_2 in 2022 exceeded 415 ppm and is expected to increase further in the coming years (Dubey and Arora, 2022; Lee et al., 2023). Because of the adverse effects of increasing global warming owing to an increase in the atmospheric CO_2 concentration, it is necessary to use appropriate carbon capture and storage (CCS) technologies to reduce CO_2 emissions from various emission sources. To meet the objectives of COP 26, worldwide annual CO_2 capture is to be increased to 715 million tons by 2030 and 4200 million tons by 2050 (Aminnaji et al., 2024a,b). Many countries have been engaged in CO_2 capture projects in response to the global climate change challenge. India has launched numerous initiatives to encourage sustainable energy technologies, including carbon capture for a 45% CO_2 emission reduction by 2030 and net zero by 2070 (Pandey et al., 2022).

Besides the intensive deployment of renewable energy, carbon capture and storage (CCS) is still projected to contribute a cumulative reduction of 15% to achieve net zero carbon emission (Intergovernmental Panel on Climate Change, 2018). The process involves CO_2 capture, transportation, and storage deep underground or its utilization, and could be a viable technique to meet the objective without impacting the economic development of countries (Bui et al., 2018). A detailed study on technical and non-technical challenges, societal and market acceptability of CCS technology shows it promising in the long run, to mitigate CO_2 emissions significantly to 70.0% by 2050 (Shirmohammadi et al., 2020). Nevertheless, a study on CCS technologies has revealed certain obstacles, such as technological innovation, financial resources, and regulatory frameworks, which necessitate the formation of alliances between various sectors to reduce deployment barriers (Zhang, 2021; Karaszova et al., 2020). The CCS-integrated conventional power plants have been reported to be cost-effective

* Corresponding author.
E-mail addresses: profakhilesharora@gmail.com, akhilesharora@dce.ac.in (A. Arora).

https://doi.org/10.1016/j.jgsce.2024.205459
Received 2 April 2024; Received in revised form 11 September 2024; Accepted 18 September 2024
Available online 20 September 2024
2949-9089/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

despite an increase in energy penalties ranging from 15 to 40% (Wang et al., 2020). CO_2 capture could be undertaken in pre-combustion, post-combustion, or oxy-fuel combustion, depending on CO_2 generation and how it gets captured. The terrell project, erected in the USA in 1972 with an annual CO_2 capture capacity of 1.7 Mt, was the earliest (Orujov et al., 2023). Between 1914 and 2020, worldwide 19 projects (nine pre-combustion and ten post-combustion) primarily integrated with power generation (i.e., US-6, UK-4, China-4, Canada-2, Korea-2, Norway-1), while three projects using oxyfuel combustion (i.e., US-1, UK-1, China-1), have been operational to capture CO_2 (Vega et al., 2020). Don Valley project (UK) has been the largest CO_2 capture capacity project (5 Mtpa). Moreover, the largest commercial-scale project applied to coal-fired power plant with post-combustion capture and CO_2 storage has been the Sask power boundary dam project (Canada). D'Alessandro et al. (2010) reported challenges and advancements in carbon capture using specialized materials which perform the separation, and found chemical looping, gas hydrate crystallization, and hybrid materials as promising technologies. A diverse range of promising new materials exist for CO_2 capture/separation using post-combustion, pre-combustion, and oxyfuel processes, as depicted in Fig. 1.

The conventional methods for CO_2 separation include physical and chemical absorption, adsorption, cryogenic separation, and membrane separation, which suffer from high energy penalty of regeneration, high process cost, low recovery efficiency, and impacts environmentally (Feyzi and Mohebbi, 2020). Meyer and Jouenne (2022) reviewed the work on industrial carbon capture by aqueous amine absorption methods for the post-combustion route. The researchers emphasized addressing the solvent's environmental concern for economic assessment, indicating higher costs for solvent management, and suggested to use solar-assisted carbon capture technology for reducing the energy penalty. Further, Sheng et al. (2023) summarized the work on gas-liquid microdispersion devices and microflow regulation technology for mass transfer enhancement for CO_2 absorption, identifying gas–liquid micro-reaction methods as an emerging process for enhancing CO_2 capture and utilization. The authors concluded that for actual chemical systems using micro dispersion technology, the controllable operation in a micro-reactor remains a challenge, further focusing on developing surfactant-free micro dispersion technology using a multistage counter-current flow micro-reactor.

Among various strategies, gas hydrate based CO_2 separation and capture technologies have exhibited tremendous potential in terms of physical feasibility, cost, and lower energy utilization (Koh et al., 2011). A one m^3 of hydrate can release ~160–180 m^3 of gas with an energy consumption of 3 MJ/kg, costing 8.75 $/t$CO_2$ (Cheng et al., 2021). Gas hydrates are non-stoichiometric water based gas storage media with a crystalline structure resembling ice. In gas hydrate, hydrogen-bonded water molecules 'hosts' form special cavities through hydrogen bonds. The gas molecules 'guests' are trapped inside the cavities under specific low temperature and high-pressure conditions, owing to van der Waals force (Dashti et al., 2015). The hydrate structure is stabilized by the repulsive interactions between guest and host molecules (Sloan and Koh, 2007). Gas hydrates can be used to store energy carriers such as methane, ethane, and hydrogen, as well as greenhouse gases such as carbon dioxide, nitrous oxide, and sulfur hexafluoride. Recently, Amminaji et al. (2024a-b) reviewed the work on CO_2 gas hydrate for CCS, including carbon management, and found CO_2 hydrate more thermodynamic stable compared to other industrial gases due to the difference in phase equilibrium, suggesting CO_2 hydrates as a viable technology. They have further highlighted a significant challenge in the micro and meso-scale studies of CO_2 hydrates due to the limitations of the current experimental and computational techniques and focused on multidisciplinary research collaborations. Fig. 2 shows the schematic of hydrate formation/decomposition. Exothermic hydrate formation occurs when CO_2 molecules combine with H_2O molecules at equilibrium. Moreover, during endothermic dissociation in the reactor, trapped gas hydrates are dissociated, separating trapped gas to capture, and the water/binary mixture is left behind in the dissociation reactor for further re-circulation.

The slow rate of hydrate formation and high pressure requirement limits the application of hydrate-based technology for commercialization (Xia et al., 2022). Most gas hydrates require relatively higher pressure and lower temperature to hold gaseous guests in the hydrate phase. The hydrate forming pressure at 273.0 K for H_2, N_2, CH_4, and CO_2 gases has been reported to be 2130, 163, 265, 111, and 122 bar, respectively (Wang et al., 2013a). Many thermodynamic and kinetic promoters, including mechanical methods, have been extensively investigated to modulate temperature and pressure conditions for practical application and to promote hydrate formation (Li et al., 2019). The agitation is found to be effective but increases energy consumption by ~0.541 kWh/mol CO_2 captured (Wang et al., 2020). Thermodynamic promoters alter the thermodynamic conditions, allowing hydrate formation at lower pressures and higher temperatures, depending on the type of additives and their concentration (Eslamimanesh et al., 2012). In recent years, thermodynamic and kinetic additives have been proposed as a potential solution for CO_2 hydrates. Adding promoters such as

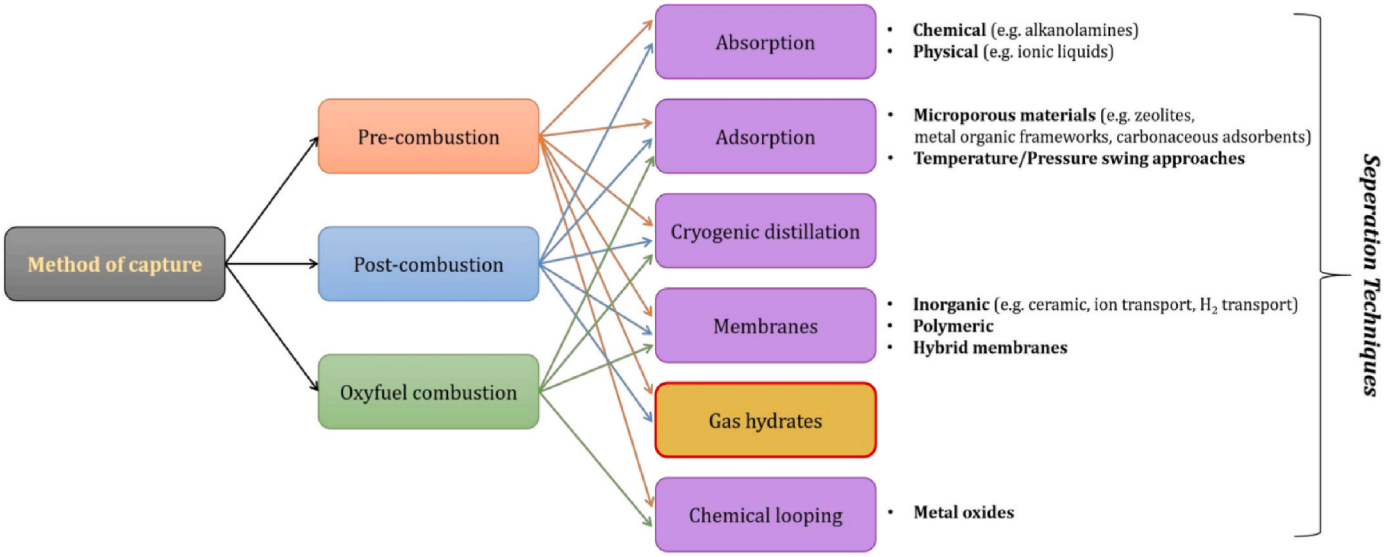


Fig. 1. Various physical and chemical process used in CO_2 capture/separation (D'Alessandro et al., 2010).

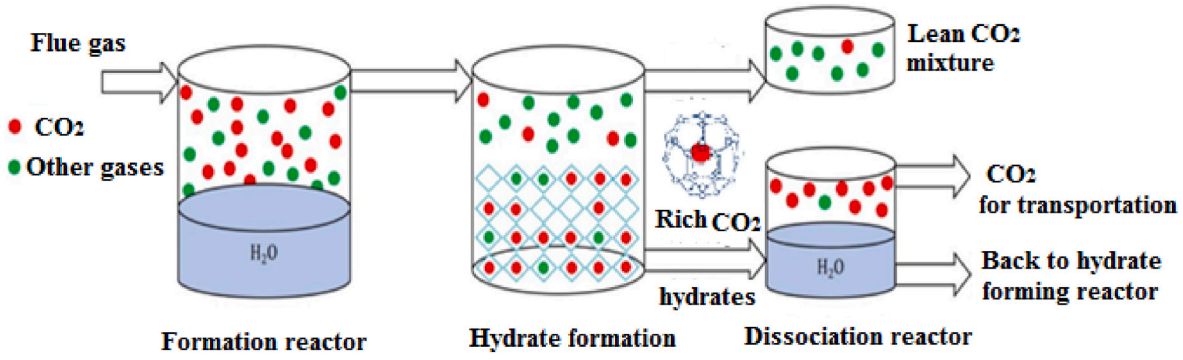


Fig. 2. Schematic of CO_2 hydrate formation/separation.

tetrahydrofuran (THF), tetrabutylammonium bromide (TBAB), cyclopentane (CP), sodium dodecyl sulfate (SDS), sodium dodecylbenzene sulfonate (SDBS), etc. have increased the hydrate based efficiency of CO_2 capture (Muromachi et al., 2014; Pahlavanzadeh et al., 2019). In the presence of tetrahydrofuran (THF), the appearance of hydrate was reported at a significantly lower pressure, reducing induction time, increasing the rate of hydrate growth, and decreasing the carbon capture cost from 75 to 53% (Zheng et al., 2016). Babu et al. (2016) studied the effect of THF composition (0.5, 1.0, and 5.6 mol%) under different driving force conditions for pre-combustion CO_2 capture, identifying the optimum 5.56 mol% THF concentration for average gas uptake of 13.9 mmol of gas/mol of water at optimum operating conditions (60 bar and 282.2 K). Xie et al. (2017) reported a considerable decrease in energy consumption from 8.8 to 3.89 MJ/kg CO_2 when the hydrate-based system (1.0 mol% THF) was integrated with membrane separation. Wang et al. (2013b) explored a new pathway for storage and transportation of CH_4 and CO_2 in 'bioclathrate' form by using pre-structured biological 'sorbents' such as plants and fungi. The accelerated hydrate formation kinetics was reported, along with the advantage of avoiding the use of energy-intensive technologies. The effect of adding porous materials has shown an increased performance of CO_2 capture due to the availability of pores and affects the phase equilibrium of hydrate formation, thus improving the gas storage performance (Wang et al., 2022; Aminnaji et al., 2024a,b). The synergistic effect of improving the kinetics of gas hydrate formation by adding silica gel (100 nm pore size) has been identified, which increased further by mixing zeolite (Arora et al., 2021).

The utilization of gas hydrates as a highly promising technology for CO_2 capture and separation has several advantages, such as (i) water-based process, (ii) tolerance to impurities in the feed gas, (iii) low energy intensive via integrating hydrate formation/dissociation enthalpies, and (iv) exhibit tremendous potential in terms of physical feasibility and lower energy utilization criteria. Moreover, it is yet to be commercialized owing to low hydrate formation rate, low gas consumption, slow kinetics, lack of industrial experience, and interfacial hindrance (Nguyen et al., 2022). Therefore, promoters that can modulate thermodynamic conditions for practical application and enhance gas consumption are extensively investigated. The kinetic promoters that accelerate the nucleation and growth of hydrate are also investigated. In the present work, an in-depth exploration of hydrate based CO_2 capture technology, highlighting the use of various promotional additives, is carried out for the CO_2 separation from gas mixture. The hybrid based hydrate technologies, development opportunities, and approaches for the commercialization of the technology are also presented. The review provides significant insights into the advancements made in gas hydrate technologies and identifies the challenges. The overview of the advancements associated with hydrate promoters will aid researchers in comprehending and selecting the most suitable promoters by combining the advantages of each to overcome the obstacles to the commercialization of this novel technology.

2. Thermodynamics of gas hydrate formation

Like other hydrates, CO_2 hydrate is an ice-like lattice structure formed by water and gas molecules at low temperatures and under high pressure, with one gas molecule of appropriate size trapped within each cage to form a stabilized hydrate containing 85.0 mol% water (Sloan and Koh, 2007). The van der Waals force stabilizes the hydrate formation and lump formed at a pressure of 30.0–100.0 bar and temperature of 275.0–285.0 K (Qasim et al., 2019). Thoutam et al. (2019) found that hydrate equilibrium pressure depends on the gas composition and temperature. The CO_2 has been more susceptive to hydrate formation (~4 times hydrate concentration) at low pressures, compared to other gases, and captured easily under similar conditions (Duc et al., 2007). The phase behavior of gas hydrates considerably changes depending on the size, shape, and chemical nature of guest molecules (Lee et al., 2023). The gas hydrate equilibrium curve controls the stability of gas hydrates, indicating the temperature and pressure at which hydrates can be formed. Cheng et al. (2021) presented a phase equilibrium shift on adding varying mass fractions of TBAB thermodynamic promoter in the range of 5–32 wt%, and found more shift towards right at higher mass fraction of additives (Fig. 3). Mass fraction of thermodynamic promoters indicates the proportion of these promoters in a mixture or system relative to the total mass. Additive molecules are trapped inside the hydrate cages at lower pressures and/or higher temperatures, and thus, milder conditions are needed to form gas hydrates compared to the system without promoter. If the temperature exceeds the equilibrium, there would be no hydrate formation.

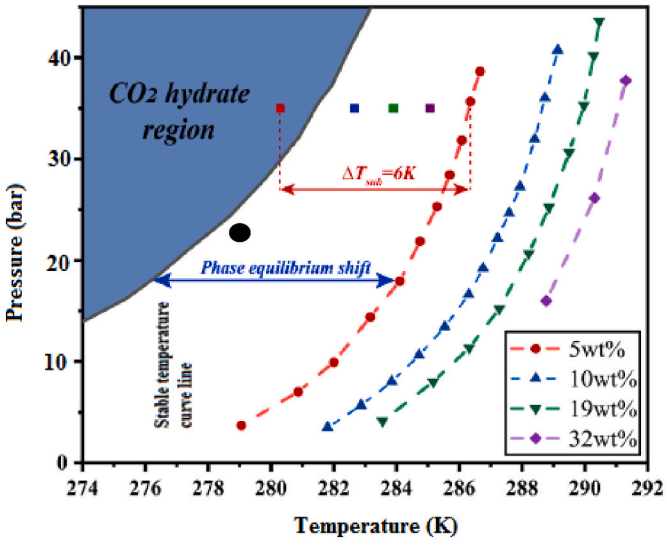


Fig. 3. Hydrate equilibrium curve and phase equilibrium shift on addition of TBAB promoter at various mass fractions of 5–32 wt% (Cheng et al., 2021).

Based on the unit-cell configurations of clathrate hydrates, they can have the three most communally crystalline hydrate structures, namely I (sI), II (sII), and H (sH) based on van der Waals radii of the confined molecules. These structures exhibit distinct shapes, including variations in the types and numbers of cavities they possess, and their formation conditions are determined by the type of gas molecules they contain (Dashti et al., 2015). The summary of hydrate structures formed in combination with related cages is shown in Fig. 4. These structures are important in understanding the behavior of clathrate hydrates, including the storage and release of guest molecules such as CO_2. The structure transition from sI to sII and sH increases the overall storage capacity of hydrates. Pure CO_2 and CH_4 hydrate belong to structure I due to the small molecule sizes. Structure II hydrates form with larger gas molecules such as nitrogen (N_2) and propane. With the mixture of small and large gas molecules, such as methane and cycloheptane, the structure H may form (Sloan and Koh, 2007; Dashti et al., 2015). Each unit cell of the sI structure is dodecahedral, formed using purified gas, and can accommodate 46 H_2O molecules per unit cell and consists of 2 small 5^{12} pentagonal dodecahedral cages (containing 12 pentagonal faces on the cage) and 6 large $5^{12}6^2$ tetrakaidecahedral cages (contains 12 pentagonal and 2 hexagonal faces on the cage). In contrast, the unit cell of the sII structure is tetrahedral in shape, comprised of 136 H_2O molecules consisting of 16 small 5^{12} cages and 8 large $5^{12}6^4$ hexacaidecahedral cages (containing 12 pentagonal and 4 hexagonal faces on the cage), and formed in the presence of promoters. The larger cages in sII hydrate can accommodate more guest molecules, including CO_2. The sH structure can accommodate 34 water molecules per unit cell and consists of three small 5^{12} cages, two medium $4^3 5^6 6^3$ mid cages (containing 3 square, 6 pentagonal and 3 hexagonal cage faces), and one large $5^{12}6^8$ icosahedral cages (containing 12 pentagonal and 8 hexagonal faces on the cage) and form a stable structure in the presence of gases (Koh et al., 2011; Sayani et al., 2022). The mean radius of the cavity of 5^{12} in sI, sII, and sH structures is 0.395, 0.391, and 0.391 nm, respectively (Sloan and Koh, 2007). The type of structure formed by the gas depends primarily on the diameter of the guest molecular (Xia et al., 2022). Molecules of gases such as CH_4, C_2H_6, and CO_2 having a diameter (d) in the range of 0.42–0.6 nm and form sI structure, while molecules of H_2, N_2 ($d < 0.42$ nm, and 0.6 nm $< d <$ 0.7 nm) form sII structure. When molecules of some gases such as cyclopentane and neohexane (0.7 nm $< d <$ 0.9 nm) are combined with small guest molecules such as CH_4, H_2S to form sH hydrates structure (Chaturvedi et al., 2021). In sH structure, CO_2 molecules can occupy small and large cages (Majid et al., 2021).

Compared to various CCS methods, the main advantage of hydrate based CO_2 capture is its large storage capacity (i.e. one volume of pure CO_2 hydrate can capture 156 vol of CO_2 gas) at standard temperature and pressure (Wang et al., 2020). The water molecule forms hydrate lattices involving subsequent reactions during capturing and dissociation.

$$CO_2 + nH_2O \rightarrow CO_2.\, nH_2O + heat \quad (1)$$

$$CO_2.\, nH_2O + heat \rightarrow CO_2 + nH_2O \quad (2)$$

2.1. Factors influencing gas consumption

Gas consumption during the hydrate capture process is influenced by temperature, pressure, driving forces, gas concentration, water-to-gas

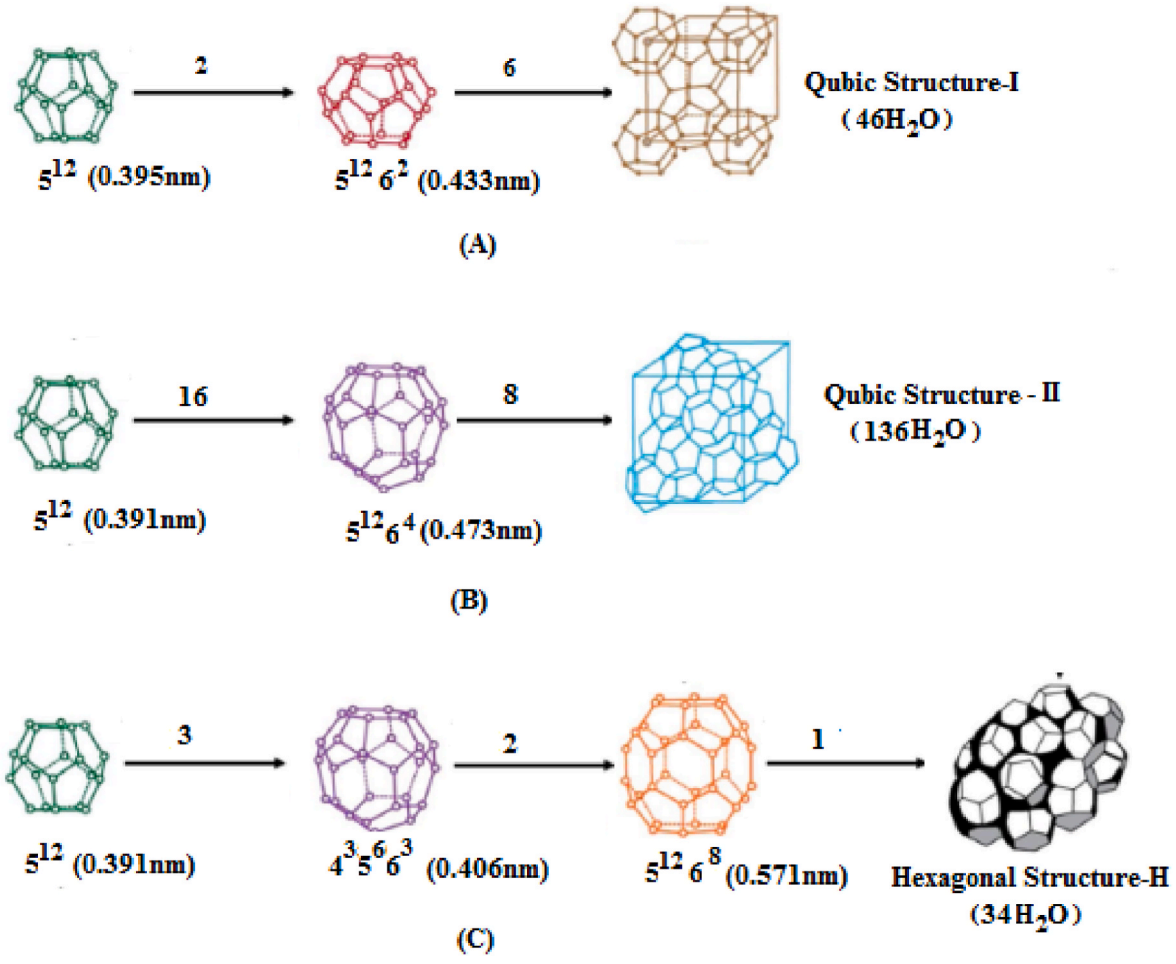


Fig. 4. Common hydrate structures, cavities, and average cavity radius (nm) (Xia et al., 2022).

ratio, hydrate formation kinetics, hydrate density, reactor design, and gas-water contact area. Gas diffusion in the hydrate is an essential factor affecting gas consumption and is highly dependent upon the properties of the media. The higher the hydrate density, the more complex is the gas diffusion (Dashti et al., 2015; Xu et al., 2020). The amount of gas molecules captured is proportional to total gas consumption and is the change in the gas quantity in the gaseous phase of the crystallizer. Gas consumption can be estimated (Zheng et al., 2016);

$$\triangle n_{g,t} = n_{g,o} - n_{g,f} = \left(\frac{PV}{ZRT}\right)_o - \left(\frac{PV}{ZRT}\right)_f \tag{3}$$

where $\triangle n_{g,t}$ is the number of moles of gas consumed to form hydrates in time 't'. P, V, T, R and Z are the gas phase pressure, volume, temperature, ideal gas constant and the compressibility factor, respectively. The subscript 'g', 'o' and 'f' represents the gas phase, initial and final value of gas moles, respectively, in the system after time 't'.

For a composition of the gas mixture, the number of moles of the individual gas (consisting of component 'i' with mole fraction 'x' in the mixture) formed hydrate (consumed) at any given time can be estimated as (Linga et al., 2007);

$$\triangle n^i_{g,t} = n^i_{g,o} - n^i_{g,f} = \left(\frac{x^i PV}{Z^i RT}\right)_o - \left(\frac{x^i PV}{Z^i RT}\right)_f \tag{4}$$

and the normalized gas uptake/consumption (gas mole/mole of water) can be expressed as (Zheng et al., 2016);

$$\triangle N_{g,tw} = \frac{\triangle n_{g,t}}{n_w} \tag{5}$$

where $\Delta N_{g,t}$ represents the gas moles consumed in water at time 't' and 'n_w' represents the presence of water moles in the solution. At the time of hydrate formation in a batch reactor, there is no gas and water flow across the boundary, and gas uptake depends on the initial input. In a semi-batch reactor, however, gas and water are supplied continuously. Normalized rate (NR) compares the rate of hydrate formation (gas moles/min.m^3) and can be estimated using eqn. (6) (Linga et al., 2010).

$$NR_t = \frac{\triangle n_{g,t}}{V_w \times t} \tag{6}$$

where 'V_w' is volume of water consumed for hydrate growth in time 't' from the induction point. The induction time is the time required for small hydrate nuclei to grow until the critical size is reached for favorable hydrate growth. Induction time depends on factors like driving force, experimental conditions, reactor configuration, gas solubility, promoter concentration, and solution state. Increasing driving force generally decreases the induction time (Babu et al., 2015).

2.2. Conversion of water into hydrate

Conversion of water to hydrate can be estimated as (Linga et al., 2010)

$$\textit{Conversion of water to hydrates}\ (mol\%) = \frac{\triangle n_{g,t} \times \textit{Hydration number}}{nH_2o} \times 100 \tag{7a}$$

where nH_2o are the total number of water molecule in the system. The hydration number is defined as the number of water molecules per guest molecule and can be estimated using equations proposed by Babu et al. (2015). In the presence of a thermodynamic promoter, eqn. (7a) can be expressed as;

$$\textit{Conversion of water to hydrates}\ (mol\%) = \frac{(\triangle n_{g,t} + \triangle n_{TH,t}) \times \textit{Hydration number}}{nH_2o} \times 100 \tag{7b}$$

where $\triangle n_{TH,t}$ is the number of moles of thermodynamic promoters consumed for hydrate

formation at the end of time 't', and can be estimated as;

$$\triangle n_{TH,t} = \textit{Hydration number} \times \frac{\textit{ratio of large to small cavity}}{\textit{Hydration numberfor full occupancy}} \tag{7c}$$

2.3. Factors influencing gas separation

The higher gas consumption does not always mean high CO_2 consumption because the gas consumed might contain gases other than CO_2. Since the gas molecules can selectively enter the hydrate cage, it is possible to separate mixed gases by adjusting the hydrate temperature and pressure (Liu et al., 2022). The efficiency of gas separation can be described by gas consumption, purity of gas, hydrate equilibrium pressure and temperature, separation factor (S.F.), and split fraction (S.Fr.) (Dashti et al., 2015). The separation efficiency of the hydrate based gas separation process is indicated by separation factor and split fraction or CO_2 recovery.

The separation factor (S.F.) quantifies the proportion of CO_2 gas captured in the hydrates rather than other gases in the mixture, and is required to be higher. The separation factor provides a measure of preferential incorporation of CO_2 in hydrate compared to other gases in the mixture. It may be infinity if no H_2/N_2 is accommodated in cages, and it is influenced by the promoter concentration, operational conditions, and driving force. In the case of hydrate formation from fuel gas mixture, an increase in driving force from 24 to 34 bar decreases the separation factor from 98.7 to 32.40 due to the enclathration of H_2 molecules into semiclathrate hydrate (Linga et al., 2007). For a flue gas mixture containing CO_2 and another gas, the separation factor can be estimated as (Song et al., 2013);

$$S.F = \frac{n^h_{CO_2} \times n^g_A}{n^g_{CO_2} \times n^h_A} \tag{8}$$

where $n^g_{CO_2}$ and n^g_A represent CO_2 and other gas molecules in the residual gas at the end of experiment, while $n^h_{CO_2}$ and n^h_A represent the presence of molecules of CO_2 and other gases in the hydrate phase. A decrease in separation factor has been reported in general with the addition of promoters (Babu et al., 2015).

The split fraction or CO_2 recovery indicates the amount of CO_2 separated from the feed gas during hydrate crystallization. Split fraction can vary from zero to one. A split fraction of 1 is highly desirable, which indicates that all the CO_2 from the feed gas has ended in the solid hydrate phase. The split fraction (S. Fr.) can be estimated as (Song et al., 2013);

$$S.Fr = \frac{n^h_{CO_2}}{n^f_{CO_2}} \tag{9}$$

where $n^h_{CO_2}$ and $n^f_{CO_2}$ are the respective CO_2 moles in hydrate phase and in feed gas. A maximum split factor of 0.42 and 0.66 was reported

without and with the promoter (0.29 mol% TBAB) using CO_2 (40 mol %)/H_2 (60mol %) mixed feed gas (Xu et al., 2012). It is preferable to have a reduced induction time, higher gas consumption, and higher separation factor (Sloan and Koh, 2007). A significant temperature difference from equilibrium at constant pressure decreases induction time and enhances the rate of nucleation and gas absorption with minimal influence on the rate of hydrate growth and separation efficiency (Wang et al., 2020). Li et al. (2012b) investigated the separation of flue gas (CO_2/N_2) in aqueous solutions with the additives, and found the separation coefficients as 15.54 and 14.06 using $TBANO_3$ and TBPB, respectively, at 33 bar and 275.15 K temperature.

3. Effects of hydrate promoters

Molecular gas transfer to the surface of the hydrate is the critical mechanism in hydrate formation. Three main factors, namely the physical barrier of the hydrate layer, exothermic heat generation, and heat distribution in quiescent systems, hinder hydrate formation (Thoutam et al., 2019). Various possible mechanisms for water transport within gas hydrate phases have been proposed, including interstitial water (Buch et al., 2009), lattice vacancy (Demurov et al., 2002), and Bjerrum defect (Kirschgen et al., 2003) assisted diffusion mechanisms. Moreover, Liang et al., (2016) observed unclear molecular mechanisms in gas diffusion. The researchers presented molecular mechanism of CO_2 cage to cage transport in gas hydrates using transition state theory, suggesting interstitial diffusion could be a dominant mechanism for water diffusion in CO_2 hydrates. The work further revealed that the linear molecular shape of CO_2 molecules helps to reduce the barrier to pass through water rings, suggesting the presence of empty cages for the CO_2 cage to cage transport in gas hydrates. For the CO_2 diffusion at 270 K, the activation of $\sim 44 \pm 6$ kJ/mol, and the diffusion coefficient in the range of $1.0 \times 10^{-16} \sim 2.0 \times 10^{-14}$ m^2/s was reported. The transport properties of water and guest molecules in dense hydrate structures are pivotal in controlling hydrate formation and stability; this is nevertheless hindered by challenges in characterizing the time-varying diffusion and permeation process (Liang et al., 2022). Furthermore, Liu et al. (2022) have reviewed various kinetic parameters of CO_2 hydrate formation such as CO_2 gas dissolution, CO_2 hydrate nucleation and growth. The author further summarized various models of formation mechanisms of CO_2 hydrate nucleation and growth, with or without additives, proposing a two-step hydrate nucleation mechanism to improve the accuracy of hydrate nucleation. The mechanism of hydrate formation is tabulated in Table 1. The hydrate nucleation depends on sub-cooling, pressure, gas-to-water ratio, promoters and their concentration, agitation, guest molecular size, etc. (Li et al., 2010; Majid et al., 2021; Sayani et al., 2022). The CO_2 molecules can efficiently occupy hydrate cages under moderate temperature and pressure conditions by employing thermodynamic promoters, boosting the diffusion coefficient and gas solubility (Zheng et al., 2016).

The promoters are classified as thermodynamic (THP) and kinetic promoters (KHP). The molecules of THP participate in the hydrate structure and are trapped inside the large cavity at lower pressures and/or higher temperatures. The phase-equilibrium boundaries of the THP gas hydrates are located in a more stable region than that required for pure gas hydrates, implying that the formation of binary hydrates may occur under the conditions of relatively low pressure and high temperature owing to the stabilizing effect (Shin et al., 2009). Moreover, to overcome the metastability of hydrate formation in a stable region, the driving force for formation needs to be sufficient (Majid et al., 2021). Fig. 5 depicts the structures of a pure CO_2 hydrate and the binary THF + CO_2 hydrate. When these promoters are present in a stoichiometric concentration (5.56 mol%), all the large cages of gas hydrates get filled with thermodynamic promoters by forming sII structure.

Table 1
Mechanisms of gas hydrate nucleation and growth (Liu et al., 2022).

Hydrate nucleation	Hydrate growth
• Classical nucleation theory • Labile cluster nucleation hypothesis • Interface nucleation • Local structuring nucleation • Blob nucleation hypothesis • Two-step hydrate nucleation mechanism	• Adsorption on gas hydrate surface • Destruction of gas hydrate cage structure • Perturbation of water surface • Capillary adsorption • Critical micelle structure • Water gas attraction at interface

Unlike THPs, the molecules of KHPs do not occupy the hydrate cages, and no shift occurs in the hydrate equilibrium curve. The THPs stabilize the structure, while the KHPs enhance the formation rate of hydrates by decreasing the induction time, without affecting structural transitions (Veluswamy et al., 2018). Adding KHPs changes the properties of the system such as interfacial tension and surface area. The kinetics of hydrate formation is closely linked to the process mechanism. When the gas is pressurized, it initially dissolves in water, resulting in supersaturation and subsequent nucleation, during which the gas is clathrated, and leads to a growth in the size of the hydrate crystal. KHPs improve gas consumption by increasing gas solubility in water and the permeation rate during the process. Fig. 6 illustrates the mechanism of hydrate formation using KHP. The comparative performance of these promoters is based on several factors, such as gas mixture, water quality, operating conditions, and the specific implementation to minimize the cost by decreasing induction time and enhancing hydrate formation rate (Sloan and Koh, 2007).

3.1. Thermodynamic hydrate promoters (THP)

Notably, the conditions required for forming clathrate hydrates and the structure of the hydrates can be changed by using appropriate thermodynamic promoters. The THPs alter the pressure equilibrium curve toward the lower side to provide a powerful driving force that promotes nucleation and further hydrate growth. Typical chemical THPs are tetrahydrofuran (THF), cyclopentane (CP), methyl-cyclopentane (MCP), tetra-n-butyl ammonium bromide (TBAB), tetrabutylphosphonium bromide (TBPB), tetra-n-butyl ammonium chloride (TBAC), tetra-n-butyl ammonium fluoride (TBAF), etc., and their mixtures (Wang et al., 2024). If the concentration of the THP is less than the stoichiometric concentration, the small guest molecules can occupy a part of such large cages in s(II) structure (Lee et al., 2023). Among the variety of thermodynamic additives, Tetrahydrofuran (THF) and tetra-n-alkyl ammonium salts such as TBAB are the most preferred thermodynamic promoters (Lin et al., 2008; Arjmandi et al., 2007; Babu et al., 2015). TBAB has the advantage of being stable at atmospheric pressure (Kamata et al., 2004), more eco-friendly, and easy to handle. Other features of using ammonium salts include an overall increase in the CO_2 capture rate of up to 90 % (Figueroa et al., 2004). The equilibrium conditions using CP and TBAB as additives produce better hydrate formation than mixing THF (Choi et al., 2021).

Linga et al. (2007) reported minimum hydrate forming equilibrium pressure of 16 and 3660 bar at 275.5 K using pure CO_2 and H_2 hydrate, respectively. Moreover, using a typical gas mixture of 39.2 mol% CO_2 and 60.8% H_2, the minimum hydrate equilibrium pressure is 68.4 bar at 275.6 K, as depicted in Fig. 7 (Babu et al., 2015). The decrease in CO_2/H_2 ratio increases the hydrate forming pressure due to the ability of H_2 and CO_2 to occupy more cages at increased driving potential. Fakharian et al. (2017) reported stable CO_2 hydrate formation at pressures below 35 bar but unstable at pressures above 55 bar. In addition, the authors reported dissociation pressures of 16 and 35 bar for CO_2 and CH_4 hydrates, respectively, at a temperature of 274.15 K with 3.5 wt% water salinity. The increase in concentration of promoters decreases the equilibrium pressure, except when using tetra-n-butyl ammonium bromide (TBAB) beyond the typical concentration. Kim et al. (2011) investigated the effect of TBAB concentration on the separation of CO_2 from a gas mixture of CO_2 (40%)/H_2 (60%) at the integrated gasification combined cycle (IGCC). The phase equilibrium conditions shifted to

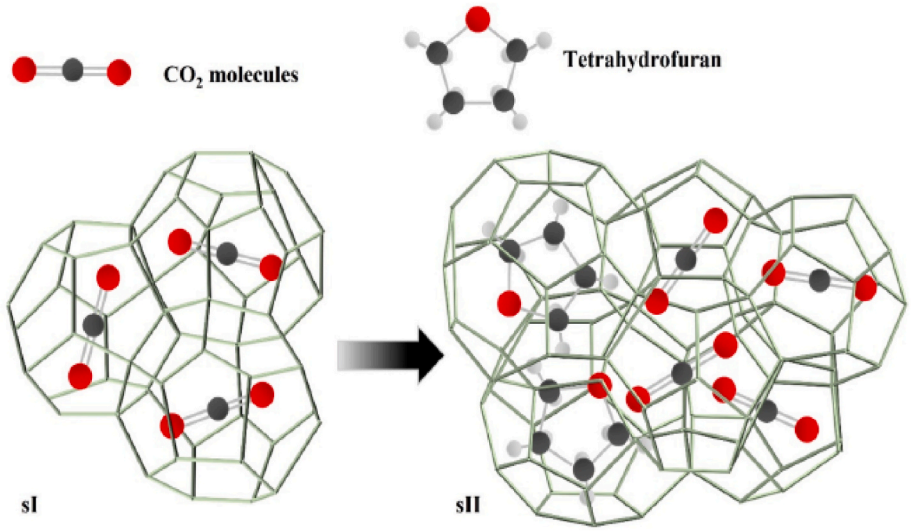


Fig. 5. Schematic of pure CO_2 hydrate (sI) and THF + CO_2 hydrate (sII) (Lee et al., 2023).

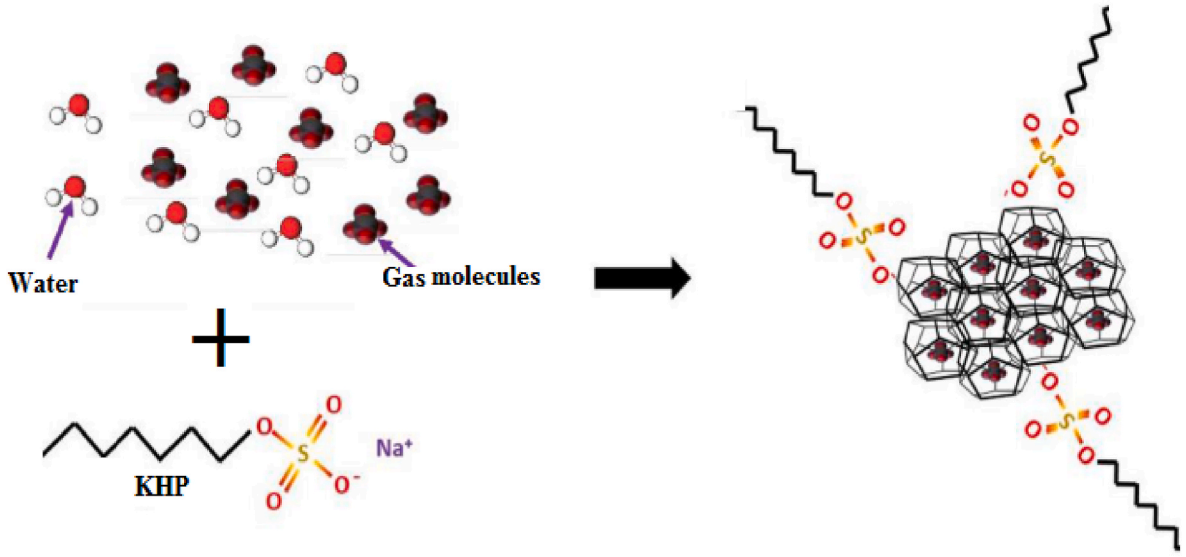


Fig. 6. Mechanism of gas hydrate in presence of KHP.

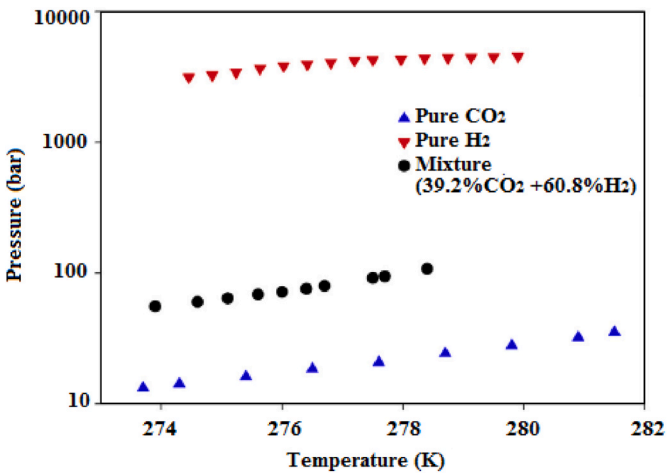


Fig. 7. Requirement of hydrate formation pressure for CO_2, H_2 and mixture (Babu et al., 2015).

milder conditions as the critical concentration of TBAB reached up to 3.0 mol%. However, the phase equilibrium temperature and pressure was increased at TBAB concentration higher than 3.0 mol%, which attributed to the presence of TBAB clathrated ions. The hydrate formation rate in 1.0 mol% TBAB solution showed the highest value (S.Fr. = 0.241, S.F. = 20.23) while lowest in 3.0 mol% TBAB solution (S.Fr. = 0.101, S.F. = 25.99) considering CO_2 molecules enclathration in the experimental range of 283–290 K and pressure range of 25–50 bar. The decreased hydrate formation rate above 1.0 mol% TBAB concentration is probably due to the decreased relative amounts of water molecules. Using TBAC, the maximum stabilization was observed at a stoichiometric concentration of 3.3 mol% of the TBAC (Kim et al., 2015a).

The effects of the most widely used promoters are discussed in this section, along with the shifting of equilibrium pressure on mixing as depicted in Fig. 8. The significant drop in equilibrium pressure has been found by mixing 1.0 mol% THF compared to the pure water system for the gas mixture consists of CO_2/H_2, and appears the optimal concentration for most promoters. The equilibrium pressure increases by mixing silica gel and reticent hydrate formation with tiny pores. The various studies using the CO_2/H_2 mixture show some deviation in equilibrium

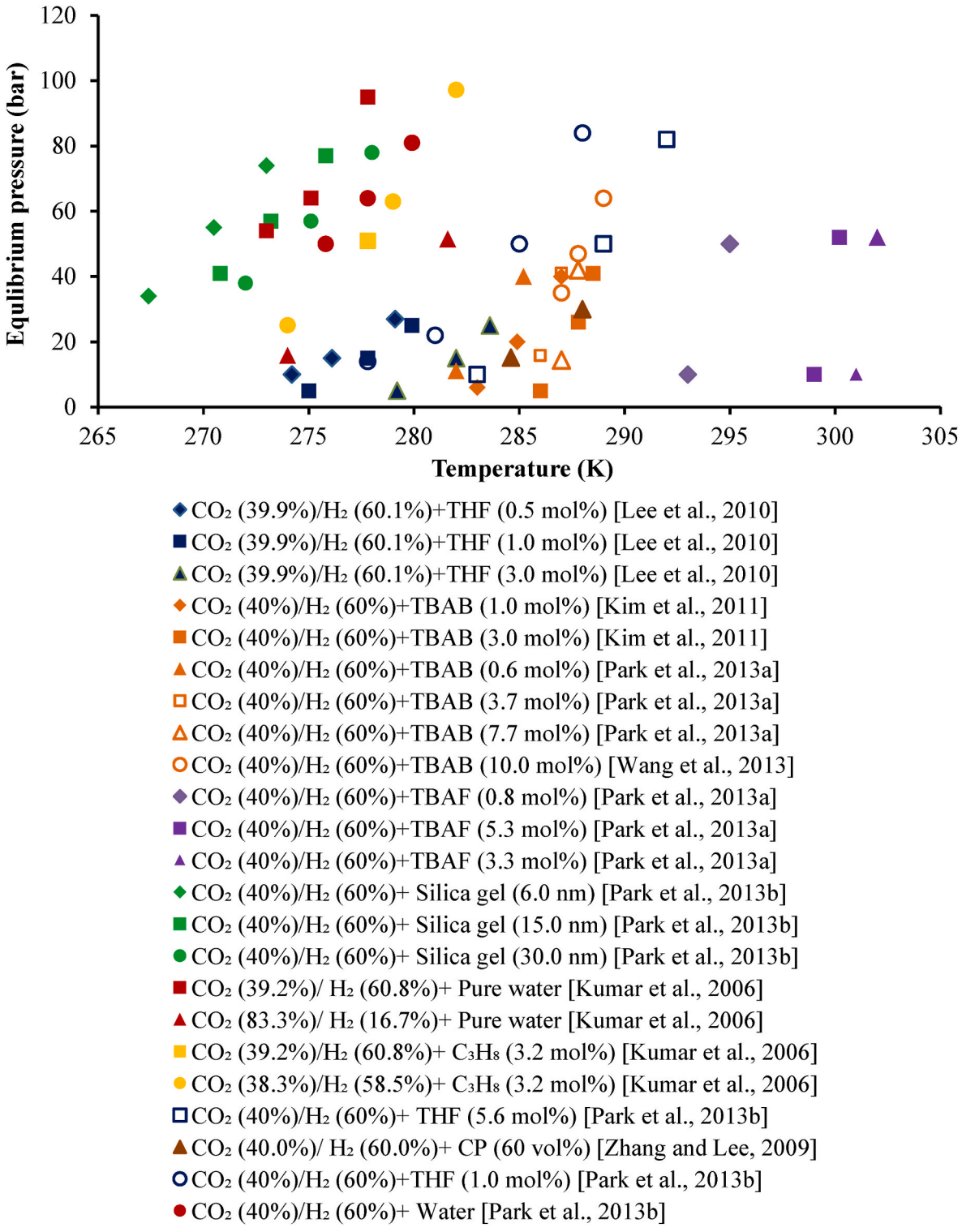


Fig. 8. Effect of promoters on equilibrium pressure in a mixture of CO_2/H_2.

conditions, attributing to variable measurement approaches and accuracy.

3.1.1. Propane

Propane is used as a THP and forms hydrate crystal (sII) at relatively low pressures compared to THF and cyclopentane (CP). Propane is a potential additive that can reduce the hydrate formation pressure at any given temperature, and it has the advantage of not being toxic. Kumar et al. (2006) studied the initial equilibrium hydrate formation conditions using CO_2/H_2 and $CO_2/H_2/C_3H_8$ mixtures. A significant decrease in hydrate-forming pressure was reported using a gas mixture rather than H_2 alone, forming sI structure. Adding 3.2 mol% C_3H_8 for a gas mixture (39.2 mol% CO_2/60.8 mol% H_2), equilibrium hydrate formation pressure and temperature was found lowered to 51 bar and 277.8K (i.e. ~50% reduction in pressure). Sloan and Koh (2007) observed a shift in the CO_2 hydrate equilibrium curve in the range of 1–10 % by mixing 6–20 mol% propane (C_3H_8) at pressures between 3 and 40 bar compared to a system without additives. The effect of silica

sand (porous material) with propane has shown an improved kinetics (Daraboina et al., 2013). In order to reduce the operating pressure, 2.5 mol% propane was mixed in a two-stage clathrate hydrate/membrane process of CO_2 separation. The equilibrium pressure of the first and second stage was reduced to 38 and 35 bar, respectively, at 273.7 K. The separation factor of 27.84 and 91.19 and split fractions of 0.47 and 0.32 were reported for the first and second stages, respectively (Linga et al., 2007). The propane mixing did not compromise the separation efficiency but decreased the hydrate formation rate. Babu et al. (2015) reviewed the works on promoters and found propane composition 2.5–3.2 mol% with a mixture of CO_2 (38.2 mol%)/H_2 (59.2 mol%) effective in reducing the equilibrium pressure and significantly forming sII structure, while adding 1.2 mol% C_3H_8 with CO_2 (80.0 mol%)/H_2 (18.8 mol%) mixture formed sI structure. In the sII structure, 43% of the large cages were occupied by propane, while CO_2 occupied 57% of the large and 34% of the small cavities. The equilibrium temperature shift by adding various promoters at a stoichiometric concentration in a fuel gas mixture has been reported in the order of C_3H_8< TBANO$_3$ <TBAB < THF < CP < TBAF. Table 2 lists the results of mixing propane as a promoter.

3.1.2. Cyclopentane (CP)

Cyclopentane is immiscible in water, has an equilibrium hydrate-forming temperature of −279.8 K at 1.0 bar, and forms an sII structure. Adding CP produces more stabilized sII hydrates than s(I) hydrates. Shifeng et al. (2010) used a 1000 cc stirred reactor at 281.1 K and pressure ranging from 27.9 to 39.5 bar to capture CO_2 from the CO_2/N_2 mixture. The remarkable hydrate formation rate with enrichment of 43.97% and 35.29% using cyclopentane and cyclopentane/water blend was estimated, with respective separation factors of 6.18 and 4.01. In cyclopentane at 273.15 K and 45 bar, gas uptake was increased to a maximum at 3 vol% of CP without further increment (Lim et al., 2013). However, immiscibility is a significant barrier to diffusing hydrate-forming gases through the layer of CP for hydrate growth. Table 3 illustrates the results of using cyclopentane as a promoter. Benefits of the CP include availability, low concentration, and excellent efficiency. The most significant drawbacks, however, are flammability, environmental risks, and equipment corrosion.

Table 2
Effect of adding propane on CO_2 hydrate formation.

Reference	System	Outcome
Kumar et al. (2006)	$CO_2/H_2/C_3H_8$	• lower pressures required to form hydrates by mixing C_3H_8.
Sloan and Koh (2007)	CO_2 (39.2 mol%)/H_2 (60.8 mol.%)/C_3H_8 (2.5 mol%)	• decrease in equilibrium pressure from 104 to 35 bar at 278.5K.
Kumar et al. (2009)	CO_2 (38.1 mol%)/H_2 (59.4 mol%), CO_2 (80.0 mol%)/H_2 (18.8 mol%) and C_3H_8 (1.2 and 2.5 mol%)	• 49% decrease in equilibrium pressure at 273.7 K
Linga et al. (2010)	CO_2 (38.1 mol%)/H_2 (59.2 mol%)/C_3H_8 (2.5 mol%)	• CO_2 gas consumption 0.010 mol/mol water in 8.7 min at 38 bar and 273.7 K
Babu et al. (2013)	CO_2 (38.1 mol%)/H_2 (59.4 mol%)/C_3H_8 (2.5 mol%) and silca sand/PU foam	• increases CO_2 uptake and decrease equilibrium pressure from 90 to 60 bar at 274.2 K. • 2.5 mol% of C_3H_8 is optimum conditions for gas maximum uptake. • propane as a co-guest has the ability to draw dispersed water from silica sand bed towards the gas phase for hydrate formation. • gas uptake of 0.0691 mol/mol H_2O in 2 h at optimum silica sand bed height 1.5 cm.
Majid et al. (2021)	CH_4/C_3H_8, CO_2/C_3H_8	• equilibrium curve shifted from 2 to 20% by varying C_3H_8 concentration (1–12%) with CH_4. • equilibrium curve shifted from 1 to 10% by varying C_3H_8 concentration (6–20%) with CO_2 at pressure of 3 and 40 bar

Table 3
Effect of adding CP on CO_2 consumption CO_2/H_2 (40/60 mol%), CO_2/N_2 (16.6/84.4 mol%).

Reference	System	Outcome
Linga et al. (2007)	$CO_2/H_2/H_2O$	• equilibrium pressure 51 bar. • decreased CO_2 mole fraction from 0.392 to 0.239 in 5 h at 273.7 K and 85 bar. • nucleation rate is three times higher than CO_2/N_2 system.
Zhang and Lee (2009)	CO_2/CP (5.0 vol%)	• induction time less than 0.2 h.
Zhang et al. (2009)	CO_2/H_2/CP (1.5 vol%)	• equilibrium pressure reduced from 53 to 13 bar.
Shifeng et al. (2010)	CO_2/N_2/ CP	• S.F. 6.18 with CP, higher than CP/water blend (4.01) at 281.8 K. • higher CO_2 uptake (43.97 mol%) than CP/water blend (35.29 mol%).
Ho et al. (2013)	CO_2/H_2/CP (7.5, 15 and 22 ml)	• S.F of 17.82 at optimal CP (15 ml). • No improvement by adding SDS with CP.
Ricaurte et al. (2014)	CO_2/CH_4 (75/25 mol %)/CP/1,3-Dioxolane (4 wt%)	• S.F. = 4.6 and S. Fr. = 52.5% higher than mixing 1,3-Dioxolane (4.2 and 52.2%).
Zhou et al. (2020)	CP/CO_2	• increased CO_2 uptake up to 0.25 mass fraction. • increased uptake from 0.020 to 0.029 mol with increase pressure (15–25 bar).

3.1.3. Tetrahydrofuran (THF)

Tetrahydrofuran (THF), a thermodynamic promoter, is a self-forming agent of sII hydrates by occupying large hydrate cages. In sI hydrates, CO_2 molecules fill the large ($5^{12}6^2$) and small (5^{12}) cages. In contrast, in binary systems such as the binary THF/CO_2 hydrate, the THF molecules fill the large cages, and CO_2 molecules fill the small cages at a molar composition of THF/water ration 1:17 (Majid et al., 2021). Mixing THF decreases the equilibrium pressure and increases dissociation enthalpy, affecting the gas uptake. At higher THF concentrations, the induction time also decreases, probably due to the shift of equilibrium pressure to the lower side. Adding a small amount of THF causes a significant depression of equilibrium pressure and increases further with an increase in THF concentration, reaching a stoichiometric (5.6 mol%). With the increase in THF concentration above the stoichiometric ratio, the equilibrium pressure may not be decreased anymore (Hashimoto et al., 2006). By making the system more thermodynamically stable, THF makes the clathrate of CO_2 molecules possible at relatively low temperatures and pressures (Lee et al., 2023). Mixing 3.0 mol% THF in CO_2/H_2 fuel gas at 280.1 K reduces the hydrate formation equilibrium pressure significantly from 91 to 10 bar (Hashimoto et al., 2006). Zhang et al. (2014) investigated the effect of THF concentration on hydrate phase equilibrium above 20 bar, while the difference in hydrate phase equilibrium is marginal below 20 bar. Compared with hydrate phase equilibrium conditions of pure water at 33 bar, the hydrate phase equilibrium temperature increases by ~4.50 K in the presence of 1.0 mol % THF and increases further with the increase in THF concentration up to 3.0 mol%. When the hydrate phase equilibrium pressure is lower than 20 bar, the hydrate phase equilibrium curves approach each other with the pressure decrease, showing the negligible effect of increased THF concentration. As a general finding, depending on the concentration ratio of THF, the kinetics of hydrate formation enhances and raises the equilibrium temperature by 2–20% at constant pressure.

Lee et al. (2010) investigated the effect of adding THF and reported

the higher CO_2 consumption (0.0092–0.0178 mol/mol of water) at the pressure range of 15–25 bar and found economical. The increase in THF concentration (1.0–3.0 mol%) at 5 and 15 bar pressures and 273.17 K increases the CO_2 captured proportionally. However, at 25 bar, the captured ratio decreases. Furthermore, Lee et al. (2010) discovered that adding THF to water dramatically reduced the equilibrium formation conditions and 1.0 mol% of THF has been the ideal composition to accelerate hydrate production. The decrease in the hydrate formation rate above 1.0 mol% concentration is owing to the pre-occupied cavities by THF molecules and less availability for CO_2/H_2 molecules, which is advantageous even at higher driving potential. The decrease in formation rate below 1.0 mol% THF can be due to the limited L-cage and comparatively small amount of S-cage hydrate formation. Kim et al. (2017) reported optimal conditions for CO_2 hydrate formation and dissociation by adding 1.5 mol% THF at 15 bar to use hydrate slurry as the working fluid for district cooling application. Linga et al. (2010) reported enhancement of CO_2 hydrate and separation efficiency using a stirred tank reactor with an increase in CO_2 recovery using promoters, as shown in Fig. 9. The consumption of 0.14–0.32 mol at a pressure of 34–73 bar was investigated using TBAB and 0.063–0.191 mol using TBAF at a pressure of 22–68 bar. The minimum operating pressure of 25 bar in the presence of THF may result in a better CO_2 recovery from the flue gas mixture from the economic viewpoint than TBAB and TBAF. Table 4 shows the results of adding THF to some systems with variable concentrations.

3.1.4. Quaternary ammonium salts

The requirement of low separation temperature (0.6 °C), environmental hazards, and slow hydrate formation rate are the main disadvantages of THF. Compared with THF, propane, cyclopentane, and quaternary ammonium salts form a semi clathrates in which the promoter molecules not only occupy the cages but also constitute a part of the host framework and have shown significant interest (Zheng et al., 2018). Moreover, limiting operating pressure and temperature range, compatibility issues, environmental effects, solubility, toxicity, and inhibition effects are the shortcomings of quaternary ammonium salts. Due to empty cages that can capture small-sized gas molecules at low-pressure and high-temperature conditions, quaternary ammonium salts have been considered a new potential in separation and have significantly stabilized hydrate dissociation conditions (Park et al., 2013a). The tetra-butylammonium bromide (TBAB) and tetra-butylammonium fluoride (TBAF) are the most common promoters to capture CO_2 by forming semi-clathrate at lower pressure. TBAB offers the lowest hydrate equilibrium pressures at moderate temperatures (typically below 286 K), which increases significantly above 286 K, resulting in higher compression costs than using THF (Linga et al., 2010). However, due to dodecahedral cages that are partially filled with

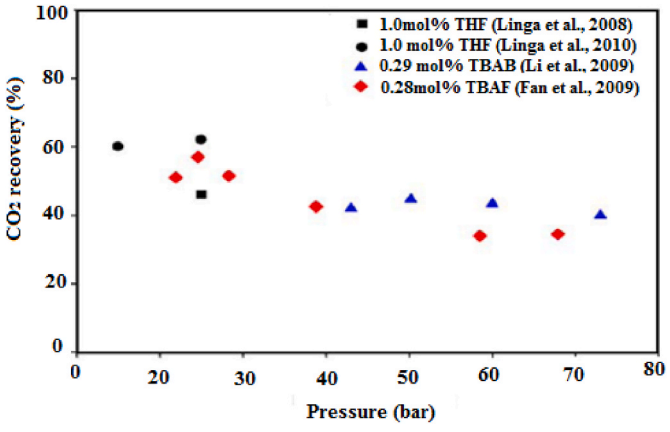


Fig. 9. Comparative CO_2 recovery using THF, TBAB, and TBAF (Linga et al., 2010).

Table 4
Summary of adding THF on the hydrate formation.

Reference	System	Outcome
Linga et al. (2007)	CO_2/N_2/THF (1.0 mol%)	• decreased pressure from 77 to 3.45 bar. • CO_2 mol fraction decreases from 0.169 to 0.104 in 5 h at 273.7 K and 15 bar.
Lee et al. (2010)	CO_2/H_2/THF (0.5–3.0 mol%)	• 1.0 mol% THF is found optimum. • pressure decreases from 110 to 18.7 bar at 278.7K.
Linga et al. (2010)	CO_2/N_2/THF (1.0–1.5 mol%)	• The higher CO_2 uptake and recovery by adding THF than adding TBAB and TBAF. • CO_2 recovery factor of 0.63 with water to hydrate conversion 14.9% using 1.0 mol% THF at 25 bar. • with concentration to 1.5 mol%, decrease induction time and gas consumption from 0.0098 to 0.0163 mol/mol. H_2O was obtained at 273.7 K.
Lee et al. (2012)	THF/CH_4, THF/CO_2, CH_4/CO_2, THF/CO_2/CH_4	• with 1.0–1.5 mol% THF a significant shift in equilibrium conditions has been found.
Zhong et al. (2015)	CO_2/CH_4/THF/(THF + SDS)	• S.F. = 11.2 and S.Fr. = 48.7% at 28 bar using pure water. • THF (500 ppm) addition decrease S.F. (4.0) and increase S.Fr. (50.7%) but decrease with increase pressure to 43 bar (SF = 2, RF = 24.4%). • 1.0 mol% SDS/THF (500 ppm) decreases S. F. (5.4) and increases S. Fr. (51.4%) at 28 bar.
Hassan et al. (2021)	CO_2/THF (3 mol %)	• 1.16 mmol CO_2/g H_2O uptake in 20 h at 288K and 36 bar.
Phan et al. (2021)	CO_2/THF	• unstable CO_2 hydrates at 25.5 bar and temperature ≥279.1 K. • stable hydrates at temperature < ~288.5 K and mixing ~1.42 mol% THF. • ~1.42 mol% THF yielded benefits of both THP and KHP at 274.1 K and 25.5 bar.
Wang et al. (2021)	CO_2/THF (1–3 mol%)	• about equal CO_2 consumption (0.02310 and 0.02335 mol CO_2/mol H_2O) at 1.0 and 3.0 mol% THF, but ~ 14.3% reduction compared to pure water system.

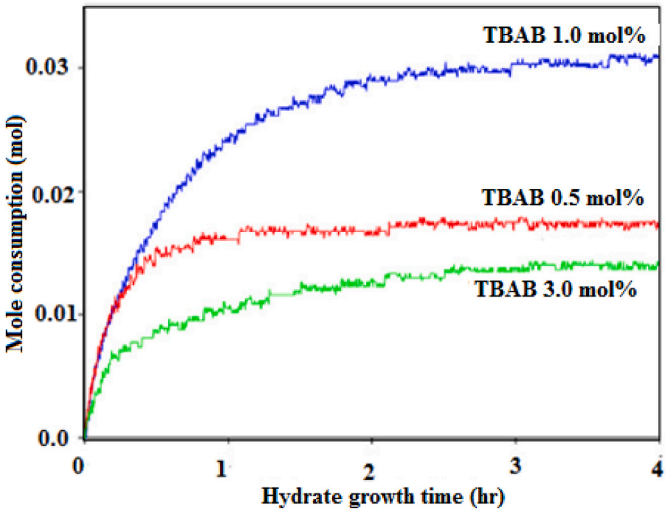


Fig. 10. Gas hydrate kinetics with TBAB blend (Kim et al., 2011).

water molecules, TBAF showed a relatively lower gas uptake than TBAB. The research indicates that varying concentrations of TBAB between 0.1 and 4.0 mol% at operating pressures between 14 and 45 bar alter the equilibrium hydrate forming temperature between 1 and 14 K. The gas uptake was found to be increased by 20% on mixing TBAB (0.3 mol%) compared to a system without a promoter (Chen et al., 2017). At a stoichiometric concentration of TBAB (0.29 mol%), the CO_2

consumption increases and induction time decreases (Park et al., 2013a).

The effect of mixing TBAB on CO_2 mole consumption is shown in Fig. 10. The rate of hydrate formation accelerates initially with an increase in TBAB concentration (0.5–1.0 mol%). The decreased formation rate at low concentrations is because of limited initial nucleation at 0.5 mol%. The reduced rate of hydrate formation above a concentration of 1.0 mol% is likely due to the relative decrease in the presence of host molecules. The gas consumption was found to be ~1.7 times with an increase in TBAB (0.61 mol%) compared to adding 0.29 mol% TBAB under similar conditions. In addition, the separation and split fractions obtained at the respective TBAB concentrations were 15.7 and 0.41 and 28.0 and 0.48. Gholinezhad et al. (2011) found an increase in CO_2 concentration up to 96% using a two-stage hydrate process. The increase in separation factor was reported to be 50% lower with an increase in TBAB concentration from 5 to 10 wt% from CO_2/N_2 gas mixture (Chen et al., 2017), while Gholinezhad et al. (2011) reported higher gas uptake with increased concentration on the same amount, owing to deferent operating pressure. Cyclopentene improves CO_2 solubility and increases CO_2 contact surface when hybridized with TBAB, resulting in a six fold rise in CO_2 uptake at 0.29 mol% TBAB than the TBAB system alone.

The employment of TBAF as a promoter forms a semi clathrate hydrate consuming lower gas at higher thermal stability compared to TBAB promoter owing to the availability of less partially water-filled dodecahedral cages to capture CO_2 and H_2 molecules (Park et al., 2013a). At similar operating pressure (0–80 bar) and concentration of TBAB and TBAF, the respective shift in equilibrium temperature occurs in the range of 282–292 K and 293–304 K, and is found minimum at 0.6 mol% and 0.8 mol% compositions, and maximum at 3.7 mol% and 3.3 mol% of respective promoters. At concentrations >3.7 mol%, the excess TBAB is not captured in cages, and free ions act as inhibitors. In addition, by adding TBAF promoter, hydrate forming conditions had been shifted toward high temperature and low-pressure sides, reaching a maximum at 3.3 mol%, compared to water alone (Park et al., 2013a). The comparative gas consumption by adding TBAB and TBAF with variable concentration is shown in Fig. 11, indicating that 3.7 mol% and 3.3 mol % as a stoichiometric concentration of TBAB and TBAF, respectively.

Zheng et al. (2016) studied the effect of pressure (20–60 bar) and temperature (28–298 K) using a CO_2/H_2 system added with 3.38 mol% TBAF. They found a decrease in gas consumption with a reduction in temperature, possibly due to interfacial mass transfer resistance, an increase in induction time over 0.33 min, and a reduction in driving force. The CO_2 solubility using 3.38 mol% TBAF at temperature 298K and pressures of 60, 40, and 20 bar were found to be 0.344, 0.260, and 0.139 mol/L, respectively, lower than those obtained using pure water (0.371, 0.291, and 0.167 mol/L) as depicted in Fig. 12. The comparative results

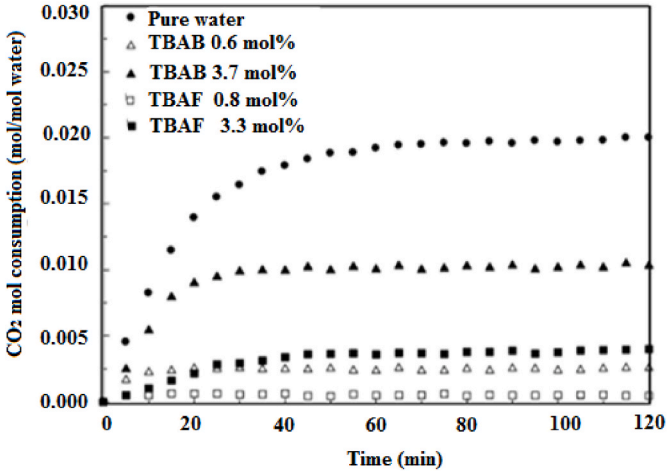


Fig. 11. Gas consumption using TBAB/TBAF at 80.0 bar (Park et al., 2013a).

of gas consumption using promoters at their stoichiometric concentrations, namely TBAF (3.38 mol%), TBAB (3.70 mol%), $TBANO_3$ (3.70 mol %), and THF (5.56 mol%), were found in order of TBAF > TBAB > $TBANO_3$ at 60 bar (except TBAB at 80 bar). However, due to the formation of distinct structures, gas uptake was lower when THF was used. The equilibrium temperature using fuel gas and TBAF additive is found to be 302.1 K, higher than the pure TBAF hydrate formation temperature (300.8 K). Therefore, the equilibrium temperature below the formation of pure TABF may lead to low CO_2 enclathration. In contrast, due to the requirement of low equilibrium temperature (277.5 K) for THF hydrate formation at 60 bar avoids the formation of THF hydrate.

Most recently, Babu et al. (2024) studied the promotional effect on the kinetics of TBAF based CO_2 capture from post-combustion flue gas mixture (5.2% CO_2)/74.8% N_2 at 80 bar and temperatures of 293.2 K and 298.2 K. The impact of three amino acids (L-methionine, L-tryptophan, and L-leucine) in a concentration ranging from 0.1 to 2.0% on TBAF mixed semi-clathrates hydrate formation were examined. The mixing of 2 wt% L-methionine in 10% TBAF solution provided the best promotional effect. The average gas consumption was increased from 2.43 to 9.96 mmol/mol of water with increased TBAF concentration from 10 to 33.7 wt% at 80 bar and 298 K, with enhanced CO_2 selectivity (~70%) at the concentration range 10–20 wt% and highest CO_2 separation at 20 wt% TBAF. The outcome of adding TBAB/TBAF is shown in Table 5.

3.2. *Kinetic hydrate promoters (KHP)*

Kinetic promoters enhance hydrate formation's kinetics and have received increased attention due to the improved hydrate formation rate upon addition in small concentration without affecting the gas storage capacity and thermal equilibrium. During hydrate formation, the kinetic promoters enhance the gas permeation rate in the hydrate layer, thus increasing the solubility of the gas in water. There are several types of KHPs, including surfactants, biosurfactants, amino acids, nanoparticles, polymers, and cellulose. However, most of them are not eco-friendly (Cheng et al., 2021; Chaturvedi et al., 2021; Majid et al., 2021).

3.2.1. *Surfactants*

Surfactants contain hydrophilic and hydrophobic groups to decide the behavior towards hydrate formation and may be ionic (anion and cation) or non-ionic. In the presence of surfactants, gas and liquid's interfacial tension and contact angle decrease. Chaturvedi et al. (2021) reported that some anionic surfactants have better effects than non-ionic, while cationic surfactants produced better performance at low concentrations, based on critical micelle concentration (CMC) and kraft point of surfactant. To avoid the inhibitory effect of a surfactant, the optimal concentration must be used, along with anti-foaming additives, to prevent foam formation, which occurs at high concentrations, resulting in decreased interfacial contact. Sodium dodecyl sulfate (SDS) is the generally used anionic surfactant for enhancing the kinetics of gas hydrates at low concentrations (Kumar et al., 2013). The SDS forms micelles and increases the gas-liquid interface's surface area to increase gas mass transfer in the hydrate. The increase in gas uptake has been reported tenfold with an increased concentration of SDS from 0.05 to 1.0 wt%.

Zhang et al. (2014) investigated the impact of adding THF/SDS on the formation of hydrates using CO_2/N_2 at temperatures from 277.45 to 295.95 K and pressures from 1.4 to 83.1 bar. They found a substantial decline in equilibrium pressure in conjunction with an enhanced THF concentration (from 1.0 to 5.0 mol%) and found the optimal concentration being 3.0 mol% THF when mixed with 1000 mg/L SDS. Thoutam et al. (2019) experimented with a stirred reactor at 35 bar and 274.1 K and found a decrease in CO_2 concentration by >70% using saline water (salinity 0.9 mol%). Moreover, a further 30% decrease was investigated in the presence of CH_4 (4.9 mol%). Furthermore, mixing SDS (1000 ppm) as a kinetic promoter enhanced CO_2 loading due to the recovery of

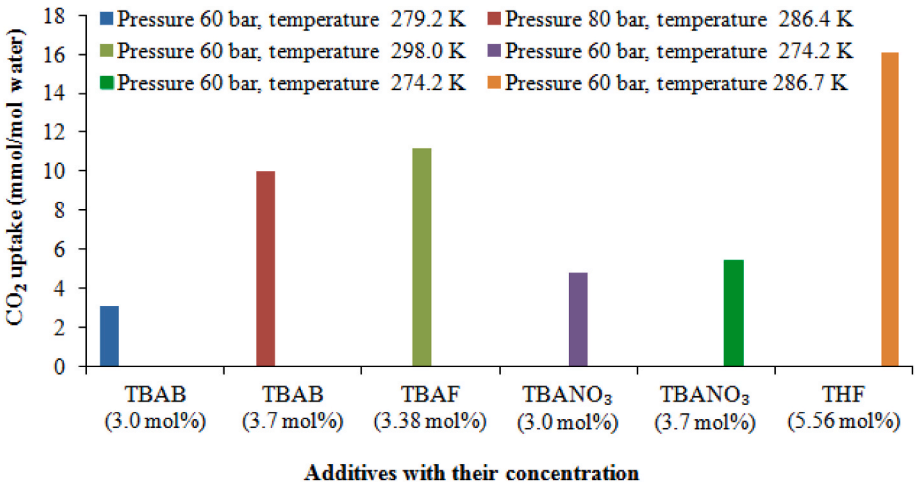


Fig. 12. Comparative CO_2 uptake using THP with operating conditions, derived from Zheng et al. (2016).

Table 5
Summary of adding TBAB/TBAF on CO_2 hydration.

Reference	System	Outcome
Li et al. (2010)	CO_2/H_2/TBAB (0.4–1.0 mol %)	• increase TBAB concentration above 0.29 mol % increase gas consumption. • at pressure >25 bar, H_2 likely to form hydrate phase compared to CO_2.
Kim et al. (2011)	CO_2/H_2/TBAB (0.5–6.0 mol %)	• at TBAB concentration >3.0 mol %, equilibrium temperature and pressure increases, while induction time decreases.
Li et al. (2012a)	CO_2/H_2/TBAB (0.29 mol%), CP (5 vol%)	• CO_2 concentration increases from 40 to 84 mol% with gas uptake 0.022 mol/mol H_2O. • significant increase in gas uptake by adding CP and TBAB (>2 times) compared to THF.
Li et al. (2012b)	CO_2/N_2/TBAB (0.29–1.0 mol %), TBPB (0.29–1.0 mol%), $TBANO_3$ (0.65–1.0 mol%)	• $TBANO_3$ and TBPB are better promoters. • Higher S. Fr. (67%) and S.R. (15.54) were obtained using $TBANO_3$.
Yang et al. (2015)	CO_2/1.0 wt% THF and TBAB (0.5–5 wt%)	• Mixture of 5 wt% THF/5 wt% TBAB significantly decrease the induction time and is better combination for higher gas uptake. • TBAB between 0.5 and 3 mol% has a limited influence on equilibrium.
Nguyen et al. (2016)	CO_2/TBA (3 wt%)	• 0.125 wt% TBAB decreases the gas uptake while at 1.0 wt% gas uptake increases by 14%.
Zheng et al. (2018)	CO_2/H_2/TBAF (0.8,1.5, 2, 2.5, 2.96 and 3.38 mol%)	• The rate of hydrate formation was maximum at 0.8 mol% TBAF system.
Li et al. (2019)	CO_2 (33.00 mol %)/CH_4 + TBAB (5.0 wt %)	• CO_2 separation factor reached to 24.9 with increased separation efficiency.
Hassan et al. (2021)	CO_2/TBAB (0.29 mol%)	• 2.1 mmol CO_2/g H_2O uptake in 20 h at 275 K and 36 bar.

34% loss that occurred by using saline H_2O and CH_4 systems, which improved further to 64% using a stirring system, as can be depicted from Fig. 13. The CO_2/distilled water system converts 57.7% of the water into hydrate, followed by the CO_2/CH_4/saline water/SDS system (35.96%) and the CO_2/CH_4/saline water system (6.67%), respectively. Wang et al. (2016) added 3.0 mmol/L SDS in a mixture of CO_2/CH_4 (27.76/72.74%) at 60 bar and obtained separation and recovery factors of 1.2 and 52.2%, respectively. They also investigated the increase in SF (5.73), which occurred with a decrease in pressure to 40 bar.

3.2.2. Amino acids

As a result of their hydrophilic nature, amino acids exhibit powerful electrostatic interactions with water and form strong hydrogen bonds, rendering them inhibitors (Sa et al., 2011). Furthermore, these acids have been reported as KHP under stirring and non-stirring conditions, making them a potential future choice to enhance CO_2 uptake. Due to foam formation during the process, the conventional THF and SDS promoters have a low hydrate formation rate and are less environmentally favorable. However, certain amino acids have demonstrated an effective kinetic promotion/inhibition effect that depends on the presence of guest molecules, making their application in CO_2 capture dependent on their side chain length and hydropathy index (Sa et al., 2015). Cai et al. (2017) investigated the mixing of L-methionine preferable to other amino acid promoters at an optimal concentration of 0.2 wt%. Prasad and Kiran (2018a, 2018b) investigated the impact of amino acids and found that L-valine, L-cysteine, and L-methionine increased CO_2 consumption by approximately 20%. Bavoh et al. (2019) reviewed the applicability of amino acid-based additives and discussed their inhibition/promotion effect and other performance influencing factors. Depending on the alkyl group in the chain, the researchers also suggested amino acids as potential promoters for future hydrate-based applications, decreasing induction time and enhancing hydrate growth by 52% (Chaturvedi et al., 2021).

Srivastava et al. (2022) used cysteine, valine, leucine, methionine, and lecithin as a surfactant and found that leucine and methionine acids were more efficient and adding promoters such as leucine, methionine, and lecithin increase the CO_2 consumption. The amino acid combinations enhanced CO_2 gas uptake by 21.53% (leucine/methionine), 16.71% (leucine/lecithin), 19.71% (methionine/lecithin), and 25.67% (leucine/methionine/lecithin) at 263.7 K with 0.5 wt% of respective combination.

3.2.3. Nanofluids

Metal nanoparticles have found extensive applications in promoting gas hydrate formation by improving heat and mass transfer. Using nanoparticles such as Cu, Ag, Zn, SiO_2, Al_2O_3, ZnO, CuO, MgO, and Fe_3O_4, the interfacial mass and heat transmission can be increased to speed up the hydrate formation and boost the solubility. Li et al. (2006) investigated the nanoparticles ranging from 5 to 55 nm in diameter, and the solutions were 0.1–1.0 wt% nano-copper suspension. Arjang et al. (2013) synthesized silver nanoparticles from 50 to 75 nm and found an 85% and 73.9% reduction in induction time and a 33.7% and 7.4%

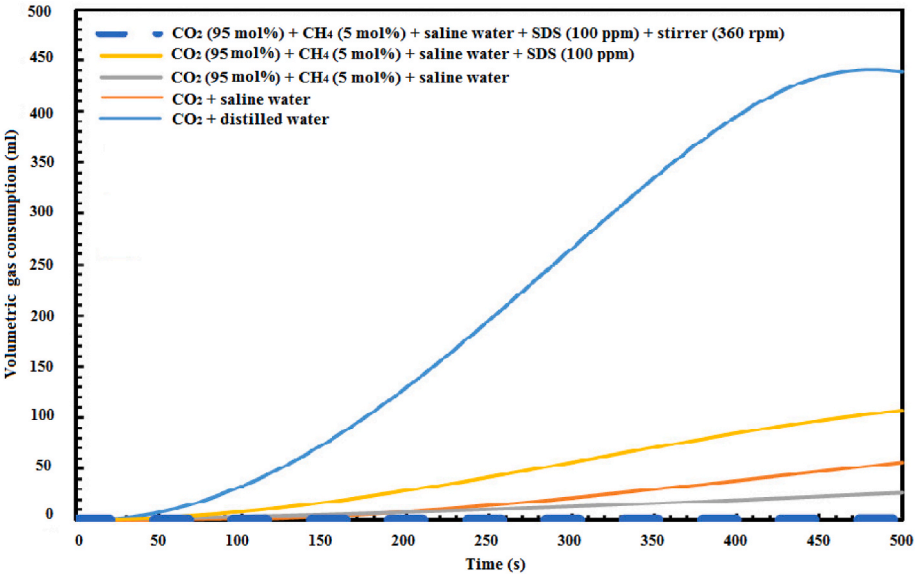


Fig. 13. Effect of impurities on CO_2 consumption with time at 35 bar and 274.15 K (Thoutam et al., 2019).

increase in trapped methane in hydrate at pressures of 47–57 bar, respectively. As a hydrate additive, nanofluids reduce interfacial tension and enhance nucleation rate due to increased surface area, thermal conductivity, and heat and mass transfer to accommodate more molecules and increase gas consumption (Wang et al., 2018). Numerous researchers have reported that the combined effect of nanoparticles and surfactants boosts performance because surfactants increase the stability of colloidal solutions. The application of multiwalled carbon nanotubes improved the gas uptake threefold compared to using a purified water system owing to the catalytic behavior of nanotubes (Park et al., 2010). Nesterov et al. (2015) reported that adding carbonate groups at the surface along with metal oxide nanoparticles decreases the induction time owing to the adsorption of CO_2 at its surface.

Further, Mohammadi et al. (2016) investigated the impact of adding zinc oxide (ZnO) (0.1 wt %) and SDS (0.1 wt%) and found a reduction in hydrate growth time with improved CO_2 storage compared to purified water. By combining SDS/ZnO and ZnO/pure water solutions, the respective increase in gas storage capacity of up to 5% and 12% was obtained. On adding graphite nanomaterial, Yu et al. (2016) found a substantial shift in the hydrate formation pressure without the remarkable effect of the concentration (0.4–1.2 wt%) (Fig. 14). The higher equilibrium pressure of CO_2 hydrate formation has been obtained by mixing graphite nanoparticles owing to the increasing system entropy/porous effect of big size nanoparticles. Moreover, graphite nanoparticles can promote hydrate formation due to high heat transfer coefficient and specific surface area.

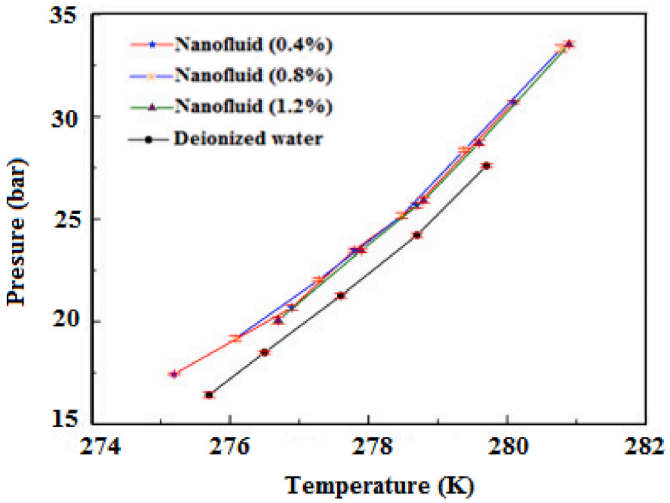


Fig. 14. Graphite nanoparticles influence on the CO_2 equilibrium curve (Yu et al., 2016). Adapted with permission from Elsevier.

Sarlak et al. (2018) reported the hydrate kinetics using a combination of CO_2/graphene oxide (GO) and CO_2/GO/SDS water solutions and found an increase in CO_2 gas storage capacity and molar consumption of 4.2% and 3.8%, respectively, with an increase in nanoparticle concentration from 0.05 to 0.1% at 34 bar. The increase in CO_2 storage capacity by 38% and the consumption by 26% was reported using 400 mg/l.SDS. Mohammadi et al. (2014) found increased CO_2 consumption by 93.9% by adding SDS (500 ppm) and silver nanoparticles (0.000045 M) compared to the purified water system at 30 bar and 273.65 K. Aliabadi et al. (2015) reported a 92.7% reduction in induction time by combining SDS and CuO compared to the pure water system and suggested an optimal concentration <1.0 wt% CuO to accelerate hydrate formation. Wang et al. (2018) fixed SDS on the surface of polymer nanospheres and grafted nano-Ag particles size 2–5 nm onto it, indicating that the concentration of 0.1 mmol/L of SDS with Ag (5%–50 mol%) reduced hydrate formation time from 422.7 ± 63.5 min to 156.7 ± 24.4–273 ± 31.5 min. Said et al. (2016) reported the effectiveness of gas uptake using various nanomaterials in order of SiO_2> Cu > Al_2O_3.

3.3. Synergists

Due to the restricted benefits of a sole promoter, combinations of promoters have been employed to boost efficiency, hydrate kinetics, and gas consumption. Synergists play a crucial role in enhancing the formation of gas hydrates. Typically, kinetic parameters such as gas uptake capacity, gas uptake rate, hydrate formation, nucleation rate, and growth rate of hydrate crystals can be utilized to measure the synergistic effect. The physicochemical mechanisms involved with the use of synergists are decreased surface tension of water (promotes nucleation), enhanced surface area, template effect, lower energy barrier, alteration of the hydrophobic or hydrophilic nature of the gas water interface, and influenced crystal morphology (Pandey and von Solms, 2023). According to Xia et al. (2017), the gas storage capacity and hydrate formation rate were increased by 84.0% and 149.0%, respectively, when tetramethylene sulfone and TBAB were combined, while equilibrium pressure decreased by 90% compared to the system consisting of pure water.

Nesterov et al. (2017) reported that combining thermodynamic and kinetic promoters or other compounds with SDS accelerated crystallization due to the formation of sII structure, which increased the gas storage capacity and rapid nucleation. Gas consumption was reported to be doubled when SDS and THF were combined compared to mixing THF alone (Dashti et al., 2015). Similarly, adding surfactants to CP increases the surface area of contact between the gas and liquid, thereby enhancing the diffusivity of gas through the interface and promoting the growth of hydrates.

Yang et al. (2015) investigated the effects of additive concentrations on the equilibrium conditions by mixing THF/TBAB and using the isochoric method. A significant decrease in equilibrium pressure was observed by combining 5%THF compared to mixing THF/TBAB or THF/SDS promoters. Mixing 5.0 mol% of THF and TBAB each decreases the induction time significantly compared to other concentrations. Among quaternary ammonium salts, 0.29 mol% TBAB has the maximum normalized gas consumption of 0.0187 mol/mol H_2O at 278.15 K and 50 bar; however, this value varies with experimental conditions and is found to be lowest for TBAF (0.62 and 3.32 mol%). The dispersion of TBAB in silica sand bed increases the normalized gas uptake, and 0.3 mol% TBAB was found to be optimal in a reactor with a fixed bed at 60 bar, yielding 0.0197 mol/mol H_2O. At stoichiometric concentration, promoters occupied the large cages, decreasing the normalized CO_2 uptake to 0.118 mol/mol H_2O, which exceeds the possibility of CO_2 occupancy in hydrate cages. Fig. 15 depicts the results of Mohammadi et al.'s (2018) investigation into the effect of adding SDS (400 ppm) and TBAF (5.0 wt%) at 38 bar and 278 K temperature for CO_2 capture. The addition of SDS at low TBAF composition impacts CO_2 consumption, which decreases with an increase in TBAF concentration.

Adding SDS increases the interfacial between the gas-water, whereas mixing of THF/SDS functions as a CO_2 hydrate catalyst (Hassan et al., 2020). The effect of using SDS (0.01 mol%)/THF (5.60 mol%) shows a higher equilibrium pressure of ~36.5 bar for forming hydrate with silica gel. Hassan et al. (2021) utilized 0.10 mol% ethylene glycol monoethyl ether (EGME) as a replacement for THF under comparable operating conditions and found a consumption of 5.45 mmol/g H_2O using a fixed bed reactor (FBR), as well as an increase in CO_2 consumption and operating capacity. However, maximum CO_2 consumption of ~6.0 mmol/gH_2O was obtained by mixing 0.01 mol% SDS and 5.6 mol% THF. They investigated further the effect of FBR height on CO_2 consumption and found an increment of 1.54 mmol of CO_2/g H_2O with decreased FBR height from 0.03 to 0.03 m and using TBAB (0.1 mol%)/SDS 0.01%) system at 283 K and 70 bar. The decrease in TBAB concentration to 0.1 mol% led to a 38% increase in CO_2 consumption compared to baseline levels. The inhabitation effect was also investigated using 0.29 mol% TBAB/3 mol% THF blend and found 31% lower consumption than the TBAB alone, owing to the combined effect of semiclathrate (TBAB) and hydrate (THF). The CO_2 uptake using various systems at operating conditions of 273.4–274.4 K and pressure (25–37.5 bar) using pure CO_2 system is shown in Fig. 16, and found dependent on concentrations of promoters and operating conditions and found maximum on mixing SDS with ZnO nanoparticles. The mixture of CP/TBAB aids in the smallest induction time (0.15 min), however, results are found to be conflicting.

The summary of adding synergist on the performance is tabulated in Table 6.

An increase in CO_2 uptake from 0.0126 to 0.0159 and from 0.022 to 0.026 mol/mol H_2O was reported using silica sand (329 mm) and silica gel solution, respectively, compared to the water system at ~80 bar (Kang et al., 2013). A decrease in induction time by 80.8% and an increase in CO_2 absorption by 12.8% was found by mixing 4.0 wt% graphite (Li et al., 2017) and found optimal using 0.15 wt% Fe_3O_4 at 274.15 K and 40 bar to enhance gas uptake by about 4 times with decreased induction time by 96.6% (Firoozabadi et al., 2018). Khanmohammadian et al. (2023) investigated the impact of pressures (30 and 40 bar), gas-liquid ratio, and nanoparticle concentration for CO_2 separation using a mixture of CO_2 (26 mol%)/CH_4 (74 mol%) with silica,

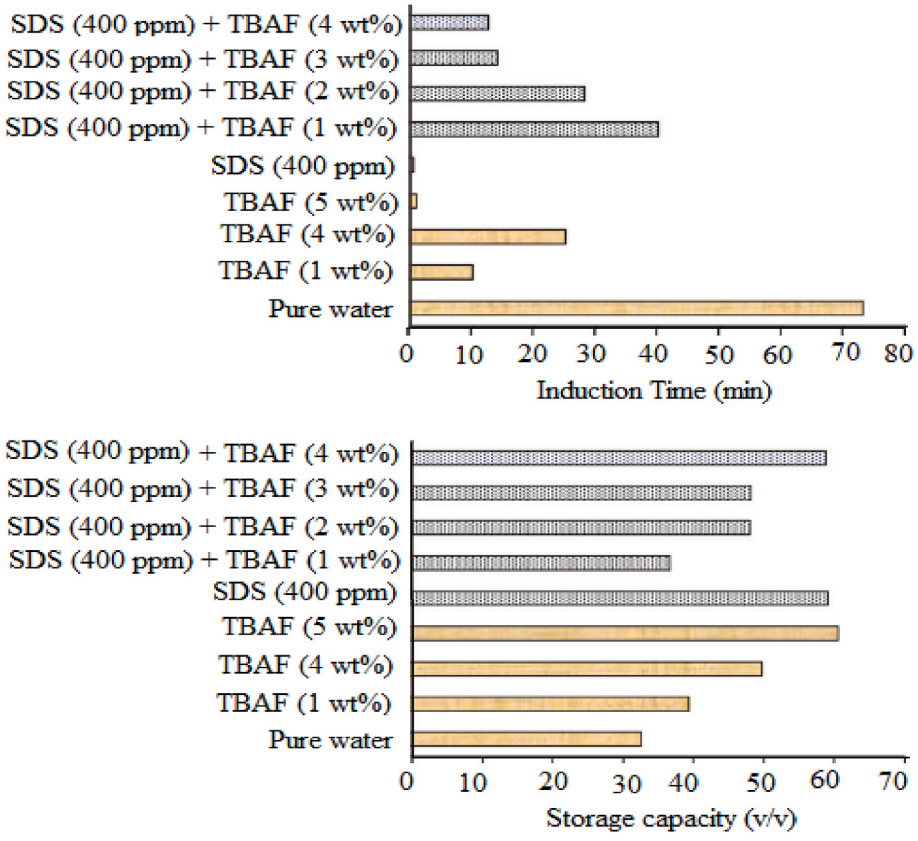


Fig. 15. Effect of SDS and TBAF on CO_2 induction time and storage capacity at 38 bar, 278 K with pure CO_2 (Mohammadi et al., 2018).

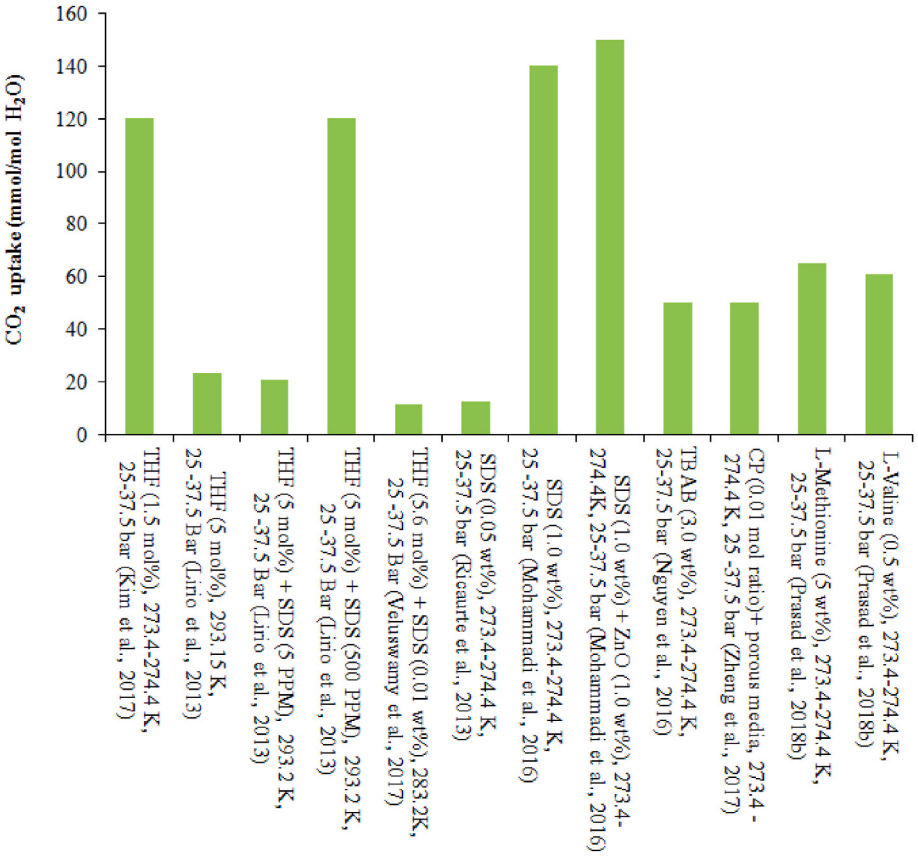


Fig. 16. CO_2 uptake using pure CO_2 with different systems.

3-aminopropyl teriethoxysilane (APTES), and potassium hydroxide (KOH) as additives at 277.15 K and found an increase gas consumption, SF, and S.Fr. by 36, 29, and 38%, respectively, on mixing KOH, compared to the water system alone at the optimal nanoparticle concentration (0.1 wt%). With an increase in pressure from 30 to 40 bar at 277.15 K at 0.71 gas/liquid ratio, the gas consumption was enhanced while the S.F. and S. Fr. decreased from 8.81 to 8.44 and 59.4%–57.2%, respectively. In comparison to nanoparticles, the addition of KOH acts as a synergist to enhance the performance of the system, and the results are depicted in Fig. 17. The addition of microporous zeolitic imidazolate frameworks (ZIF-8) decreased the induction time by 52% and enhanced hydration factor by 14.0% was reported (Chaturvedi et al., 2021; Denning et al., 2021), suggesting that these frameworks may have future applications.

A combination of porous materials, surfactants, and nanomaterials can lead to synergistic effects and can be utilized to enhance the hydrate formation. High gas uptake of water confined in nanopores of Metallic organic frameworks (MOFs) under hydrate formation conditions was identified (Pandey and von Solms, 2023). Numerous studies have been carried out to identify the promotional effect of MOFs, hydrophobic solid surfaces, and metallic surfaces on gas hydrate formation. Pandey and von Solms (2023) have identified the properties of MOFs materials due to their large surface area and a high pore volume and found improved efficiency of gas hydrate formation. In some studies, MOFs have been used to inhibit CO_2 hydrate in gas-dominated pipelines while being used to improve CO_2 storage in the hydrate phase (Wang et al., 2024). Strong CO_2 dissolution was reported in interstitial meso/-macropores of MIL-53 of pore diameters ~0.6 nm; however, above 10 nm pore size inhibited the thermodynamic behavior (Kim et al., 2015b). Confined hydrates in the meso/macropores are much higher than that of bulk hydrate. Effects of pore geometries, hydrophilicity, low heat conductivity, metal nodes, stability, and high production cost of MOFs are unknown parameters, and pore geometry has shown minimal influence on the formation kinetics (Denning et al., 2022). Open hydrophobic solids enhance gas hydrate formation by promoting the tetrahedral structuring of water (Li et al., 2020). Wang et al. (2015) reported that hydrophobic sand coated with silica sand with octadecyl trichlorosilane (OTS) decreased the induction time for gas hydrate formation from 576 to 71 s in a hydrophobized sand bed. A significant acceleration of nucleation is due to interfacial gas enrichment. However, hydrophobic surfaces are toxic and less durable. The effect of some porous material on gas uptake for pure CO_2 and a gas mixture of CO_2/N_2 is depicted in Fig. 18, and significant improvement is found on mixing porous material, with maximum using UTSA-120a > pumice > HKUST-1.

3.4. Hybrid integration

Linga et al. (2008) investigated the post-combustion CO_2 capture with three-stage hydrate-membrane approaches using THF (1.0 mol%) promoter at each reactor. The experiment was performed at the initial composition of CO_2 (16.9%)/N_2 (83.1%) at 273.7 K and 25 bar with a driving force of 22 bar. The CO_2 composition was enhanced to 38.0% and 70.0% during the second and third stages, respectively, in the rector under similar operating conditions. The induction time (0.7 min, 0.7 min, and 0.3 min), split fraction (0.46, 0.47, and 0.37) and separation factors (7.6, 0 6.8, and 7.7) under the first, second, and third stages of hydrate formation were estimated, respectively. The 53.0% decrease in the compression cost was reported by coupling the system with lean mixture membrane separation with the recovery of 50% CO_2 in the first and second stages and 37.0% in the third stage. The combination of hydrate-based technology with other traditional techniques, as shown in Fig. 19, nearly captures complete CO_2 with a decrease in capture cost (Yu et al., 2021).

The hybrid hydrate/membrane-based technology combines the

Table 6
Summary of adding synergist on the hydrate formation.

Reference	System	Outcome
Ricaurte et al. (2013)	CO_2/CH_4/SDS (3000 ppm)/ THF (4 wt%)	• 4.1 S.F. and 44.1% R.F. was obtained at 40 bar.
Park et al. (2013b)	CO_2/H_2/THF (1.0–5.6 mol %)/silica gel (6–100 nm)	• most stable gas hydrates at 5.6 mol% THF.
Lirio et al. (2013)	CO_2/THF(5 mol%), SDS (5–500 ppm)	• 37% storage capacity with 500 ppm SDS and increases to 121% by adding THF (5 mol %).
Choi et al. (2014)	CO_2/THF (7.8–20 wt%), SDS (0.05–0.8 wt%)/Al_2O_3 (0.1–0.6 wt, silicon dioxide (SiO_2) (0.1 wt%) etc.	• 3.74 times increased rate of hydrate formation by combining SDS (0.6 wt %) and Al_2O_3 (0.2 wt%). • adding THF (10 wt%), SDS (0.6 wt%) and Al_2O_3 (0.2 wt %) produced optimal performance.
Nashed et al. (2018, 2019)	CO_2/SDS (0.03 wt %)/Multiwall carbon nanotubes	• highest initial hydrate formation rate and CO_2 uptake at 0.01 and 0.05 wt% nanotubes and SDS (0.03 wt %).
Mohammadi et al. (2016)	CO_2/ZnO (0.001 mass fraction)/SDS (1.0 wt%)	• mixture of SDS/H_2O enhances 6% gas consumption than pure H_2O. • ZnO enhances 8% gas consumption than pure H_2O. • ZnO/SDS enhances 12.6% gas consumption. • The gas uptake increases by 5% and 12% with SDS compared with mixing of ZnO and pure water, respectively.
Veluswamy et al. (2017)	CO_2/THF (5.6 mol%)/SDS (0.01–0.5 wt%)	• enhanced gas consumption ~50% by mixing SDS (0.05 wt%), but decreases by ~50% by mixing THF. • 10 times increase in gas uptake estimated with increase of SDS composition from 0.05 to 1.0 wt%.
Mohammadi et al. (2018)	CO_2/TBAF (5.0 wt%)/SDS (400 ppm)	• hydrate storage capacity increases by 86.1% mixing TBAF and by 81.6% with SDS, respectively, compared to pure water without considerable increase in gas uptake.
Xia et al. (2019)	CO_2/H_2/TBAB (11.0 wt%), carbon nanotubes (0.5 wt%), Al_2O_3	• better effects are obtained using TBAB–CNT than TBAB–Al_2O_3 on hydrate formation.

advantages of both approaches and can be suited to meet the large emission loads of plants by selecting a suitable membrane. The successful integration of membrane with gas hydrate process depends on permeability and selectivity of material. Kumar et al. (2009) studied the two-stage hydrate capture process operating at 38 and 35 bar at 273.7 K temperature by mixing 2.5 mol% C_3H_8. The lean CO_2 mixture passed through the membrane in second stage, reducing the energy penalty to 20%. The authors suggested adding suitable promoters to decrease the operating pressure in the first stage without diminishing the separation efficiency. Li et al. (2011a,b) presented the one-stage hydrate/membrane process under pre-combustion CO_2 capture using 0.29 mol % TBAB at 30.0 bar and suggested using an alternative approach as membrane separation in conjunction with gas hydrate separation. The high purity ~99 mol % CO_2 and H_2 gas from residue CO_2/H_2 lean mixture was obtained, decreasing compression energy needed for storage. The purity of CO_2 obtained using hybrid mode was higher than that obtained without membrane (~97 mol%).

The effect of heat integration using hydrate-based membrane separation techniques has been studied by Feyzi and Mohebbi (2020) as shown in Fig. 20. They found a considerably reduced energy consumption and CO_2 capture cost by mixing TBAB and dodecyl trimethyl ammonium chloride (DTAC) using the hybrid technology, as depicted in Fig. 21. Combining hydrate based gas and membrane separation techniques is a novel process that recovers more than 90% of CO_2 at higher than 95% concentration. When the promoters are added, the requirement for the membrane size decreases. It has also been concluded that by adding hydrate promoters, the exergy destruction in the cooling and membrane unit decreases, and the investment cost decreases. Petukhov et al. (2022) studied the continuous CO_2 hydrate separation from flue gas consisting of CO_2/N_2 (83.0/17.0 mol%) mixture and using a stage hybrid hydrate-membrane approach adding TBAB promoter and found a reduction of CO_2 capture cost to 35.0$/ton and 92.0% recovery. An increase in CO_2 recovery and decreased environmental emissions using membrane-assisted hydrate technology were reported.

4. Comparative economical assessment

The use of CO_2 capture using conventional methods has reduced the power generation efficiency from 38 to 26%, necessitating a less energy-intensive separation process (Song et al., 2013). Spencer (1999) reported reduced energy costs by 50–75% using gas hydrate CO_2 capture compared to absorption methods, demonstrating its superiority over the absorption method. Using hydrate-based technology, the minimal energy requirement was reported to be 3.0 MJ/kgCO_2, which is substantially lower than the absorption method of CO_2 capture (4.0–6.0 MJ/kgCO_2). Cost of hydrate-based technology has varied from 20 to 40 $/t$CO_2$ avoided with an associated 15% energy penalty (Aminnaji et al., 2024a), depending on the quality of feed gas and operation conditions, which have been significantly lower than conventional absorption-based capture (40–100 $/t$CO_2$ avoided) with 30% energy penalty (Spencer, 1999). The energy penalty of hydrate based technology has been reported to be decreased by 50–75 % compared to conventional absorption methods, decreasing the capturing cost of CO_2 (Duc et al., 2007). Low CO_2 separation energy (0.57kWh/kg CO_2) and low energy penalty (6.0–8.0%) have been reported, declaring hydrate based CO_2 separation as a potential technology under moderate temperature (273.17–277.17 K) and pressure (7.0–20.0 bar) (Seo et al., 2005; Fan et al., 2011). Nguyen et al., (2022) has developed equations for the economic assessment of hydrate based technology. They estimate energy penalty of 15 %, making hydrate based technology potentially competitive with conventional absorption methods. Linga et al. (2007) estimated an $18/t$CO_2$ capturing cost using a two-stage hybrid hydrate-membrane process, showing significant capital cost connected with high-pressure compression. A comparison of CO_2 capture cost using various methods is given in Table 7. However, the variation in capturing cost is expected owing to many system components, chemical additives, pressure vessels, agitation, auxiliaries etc. Aminnaji et al., (2024b) identified a total cost of $29.47/t$CO_2$ captured with the hybrid hydrate-membrane separation method using TBAB as a promoter. Despite increasing energy consumption using TBAB promoter, the process was most cost-effective because of high separation efficiency, significantly reducing the overall cost. Zheng et al. (2016) found that the mixing of thermodynamic promoters such as THF, tetrahydropyran, propane (C_3H_8), cyclopentane (CP), cyclohexane, neohexane, and salts (ammonium/phosphonium) decreased the energy cost and gas consumption depends on the promoter's properties and reported decrease in carbon capture cost from 75 to 53% using THF promoter.

5. Challenges and prospects

Using gas hydrates for carbon capture is a promising method for reducing greenhouse gas emissions. Hydrate-based technology can obtain a CO_2 consumption of 9.67 mmol/g.H_2O from a fuel gas mixture. However, due to the occupancy of large cages by the promoter

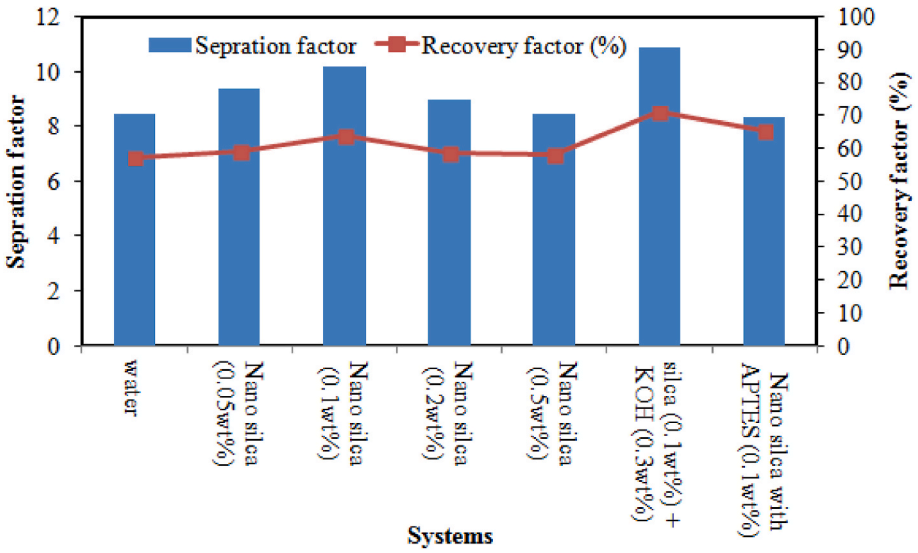


Fig. 17. Effect of adding KOH and APTES with silica at various concentrations.

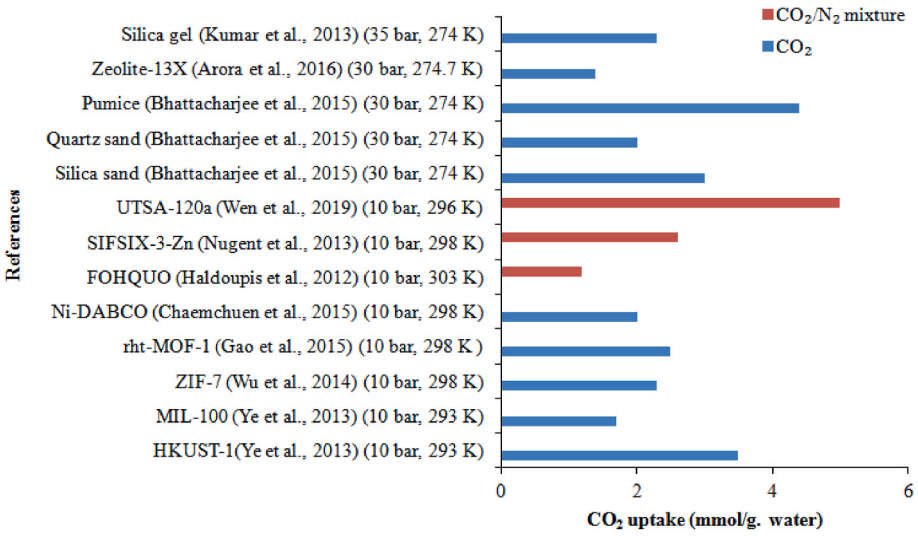


Fig. 18. CO_2 uptake using various porous materials.

molecules, the maximum capacity of 6.56 mmol/g water can only be attainable. Moreover, it also faces barriers such as the requirement of high pressure and slow hydrate formation kinetics. Adopting gas hydrate technology necessitates the resolution of crucial challenges. The challenges associated with equilibrium find no agreed promoter selection criteria and need more insights into the mechanism of promoters. Unlike other conventional techniques of CO_2 capture, hydrate-based CO_2 capture is under the research stage, and operational knowledge at an industrial scale does not exist. The designs of the reactor required to be consistent optimization for future research. The most significant industrial and commercial issues have been flow blockage due to hydrate crystals and the need for improved designs of absorber reactors. The contradictory results regarding gas uptake and separation factor, CO_2 solubility kinetics, and low mass transfer rates during hydrate formation are not adequately addressed. The other challenge is avoiding environmental hazards; long-term stability and potential environmental impacts must be meticulously evaluated. Challenges regarding slow nucleation and hydrate growth rate at milder operating conditions require overcoming to make gas hydrate-based technology viable. The critical challenge is using an optimized concentration of promoters and deserved attention for promising promoters with less environmental hazard and huge gas storage capacity. The sluggishness in the kinetics can potentially hinder the time effectiveness of the process. The formation of gas hydrates and their stability are complex and challenging to control, requiring specific temperature and pressure conditions to develop a reliable process. Technology's readiness and commercial viability are still uncertain. As such, the economic assessments of this process mainly rely on theoretical or laboratory-scale studies. To make gas hydrate CO_2 capture economically viable, significant research and development initiatives were needed to improve its cost-effectiveness.

The gas hydrate method has been recognized as a novel and promising technology with the potential to attain carbon neutrality if implemented effectively. Aligned research, industrial collaboration, and policy are needed to connect the full potential of hydrate-based technologies for a greener future. Hydrate-based carbon capture represents a chance for continued research to receive acceptance from the public, industry, and government for its future acceptability by developing crucial hydrate promoters. Developing novel materials and processes that can facilitate hydrate formation under milder conditions is a crucial area of future research. Integrated heat utilization technologies should be studied to make hydrate-based technologies more energy-efficient and cost-effective. To accelerate the kinetics, the improved designs of

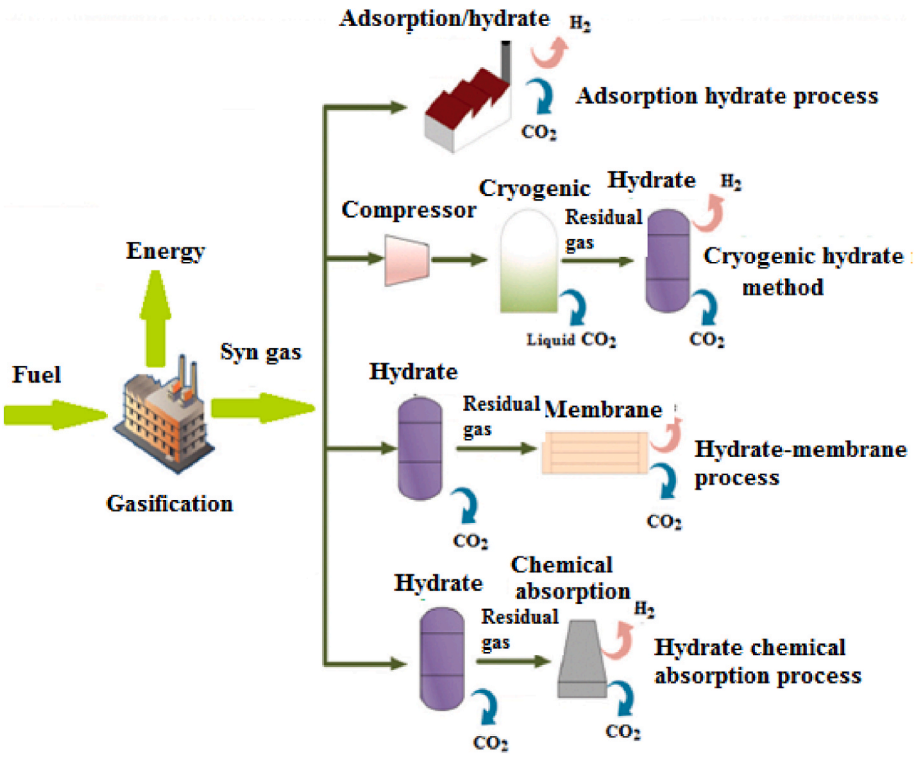


Fig. 19. Hydrate based hybrid CO_2 capture technologies.

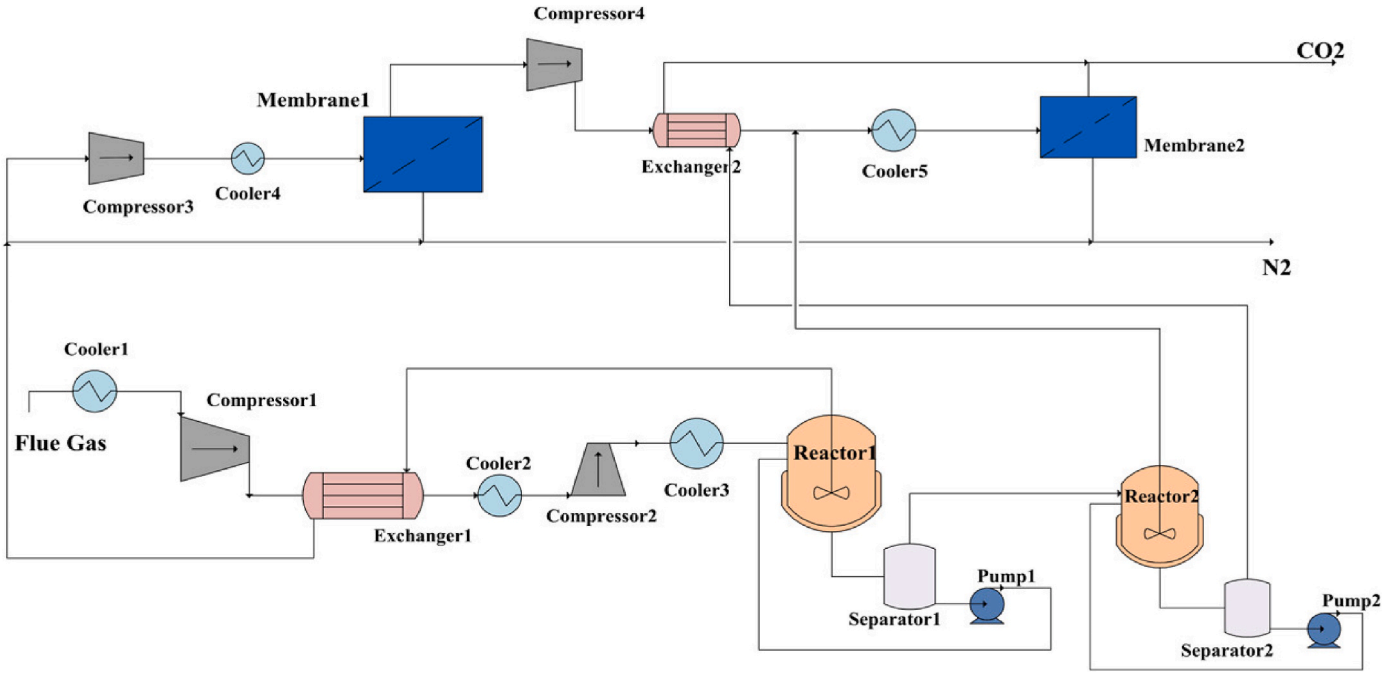


Fig. 20. Two stage hydrate-membrane based technology with heat extraction (Feyzi and Mohebbi, 2020).

reactors are required critical consideration in future research. Gas hydrate CO_2 capture technology will be made more feasible and effective due to advancements in material science, process engineering, and monitoring technologies, which will aid in overcoming current obstacles and enhancing its viability. While hydrate-based technology is effective on a laboratory scale, it needs to be estimated when it would scale up.

6. Conclusions and recommendations

6.1. Conclusions

The advancement of CO_2 capture technologies is encouraging, and hydrates-based CO_2 capture has an auspicious future in capturing CO_2 from highly concentrated gases, which attracted researchers due to its vast potential for achieving sustainable development goals. The various

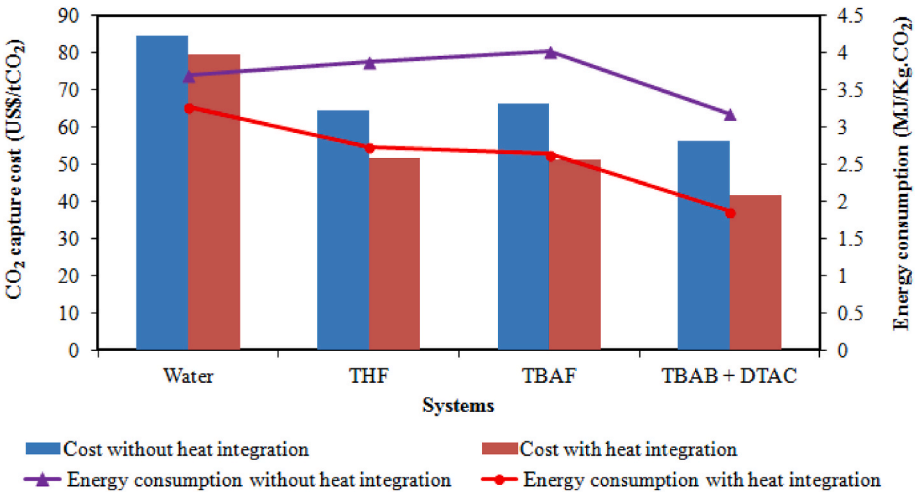


Fig. 21. Energy consumption and CO_2 capture cost using hydrate-based membrane separation technology, derived from Feyzi and Mohebbi (2020).

Table 7
CO_2 capturing cost, energy penalty, and energy consumption using various CCS technologies (Spencer, 1999; Duc et al., 2007; Linga et al., 2007; Nguyen et al., 2022).

Parameters	Hydrate based	Membrane	Adsorption	Absorption	Cryogenic
Energy consumption (MJ/kg CO_2)	1.3–4.7	0.5–11	1.5–2.0	4–6	6–10
Energy penalty (%)	15	16–24	21	30	–
CO_2 capture cost ($/t$CO_2$)	20–40	23–75	–	40–100	52.8

promoters, phase equilibrium conditions, synergists, porous materials, and nanomaterials have positively impacted hydrate formation to accelerate the kinetics. Additionally, the possibility of enhancing CO_2 absorption capacity at optimal promoter concentrations, experimental conditions, and heat integration with hybrid hydrate-membrane technology has been found emerging to reduce the operating cost and energy penalty. The following conclusions have been deduced, and recommendations are made to conduct further research.

a. In the presence of promoters, the degree of occupancy by CO_2 in the hydrate depends on operational conditions, reactor design, and promoter concentration. Cyclopentane (CP) has performed the best in terms of gas uptake, followed by tetrahydrofuran (THF) and Tetra-n-butyl salts. The CP/TBAB hybrid system decreases the induction time to 0.15 min. However, contradictory results are noticed.
b. The porous materials with high thermal conductivity significantly improve the performance of the gas hydrate process, which can be accelerated further by adding kinetic promoters. Depending on their structure, the zeolitic imidazolate framework system has been found to decrease nucleation time by 52–92%.
c. The effect of nanomaterial is insignificant; however, when added with surfactants, the hydrate formation rate has been enhanced remarkably with decreased induction time. Depending on the concentration, a combination of SDS (500 ppm) and CuO (0.01 wt%) nanoparticles has decreased the induction time by 20% compared to a system without nanoparticles.
d. The use of hybrid integration is advantageous compared to using single hydrate based technology. Integrating a membrane-hydrate based process reduces energy consumption, compressor work, and cost compared to a standalone hydrate based system with more CO_2 capture and is highly attractive.
e. Properties of activated carbon and MOF materials can be tailored (e.g., surface area, pore size/volume, and chemical functionalization) to fit the needs of gas hydrate pressure and temperature conditions to reach high storage densities and conversion rates.

6.2. *Recommendations*

- Research on fundamental mechanisms of hydrate formation and CO_2 capture at a molecular level to investigate dual inhibition-promotion behavior and refining existing models to simulate hydrate behavior under various conditions can be carried out to optimize the performance.
- To explore novel materials and develop more efficient additives, including bio-based or nanomaterial-enhanced hydrates, environmentally friendly solutions, and supportive policies.
- With the potential advancements in materials science, the hybrid collaborations and process optimization can be explored to increase the performance and decrease the CO_2 capturing cost.

CRediT authorship contribution statement

Aseem Dubey: Writing – original draft, Software, Methodology, Investigation, Data curation. **Akhilesh Arora:** Writing – review & editing, Visualization, 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

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

Acknowledgements

The authors express heartfelt thanks to Delhi Technological University (DTU), Delhi to provide required facilities for carrying out this work.

References

Aliabadi, M., Rasoolzadeh, A., Esmaeilzadeh, F., Alamdari, A., 2015. Experimental study of using CuO nanoparticles as a methane hydrate promoter. J. Nat. Gas Sci. Eng. 27, 1518–1522.

Aminnaji, M., Qureshi, M.F., Dashti, H., Hase, A., Mosalanejad, A., Jahanbakhsh, A., Babaei, M., Amiri, A., Maroto-Valer, M.M., 2024a. CO_2 Gas hydrate for carbon capture and storage applications – Part 1. Energy 300, 131579.

Aminnaji, M., Qureshi, M.F., Dashti, H., Hase, A., Mosalanejad, A., Jahanbakhsh, A., Babaei, M., Amiri, A., Maroto-Valer, M.M., 2024b. CO_2 gas hydrate for carbon capture and storage applications – Part 2. Energy 300, 131580.

Arjmandi, M., Chapoy, A., Tohidi, B., 2007. Equilibrium data of hydrogen, methane, nitrogen, carbon dioxide, and natural gas in semi-clathrate hydrates of tetrabutyl ammonium bromide. Chem Eng Data. 52 (6), 2153–2158.

Arjang, S., Manteghian, M., Mohammadi, A., 2013. Effect of synthesized silver nanoparticles in promoting methane hydrate formation at 4.7 MPa and 5.7 MPa. Chem. Eng. Res. Des. 91 (6), 1050–1054.

Arora, A., Kumar, A., Bhattacharjee, G., Balomajumder, C., Kumar, P., 2021. Hydrate-based carbon capture process: assessment of various packed bed systems for boosted kinetics of hydrate formation. J. Energy Resour. Technol. 143 (3), 033005.

Babu, P., Yang, T., Veluswamy, H.P., Kumar, R., Linga, P., 2013. Hydrate phase equilibrium of ternary gas mixtures containing carbon dioxide, hydrogen and propane. Chem Thermo 61, 58–63.

Babu, P., Linga, P., Kumar, R., Englezos, P., 2015. A review of the hydrate-based gas separation (HBGS) process for carbon dioxide pre-combustion capture. Energy 85, 261–279.

Babu, P., Ong, H.W.N., Linga, P., 2016. A systematic kinetic study to evaluate the effect of tetrahydrofuran on the clathrate process for pre-combustion capture of carbon dioxide. Energy 94, 431–442.

Babu, A., Viswanadhan, S.K., Veluswamy, H.P., 2024. Investigation of semi-clathrate formation kinetics for post-combustion CO_2 capture in the presence of amino acids. J. Gas Sci. Eng. 125, 205292.

Bavoh, C.B., Lal, B., Osei, H., Sabil, K.M., Mukhtar, M., 2019. A review on the role of amino acids in gas hydrate inhibition, CO_2 capture and sequestration, and natural gas storage. J. Nat. Gas Sci. Eng. 64, 52–71.

Buch, V., Devlin, J.P., Monreal, I.A., Jagoda-Cwiklik, B., Uras-Aytemiz, N., Cwiklik, L., 2009. Clathrate hydrates with hydrogen-bonding guests. Phys. Chem. Chem. Phys. 11, 10245–10265.

Bui, M., Adjiman, C.S., Bardow, A., Anthony, E.J., Boston, A., Brown, S., et al., 2018. Carbon capture and storage (CCS): the way forward. Energy Environ. Sci. 11 (5), 1062–1176.

Cai, Y., Chen, Y., Li, Q., Li, L., Huang, H., Wang, S., Wang, W., 2017. CO_2 hydrate formation promoted by a natural amino acid L-Methionine for possible application to CO_2 capture and storage. Energy Technol. 5 (8), 1195–1199.

Chaturvedi, E., Laik, S., Mandal, A., 2021. A comprehensive review of the effect of different kinetic promoters on methane hydrate formation. Chin. J. Chem. Eng. 32, 1–16.

Chen, S.J., Zhu, M., Fu, Y., Huang, Y.X., Tao, Z.C., Li, W.L., 2017. Using 13X, LiX, and LiPdAgX zeolites for CO_2 capture from postcombustion flue gas. Appl. Energy 191, 87–98.

Cheng, Z., Xu, H., Wang, S., Liu, W., Li, Y., et al., 2021. Effect of nanoparticles as a substitute for kinetic additives on the hydrate-based CO_2 capture. Chem. Eng. 424, 130329.

Choi, J.W., Chung, J.T., Kang, Y.T., 2014. CO_2 hydrate formation at atmospheric pressure using high efficiency absorbent and surfactants. Energy 78, 869–876.

Choi, S., Park, J., Park, J.H., Kim, S.C., Won, S.O., Kang, Y.T., 2021. Study on CO_2 hydrate formation characteristics with promoters for CO_2 capture and cold thermal energy transportation Jr. of Cleaner Production 295, 126392.

D'Alessandro, D.M., Smit, B., Long, J.R., 2010. Carbon dioxide capture: prospects for new materials. Angew. Chem., Int. Ed. 49 (35), 6058–6082.

Daraboina, N., Ripmeester, J., Englezos, P., 2013. The impact of SO_2 on post combustion carbon dioxide capture in bed of silica sand through hydrate formation. Int. J. Greenh. Gas Control 15, 97–103.

Dashti, H., Yew, L.Z., Lou, X., 2015. Recent advances in gas hydrate-based CO_2 capture. J. Nat. Gas Sci. Eng. 23, 195–207.

Demurov, A., Radhakrishnan, R., Trout, B.L., 2002. Computations of diffusivities in ice and CO_2 clathrate hydrates via molecular dynamics and Monte Carlo simulations. J. Chem. Phys. 116, 702–709.

Denning, S., Majid, A.A.A., Lucero, J.M., Crawford, J.M., Carreon, M.A., Koh, C.A., 2021. Methane hydrate growth promoted by microporous zeolitic imidazolate frameworks ZIF-8 and ZIF-67 for enhanced methane storage. ACS Sustainable Chem. Eng. 9 (27), 9001–9010.

Denning, S., Majid, A.A., Koh, C.A., 2022. Stability and growth of methane hydrates in confined media for carbon sequestration. J. Phys. Chem. C 126 (28), 11800–11809.

Dubey, A., Arora, A., 2022. Advancements in carbon capture technologies: a review. J. Clean. Prod. 373, 133932.

Duc, N.H., Chauvy, F., Herri, J.H., 2007. CO_2 capture by hydrate crystallization−A potential solution for gas emission of steelmaking industry. Energy Convers. Manage. 48, 1313–1322.

Eslamimanesh, A., Gharagheizi, F., Illbeigi, M., Mohammadi, A.H., Fazlali, A., Richon, D., 2012. Phase equilibrium modeling of clathrate hydrates of methane, carbon dioxide, nitrogen, and hydrogen+ water soluble organic promoters using Support Vector Machine algorithm. Fluid Phase Equilib 316, 34–45.

Fakharian, H., Ganji, H., Naderifar, A., 2017. Saline produced water treatment using gas hydrates. J. Environ. Chem. Eng. 5 (5), 4269–4273.

Fan, S., Wang, Y., Lang, X., 2011. CO_2 capture in form of clathrate hydrate-problem and practice. In: Proceedings of the 7th International Conference on Gas Hydrates (ICGH 2011), UK.

Feyzi, V., Mohebbi, V., 2020. Hybrid hydrate-membrane post-combustion CO_2 capture: a conceptual process design and analyses. Ind. Eng. Chem. Res. 59, 13132–13142.

Figueroa, J., Currier, R., Anderson, G., Tam, S., Deppe, G., Spencer, D., 2004. Developments in the SIMTECHE process CO_2 capture by formation of hydrates. In: Proc. Third Annual Conference on Carbon Capture and Sequestration. Washington DC, USA. May 3-6.

Firoozabadi, S.R., Bonyadi, M., Lashanizadegan, A., 2018. Experimental investigation of Fe_3O_4 nanoparticles effect on the carbon dioxide hydrate formation in the presence of magnetic field. Nat Gas Sci Eng 59, 374–386.

Gholinezhad, J., Chapoy, A., Tohidi, B., 2011. Separation and capture of carbon dioxide from CO_2/H_2 syngas mixture using semi-clathrate hydrates. Chem. Eng. Res. Des. 89 (9), 747–751.

Hashimoto, S., Murayama, S., Sugahara, T., Ohgaki, K., 2006. Phase equilibria for H_2 plus CO_2 plus tetrahydrofuran plus water mixtures containing gas hydrates. J. Chem. Eng. Data 51 (5), 1884–1886.

Hassan, M.H.A., Sher, F., Zarren, G., Suleiman, N., Tahir, A.A., Snape, C.E., 2020. Kinetic and thermodynamic evaluation of effective combined promoters for CO_2 hydrate formation. J. Nat. Gas Sci. Eng. 78, 103313.

Hassan, M.H.A., Sher, F., Fareed, B., Ali, U., Zafar, A., Bilal, M., Iqbal, H.M.N., 2021. Sustainable hydrates for enhanced carbon dioxide capture from an integrated gasification combined cycle in a fixed bed reactor. Ind. Eng. Chem. Res. 60 (30), 11346–11356.

Ho, L.C., Babu, P., Kumar, R., Linga, P., 2013. HBGS (hydrate based gas separation) process for carbon dioxide capture employing an unstirred reactor with cyclopentane. Energy 63, 252–259.

IEA, 2019. Global Energy & CO_2 Status Report 2019 IEA. https://www.iea.org/reports%20/global-energy-co2-status-report-2019.

IEA. Global Energy Review: CO_2 Emissions in 2021. https://www.iea.org/reports/global-energy-review-co2-emissions-in-2021-2.

Intergovernmental Panel on Climate Change. https://www.ipcc.ch/2018/10/08/summary-for-policymakers-of-ipcc-special-report-on-global-warming-of-1-5c-approved-by-governments, 2018.

Kamata, Y., Oyama, H., Shimada, W., Ebinuma, T., Takeya, S., 2004. Gas separation method using tetra-n-butyl ammonium bromide semi-clathrate hydrate. Jpn. J. Appl. Phys. 43, 362–365.

Kang, S.P., Lee, J., Seo, Y., 2013. Pre-combustion capture of CO_2 by gas hydrate formation in silica gel pore structure. Chem Eng. J. 218, 126–132.

Karaszova, M., Zach, B., Petrusova, Z., Červenka, V., Bobak, M., Šyc, M., Izak, P., 2020. Post-combustion carbon capture by membrane separation, review. Sep. Purif. Technol. 238, 116448.

Khanmohammadian, E., Mohammadi, M., Hashemi, R., Eslami, S., Ehsani, M.R., 2023. Improvement of gas hydrate-based CO_2 capture from CH_4/CO_2 mixture using silica and modified silica nanoparticles in the presence of potassium hydroxide. Fuel 334 (1), 126458.

Kim, S.M., Lee, J.D., Lee, H.J., Lee, E.K., Kim, T., 2011. Gas hydrate formation method to capture the carbon dioxide for pre-combustion process in IGCC plant. Int. J. Hydrog. Energy 36 (1), 1115–1121.

Kim, S., Baek, I., You, J.K., Seo, Y., 2015a. Guest gas enclathration in tetra-n-butyl ammonium chloride (TBAC) semiclathrates: potential application to natural gas storage and CO_2 capture. Appl. Energy 140, 107–112.

Kim, D., Ahn, Y.H., Lee, H., 2015b. Phase equilibria of CO_2 and CH_4 hydrates in intergranular meso/macro pores of MIL-53 metal organic framework. J. Chem. Eng. Data 60 (7), 2178–2185.

Kim, S., Lee, S.H., Kang, Y.T., 2017. Characteristics of CO_2 hydrate formation/dissociation in H_2O + THF aqueous solution and estimation of CO_2 emission reduction by district cooling application. Energy 120, 362–373.

Kirschgen, T.M., Zeidler, M.D., Geil, B., Fujara, F., 2003. A deuteron nmr study of the tetrahydrofuran clathrate hydrate - part ii: coupling of rotational and translational dynamics of water. Phys. Chem. Chem. Phys. 5, 5247–5252.

Koh, C.A., Sloan, E.D., Sum, A.K., Wu, D.T., 2011. Fundamentals and applications of gas hydrates. Ann. Rev. Chem. Biomol. Eng. 2, 237–257.

Kumar, A., Sakpal, T., Linga, P., Kumar, R., 2013. Influence of contact medium and surfactants on carbon dioxide clathrate hydrate kinetics. Fuel 105, 664–671.

Kumar, R., Wu, H.J., Englezos, P., 2006. Incipient hydrate phase equilibrium for gas mixtures containing hydrogen, carbon dioxide and propane. Fluid Phase Equilib. 244 (2), 167–171.

Kumar, R., Linga, P., Ripmeester, J.A., Englezos, P., 2009. A two-stage clathrate hydrate/membrane process for pre-combustion capture of carbon dioxide and hydrogen. J. Environ. Eng. 135 (6), 411–417.

Lee, H.J., Lee, J.D., Linga, P., Englezos, P., Kim, Y.S., Lee, M.S., Kim, Y.D., 2010. Gas hydrate formation process for pre-combustion capture of carbon dioxide. Energy 35 (6), 2729–2733.

Lee, Y.J., Kawamura, T., Yamamoto, Y., Yoon, J.H., 2012. Phase equilibrium studies of tetrahydrofuran (THF) + CH_4, THF + CO_2, CH_4+ CO_2, and THF + CO_2+ CH_4 hydrates. Chem. Eng. Data 57 (12), 3543–3548.

Lee, Y., Kim, H., Lee, W., Kang, D.W., Lee, J.W., Ahn, Y.H., 2023. Thermodynamic and kinetic properties of CO_2 hydrates and their applications in CO_2 capture and separation. J. Environ. Chem. Eng. 11, 110933.

Li, J., Liang, D., Guo, K., Wang, R., Fan, S., 2006. Formation and dissociation of HFC134a gas hydrate in nano-copper suspension. Energy Convers. Manage. 47 (2), 201–210.

Li, S., Fan, S., Wang, J., Lang, X., Lang, X., 2010. Semiclathrate hydrate phase equilibria for CO_2 in the presence of tetra-n-butyl ammonium halide (bromide, chloride, or fluoride). J. Chem. Eng. Data 55 (9), 3212–3215.

Li, X.S., Xu, C.G., Chen, Z.Y., Wu, H.J., 2011a. Hydrate-based pre-combustion carbon dioxide capture process in the system with tetra-n-butyl ammonium bromide solution in the presence of cyclopentane. Energy 36 (3), 1394–1403.
Li, X.S., Xia, Z.M., Chen, Z.Y., Wu, H.J., 2011b. Pre combustion capture of carbon dioxide and hydrogen with a one-Stage hydrate/membrane process in the presence of tetra-nbutylammonium bromide (TBAB). Enr. Fuels 25, 1302–1309.
Li, X.S., Xu, C.G., Chen, Z.Y., Cai, J., 2012a. Synergic effect of cyclopentane and tetra-n-butyl ammonium bromide on hydrate-based carbon dioxide separation from fuel gas mixture by measurements of gas uptake and X-ray diffraction patterns. Int. J. Hydrog. Energy 37 (1), 720–727.
Li, X.S., Zhan, H., Xu, C.G., Zeng, Z.Y., Lv, Q.N., Yan, K.F., 2012b. Effects of tetrabutyl-(ammonium/phosphonium) salts on clathrate hydrate capture of CO_2 from simulated flue gas. Energy Fuels 26, 2518–2527.
Li, D.L., Peng, H., Liang, D.Q., 2017. Thermal conductivity enhancement of clathrate hydrate with nanoparticles. Int. J. Heat Mass Tran. 104, 566–573.
Li, A., Wang, J., Bao, B., 2019. High-efficiency CO_2 capture and separation based on hydrate technology: a review. Greenh. Gases: Sci. Technol. 9 (2), 175–193.
Li, Y., Chen, M., Song, H., Yuan, P., Liu, D., Zhang, B., Bu, H., 2020. Methane hydrate formation in the stacking of kaolinite particles with different surface contacts as nanoreactors: a molecular dynamics simulation study. Appl. Clay Sci. 186, 105439.
Liang, H., Guan, D., Shi, K., Yang, L., Zhang, L., Zhao, J., Song, Y., 2022. Characterizing mass-transfer mechanism during gas hydrate formation from water droplets. Chem. Eng. J. 428, 132626.
Lin, W., Delahaye, A., Fournaison, L., 2008. Phase equilibrium and dissociation enthalpy for semi-clathrate hydrate of CO_2 + TBAB. Fluid Phase Equilib 264 (1–2), 220–227.
Liang, S., Liang, D., Wu, N., Yi, L., Hu, G., 2016. Molecular mechanisms of gas diffusion in CO_2 hydrates. J. Phys. Chem. 120 (30), 16298–16304.
Linga, P., Adeyemo, A., Englezos, P., 2008. Medium-pressure clathrate hydrate/membrane hybrid process for post combustion capture of carbon dioxide. Environ. Sci. Technol. 42 (1), 315–320.
Linga, P., Kumar, R., Englezos, P., 2007. Gas hydrate formation from hydrogen/carbon dioxide and nitrogen/carbon dioxide gas mixtures. Chem. Eng. Sci. 62 (16), 4268–4276.
Linga, P., Kumar, R., Lee, J.D., Ripmeester, J., Englezos, P., 2010. A new apparatus to enhance the rate of gas hydrate formation: application to capture of carbon dioxide. Int. J. Greenh. Gas Control 4 (4), 630–637.
Lim, Y., Kim, J., Jung, J., Lee, C.S., Han, C., 2013. Modeling and Simulation of CO_2 Capture Process for Coalbased Power Plant using Amine Solvent in South Korea. Energy Proc 37, 1855–1862.
Lirio, C.F.D.S., Pessoa, F.L.P., Uller, A.M.C., 2013. Storage capacity of carbon dioxide hydrates in the presence of sodium dodecyl sulfate (SDS) and tetrahydrofuran (THF). Chem. Eng. Sci. 96, 118–123.
Liu, F.P., Li, A.R., Qing, S.L., Luo, Z.D., Ma, Y.L., 2022. Formation kinetics, mechanism of CO_2 hydrate and its applications. Renew. and Sustanable Ener. Reviews. 159, 112221.
Majid, A.A.A., Worley, J., Koh, C.A., 2021. Thermodynamic and kinetic promoters for gas hydrate technological applications. Energy Fuels 35 (23), 19288–19301.
Meyer, F.D., Jouenne, S., 2022. Industrial carbon capture by absorption: recent advances and path forward. Curr. Opin. Chem. Eng. 38, 100868.
Mohammadi, A., Manteghian, M., Haghtalab, A., Mohammadi, A.H., Rahmati-Abkenar, M., 2014. Kinetic study of carbon dioxide hydrate formation in presence of silver nanoparticles and SDS. Chem. Eng. J. 237, 387–395.
Mohammadi, M., Haghtalab, A., Fakhroueian, Z., 2016. Experimental study and thermodynamic modeling of CO_2 gas hydrate formation in presence of zinc oxide nanoparticles. Chem. Thermo. 96, 24–33.
Mohammadi, M., Pakzad, M., Mohammadi, A.H., Jahangiri, A., 2018. Kinetics of (TBAF + CO_2) semi-clathrate hydrate formation in the presence and absence of SDS. Petr. Sci. 15, 375–384.
Muromachi, S., Udachin, K.A., Shin, K., Alavi, S., Moudrakovski, I.L., Ohmura, R., et al., 2014. Guest-induced symmetry lowering of an ionic clathrate material for carbon capture. Chem. Commun. 50, 11476–11479.
Nashed, O., Partoon, B., Lal, B., Sabil, K.M., Shariff, A.M., 2018. Review the impact of nanoparticles on the thermodynamics and kinetics of gas hydrate formation. J. Nat. Gas Sci. Eng. 55, 452–465.
Nashed, O., Partoon, B., Lal, B., Sabil, K.M., Shariff, A.M., 2019. Investigation of functionalized carbon nanotubes' performance on carbon dioxide hydrate formation. Energy 174, 602–610.
Nesterov, A.N., Reshetnikov, A.M., Manakov, A.Y., Rodionova, T.E., Paukshtis, E.A., Asanov, I.P., Bardakhanov, S.P., Bulavchenko, A.I., 2015. Promotional and inhibition of gas hydrates formation by oxide powders. J. Mol. Liq. 204, 118–125.
Nesterov, A.N., Reshetnikov, A.M., Manakov, A.Y., Adamova, T.P., 2017. Synergistic effect of combination of surfactant and oxide powder on enhancement of gas hydrates nucleation. J. Energy Chem. 26, 808–814.
Nguyen, N.N., La, V.T., Huynh, C.D., Nguyen, A.V., 2022. Technical and economic perspectives of hydrate-based carbon dioxide capture. Applied Energy 307, 118237.
Nguyen, N.N., Nguyen, A.V., Nguyen, K.T., Rintoul, L., Dang, L.X., 2016. Unexpected inhibition of CO_2 gas hydrate formation in dilute TBAB solutions and the critical role of interfacial water structure. Fuel 185, 517–523.
Orujov, A., Coddington, K., Aryana, S.A., 2023. A review of CCUS in the context of foams, regulatory frameworks and monitoring. Energies 16, 3284.
Pahlavanzadeh, H., Khanlarkhani, M., Rezaei, S., Mohammadi, A.H., 2019. Experimental and modelling studies on the effects of nanofluids (SiO_2, Al_2O_3, and CuO) and surfactants (SDS and CTAB) on CH_4 and CO_2 clathrate hydrates formation. Fuel 253, 392–405.
Pandey, G., Poothia, T., Kumar, A., 2022. Hydrate based carbon capture and sequestration (HBCCS): an innovative approach towards decarbonization. Appl. Energy 326, 119900.
Pandey, J.S., von Solms, N., 2023. Metal–organic frameworks and gas hydrate synergy: a pandora's box of unanswered questions and revelations. Energies 16, 111.
Park, S.S., Lee, S.B., Kim, N.J., 2010. Effect of multi-walled carbon nanotubes on methane hydrate formation. Ind. Eng. Chem. 16, 551–555.
Park, S., Lee, S., Lee, Y., Seo, Y., 2013a. CO_2 capture from simulated fuel gas mixtures using semiclathrate hydrates formed by quaternary ammonium salts. Environ. Sci. Technol. 47 (13), 7571–7577.
Park, S., Lee, S., Lee, Y., Seo, Y., 2013b. Hydrate-based pre-combustion capture of carbon dioxide in the presence of a thermodynamic promoter and porous silica gels. Int. J. Greenh. Gas Control 14, 193–199.
Petukhov, A.N., Atlaskin, A.A., Kudryavtsev, M.S., Kryuchkov, S.S., et al., 2022. CO_2 capture process through hybrid gas hydrate-membrane technology: complex approach for the transition from theory to practice. Enviro. Chem. Eng. 10, 108104.
Phan, A., Schlösser, H., Striolo, A., 2021. Molecular mechanisms by which tetrahydrofuran affects CO_2 hydrate growth: implications for carbon storage. Chem. Eng. J. 418, 129423.
Prasad, P.S.R., Kiran, B.S., 2018a. Are the amino acids thermodynamic inhibitors or kinetic promoters for carbon dioxide hydrates? J. Nat. Gas Sci. Eng. 52, 461–466.
Prasad, P.S.R., Kiran, B.S., 2018b. Clathrate hydrates of greenhouse gases in the presence of natural amino acids: storage, transportation and separation applications. Sci. Rep. 1–10.
Qasim, A., Khan, M.S., Lal, B., Shariff, A.M., 2019. A perspective on dual purpose gas hydrate and corrosion inhirbitors for flow assurance. J. Pet. Sci. Eng. 183, 106418.
Ricaurte, M., Dicharry, C., Broseta, D., Renaud, X., Torré, J.P., 2013. CO_2 removal from a CO_2– CH_4 gas mixture by clathrate hydrate formation using THF and SDS as water-soluble hydrate promoters. Ind. Eng. Chem. Res. 52 (2), 899–910.
Ricaurte, M., Dicharry, C., Renaud, X., Torré, J., 2014. Combination of surfactants and organic compounds for boosting CO_2 separation from natural gas by clathrate hydrate formation. Fuel 122, 206–217.
Rodhe, H., 1990. A comparison of the contribution of various gases to the greenhouse effect. Science 248, 1217–1219.
Sa, J.H., Lee, B.R., Park, D.H., Han, K., Chun, H.D., Lee, K.H., 2011. Amino acids as natural inhibitors for hydrate formation in CO_2 sequestration. Environ. Sci. Technol. 45, 5885–5891.
Sa, J.H., Kwak, G.H., Han, K., Ahn, D., Lee, K.H., 2015. Gas hydrate inhibition by perturbation of liquid water structure. Sci. Rep. 5, 11526.
Said, S., Govindaraj, V., Herri, J.M., Ouabbas, Y., Khodja, M., Belloum, M., et al., 2016. A study on the influence of nanofluids on gas hydrate formation kinetics and their potential: application to the CO_2 capture process. J. Nat. Gas Sci. Eng. 32, 95–108.
Sarlak, H., Azimi, A., Mostafa, S., Ghomshe, T., Mirzaei, M., 2018. Thermodynamic study of CO_2 hydrate formation in the presence of SDS and graphene oxide nanoparticles. Adv. in Envir. Tech. 4, 233–240.
Sayani, J.K.S., Lal, B., Pedapati, S.R., 2022. Comprehensive review on various gas hydrate modelling techniques: prospects and challenges. Arch. Comput. Methods Eng. 29, 2171–2207.
Sheng, L., Wang, K., Deng, J., Chen, G., Luo, G., 2023. Gas–liquid microdispersion and microflow for carbon dioxide absorption and utilization: a review. Curr. Opin. Chem. Eng. 40, 100917.
Seo, Y.T., Moudrakovski, I.L., Ripmeester, J.A., Lee, J.W., Lee, H., 2005. Efficient recovery of CO_2 from flue gas by clathrate hydrate formation in porous silica gels. Environ. Sci. Technol. 39 (7), 2315–2319.
Shin, H.J., Lee, Y.J., Im, J.H., Han, K.W., et al., 2009. Thermodynamic stability, spectroscopic identification and cage occupation of binary CO_2 clathrate hydrates. Chem. Eng. Sci. 64 (24), 5125–5130.
Shirmohammadi, R., Aslani, A., Ghasempour, R., 2020. Challenges of carbon capture technologies deployment in developing countries. Sustain. Energy Technol. Assess. 42, 100837.
Shifeng, L., Shuanshi, F., Jinqu, W., Xuemei, L., Yanhong, W., 2010. Clathrate hydrate capture of CO_2 from simulated flue gas with cyclopentane/water emulsion. Chin. J. Chem. Eng. 18, 202–206.
Sloan, E.D., Koh, C.A., 2007. Clathrate Hydrates of the Natural Gases, third ed. CRC Press, Boca Raton, FL.
Song, Y., Wang, X., Yang, M., Jiang, L., Liu, Y., Dou, B., Zhao, J., Wang, S., 2013. Study of selected factors affecting hydrate-based carbon dioxide separation from simulated fuel gas in porous media. Energy Fuels 27 (6), 2859–3602.
Spencer, D.F., 1999. Integration of an Advanced CO_2 Separation Process with Methods for Disposing of CO_2 in Oceans and Terrestrial Deep Aquifers. ABB Corporate Research Ltd.: Baden-Daettwil, Switzerland.
Srivastava, S., Kollemparembil, A.M., Zettel, V., Claben, T., Gatternig, B., Delgado, A., Hitzmann, B., 2022. Experimental investigation of CO_2 uptake in CO_2 hydrates formation with amino acids as kinetic promoters and its dissociation at high temperature. Nature portfolio 12, 8359.
Thoutam, P., Gomari, S.R., Chapoy, A., Ahmad, F., Islam, M., 2019. Study on CO_2 hydrate formation kinetics in saline water in the presence of low concentrations of CH_4. ACS Omega 4 (19), 18210–18218.
Vega, F., Baena-Moreno, V.M., Gallego Fernandez, L.M., Portillo, E., Navarrete, B., Zhang, Z., 2020. Current status of CO_2 chemical absorption research applied to CCS: towards full deployment at industrial scale. Appl. Energy 260, 114313.
Veluswamy, H.P., Premasinghe, K.P., Linga, P., 2017. CO_2 Hydrates – effect of additives and operating conditions on the morphology and hydrate growth. Ener. Procedia. 105, 5048–5054.

Veluswamy, H.P., Kumar, A., Seo, Y., Lee, J.D., Linga, P., 2018. A review of solidified natural gas (SNG) technology for gas storage via clathrate hydrates. Appl. Energy 216, 262–285.
Wang, Y., Lang, X., Fan, S., 2013a. Hydrate capture CO_2 from shifted synthesis gas, flue gas and sour natural gas or biogas. J. Ener. Chem. 22 (1), 39–47.
Wang, W., Ma, C., Lin, P., Sun, L., Cooper, A.I., 2013b. Gas storage in renewable bioclathrates. Energy Environ. Sci. 6, 105–107.
Wang, P., Teng, Y., Zhu, J., Bao, W., Han, S., Li, Y., Zhao, Y., Xie, 2022. Review on the synergistic effect between metal-organic frameworks and gas hydrates for CH_4 storage and CO_2 separation applications. Renew Sustain Energy Rev 167, 112807.
Wang, J., Wang, R., Yoon, R.-H., Seol, Y., 2015. Use of hydrophobic particles as kinetic promoters for gas hydrate formation. J. Chem. Eng. Data 60 (2), 383–388.
Wang, F., Fu, S., Guo, G., Jia, Z., Luo, S., Guo, R., 2016. Experimental study on hydrate-based CO_2 removal from CH_4/CO_2 mixture. Energy 104, 76–84.
Wang, F., Song, Y.-M., Liu, G.-Q., Guo, G., Luo, S.-J., Guo, R.-B., 2018. Rapid methane hydrate formation promoted by Ag & SDS-coated nanospheres for energy storage. Appl. Energy 213, 227–234.
Wang, X., Zhang, F., Lipiński, W., 2020. Research progress and challenges in hydrate-based carbon dioxide capture applications. Appl. Energy 269, 114928.
Wang, H., Wu, Q., Zhang, B., 2021. Influence of THF and THF/SDS on the kinetics of CO_2 hydrate formation under stirring. Front. Energy Res. 9, 633929.
Wang, X., Zhang, Y., Wang, F., Yin, Z., Zhang, Z., Ting, V.P., 2024. Unlocking the potential of hydrate-based carbon capture: a review of passive techniques for CO_2 hydrate formation promotion. Gas Sci. Eng. 126, 205323.
Xia, Z.M., Li, X.C., Chen, Z.Y., Li, G., Cai, J., Wang, Y., Yan, K.F., Xu, C.G., 2017. Hydrate-based acidic gases capture for clean methane with new synergic additives. Appl. Energy 207, 584–593.
Xia, Z., Li, Z., Chen, Z., Li, X., Zhang, Y., Yan, K., et al., 2019. $CO_2/H_2/H_2O$ hydrate formation with TBAB and nanoporous materials. Energy Proc. 158, 5866–5871.
Xia, Z., Zhao, Q., Chen, Z., Li, X., Zhang, Y., Xu, C., Yan, K., 2022. Review of methods and applications for promoting gas hydrate formation process. J. Nat. Gas Sci. Eng. 101, 104528.
Xie, N., Chen, B., Tan, C., Liu, Z., 2017. Energy consumption and exergy analysis of MEA-based and hydrate-based CO_2 separation. Ind. Eng. Chem. Res. 56, 15094–15101.
Xu, C.G., Li, X.S., Lv, Q.N., Chen, Z.Y., Cai, J., 2012. Hydrate-based CO_2 (carbon dioxide) capture from IGCC (integrated gasification combined cycle) synthesis gas using bubble method with a set of visual equipment. Energy 44 (1), 358–366.
Xu, C.G., Xie, W.J., Chen, G.S., Yan, X.X., Cai, J., Chen, J.Y., Li, X.S., 2020. Study on the influencing factors of gas consumption in hydrate-based CO_2 separation in the presence of CP by Raman analysis. Energy 198, 117316.
Yang, M., Jing, W., Wang, P., Jiang, L., Song, Y., 2015. Effects of an additive mixture (THF+ TBAB) on CO_2 hydrate phase equilibrium. Fluid Ph. Equilib. 401, 27–33.
Yu, Y., Zhou, S., Li, X., Wang, S., 2016. Effect of graphite nanoparticles on CO_2 hydrate phase equilibrium. Fluid Ph. Equilib. 414, 23–28.
Yu, Y.S., Zhang, X., Liu, J.W., Lee, Y., Li, S.C., 2021. Natural gas hydrate resources and hydrate technologies: a review and analysis of the associated energy and global warming challenges. Energy Environ. Sci. 14, 5611–5668.
Zhang, J.S., Lee, J.W., 2009. Equilibrium of hydrogen + cyclopentane and carbon dioxide + cyclopentane binary hydrates. Chem. Eng. Data 54 (2), 659–661.
Zhang, Y., Yang, M., Song, Y., Jiang, L., Li, Y., Cheng, C., 2014. Hydrate phase equilibrium measurements for (THF+ SDS+ CO_2+ N_2) aqueous solution systems in porous media. Fluid Ph. Equilib. 370, 12–18.
Zhang, H., 2021. Regulations for carbon capture, utilization and storage: comparative analysis of development in Europe, China and the Middle East. Resour. Conserv. Recycl. 173, 105722.
Zheng, J., Zhang, P., Linga, P., 2016. Semiclathrate hydrate process for pre-combustion capture of CO_2 at near ambient temperatures. Appl. Energy 194, 267–278.
Zhang, J., Yedlapalli, P., Lee, J.W., 2009. Thermodynamic analysis of hydrate-based precombustion capture of CO_2. Chem. Eng. Sci. 64 (22), 4732–4736.
Zheng, J., Bhatnagar, K., Khurana, M., Zhang, P., Zhang, B.Y., Linga, P., 2018. Semiclathrate based CO_2 capture from fuel gas mixture at ambient temperature: effect of concentrations of tetra-n-butylammonium fluoride (TBAF) and kinetic additives. Appl. Energy 217, 377–389.
Zhong, D.L., Li, Z., Lu, Y.Y., Wang, J.L., Yan, J., 2015. Evaluation of CO_2 removal from a CO_2 + CH_4 gas mixture using gas hydrate formation in liquid water and THF solutions. Appl. Energy 158, 133–141.
Zhou, X., Zhang, Y., Zang, X., Liang, D., 2020. Formation kinetics of the mixed cyclopentane-carbon dioxide hydrates in aqueous sodium chloride solutions. Energies 13 (17), 4388.

Journal of Physics D: Applied Physics

PURPOSE-LED PUBLISHING™

ACCEPTED MANUSCRIPT

Emergence of Topological Phase and Non-Trivial Surface States in Rare-Earth Semimetal GdSb with Pressure

To cite this article before publication: Nidhi . *et al* 2024 *J. Phys. D: Appl. Phys.* in press https://doi.org/10.1088/1361-6463/ad7da0

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. All rights, including for text and data mining, AI training, and similar technologies, are reserved.**.

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 04/10/2024 at 06:31

Emergence of Topological Phase and Non-Trivial Surface States in Rare-Earth Semimetal GdSb with Pressure

Nidhi[1], Ramesh Kumar[1], Ramesh K. Bibiyan[2], Mukhtiyar Singh[1*]

[1]Computational Quantum Material Design (CQMD) Lab, Department of Applied Physics, Delhi Technological University, Delhi (INDIA)-110042

[2]Department of Physics, Guru Jambheshwer University of Science & Technology, Hisar (INDIA)-125001

[*]mukhtiyarsingh@dtu.ac.in; msphysik09@gmail.com

Abstract:

We studied the evolution of band topology with external pressure in rare-earth GdSb using *first-principles* calculations. This material crystallizes in a *rocksalt-type* structure and shows the structural phase transition to *CsCl-type* structure at 26.1 GPa. The phonon dispersions are analyzed to ascertain the dynamical stability of this material. We used the hybrid density functional theory with the inclusion of spin-orbit coupling to investigate the structural, electronic, and topological phase transitions. At ambient pressure, GdSb shows a topologically trivial state which is in agreement with the experimental reports. The first topological phase transition is observed at 6 GPa of hydrostatic pressure (at the high symmetry *X-point*) verified with the help of single band inversions and surface states analysis along the (001) plane. The non-zero value of the first Z_2 topological invariant and the presence of the Dirac cone also confirm the topological phase of this material. Further, an increase in pressure to 12 GPa, two band inversions, at *Γ*- as well as *X-points*, are observed which corresponds to the trivial nature of GdSb. The same is verified with (0;000) values of Z_2 topological invariants and a pair of Dirac cones in surface states. It is noted that the crystal symmetries are preserved throughout the study and topological phase transition values are much lower than structural phase transition pressure i.e., 26.1 GPa.

Keywords: first-principles calculations, hybrid functional, topological semimetal, GdSb, rare-earth monopnictides

1. Introduction

Accepted Manuscript

Over the last two decades, Topological Materials (TMs) have been at the center stage of research in condensed matter physics due to their peculiar electronic transport at the surface/edge [1-4]. These unusual properties make TMs an ideal candidate for various futuristic technological domains such as spintronics [5], quantum computing [6] and waste-to-useful energy harvesting [7]. In the presence of inherent spin-orbit coupling (SOC), a typical TM possesses exotic metallic surface states in *(d-1)*-dimensions but has a *d*-dimensional bulk insulator [8] by virtue of spin-momentum locking with time-reversal symmetry (TRS) as well as inversion symmetry (IS). These surface states are robust against impurities and prevent backscattering [8-10]. TMs can be broadly categorized as topological insulators (TIs) [4, 9-11] and topological semimetals (TSM) [12-15]. The TSM exhibit many novel phenomena such as chiral anomaly [16], Weyl fermion quantum transport [17] etc. These TSM can be further divided into Weyl, Dirac and nodal-line semimetals [14-15, 18-20]. These TSM show linear dispersion around the nodes, known as Dirac and Weyl points, and can be identified with the Fermi arc on the surface [21-24]. There is another class known as Z_2 topological semimetal which shows non-linear dispersion and have no bulk band gap but their surface states are protected by TRS and IS [25-26]. This category includes recently discovered rare-earth based monopnictide family [27-31] RE-V (RE: rare earth, V = P, As, Sb, and Bi). These materials have applications in solar cells [32], tunnel junctions [33], terahertz detectors etc [34]. The angle-resolved photoemission spectroscopy (ARPES) has been employed to examine the topological properties of CeSb and CeBi semimetals [35, 36], also Guo et al have discovered a potential Weyl fermion in ferromagnetic CeSb. The first-principles studies of CeX [36-39], PrX (X=Sb, Bi) [38-39], SmBi [39] and GdBi [39] have confirmed the band inversions and non-zero Z_2 topological invariants in these materials.

An increase in spin-orbit interaction (SOI) via pressure or strain can be an effective tool to obtain a topological phase transition (TPT) as it maintains the charge neutrality in the system. On the other hand, unwanted defects and inhomogeneity occur in the case of doping [25, 26, 40-42]. It was reported that RE-V materials such as LaY (Y=As, Sb, Bi) [25, 26, 43], YX (X=Sb, Bi) [44, 45], TmSb [46], YbAs [47], etc., show TPT with the applied hydrostatic pressure.

A potential RE-V material Gadolinium mono-Antimonide (GdSb) has an electron-hole compensated semimetallic nature which may result in extremely large magnetoresistance (XMR) in this material [48]. An ARPES study has revealed quantum confinement effects in this material and an electron-rich Hall coefficient is observed which has been attributed to high electron Hall

Accepted Manuscript

mobility [49]. A recent experimental and theoretical study has established GdSb as a topologically trivial semimetallic at ambient conditions and converted it into a topologically non-trivial state with biaxial strain. The epitaxial growth of this material on semiconducting III–V buffer layer structure has suggested that the electron-hole carrier ratio can be tuned with compressive strain and can be useful as another degree of freedom to tune magnetoresistance such materials [39, 50]. These studies motivate us to investigate the TPT in GdSb under hydrostatic pressure. Existence of topological non-trivial phase in this material can enhance the electron-hole concentration in the vicinity of Fermi level. Therefore, we examined the electronic, structural stability and topological properties of GdSb semimetal at ambient and elevated hydrostatic pressure conditions. The TPT is observed in this material at 6 GPa and becomes trivial again at 12 GPa with an even number of band inversions. The topological non-trivial nature at ambient and applied pressure is confirmed with the Z_2 topological invariant at time reversal invariant momenta (TRIM) points. We observed the Dirac cone in surface density of states (SDOS) projected at the (001) plane.

2. Computational details

All the structural, electronic, and topological properties of GdSb were obtained using first-principles calculations [51-53] within the framework of Projected Augmented Wave (PAW) [53] potentials as implemented in the Vienna Ab initio Simulation Package (VASP) [54]. The PAW potential used for Gd and Sb had ten ($4f^7\ 5d^1\ 6s^2$) and fifteen valence electrons ($4d^{10}\ 5s^2\ 5p^3$) respectively. Initially, the exchange-correlation potentials were treated using generalized gradient approximation (GGA) of Perdew–Burke–Ernzerhof (PBE) [55] for geometrical optimizations wherein the interatomic forces were kept smaller than 0.001 eV. Later, the hybrid functional (HSE06) [56-57] was adopted for electronic band structure calculations. The optimized plane wave basis set cut-off energy was 340 eV and a 7x7x7 k-mesh was used for sampling of the Brillion zone. All calculations were carried out with consideration of SOC, except ionic optimization, and the overall energy convergence criteria was set to 10^{-6} eV. An eight-atom cubic unit cell was used to optimize the structure under volumetric pressure, from which a two-atom primitive unit cell was taken out for further calculations. The dynamical stability of the material was confirmed by phonon band structure calculations using the PHONOPY code [58]. The Wannier90 was used to calculate the maximally localized Wannier functions (MLWFs) using the tight-binding model [59]. Green's function approach, as implemented in the Wanniertool package [60], was employed to calculate

Accepted Manuscript

the surface states of the material along (001) plane. The product of parities at TRIM points was used to calculate the Z_2 topological invariants in accordance with the Kane and Mele criteria [1, 2].

3. Results and discussions:

Similar to the other rare earth monopnictides, GdSb also crystallizes in a *NaCl-type* crystal structure at ambient conditions with space group $Fm\bar{3}m$. This material has the atomic positions (0, 0, 0) for Sb and (0.5, 0.5, 0.5) for Gd, as illustrated in figure 1(a). Our optimized lattice parameter (table 1) is consistent with other experimental and theoretical reports. The primitive unit cell is extracted from the conventional unit cell to prevent band folding. The truncated octahedron Brillion zone (BZ) at the equilibrium condition of *NaCl-type* structure is shown in figure 1(c) with *Γ*-point at the center and the *M-point* in the center of squares taken in the (001) plane, as marked with yellow color in figure 1(c). The BZ preserves its truncated octahedron shape with changing volumetric pressure.

Table 1. Comparative lattice parameter and structural phase transition (SPT) of GdSb

GdSb	Previous experimental study	Previous theoretical study	Present study
Lattice parameter, a (Å)	6.21 [61]; 6.219 [50]; 6.22 [62, 63]	6.09 [64]; 6.219 [65]; 6.25 [66];	6.248
SPT in GPa	---------	22 [67]; 24.6 [61]; 26.6 [66]	26.1

Accepted Manuscript

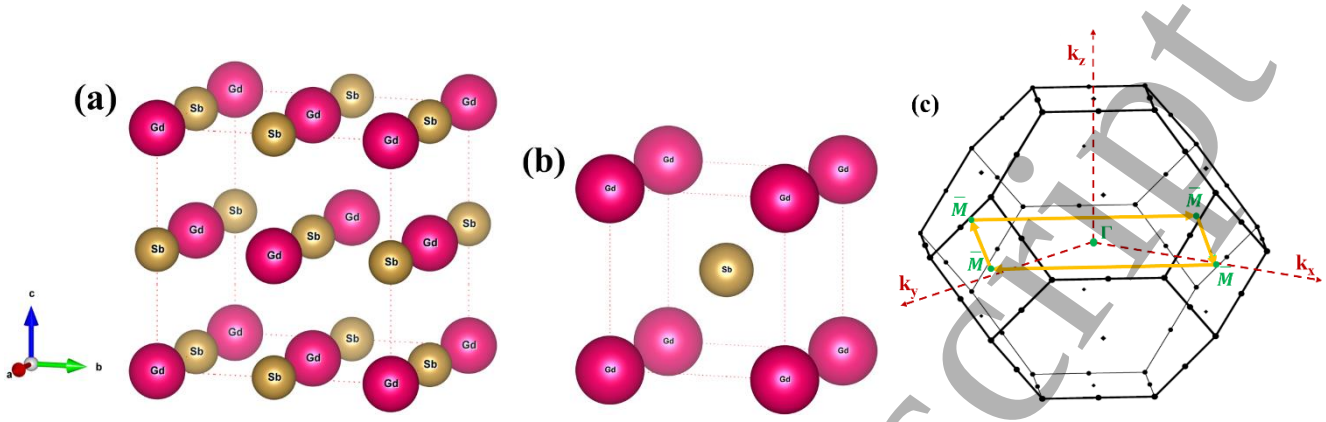


Figure 1. The Crystal structure of GdSb (a) *NaCl-type* (b) *CsCl-type* (c) Brillouin zone of GdSb. The (001) plane is represented by a yellow boundary cornered by $\bar{M}$ points.

The structural phase transition (SPT) is found to examine the stability of this material in *NaCl-type* structure. As the volumetric pressure increases, the crystal structure of GdSb transforms into *CsCl-type* from *NaCl-type* (figure 1(b)) and SPT takes place. The enthalpy of the *NaCl-type* and *CsCl-type* crystal structures are calculated for the pressure up to 30 GPa. The structure having lower enthalpy H = E+ PV (where E denotes the total energy, P = pressure and V = volume of the unit cell) at a given value of hydrostatic pressure, is more stable. At ambient pressure, figure 2(a), the enthalpy of the *NaCl-type* structure lies below the *CsCl-type* structure and it becomes more stable. When we increase pressure, the enthalpy of both types of structures increases and a crossover at a hydrostatic pressure of 26.1 GPa is observed. At 26.1 GPa the *CsCl-type* structure becomes more stable than the *NaCl-type* structure and identifies as SPT pressure. figure 2(b) illustrates the first-order thermodynamical phase transition (FTPT), which corresponds to a change in crystal symmetry, as shown by the abrupt change in volume at SPT. Moreover, phonon dispersions in figure 2(c-d) at ambient and elevated pressure have verified the dynamic stability of the *NaCl-type* structure of GdSb within the range of SPT.

Accepted Manuscript

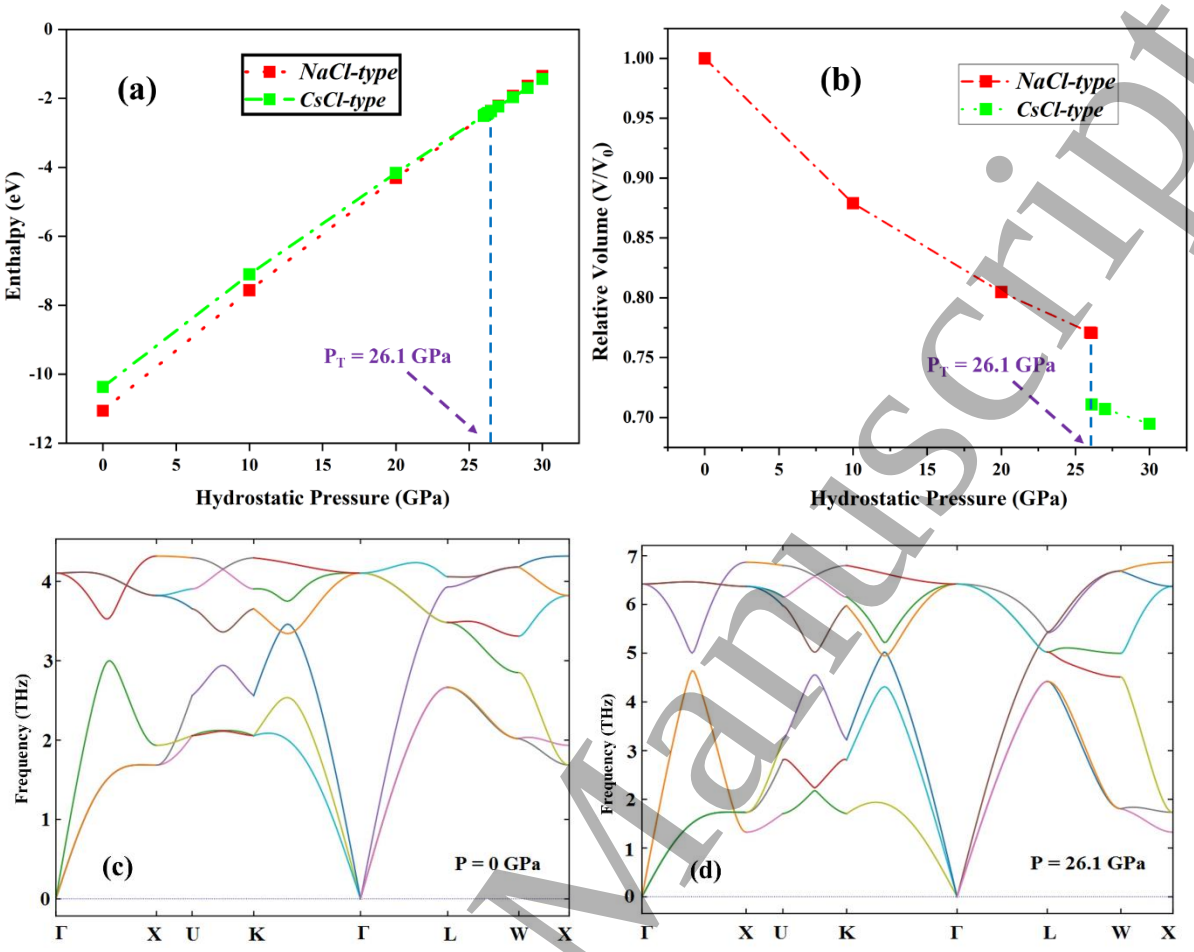


Figure 2. (a) The Enthalpy of GdSb as a function of pressure for *NaCl-type* and *CsCl-type* structures, (b) variation of the relative volume of GdSb as a function of pressure. The phonon band structures of GdSb at (c) 0 GPa, (d) 26.1 GPa.

After determining the stability of the GdSb, we have investigated the band structure of the material at ambient conditions using GGA-PBE and HSE06 functionals. The band structure calculation is done along the high symmetry k-path, i.e., *X* to *Γ*, *Γ* to *L*, *L* to *X*, and *X* to *W*. This chosen path contains the eight TRIM points *Γ, L, X* of the bulk *NaCl-type* structure of the material. The band structures of the material using GGA-PBE+SOC and HSE06+SOC are shown in figure 3(b-c).

Accepted Manuscript

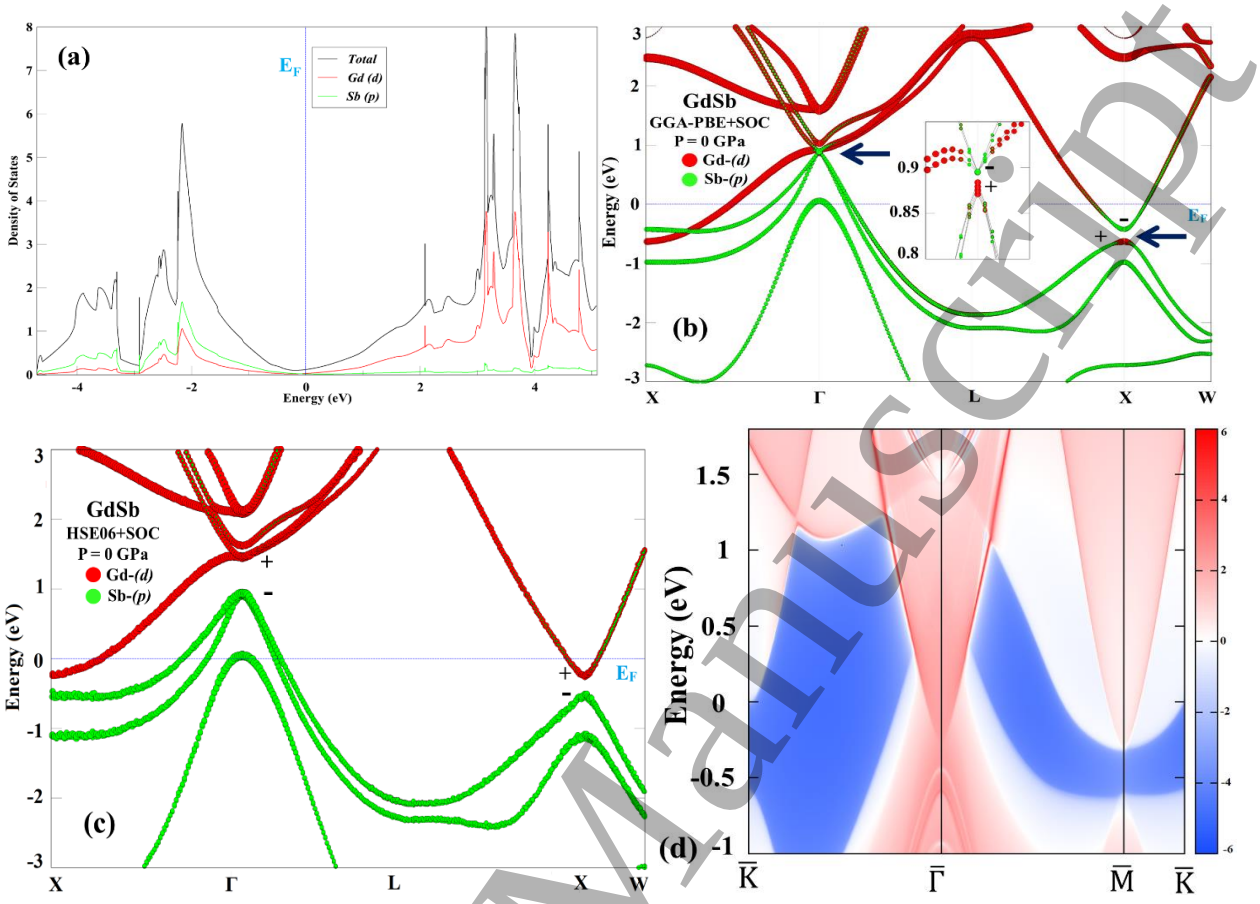


Figure 3. (a) The projected density of state (PDOS) of the material. The band structure of GdSb using (b) GGA-PBE+SOC (c) HSE06+SOC (d) The surface states of the material showing the absence of Dirac cone. The Fermi level is set to 0 eV.

The band structure, figure 3(b), using GGA-PBE+SOC functional shows an overlapping of the bands that occurred at *X*- and *Γ-points* (inset) at ambient conditions. It is observed that the GGA-PBE functional predicts two band inversions at *X*- and *Γ-points* at ambient conditions which is in contrast to the experimental observations [68]. Therefore, to determine the accurate bandstructure, we used HSE06 functional which predicts an energy gap at *X*- and *Γ-points* at ambient conditions as shown in figure 3(c). There is no band inversion observed in the bulk electronic band structure and the topologically trivial nature is identified which is in line with the experimental findings [56]. The valence band (VB) of the material has mainly constituted by the *p-orbital* of Sb and the conduction band (CB) contributed by the *d-orbital* of Gd. The small overlapping of the VB and CB at *Γ*- and *X-point* in figure 1(c) signifies the semimetallic nature of the material. This character of the material can also be confirmed by the projected density of states (PDOS) in figure 3(a).

Accepted Manuscript

To verify the TPT in the material, we have computed the Z_2 topological invariant using the product of parities at all TRIM points. In 3D systems, there are four Z_2 topological invariants (ν_0; $\nu_1\nu_2\nu_3$) which can be calculated according to the Kane and Mele model [1, 2] using,

$$(-1)^{\nu_0} = \prod_{n_j=0,1} \delta_{n_1 n_2 n_3} \quad (1)$$

$$(-1)^{\nu_{i=1,2,3}} = \prod_{n_i=1; n_{j\neq i}=0,1} \delta_{n_1 n_2 n_3} \quad (2)$$

Here, δ represents the parities of all the filled occupied bands at TRIM points. The value of ν_0 =1 corresponds to the topological non-trivial phase of the system and ν_0 = 0 signifies the weak or trivial nature of the system. If the any of remaining of them is 1 then the system corresponds to a weak TI.

Table 2. The parities at all TRIM points in BZ at various pressures.

Hydrostatic Pressure	No. of band inversion	3X	4L	Γ	Z_2 invariants
0 GPa	No inversion	-	+	-	(0;000)
6 GPa	One inversion	+	+	-	(1;000)
12 GPa	Two inversions	-	+	-	(0;000)

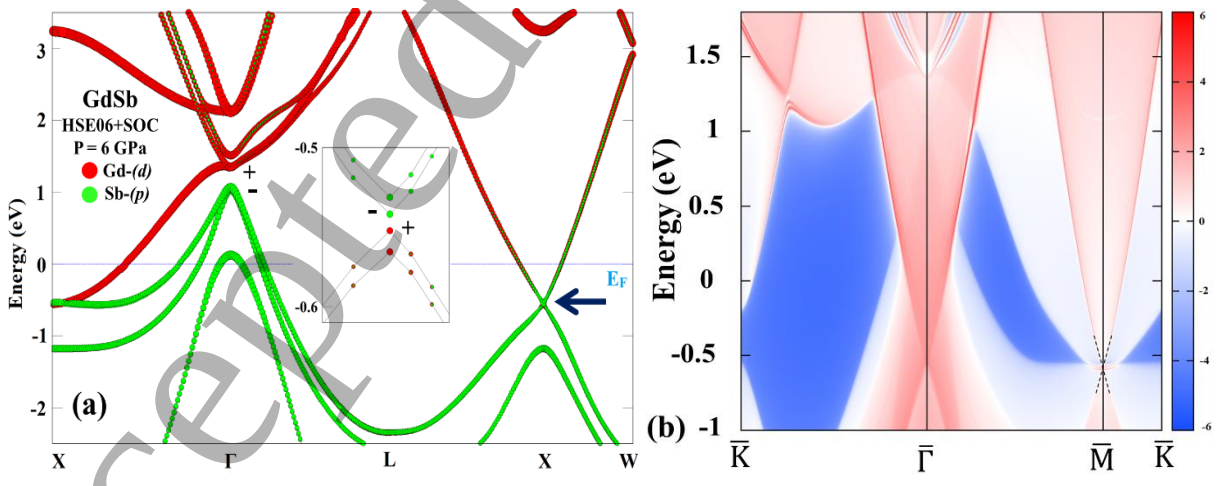
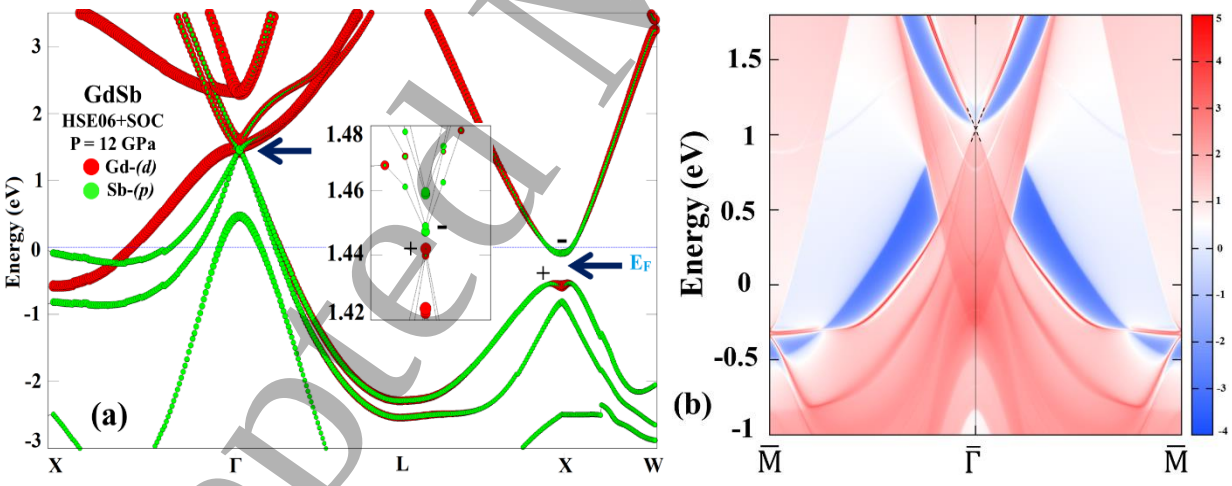


Figure 4. (a) The band structure of GdSb using HSE06+SOC functional at 6 GPa. The Fermi level is set to be 0 eV. (b) The surface states of the material show the presence of a single Dirac cone at $\bar{M}$-*point*.

Accepted Manuscript

The parities of the filled states at ambient pressure are given in table 2, where the value of ν_0 comes out to be 0 and confirms the topologically trivial state of the material. The SDOS calculated at ambient pressure, shown in figure 3(d), has no Dirac cone, which also validates the trivial nature of the material. Hence, we can say that we have anticipated the trivial nature of GdSb accurately at ambient conditions.

Next, we examined the electronic band structure of the material at elevated hydrostatic pressures. Up to 5.5 GPa, no band inversion is observed but the energy gap between the CB and VB reduces continuously. At 6 GPa, a band inversion at the *X-point* is observed as shown in figure 4(a). This band inversion can be confirmed with the inverted contribution of the *p-orbital* of Sb and the *d-orbital* of the Gd in CB and VB (inset of figure 4(a)), respectively. This band inversion results in switching the parity of the highest occupied band at the *X-point* and hence switches the value of the first Z_2 topological invariant from zero to one (see equation 1). The presence of a Dirac cone in SDOS is also observed along the (001) plane as shown in figure 3(b). This non-zero value of ν_0 and the existence of the Dirac cone in the surface states confirms our claim of the non-trivial topological phase of the material at 6 GPa.



Accepted Manuscript

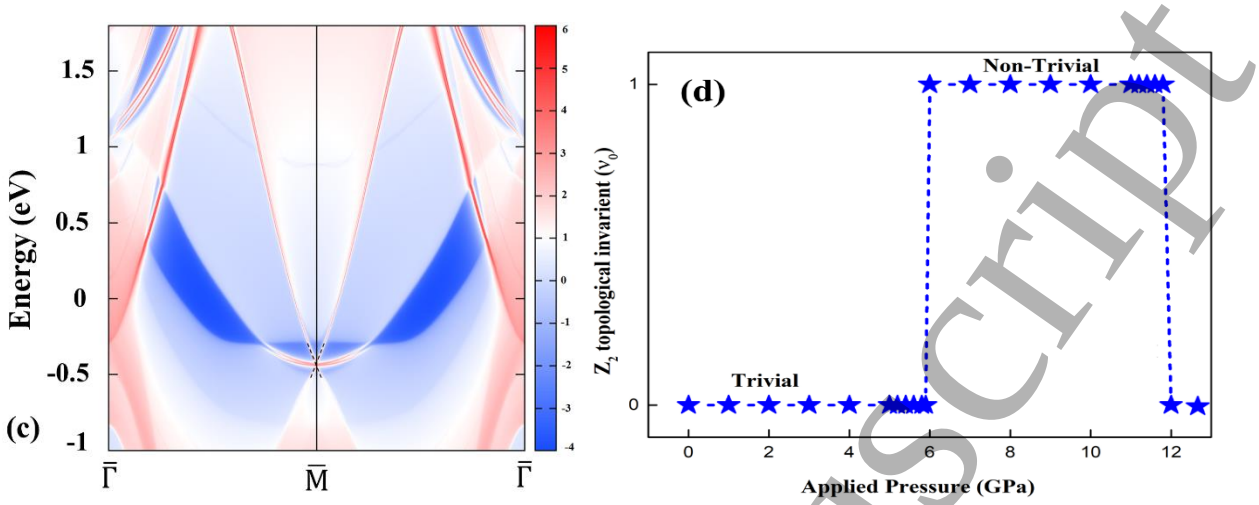


Figure 5. (a) The Band structure of GdSb using HSE06+SOC functional at 12 GPa. The Fermi level is set to be 0 eV. The surface states correspond to k-path (b) $\bar{M} - \bar{\Gamma} - \bar{M}$ and (c) $\bar{\Gamma} - \bar{M} - \bar{\Gamma}$ for (001) plane. (d) The variation of the first Z_2 topological index (ν_0) as a function of pressure.

The system remains in a non-trivial state till the hydrostatic pressure reaches up to 11.5 GPa. The further increase in pressure to 12 GPa, results in another band inversion occurring at *Γ-points* as shown in figure 4(a). The parity is switched to the corresponding highest occupied band at *X*- and *Γ*-points, which makes the first Z_2 topological invariant (ν_0) to zero as shown in table 2. Now, the bulk band structure of GdSb contains an even number (two) of band inversions that correspond to either a weak or trivial TM. This is further investigated by calculating the remaining three Z_2 topological invariants. The values of ($\nu_1\nu_2\nu_3$) are found to be (000) as per calculation using equation 2, which confirms that the system returns to its trivial state again at 12 GPa. The two Dirac cones are observed at the $\bar{\Gamma}$ and $\bar{M}$ corresponding to k-path along $\bar{M} - \bar{\Gamma} - \bar{M}$ and $\bar{\Gamma} - \bar{M} - \bar{\Gamma}$ of (001) plane as shown in figure 5(b-c). The variation of the first Z_2 topological invariant with the applied hydrostatic pressure is shown in figure 5(d). Hence, the rare earth monopnictide GdSb shows a topological non-trivial phase for the hydrostatic pressure range of 6 GPa to 12 GPa.

4. Conclusion

We have systematically investigated the structural stability, electronic, and topological properties of rare earth monopnictide GdSb utilizing the first-principles calculations with hybrid functional (HSE06) at ambient and elevated pressures. The SPT, from *NaCl-type* to *CsCl-type* structure, of GdSb has been found at 26.1 GPa and this transition is first order thermodynamical phase

Accepted Manuscript

transition. The phonon dispersion calculations show the dynamical stability of the material for the *NaCl-type* structure at ambient to structural phase transition pressure. At ambient conditions, this material shows a topological trivial nature and matches with the experimental results. The topological phase transition has been observed at 6 GPa of applied pressure with an inverted bands between *p-orbital* of Sb and *d-orbital* of Gd at *X-point*. The non-zero value of the first Z_2 topological invariant ($\nu_0;\ \nu_1\nu_2\nu_3$:1;000) and the presence of the Dirac cone in surface density of states confirmed our claim of a non-trivial nature. Further, increase in pressure to 12 GPa has shown one more band inversion which switched the parity of the highest occupied band at Γ-point. This has converted the Z_2 topological invariant ($\nu_0;\ \nu_1\nu_2\nu_3$) to (0000) again making it a weak topological at 12 GPa. A pair of Dirac cones are found at the $\bar{\Gamma}$ and $\bar{M}$*-points* corresponding to k-path along $\bar{M}-\bar{\Gamma}-\bar{M}$ and $\bar{\Gamma}-\bar{M}-\bar{\Gamma}$ of (001) plane. In conclusion, we have observed that the GdSb shows a topological non-trivial phase between 6-12 GPa of applied pressure.

ACKNOWLEDGEMENTS

One of the authors (Ramesh Kumar) would like to thank the Council of Scientific and Industrial Research (CSIR), Delhi, for financial support. The authors also acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM SEVA′ at IIT, Hyderabad, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kane C L and Mele E J 2005 *Phys. Rev. Lett.* **95** 146802.
2. Fu L, Kane C L and Mele E J 2007 *Phys. Rev. Lett.* **98** 106803.
3. Hasan M Z and Kane C L 2010 *Rev. Mod. Phys.* **82(4)** 3045.
4. Moore J E 2010 *Nature* **464** 194.
5. Yang, S A. 2016 *SPIN* **6** 1640003.

Accepted Manuscript

6. Nayak C, Simon S H, Stern A, Freedman M and Das Sarma S 2008 *Rev. Mod. Phys.* **80(3)** 1083-1159.
7. Sangeeta, Kumar R and Singh M 2022 *J. Mater. Sci.* **57(23)** 10691-10701.
8. Parameswaran S A, and Wan Y 2017 *Phys.* **10** 132.
9. Hasan M Z and Moore J E 2011 *Annu. Rev. Condens. Matter Phys.* **2(1)** 55-78.
10. Hasan M Z, Xu S-Y, Neupane M, Roche M S and Valenzuela S O 2015 Topological Insulators: Fundamentals and Perspectives ed F Ortmann, S Roche and S O Valenzuela (Wiley)
11. Qi X L and Zhang S C 2011 *Phys. Rev. Lett.* **83(4)** 1057.
12. Burkov A A 2016 *Nat. Mater.* **15(11)** 1145-1148.
13. Kong X, Li L and Peeters F M 2018 *Appl. Phys. Lett.* **112**(25).
14. Singh B, Ghosh B, Su C, Lin H, Agarwal A and Bansil A 2018 *Phys. Phys. Lett.* **121(22)** 226401.
15. Politano A, Chiarello G, Ghosh B, Sadhukhan K, Kuo C N, Lue C S and Agarwal A 2018 *Phys. Phys. Lett.* **121(8)** 086804.
16. Huang X, Zhao L, Long Y, Wang P, Chen D, Yang Z, and Chen G 2015 *Phys. Rev. X* **5(3)** 031023.
17. Vazifeh M M and Franz M 2013 *Phys. Phys. Lett.* **111(2)** 027201.
18. Wan X, Turner A M, Vishwanath A and Savrasov S Y 2011 *Phys. Rev. B* **83(20)** 205101.
19. Weng H, Fang C, Fang Z, Bernevig B A and Dai X 2015 *Phys. Rev. X*, **5(1)** 011029.
20. Liu Z K, Yang L X, Sun Y, Zhang T, Peng H, Yang H F and Chen Y L 2016 *Nat. Mater* **15(1)** 27-3.
21. Liu, Z. K., Zhou, B., Zhang Y, Wang Z J, Weng H M, Prabhakaran D and Chen Y L 2014 *Science*, **343(6173)** 864-867.
22. Xu S Y, Liu C, Kushwaha S K, Sankar R, Krizan J W, Belopolski I and Hasan M Z 2015 *Science*, **347(6219)** 294-298.
23. Xu S Y, Belopolski I, Alidoust N, Neupane M, Bian G, Zhang C, Sankar R, Chang G, Yuan Z , Lee C C, 2015 *Science*. **349** 613.
24. Yang L X, Liu Z K, Sun Y, Peng H, Yang H F, Zhang T, and Chen Y L 2015 *Nat. phys.* **11(9)** 728-732.
25. Guo P J, Yang H C, Liu K and Lu Z Y 2017 *Phys. Rev. B* **96(8)** 081112.

Accepted Manuscript

26. Khalid S, Sabino F P and Janotti A 2018 *Phys. Rev B* **98(22)** 220102.
27. Lou R, Fu B B, Xu Q N, Guo P J, Kong L Y, Zeng L K, and Wang S C 2017 *Phys. Rev B* **95(11)** 115140.
28. Duan X, Wu F, Chen J, Zhang P, Liu Y, Yuan H and Cao C 2018 Commun. phys. **1**, 71.
29. Nayak J, Wu S C, Kumar N, Shekhar C, Singh S, Fink J, and Felser C 2017 *Nat. Commun.* **8(1)** 13942.
30. Liu Z K, Yang L X, Wu S C, Shekhar C, Jiang J, Yang H F and Chen Y L 2016 Nat. *Commun.* **7(1)** 12924.
31. Zeng M, Fang C, Chang G, Chen Y A, Hsieh T, Bansil A, and Fu L 2015 *arXiv preprint arXiv* **1504** 03492.
32. Zide J M O, Kleiman-Shwarsctein A, Strandwitz N C, Zimmerman J D, Steenblock-Smith T, Gossard A C, and Stucky G D 2006 *Appl. Phys. Lett.* **88** (16).
33. Hanson M P, Gossard A C and Brown E R *Appl. Phys. Lett.* **89** (11).
34. Bjarnason J E, Chan T L J, Lee A W M, Brown E R, Driscoll D C, Hanson M and Muller R E 2004 *Appl. Phys. Lett.* **85(18)** 3983-3985.
35. Guo C, Cao C, Smidman M, Wu F, Zhang Y, Steglich F and Yuan H 2017 *npj quantum mater.* **2(1)** 39.
36. Alidoust N, Alexandradinata A, Xu S Y, Belopolski I, Kushwaha S K, Zeng M and Hasan M Z 2016 *arXiv preprint arXiv:***1604** 08571.
37. Kuroda K, Ochi M, Suzuki H S, Hirayama M, Nakayama M, Noguchi R and Kondo T 2018 *PRL* **120(8)** 086402.
38. Li P, Wu Z, Wu F, Cao C, Guo C, Wu Y and Liu Y 2018 *Phys. Rev. B* **98(8)** 085103.
39. Duan X, Wu F, Chen J, Zhang P, Liu Y, Yuan H & Cao C 2018 Commun. Phys **1** 71.
40. Pal K and Waghmare U V *Appl. Phys. Lett.* **105(6).**
41. Qi Y, Shi W, Naumov P G, Kumar N, Sankar R, Schnelle W and Medvedev S A 2017 *Adv. Mater.* **29(18)** 1605965.
42. Rajaji V, Manjón F J and Narayana C 2022 *J. Condens. Matter Phys.* **34(42)** 423001.
43. Nayak J, Wu S-C, Kumar N, Shekhar C, Singh S, Fink J, 2017 Nat. Commun. **8** 13942.
44. Wadhwa P, Kumar R, Kumar S and Shukla A 2018 *arXiv preprint arXiv:* **1806** 0059.
45. Kumar R and Singh M 2024 *J. Phys.: Condens. Matter* **36** 345601

Accepted Manuscript

46. Wadhwa P, Kumar S, Shukla A and Kumar R 2019 *J. Condens. Matter Phys.* **31(33)** 335401.
47. Singh M, Kumar R and Bibiyan R K 2022 *Eur. Phys. J. Plus* **137(5)** 633.
48. Song J J, Tang F, Zhou W, Fang Y, Yu H L, Han Z D, Qian B, Jiang X F, Wang D H, Du Y W 2018 J. Mater. Chem. C, **6** 3026-3033.
49. Inbar H S, Ho D Q, Chatterjee S, Pendharkar M, Engel A N, Dong J T, Khalid S, Chang Y H, Guo T, Fedorov A V, Lu D, Hashimoto M, Read D, Janotti A and Palmstrøm C J 2022 *Phys. Rev. Mater.* **6** L121201.
50. Inbar H S, Ho D Q, Chatterjee S, Engel A N, Khalid S, Dempsey C P and Palmstrøm C J 2023 *APL Mater.* **11** 111106**.**
51. Hohenberg P, Konn W 1964 *Phys. Rev.***136(3B)** 864.
52. Kohn W and Sham L J *Phys. Rev.***140(4A)** 1133.
53. Kresse G and Joubert D 1999 *Phys. Rev. B* **59(3)**1758.
54. Kresse G and Furthmüller J 1996 *Phys. Rev. B* **54(16)** 11169.
55. Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. B* **77(18)** 3865.
56. Heyd J, Scuseria G E and Ernzerhof M 2003 J Chem Phys **118** 8207–8215.
57. Heyd J, Scuseria G E and Ernzerhof M 2003 J Chem Phys **124** 219906
58. Togo A and Tanaka I 2015 Scripta Materialia **108** 1-5.
59. Mostofi A A, Yates J R, Lee Y S, Souza I, Vanderbilt D and Marzari N 2008 *Comput. Phys. Commun.* **178(9)** 685-699.
60. Wu Q, Zhang S, Song H F, Troyer M and Soluyanov A A 2018 *Comput. Phys. Commun* **224** 405-416.
61. Pagare G, Soni P, Srivastava V and Sanyal S P 2009 *J. Phys. Chem. Solids*. **70(3-4)** 650-654.
62. Li D X, Haga Y, Shida H, Suzuki T and Kwon Y S 1996 1996 *Phys. Rev. B* **54(15)** 10483.
63. Li D X, Haga Y, Shida H, Suzuki T, Kwon Y S and Kido G 1997 *J. Condens. Matter Phys.* **9(48),** 10777.
64. Larson P and Lambrecht W R 2006 *Phys. Rev. B* **74(8)** 085108.
65. Duan C G, Sabiryanov R F, Mei W N, Dowben P A, Jaswal S S and Tsymbal E Y 2007 *J. Phys.: Condens. Matter* **19** 315220.

Accepted Manuscript

66. Korozlu N, Colakoglu K, Deligoz E and Surucu-Khan G O 2010 *Philosophical Magazine*, **90(14)** 1833-1852.
67. Shekar N V and Sahu P C 2006 *J. Mater. Sci.* **41** 3207-3228.
68. Crowley J M, Tahir-Kheli J and Goddard W A 2015 *J. Phys. Chem. Lett.* **6(19)** 3792-3796.

Accepted Manuscript

2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) | 979-8-3503-9156-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSTSN61422.2024.10670894

Enhancing Image Translation with Conditional Generative Adversarial Networks

Shobhit Verma
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
shobhit.verma2001@gmail.com

Srishti Arora
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
srishti.arora1909@gmail.com

Tanvi Kayat
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
tkayat03@gmail.com

Anurag Goel
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
anurag@dtu.ac.in

***Abstract*—This article examines the advantages of engaging in conflictual discourse as a means to address and transform image-related issues in a revolutionary manner. Unlike earlier approaches that necessitated human customization for particular applications, Conditional Generative Adversarial Networks (cGANs) automatically produce image outputs and negative impacts. Picture painting, object reconstruction from edge maps, and picture combining from a label map are just a few activities that may be executed using a unified framework. The notion has been tested and proven to work in various ways it doesn't need to be compared to other standards in any way to be implemented or embraced, many people on the internet, including artists, have played around with Pix2pix and its related apps since it was first released, making it more flexible and easier to use. As a major step forward in traditional image interpretation, this study eliminates the need for human map generation. So, there's no requirement for overlapping sessions devoted to particular objectives the community can get satisfactory answers for various image interpretation issues in this way.**

***Index Terms*—Image Translation, Generative Adversarial Networks, Contrastive Learning, Patch-Based Framework, Cross-Domain Similarity**

I. Introduction

A source image into an outcome image can be applied to many issues in the domains of processing images, computer graphics, and computer vision. The depiction of a situation can be achieved in either the English or French language. The picture can be represented as an RGB image, a gradient field, an edge map, a semantic label map, or other comparable representations. Computerized image-to-image translation refers to the process, of converting one visual depiction of an instance into a different one, analogous to how automatic language translation converts one language into another. A few samples of image translation are depicted in Figures 1. To complete this work, a sufficient amount of training data is necessary. Historically, each of these jobs has been addressed using distinct, purpose-built machinery, even if the context remains consistent: predicting pixels based on other pixels. The objective of this study is to establish a unified framework that can be used for all of these issues [1]. Significant advancements have been made by the community in this area because of the widespread use of Convolutional neural network (CNN) as the prime method for tackling various image prediction challenges. CNNs are trained with

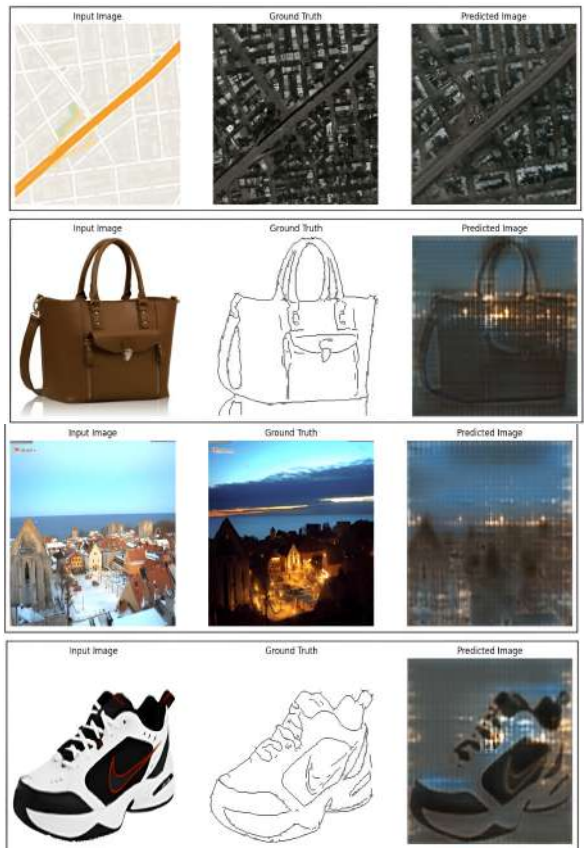


Fig. 1: Examples of Image Translation

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:41:12 UTC from IEEE Xplore. Restrictions apply.

the objective of minimizing a loss function, which serves as a metric for evaluating the quality of the outputs. While the process of learning may be automated, the creation of effective loss functions still necessitates a considerable amount of manual labor. In essence, we must communicate our desires to CNN regarding the specific aspects we wish for them to minimize. Less sharp images will probably be produced if the CNN is trained using a simple algorithm that reduces the Euclidean distance between predicted and actual pixels [2].

The reason behind this is that the process of averaging all possible outputs minimizes the Euclidean distance, resulting in a blurred effect. Developing loss functions that effectively guide the CNN to produce desired outcomes, such as generating clear and realistic images, remains a challenging task that typically necessitates expertise in the field. It would be preferable if we could provide a general goal, such as achieving output that is indistinguishable from actuality, and automatically derive a suitable loss function to achieve this purpose.

Thankfully, the recently suggested GANs achieve just this objective. GANs seek to train a generative model by minimizing a loss function that differentiates between authentic and counterfeit output images. Images that are not clear will not be accepted, as they appear artificial. Due to the adaptive nature of GANs in learning loss, they can be utilized for various applications that often necessitate diverse types of loss functions [3].

This paper explores the utilization of GANs in conditional scenarios. Conditional GANs are designed with the explicit purpose of generating a model of data that includes specific criteria. Conditional Generative Adversarial Networks (GANs) are well-suited for addressing challenges associated with image translation, wherein a source picture is used to generate an outcome image [4].

Prior research has only concentrated on certain uses, resulting in uncertainty over the general effectiveness of image-conditional GANs as a versatile approach for translating images into images. Our main accomplishment is demonstrating that conditional GANs produce satisfactory results across a wide range of challenges. Our next step is to propose a straightforward and efficient methodology that may generate desirable outcomes and evaluate the impacts of significant architectural decisions.

II. RELATED WORK

Picture-to-image translation challenges often involve tasks that need per-pixel regression or classification, which are constructed using structured loss-based image modeling [5]. These formulas consider the resultant space as unorganized because they assume that, based on the source image, each resultant pixel is conditionally detached from every other pixel [6]. Conditional GANs, otherwise, acquire knowledge of a well-organized loss function [7].

Structured losses impose a penalty on the combined arrangement of the output. Extensive research has been conducted on these types of losses, employing several methodologies including non-parametric losses, convolutional pseudo-prior, feature matching, conditional random fields the convolutional pseudo-prior, the SSIM metric, and a loss function based on corresponding correlation statistics. The key characteristic of the conditional GAN is its ability to train the loss function, which can penalize any discrepancies in structure between the output and target, in theory.

Conditional GANs have been previously utilized in the conditional context, indicating that we are not the pioneers in this application. Previous and simultaneous studies have trained GANs using discrete labels and pictures. The image-conditional models have addressed various tasks such as predicting images from a normal map, predicting future frames, generating product photos, and generating images from sparse annotations. Other studies have utilized GANs to do image-to-image mappings. However, they applied the GAN unconditionally and incorporated additional techniques, such as L2 regression, to establish a dependency between the output and input. These efforts have achieved remarkable outcomes in in-painting, future state prediction, user-constrained picture change, style transfer, and super-resolution.

For a particular purpose, each technique was designed. Unlike other frameworks, ours has no application-specific components. Our setup is significantly simpler than the majority of others. Several architectural decisions made by our method for the generator and discriminator set it apart from previous works as well [8].

The discriminator we employ utilizes a convolutional PatchGAN classifier that particularly imposes penalties on the arrangement of image patches. In contrast to earlier research, our generator employs a U-Net architecture [9]. Previous work introduced a PatchGAN architecture to accurately collect local style statistics. This study showcases the effectiveness of this strategy across a wider range of problems and investigates the consequences of modifying the patch's size.

Improving image-to-image translation through the use of pretraining approaches is the new focus of this paper. To manage different image translation tasks, the suggested framework adjusts a pretrained diffusion model, as opposed to conventional techniques that necessitate specific architectures and initial training. Framework improvements to texture synthesis and generation quality are achieved through the use of adversarial training and normalized guided sampling. In a variety of tasks, such as mask-to-image, sketch-to-image, and geometry-to-image translation, the suggested pretraining-based image-to-image translation (PITI) has been shown to effectively synthesize high-quality and realistic images through extensive empirical evaluations on challenging benchmarks [10].

Traditional image-to-image translation approaches have problems producing complicated scenes of sufficient quality and often necessitate specialized structures. Learning large-scale domain mappings and expressing tiny feature changes are two areas where StarGAN falls short, but it does allow multi-domain translation with a single generator. To solve these problems and achieve far better performance on large-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:41:12 UTC from IEEE Xplore. Restrictions apply.

scale domains, SuperstarGAN [11] employs an independent classifier with data augmentation.

III. Methodology Used

GANs are the models that understand and learn to map a random noise vector, denoted as z, into a visually perceptible final image, represented as y. The mapping is represented by the function G in the form of equation 1.

$$G : z \rightarrow y \tag{1}$$

Alternatively, cGANs acquire the ability to create a connection between observed data points [12].

The input picture (x) and random noise vector (z) are mapped to the output (y) via function G. Simply put, the given equation shows that G is a function that, when applied to the inputs x and z, yields the output y as referred in equation 1.

To correctly identify the generator's synthetic outputs, a discriminator (D) is trained via adversarial training. The objective of this training is to enable the generator (G) to produce outputs that closely resemble authentic photographs.

A. Objectives

1) Leverage the Potential of Conditional GANs Highlight how conditional GANs can outperform unpaired methods in image translation and achieve remarkable efficiency.
2) Employ Paired Data for Improved Evaluations Use paired data to keep an eye on the training process this will ensure that the input and output domains are more closely aligned, leading to better picture translations [13].
3) Attain cutting-edge performance Strive to develop a versatile image translation model that outperforms existing approaches in areas including style transfer, image editing, and semantic segmentation.
4) Facilitate Practical Applications Implementations Bridge the gap between theory and practice by providing a trustworthy cGAN-based image translation solution, applicable to a wide range of real-world contexts requiring image editing, content production, and visual content manipulation [14].

B. Applications

1) Semantic Labels - Photo: This application entails the translation of semantic labels into actual photographs and vice versa. It closely corresponds to the paper's focus on conditional adversarial networks and their capacity to acquire mappings between diverse visual representations. Both the forward and backward processes of evaluating the network's ability to generate realistic images using semantic labeling are made possible by it [15].
2) Map - Aerial Photo: This application deals with a common task in image translation, which is the conversion of maps and aerial photos, and the incorporation of it in the paper enhances its analysis of picture translations and may provide useful information regarding the network's effectiveness when dealing with geographical data.
3) Edges - Photo: An important part of image synthesis, this program primarily aims to create real images from edge maps. That way, we can see how well the network performs when given sparse input data and see if it can fill in the gaps to produce realistic textures.
4) Day - Night: The primary goal of this application is to convert images from daytime to nighttime settings, which presents its unique challenges about lighting and color distribution. Here is a great chance to observe the network's performance under demanding conditions, such as complex modifications and visually appealing output.

IV. IMPLEMENTATION

A. Data Preparation and Pre-processing

To train and assess the proposed model, various data sets are used to address different visual translation issues, depending on the task at hand. A well-known benchmark for image task translation is the Cityscapes dataset [16**?**].

TABLE I: FCN scores - Cityscape label photos

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.46	0.13	0.15
GAN	0.24	0.07	0.03
cGAN	0.58	0.24	0.17
L1+GAN	0.65	0.26	0.14
L1+cGAN	0.69	0.23	0.18
Ground truth	0.80	0.26	0.21

One thing that makes this dataset special is the abundance of cityscape annotations it contains. Many algorithms are tested using this benchmark to see how well they can convert semantic labels into visual representations. We measure these findings by employing the FCN-score on cityscape labels-photos, as shown in Table I. The process of turning maps into aerial photographs can also benefit from data retrieved from Google Maps and other sources. To make sure the data is compatible with the model's design and to get the most out of training, pre-processing activities on the data before feeding it into the model are crucial. Common preparatory tasks for ensuring reliable model evaluation include standardizing image sizes, normalizing the values of pixels to a consistent scale, and partitioning the dataset into testing, training, and validation subsets [16].

B. Model Architecture Design

1) Designing the Generator Network: We need to design the generator network that will take in input data and turn it into aesthetically appealing images, this network is essential to the image translation process. The famous U-Net [9] architecture served as an inspiration for the generator's careful construction which focuses on reaching a balance between efficiency and efficacy. Connected by skip connections, which aid in maintaining spatial information across the network, the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:41:12 UTC from IEEE Xplore. Restrictions apply.

U-Net design comprises a decoder and encoder components where input data is retrieved by the encoder in the form of abstract features, and the decoder uses these features to rebuild the output image. To enhance the accuracy of the generated images, bypass connections facilitate the transmission of low-level characteristics among the encoder and decoder layers [17]. Table II compares the U-Net against an encoder-decoder on cityscape generation.

TABLE II: FCN-scores (different generator architectures)-cityscapes label photos

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Encoder-Decoder(L1)	0.34	0.17	0.06
Encoder-Decoder(L1+cGAN)	0.24	0.06	0.08
U-net(L1)	0.44	0.16	0.19
U-net(L1+GAN)	0.53	0.24	0.15

2) Building the Discriminator Network: Working alongside the generator network, the discriminator network provides feedback on the generated image's validity and ways to enhance the synthesis process. To obtain a detailed and precise distinction between genuine and fake photographs, a Patch-GAN discriminator is employed to assess the legitimacy of the pictures at a local level. The discriminator has learned to classify input photos as real or fake and then provide the generator with critical comments to improve the generated images. Through repeated training, the discriminator and generator networks enhance the performance of picture translation.

C. Loss Function Formulation and Optimization

Defining the Generator Loss Function: The generator loss function is crucial in directing the optimization process during training. The normal composition of this is comprised of two essential elements: L1 loss and conditional GAN loss [18]. The L1 loss, also known as mean absolute error, ensures that the generated and targeted pictures are identical at the pixel level. This encourages the generator to produce images that closely resemble the actual data. Conversely, the conditional GAN loss, which is expressed as binary cross-entropy, encourages the production of lifelike images by teaching the generator to deceive the discriminator in an adversarial manner [19].

Formulating the Discriminator Loss Function: The discriminator loss function is designed to optimize the discriminator's capacity to differentiate genuine and counterfeit images, in combination with the generator loss function. The loss function, commonly denoted as binary cross-entropy loss, quantifies the discrepancy between the discriminator's forecasts and the true labels. The discriminator aims to accurately distinguish actual images as real and created images as fake, hence providing essential insight to the generator to enhance image synthesizing [20].

D. Training and Evaluation

The discriminator and generator networks are optimized iteratively using the adversarial training process. Each iteration of the process entails the discriminator verifying the genuineness of both genuine and artificial images, while the generator generates synthetic images based on input data. The generator's objective is to create increasingly realistic images to fool the discriminator, whose objective is to correctly distinguish between actual and fake images. The adversarial feedback loop boosts image translation accuracy by fortifying both networks.



Fig. 2: Results: Shoes

E. Extensive Model Assessment

After training, the model is evaluated extensively on a separate test dataset using a wide range of qualitative as well as quantitative metrics [6]. Images are assessed using a variety of quantitative metrics, such as FCN scores, perceptual similarity indices, and research on human perception. We also restore the most recent saved checkpoint for evaluation to further confirm the model's effectiveness in real-world image translation jobs. The performance Performance of photo - labels on cityscapes is represented in Table III. A thorough and careful review can help you understand the model's strengths and shortcomings. Hopefully, more research on image-to-image translation will expand upon this data.

TABLE III: Performance of photo - labels on cityscapes

Loss	Per-pixel acc.	Per-class acc.	Class IOU
L1	0.84	0.47	0.33
cGAN	0.72	0.22	0.25
L1+cGAN	0.87	0.33	0.24

V. Results

Conditional GANs have shown effective in cross-domain image translation:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:41:12 UTC from IEEE Xplore. Restrictions apply.



Fig. 3: Results: Bag

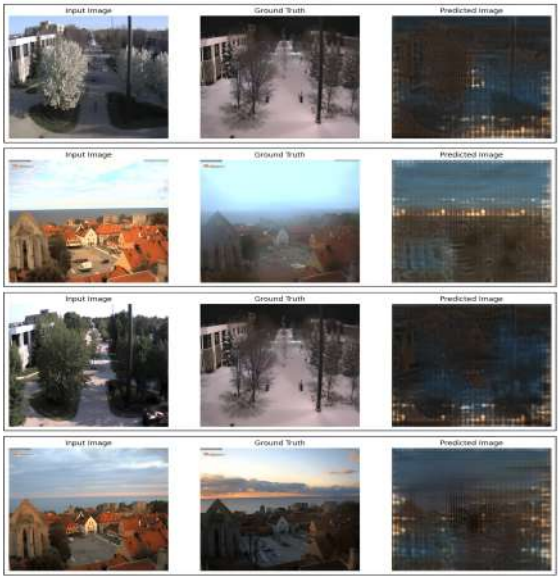


Fig. 4: Results: Night and Day

A. Semantic Labels - Photo

Conditional GANs achieved exceptional performance in the translation between real pictures and semantic labels [21] as shown in Figures 2, 3, 4. The remarkable ability of these networks to produce realistic images from abstract representations showcases their proficiency in understanding and accurately replicating complex visual situations. These observations can be reassured by this Table IV.

B. Map - Aerial Photo

using Conditional GANs, the process of converting aerial photos to maps is significantly enhanced as shown in Figure 5. Reliable map-making from aerial photos and inverse map-making are both made possible by these networks' proficiency



Fig. 5: Results: Map

TABLE IV: FCN scores (different receptive field sizes for discriminator)- Cityscapes label- photo

Discriminator receptive field	Per-pixel acc.	Per-class acc.	Class IOU
1 x 1	0.34	0.17	0.17
16 x 16	0.62	0.25	0.15
70 x 70	0.63	0.27	0.13
286 x 286	0.41	0.13	0.15

in complex mappings across several geospatial data domains. This capability opens up a lot of new options, such as accurate map production and detailed analysis of geographical data.

C. Edges - Photo

Drawing convincing pictures from edge maps is one of the most basic problems with image synthesis. But Conditional GANs show their value by completing the job, filling in missing features, and making aesthetically pleasing textures from sparse input data. In this case, Conditional GANs [21] prove their worth by reliably creating new images with less data [22].

Conditional GANs can adapt to various image-to-image translation challenges, and the findings demonstrate that they generally operate effectively [23]. Because of it's exceptional ability to learn complicated mappings between various visual representations and produce high-quality images. Conditional GANs find extensive use in many fields, including semantic segmentation, geospatial data analysis, and image synthesis. These results show that Conditional GANs [21] can revolutionize many real-world applications and open up new avenues for research and development in areas like computer vision, spatial analysis, and more.

As compared to more conventional approaches, such as bicubic interpolation or elementary image processing algorithms, Conditional Generative Adversarial Networks (cGANs) provide substantial benefits in improving picture

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:41:12 UTC from IEEE Xplore. Restrictions apply.

translation. cGANs are effective in creating realistic, high-quality images with the help of adversarial training, and cGANs provide frameworks that are flexible and adaptable enough to operate with various types of datasets. More smooth integration into bigger systems is the result of their ability to provide end-to-end learning without depending on customized features. cGANs are a formidable tool for image translation tasks due to their exceptional performance in collecting complex mappings and generating high-fidelity images.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated that our method is effective in multiple areas, including semantic labels picture translation, maps-aerial photo transformation, and edges photo-synthesis. This demonstrates that Conditional GANs may produce realistic and visually appealing images from various input sources, highlighting their potential for practical use because of their versatility, scalability, and effectiveness in different fields. In the future, we plan to extend the model's capabilities to handle multi-modal translation tasks, including translating between several domains at the same time or producing many outputs from a single input. This could greatly enhance its versatility and usefulness.

REFERENCES

[1] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *ICCV*, 2015.

[2] Philip Isola Richard Zhang and Alexei A. Efros. Colorful image colorization. *ECCV*, 2016.

[3] Arthur Szlam Emily Denton, Soumith Chintala and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *NIPS*, 2015.

[4] Li Fei-Fei Justin Johnson, Alexandre Alahi. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016.

[5] Jan Kautz Ming-Yu Liu, Thomas Breuel. Unsupervised image-to-image translation networks. *NIPS*, 2017.

[6] Sunggyun Park Anthony S. Paek In So Kweon Donggeun Yoo, Namil Kim. Pixel-level domain transfer. *ECCV*, 2014.

[7] Evan Shelhamer Jonathan Long and Trevor Darrell. Fully convolutional networks for semantic segmentation. *Conference on Computer Vision and Pattern Recognition*, 2015.

[8] Soumith Chintala Alec Radford, Luke Metz. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[9] Phillipp Fischer Olaf Ronneberger and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[10] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.

[11] Kanghyeok Ko, Taesun Yeom, and Minhyeok Lee. Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains. *Neural Networks*, 162:330–339, 2023.

[12] M. Mirza B. Xu Ian Goodfellow, J. Pouget-Abadie. Generative adversarial networks. *NIPS*, 2014.

[13] Yann LeCun Michael Mathieu, Camille Couprie. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016.

[14] Jun-Yan Zhu Andrew Tao Jan Kautz' Bryan Catanzaro Ting-ChunWang, Ming-Yu Liu. High-resolution image synthesis and semantic manipulation with conditional gans. *CVPR*, 2018.

[15] Eli Shechtman Alexei A. Efros Jun-Yan Zhu, Philipp Krähenbühl. Generative visual manipulation on the natural image manifold. *ECCV*, 2016.

[16] Hugo Larochelle Ole Winther Anders Boesen Lindbo Larsen, Søren Kaae Sønderby. Autoencoding beyond pixels using a learned similarity metric. *ICML*, 2016.

[17] Jeff Donahue Trevor Darrell Alexei A. Efros Deepak Pathak, Philipp Krahenbuhl. Context encoders: Feature learning by inpainting. *CVPR*, 2016.

[18] Timo Aila Tero Karras, Samuli Laine. A style-based generator architecture for generative adversarial networks. *CVPR*, 2019.

[19] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 33(4), 2014.

[20] Varun Agrawal Amit Raj Jingwan Lu Chen Fang Fisher Yu James Hays Wenqi Xian, Patsorn Sangkloy. Texturegan: Controlling deep image synthesis with texture patches. *CVPR*, 2018.

[21] Simon Osindero Mehdi Mirza. Conditional generative adversarial nets. *CVPR*, 2014.

[22] Samuli Laine Jaakko Lehtinen Tero Karras, Timo Aila. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018.

[23] Toward Multimodal Image to Image Translation. Jun-yan zhu, richard zhang, deepak pathak, trevor darrell, alexei a. efros, oliver wang, eli shechtman. *CVPR*, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 10:41:12 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20246-3

Evolution of transformer-based optical flow estimation techniques: a survey

Nihal Kumar[1] · **Om Prakash Verma**[1] · **Anil Singh Parihar**[1]

Received: 6 April 2024 / Revised: 9 August 2024 / Accepted: 7 September 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Optical flow plays a crucial role in computer vision, and with the rapid advancements of Transformer-based models in natural language processing (NLP) and various visual tasks, Transformer architectures are becoming increasingly prevalent in the field of optical flow. In recent years, Transformer-based optical flow techniques have demonstrated superior accuracy compared to conventional CNN-based models. However, despite the existence of survey work on traditional and CNN-based methods for flow estimation, there is a lack of research focused exclusively on the emerging trend of Transformer-based approaches. This gap in the literature hinders the understanding and further development of these state-of-the-art techniques. To address this issue, we provide a systematic review and assessment of optical flow techniques with a sole focus on Transformer-based models, while also including CNN-based models for a fair comparison. We illustrate groundbreaking research on optical flow conducted in recent years using a timeline and establish a framework to comprehend the progression of Transformer implementation. Furthermore, we provide a detailed comparison of the performance of these techniques, experimental configurations, and computational efficiency, and emphasize visual comparisons to highlight the drawbacks and benefits of the benchmark datasets (FlyingChairs, Sintel, and KITTI2015). Our analysis reveals that while Transformer-based models achieve higher accuracy, they often come with increased computational demands compared to CNN-based approaches. This survey aims to bridge the gap in the existing literature and provide a comprehensive understanding of Transformer-based optical flow techniques, facilitating further research and development in this field.

Keywords Computer vision · Optical flow · Transformer · CNN

✉ Anil Singh Parihar
parihar.anil@gmail.com

Nihal Kumar
nihalgupta84@gmail.com

Om Prakash Verma
opverma.dce@gmail.com

[1] Machine Learning Research Lab, Delhi Technological University, Delhi 110042, India

Published online: 25 September 2024

1 Introduction

Optical flow, a pivotal technique in computer vision, refers to the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and the scene. It enables prediction of pixel movements from one video frame to the next, which is crucial for dynamic scene analysis. This capability facilitates applications across a broad spectrum of computer vision tasks, including image deblurring [1], video frame interpolation [2], object segmentation [3], object detection [4], medical imaging [5, 6], motion detection [7], visual monitoring [8–13], video colorization [14–18], and action recognition [19–21].

Challenges such as occlusion, pixel displacement, motion blur, and the detection of small objects have historically impeded the estimation of optical flows. Over the course of history, researchers have explored different techniques to tackle the challenges associated with optical flow estimation. These techniques have resulted in significant advances in the field. For example, Horn and Schunck [22] approached optical flow estimation as an optimization problem and employed variational techniques. Lukas and Kanade [23], on the other hand, adopted the spatial intensity gradient technique. Subsequent techniques involved the use of coarse to fine strategies [24], feature matching, and regularization [25–28] to estimate optical flow.

With the advancement of technology and the emergence of deep learning techniques, particularly convolution neural networks (CNN), there has been a significant shift in optical flow estimation in the computer vision domain. CNN has become a fundamental component in this field due to its ability to uncover intricate patterns from large-scale datasets. Various techniques, including FlowNet 2.0, PWCNet, IRR, LiteFlowNet, RAFT, and others [29–34], have incorporated CNN into their pipelines, which typically consist of feature extraction, correlation, and flow map decoding. However, CNN has the limitation that it can only process local contiguous sections of an input space, restricting its ability to understand the broader context or relationships between distant features in complex data structures.

There are a limited number of surveys dedicated to optical flow. Beauchemin et al. [35] explored 2D motion and extended the calculation of optical flow, emphasizing its application in various computer vision tasks. Fortun et al. [36] classified optical flow estimation methods, focusing on traditional approaches. More recent surveys [37–39] offer detailed comparisons between conventional and learning-based optical flow estimation techniques. Zhai et al. [40] conducted a review of scene understanding and optical flow in parallel, while [41] examined a limited number of unsupervised techniques. Table 1 provides an overview of the existing optical flow survey studies.

Recently, the use of Transformer-based architectures [42] has increased to address CNN constraints. Initially designed for NLP tasks, Transformers leverage attention mechanisms and sequence processing for optical flow estimation. The self-attention mechanism efficiently handles long-range dependencies and global context, crucial for accurate estimation, surpassing CNN limitations. Despite their advantages, Transformers face challenges such as increased computational complexity and the need for extensive training data. Although transformer adoption in optical flow estimation is growing, comprehensive surveys and comparative analyses of these state-of-the-art techniques are lacking, as existing surveys focus mainly on conventional and CNN-based methods.

To address this research gap, we conduct a thorough analysis of Transformer-based optical flow estimation techniques. We carefully select articles based on the following criteria: (i) SoTA performance on benchmark datasets, (ii) innovative incorporation of transformer

Table 1 Summary of related surveys on optical flow estimation

Title of survey	Year	Reference	Comment
The computation of optical flow	1995	Beauchemin et al. [35]	Optical flow calculation using conventional techniques
Optical flow modeling and computation	2015	Fortun et al. [36]	Discussion of recent developments in modeling optical flow estimation
A survey of variational and CNN-based optical flow techniques	2019	Tu et al. [37]	Recent progress focusing on variational and CNN-based techniques
Optical flow estimation with deep learning, a survey on recent advances	2020	Savian et al. [38]	Focus on deep learning and hybrid methods
Optical flow estimation in the deep learning age	2020	Hur et al. [39]	Development of techniques from the classical energy-based function to the CNN-based
Optical flow and scene flow estimation : a survey	2021	Zhai et al. [40]	Techniques of scene flow and optical flow, and future challenges
Survey on unsupervised learning methods for optical flow estimation	2022	Dobricki et al. [41]	Unsupervised techniques comparison with supervised techniques

components into the flow estimation pipeline, and (*iii*) potential for future research and development. By focusing on these understandings, we aim to provide readers with a clear understanding of the current landscape and future directions of transformer-based optical flow estimation.

Our aim is to offer readers a lucid comprehension by specifically highlighting the latest and cutting-edge Transformer-based methods. These approaches show the integration of transformers into the flow estimation process more comprehensively, while also briefly mentioning other techniques that incorporate transformer components. A word art is shown in Fig. 1 to visualise the selected transformer-based techniques.

The key contributions of our study can be summarised as follows.

- To the best of our knowledge, this is the first experimental analysis to provide a comprehensive overview of Transformer-based optical flow estimation techniques. We also offer a detailed analysis of CNN-based techniques to understand the advantage of the transformer effectively.

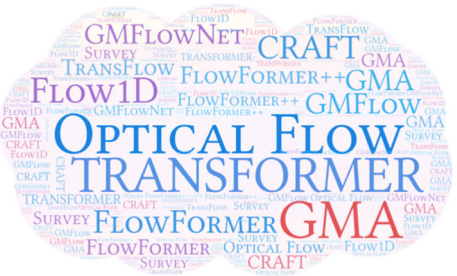


Fig. 1 Word cloud showing the Transformer-based optical flow techniques

- After reviewing numerous papers, we introduce a baseline network that utilizes learning-based methods to establish a link between the evolution of the transformer in SoTA techniques.
- We provide a comprehensive summary comparing results, computational costs, training techniques, and implementation details. Visual comparisons are also included to highlight the strengths and weaknesses of the techniques and to reveal potential future directions.

Figure 2 displays the structure of the work. Section 2 outlines the evolution of optical flow estimation techniques, divided into Conventional and Learning-based techniques. Section 3 discusses the integration of Transformer in optical flow estimation techniques and its relationship to the baseline (flow estimation pipeline). Detailed analyses of the performance of the techniques and experimental and visual comparisons are provided in Sections 4 and 5. Lastly, conclusive remarks are provided in Section 6.

2 Literature review of optical flow estimation techniques

In this section, we examine the evolution of optical flow estimation. Figure 4 presents the timeline of optical flow estimation techniques, which dates from the 1980s to the present. We divide the techniques into two categories: Conventional techniques, which are based on mathematical modelling and do not require a large amount of data, and Learning-based Techniques, which emerged with deep learning and require a large amount of data for training, are shown in Fig. 3. The techniques can be further categorised into convolution-based techniques and Transformer-based techniques, which are elaborated in the subsequent subsections.

2.1 Conventional techniques

The development of knowledge-based optical flow estimation techniques has been a comprehensive journey. Initially, Horn and Schunck [22], Lucas and Kanade [23] established the foundation with their focus on brightness constancy and local window-based techniques.

The 1980s furthered this development, with Nagel [43] and Anandan [44] introducing gradient-based techniques and computational frameworks, and Uras et al. [45] enhancing motion perception models.

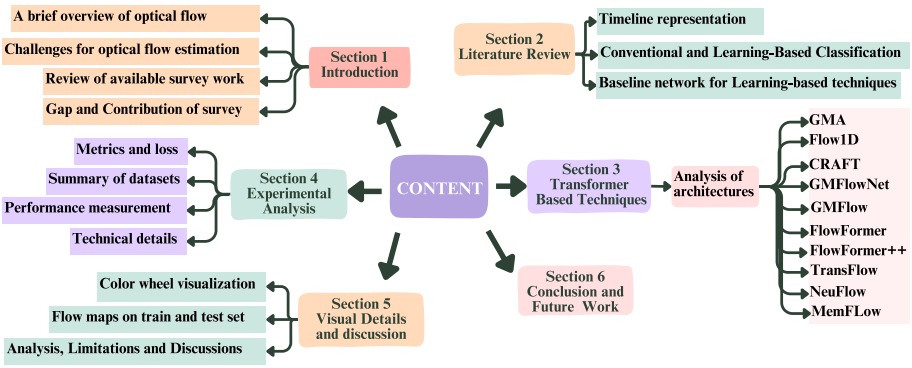


Fig. 2 Structure of paper

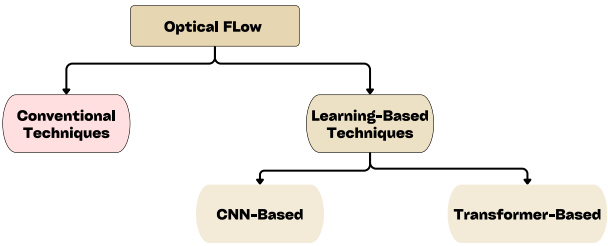


Fig. 3 Classification of the optical flow techniques

In the 1990s, the focus shifted to understanding 3D motion and enhancing robustness. Notable contributions came from Hildreth et al. [46], Adiv and Gilad [47], Black and Jepson [107], and Black and Anandan [48].

The early 2000s saw advances with Thompson [49], Weiss et al. [50], Brox et al. [25], Roth and Black [51] and Sun et al. [52], tackling discontinuities and incorporating statistical and learning-based techniques.

The 2010s continued this trend, with Brox and Malik [25] and Xu et al. [53] focussing on large-displacement optical flow and motion detail preservation. This decade also saw the introduction of innovative techniques such as DeepFlow by Weinzaepfel et al. [54], FlowFields by Bailer et al. [55], and EpicFlow by Revaud et al. [56], which emphasized dense correspondence fields and edge-preserving interpolation for large-displacement optical flow estimation.

2.2 Learning-based techniques

The blue arrow in Fig. 4 represents the Learning-based SoTA techniques from 2015 to the present. After the release of the clean version of the Sintel dataset [57] in 2009 and the KITTI2012 dataset [58] in 2012, a substantial number of data samples became available for learning-based techniques. This led to a significant increase in the use of CNN-based techniques. Researchers have investigated different techniques, such as supervised,

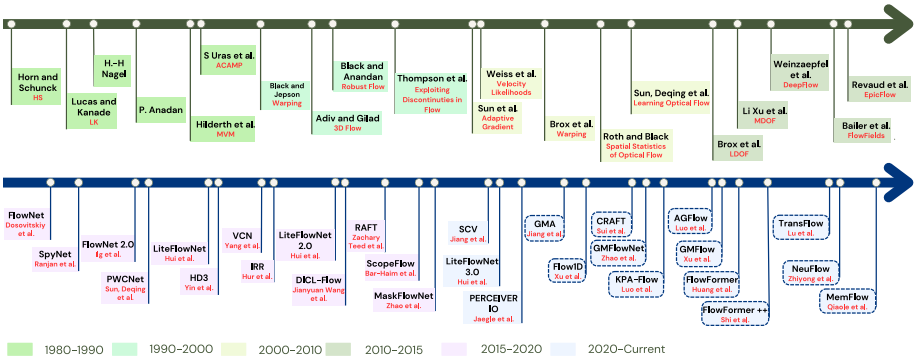


Fig. 4 Evolution of optical flow estimation techniques: this timeline showcases the progression from traditional to Transformer-based techniques, highlighting key contributions that have shaped the field. Dashed boxes are used to represent the transformer-based techniques analyzed in this work, and different colors are shown using labels to indicate different time periods

self-supervised, and unsupervised, to estimate the optical flow. In this discussion, our emphasis is on supervised and self-supervised techniques. From 2015 to now, there has been a notable improvement in the accuracy of flow estimation. In the following sections, we discuss breakthrough techniques that have achieved the best performance. These techniques aim to calculate the optical flow using the three fundamental blocks shown in Fig. 5, which involves the use of CNN and transformer. Figure 5 illustrates a comprehensive framework that outlines the stages involved in CNN or Transformer-based optical flow estimation techniques. At its core, the process is segmented into three primary stages: the extraction stage, where features and contextual information are gathered; the Similarity Stage/Cost Volume, which focuses on feature matching and estimation; and the Decoder/Refinement Stage, where the final optical flow is refined and decoded. To improve performance, researchers have suggested making changes to different aspects of the baseline framework. These modifications include improving the feature extraction stage, optimizing similarity calculations, and enhancing the decoder stage to achieve more precise optical flow estimation. Each modification aims to address specific challenges, leading to advances in the accuracy and robustness of optical flow estimation models.

2.2.1 Use of CNN in baseline

The rise of CNN revolutionized the modeling of motion patterns by taking advantage of its ability to learn hierarchical representations. CNN-based techniques can be categorized into two main structures: the coarse-to-fine technique-based technique and the iterative refinement technique. These flow pipelines have shown impressive performance. Dosovitskiy et al. [59] introduced the synthetic FlyingChairs dataset, which has 22,000 samples, along with the development of FlowNet [59]. FlowNet is the first CNN-based end-to-end optical flow estimation pipeline. Ranjan and Black [31] proposed SpyNet to address challenges related to the parameters and speed of the model. The emergence of FlowNet 2.0 [29] introduced the concept of creating a cost volume for correspondences and cascading CNN blocks to decode flow maps.

In their work on PWCNet [30], Deqing Sun et al. introduced a technique called multi-feature extraction with wrapping to address the issue of capturing motion details of small objects. Tak-Wai Hui et al. proposed the LiteFlowNet series [32, 60, 61], which aimed to reduce computational burden by reducing the baseline. Yin et al. [104] proposed HD3, a technique that takes advantage of multilevel feature correspondence to estimate optical flow and stereo matching. Several other techniques, such as VCN, DICL-Flow, ScopeFlow, and MaskFlowNet [62–65], focused on the similarity stage by incorporating multiple frames.

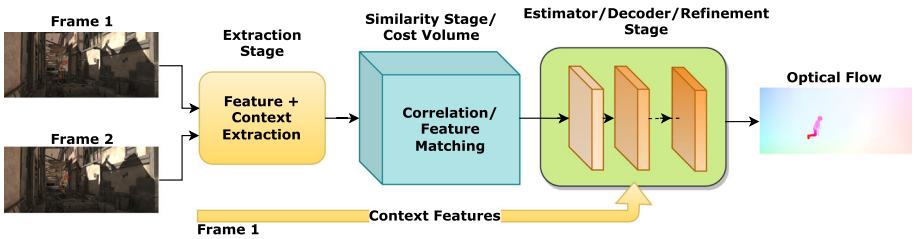


Fig. 5 Schematic overview of CNN/Transformer-based optical flow estimation techniques: the existing works, which generalize optical flow estimation, serve as a baseline for comprehending the evolution and development of these techniques

RAFT marks a significant change in the way flow estimation is addressed. The IRR-PWC pipeline has been modified to achieve remarkable performance. Teed and Deng [34] devised a cost volume with multiple dimensions and a lookup operator to verify the correspondences. In addition, they used an RNN to iteratively refine the flow maps, resulting in improved accuracy. However, it does not utilize the potential of a transformer for global learning. Luo et al. [66] introduce KPA-Flow over the RAFT pipeline by using the custom kernel function to effectively mine local relations to infer flow maps. AGFlow [67] (2022) is the pioneering technique that employs adaptive graph reasoning to understand movement in the context of the scene. Using context information effectively, AGFlow achieves enhanced accuracy.

2.2.2 Use of transformer in baseline

The emergence of Transformer models in 2017 by Vaswani et al. [42] marked a significant milestone in the field of Natural Language Processing (NLP), opening up new possibilities for language task development. Transformers have gained popularity as an alternative to CNN due to their scalability and parallel processing capabilities on GPUs. This has led to an improved model performance in computer vision tasks. Transformers possess a robust capability to identify complex patterns and contextual associations in visual data by perceiving the image as a sequence of patches. The introduction of Visual Transformer (ViT) [68] pioneered the application of the transformer directly to sequences of image patches, demonstrating its effectiveness in computer vision and achieving impressive results. There has been significant and rapid progress in computer vision with the introduction of Transformer-based models. A notable example is the DEtection TRansformer (DETR) [69], which is available in 2020 and is used to effectively analyze global relationships in data to improve object detection. Zhang et al. [70] is the upgrade of DETR based on the transformer for object detection. Subsequently, the Swin Transformer [71] applies a hierarchical transformer to non-overlapping image windows to accurately interpret local and global information. Furthermore, the Transformer iN Transformer (TNT) [72] model introduced features such as a nested transformer structure to improve the representation of visual structures by effectively modeling local and global dependencies.

The accuracy of optical flow estimation depends mainly on the ability to identify the corresponding pixels across different frames. To enable the precise understanding of motion between successive video frames and generate accurate flow estimates, transformer architectures leverage their self-attention and multi-head attention mechanisms during the feature extraction stage to capture global information. The self-attention operation relates different positions of a single sequence to compute the representations shown in (1):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

In this context, Q, K, and V represent the Query, Key, and Value vector/matrix, respectively. d_k denotes the dimension of the key vectors. The expression QK^T signifies the dot product between queries and keys, which evaluates their compatibility. Furthermore, $\sqrt{d_k}$ acts as a scaling factor to avoid excessively large dot products. softmax$(\cdot)$: Softmax function to obtain attention weights. The output is a weighted sum of the values V, where the weights are attention scores calculated by the compatibility between queries Q and keys K. Multi-head attention, on the other hand shown in (2), projects the inputs and computes self-attention in parallel from different representation subspaces, allowing the model to attend to

information from diverse perspectives jointly:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

In (3), h represents the number of attention heads, W_i^Q, W_i^K, W_i^V are the trainable projection metrices for queries, keys, and values, respectively, corresponding to the head i^{th}, Concat$(\cdot)$ denotes the concatenation operation along a suitable dimension, and W^O stands for the trainable output projection matrix. Furthermore, in the similarity computation stage, cross attention enables the relating of representations from different sequences shown in (4), allowing the decoder to gather relevant information from the encoded memory when generating outputs:

$$\text{CrossAttention}(Q, K, V) = \text{Attention}(QW^Q, MW^K, MW^V) \tag{4}$$

where M denotes encoded memory representations, and W^Q, W^K, and W^V are trainable projection metrices. For video data, temporal attention modules in (5) specifically attend over representations from neighboring frames, capturing motion cues across time for improved flow estimation available below:

$$\text{TemporalAttention}(Q, K, V) = \text{Attention}(Q, K, V) \tag{5}$$

$$K, V = (f_1, f_2, ..., f_T) \tag{6}$$

where $f_1, f_2, ..., f_T$ in (6) represents features of neighbouring video frames. Positional encodings are also added to the input embeddings to make the outputs position-aware, as transformers lack inductive biases for spatial locations:

$$\begin{aligned} \text{PE}(\text{pos}, 2i) &= \sin(\text{pos}/10000^{2i/d_{\text{model}}}) \\ \text{PE}(\text{pos}, 2i+1) &= \cos(\text{pos}/10000^{2i/d_{\text{model}}}) \end{aligned} \tag{7}$$

In the context, pos represents the position index, i stands for the channel index, d_{model} denotes the model's dimension, and PE$(\cdot)$ refers to the positional encoding function shown in (7). Positional information is encoded using sine and cosine functions with various frequencies. This leads to a major breakthrough in well-known benchmarks for optical flow estimation, such as Sintel [57] and KITTI2015 [73], allowing for a more precise understanding of motion between successive video frames and generating remarkably precise and consistent flow estimates. We have selected the SoTA techniques for in-depth analysis and dive into these techniques in the following sections.

3 Optical flow estimation based on transformer

In this section, we examine the techniques that rely on the transformer framework. An in-depth examination of the architecture of the selected techniques is available in Section 3.2. In addition, we have included a comparison table to illustrate the incorporation of the transformer. Jaegle et al. [74] proposed a generalized multi-modal encoder-decoder architecture to process multiple inputs. Perceiver IO tackles the task of estimating optical flow first using CNN as a pre-processing step and incorporating a transformer architecture. This technique allows for the separation of input and output processing from the data's size and characteristics, enabling the model to handle dense visual tasks effectively. Jiang et al. [75] introduce "COTR: Correspondence Transformer for Matching Across Images" that combines sparse

and dense matching by leveraging the inductive bias of densely connected networks. The input to the network is the concatenation of feature maps from two consecutive frames, which are processed by a shared CNN backbone. These feature maps are then fed into the encoder-decoder-based transformer, along with the query point (x), and the output of the transformer is decoded by a multi-layer perceptron to obtain the correspondence x. Liu et al. [76] jointly estimate the scene flow and the optical flow using the same pipeline. The novelty of the architecture lies in the fusion of the lidar maps and feature maps of the frames in the 2D and 3D branches with multiple bidirectional connections between specific layers using CNN and transformer. Luo et al. [66] proposed the "Learning Optical Flow with Kernel Patch Attention (KPA)" architecture over the RAFT pipeline to enhance motion features at the refinement level using the KPA module. The KPA module uses a kernel attention operator to work on each local patch and learns to mine context affinities to better infer flow fields. Luo et al. [77] introduced "GAFlow: Incorporating Gaussian Attention into Optical Flow" to optimize the motion update block of [78]. The authors used attention to enhance feature maps and formulate the Gaussian filter concept for guided attention to emphasize contextually related regions during matching. Chen et al. [79] used multiple frames to extract multiple feature maps, resulting in multiple cost volumes. These cost volumes are guided by the transformer, resulting in better occlusion handling. The FCTR module in the "Feature Correlation Transformer to Estimate Ambiguous Optical Flow" introduced by Fang et al. in 2023 [80]. This module utilizes a transformer structure with modifications in self- and cross-attention to enhance global receptive fields, along with positional embedding to establish correspondences in flow estimation. Dong et al. [81] proposed the MatchFlow model, which inherits QuadTree attention to enhance the motion feature maps extracted from the feature extractor. Weinzaepfel et al. [82] jointly estimates stereo matching and optical flow using a ViT-based encoder and decoder using pre-training over 5 million images. Shi et al. [83] calculated optical flows in both directions for several frames, using temporal clues to improve overall performance.

3.1 Selection of techniques

We examined the evolution of optical flow estimation pipelines, observing a trend towards Transformer or hybrid designs. The majority of the current state-of-the-art methods adhere to the RAFT pipeline structure. Therefore, our selection emphasizes how effectively we can integrate Transformer architecture into the pipeline. During the selection process, we identified that transformers are incorporated at three key stages of the optical flow estimation pipeline, as depicted in Fig. 5. To highlight the progression and impact of Transformer architectures within the pipeline, we selected state-of-the-art techniques that showcase the highest performance and innovative Transformer integration at each stage.

Flow1D [84] architecture utilizes the Transformer to create the cost volume from the feature maps. The GMA [78] technique employs a transformer solely in the decoder stage. In [85], Sui et al. used a CNN-based network to extract features from a pair of frames and then used a Semantic Smoothing Transformer to enhance the features of Frame 2. GMFlowNet [86] uses a Transformer exclusively in the extraction stage. Haofei Xu et al. restructured the pipeline to reduce the inference time by merging a transformer with a CNN to improve features and compute similarity, as shown in Fig. 10. FlowFormer [88] employs a transformer in each baseline stage, except in the refinement stage. Shi et al. introduced FlowFormer++ [89], an enhancement of the initial FlowFormer framework, aimed at improving performance. This enhancement is achieved by incorporating masked preprocessing into the cost-volume

encoding phase and uses the full Transformer-based flow estimation pipeline to improve optical flow. NeuFlow is specifically designed for real-time optical flow estimation. Consequently, NeuFlow employs a CNN-based shallow backbone at 1/16 and 1/8 scales and processes the features using a Transformer-based cross-attention block. Qiaole Dong et al. introduced the MemFlow [92] architecture, which leverages flash attention in the memory buffer to enhance 2D motion features. This enhancement is achieved by processing both 2D motion and context features within a memory buffer Transformer block that uses flash attention to improve the 2D motion features.

By carefully selecting these techniques, we aim to provide a comprehensive overview of the architectural progression and the significant role of transformers in advancing optical flow estimation. Each technique is presented using architectural components that include Transformers in the process. The empirical evaluation of these techniques is available in Section 4. Table 2 offers a simplified overview of how optical flow methods have been integrated into the Transformer model. The table contains three columns corresponding to the three stages shown in Fig. 5, indicating whether the techniques use CNN, Transformer, or both. This table provides a clear visualization of how the incorporation of transformers in different stages contributes to the superior performance of these techniques on the benchmark datasets.

3.2 Architecture analysis of selected techniques

- **Global Motion Aggregation (GMA)** [78] To address occlusion, GMA module defines an occluded point as one that appears in the reference frame but is absent in the subsequent frame. This definition is slightly broader than conventional interpretations, as it also encompasses points that transition out of the frame. Such scenarios introduce significant ambiguities that cannot be effectively resolved with purely local techniques. The GMA module focuses on motion as overall characteristics rather than specific local aspects, based on the assumptions that points on an object typically exhibit consistent motions. It implies that even if certain sections of an object, such as a fast-moving component, are absent in subsequent frames due to rapid transitions or occlusions, their motion can still be

Table 2 Overview of adoption of transformer in flow estimation pipeline: Three stages from baseline in Fig. 5 as shown here

	Extraction stage		Similarity stage	Decoder stage
	Feature	Context feature		
GMA [78]	CNN	CNN	CNN	CNN+TF
Flow1D [84]	CNN	CNN	CNN+TF	CNN
CRAFT [85]	CNN+TF	CNN	TF	CNN+TF
GMFlowNet [86]	CNN+TF	CNN	CNN	CNN
GMFlow [87]	CNN+TF	–	CNN+TF	TF
FlowFormer [88]	CNN+TF	CNN	TF	CNN+TF
FlowFormer++ [89]	CNN+TF	CNN	TF	CNN+TF
TransFlow [90]	TF	–	TF	TF
NeuFlow [91]	CNN	CNN	CNN+TF	CNN+TF
MemFlow [92]	TF	TF	CNN	CNN+TF

The **CNN** is for convolution neural network and **TF** stands for Transformer

accurately estimated by aggregating motion features from other parts of the same object. With this idea, the aggregator block in Fig. 6 adopts the transformer to pay attention to the finding of similarities in the motion vectors. Query vector Q and key vector K are projections of the generated context feature maps. After taking the dot product of the query and the key vector, the projections passed through a softmax layer to create a 4D attention matrix. 2D motion features through the cost volume are considered as the value vector V. The 4D attention matrix computed using Q and K is concatenated with the value vector using an aggregator block to generate aggregated motion features that the GRU decodes.

In summary, GMA presents a revolutionary technique for addressing the occlusion issue by guaranteeing the transfer of motion features from unoccluded pixels to the occluded ones.

- **Flow1D: High-Resolution Optical Flow from 1D Attention and Correlation** [84] Prior research has typically approached optical flow estimation as a problem of searching in two dimensions. However, this technique becomes impractical for large displacements, especially in high-resolution images, because of the computational complexity associated with increasing the size of the search window quadratically. To address this issue, Flow1D simplifies the 2D search space employing two consecutive 1D operations (attention and correlation) in perpendicular directions. In doing so, this technique effectively represents 2D correspondences with significantly reduced computational requirements. The transformer has been applied to emphasize the 1D splitter features extracted from the CNN backbone. The architecture diagram in Fig. 7 illustrates the process in which the 2D features extracted using CNN are divided into horizontal and vertical 1D features. The detailed information for each component used with the transformer is presented below.

  1. **1D attention and Correlation:** By combining features from both Frame 1 and Frame 2 column-wise or row-wise, this technique utilizes the transformer to apply one-directional attention, vertically or horizontally. 1D attention undergoes intricate computations and positional encoding to enhance spatial context features across frames. The resulting 1D features calculate a weighted sum of pixels within the same column or row, thereby generating new features where each vector is aware of the features in the corresponding column or row.
  2. **1D Correlation and 3D Cost Volume:** After applying 1D attention, the 1D features of the frames can be accessed independently for both horizontal and vertical directions. 1D correlation executed in the orthogonal direction among the 1D

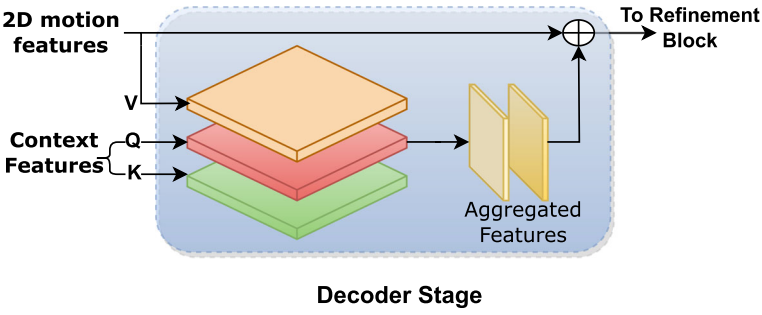


Fig. 6 Architecture of GMA [78]: with comparison to baseline; this figure shows the modifications only in the refinement stage

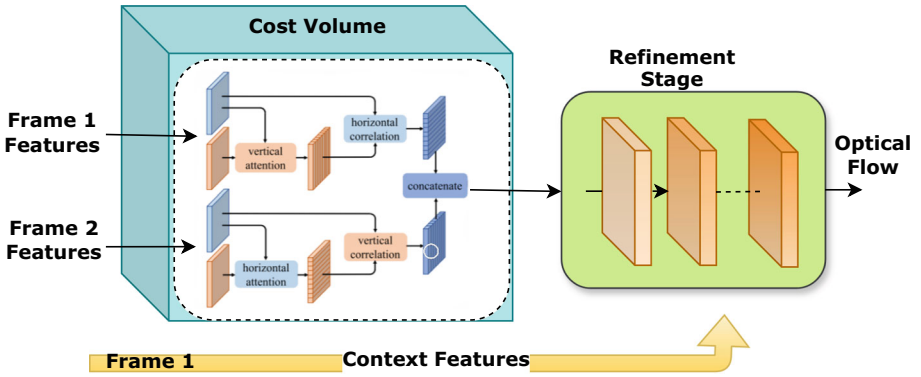


Fig. 7 Flow1D [84] Architecture: this fig. highlights the modifications and integration of the transformer technique within Flow1D architecture [84] as compared to the baseline. The extraction stage is modified with convolution and transformer to obtain more enhanced 1D features

features(horizontal/vertical) of Frame 1 and Frame 2. By alternating the directions of attention and correlation, two different 3D cost volumes are created and later concatenated for the final optical flow estimation. This 3D cost volume is used to search the 2D motion using a lookup operator in the refinement stage.

- **CRAFT: Cross-Attentional Flow Transformer for Robust Optical Flow** [85] The cost volume is the core block of the baseline. This cost volume exits the matching features between frames; however, this can also have the features of the noise existing in frames. And the localized features of the CNN can also limit the contextual information into pixel features. Both of these gaps are answered by CRAFT. CRAFT architecture [85] in Fig. 8 intricately orchestrates the use of the transformer in frame 2 to inherit the context of global characteristics with respect to Frame 1. And cross-attention [68] with relative position bias [93] is used to generate the cost volume instead of the dot product to put a constraint on the noise features. These operations of both blocks are described below:

  1. **Semantic Smoothing Transformer (SSTrans):** SSTrans uses Expanded Attention to compute multiple sets of features on extracted features from CNN backbone from Frame 2 and aggregate them using dynamic mode attention. The output features of SSTrans are a linear combination of all features. A weighted skip connection is

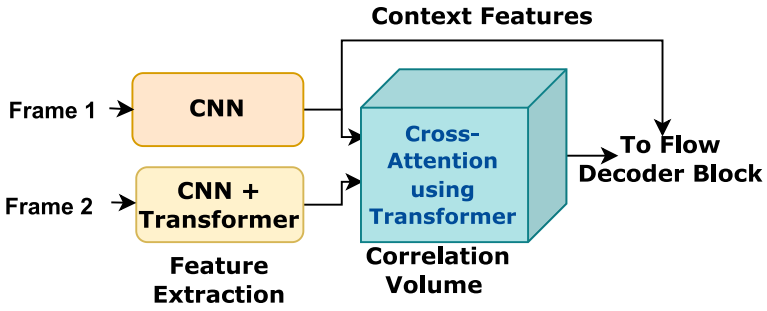


Fig. 8 Modified architecture of CRAFT [85]: in this figure, we illustrate the changes made to the stages of the baseline using the transformer

added to preserve the original frame features. Relative position bias adds spatial bias by adjusting the attention weights between pixels in the same frame.

2. **Cross-Frame Attention:** SSTrans enhances the global and semantic stability of Frame 2 features. The features from Frame 2 and Frame 1 are inputted into the transformer's cross-attention module. The query vector Q of the Frame 1 features is assigned to the key vector K of the Frame 2 features to produce correspondence vectors, which helps filter out feature noise. These calculated features are then produced with varying scale values to construct a multilevel cost volume.

- **GMFlowNet: Global Matching with Overlapping Attention for Optical Flow Estimation** [86]
  Swin series of vision transformers uses CNN and the transformer to capture feature maps from the frames. However, the optical flow estimation problem inherits the long-range motion dependencies between frames. As a result, a large displacement error occurs in optical flow maps. Motion patterns, brightness consistency, and the general dynamics of the scene are key parameters for accurate optical flow. A way to encounter these issues is to optimize the matching of the feature maps before direct regression.
  Figure 9 represents the integration of the transformer into two stages. The first stage is used for the extraction of the features of frames, which is done by combining the CNN and the transformer. The second stage uses a softmax layer, which has been used to aggregate the correspondence features, and a matching loss is used to optimize the matching features in cost volume.

  1. **Transformer-Driven Feature Extraction using POLA:** The CNN backbone initially prepares the features for the Patch-based Overlapping Attention (POLA) component, which is based on a Swin transformer [71]. Zhao et al. have modified the multi-head attention of the swin transformer with the proposed neighborhood attention. Neighborhood attention divides features into M $\times$ M non-overlapping patches and attends every patch with itself and its eight neighboring patches. By this, POLA involves inter-patch features within one block and propagates information directly with less information loss.
  2. **Global Matching in Cost-Volume:** 4D cost volume inherits the emphasized features of the POLA stage. A softmax layer with an argmax operation is then used to process the volume, resulting in the generation of an initial coarse flow. This coarse representation shown in the second stage of Fig. 9 induces global matching, which aims to capture global context information by aggregating features from multiple levels and

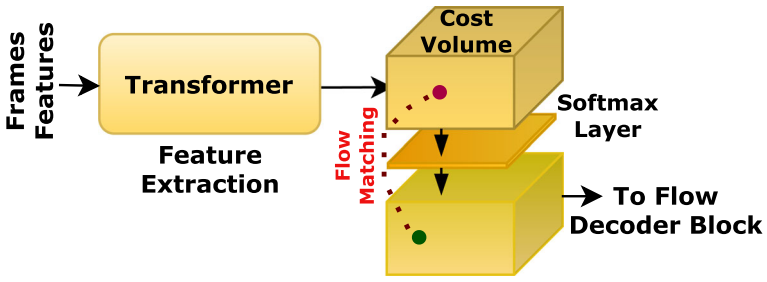


Fig. 9 The architecture of GMFlowNet [86] : the transformer has been integrated with the CNN in the extraction stage only of the baseline

scales. It helps to understand the relationships between different parts of an image and improves the overall matching accuracy. The matching loss added with the loss L_1 helps train the model to learn discriminative features and encourages accurate matching between images.

- **GMFlow: Learning Optical Flow via Global Matching** [87] The techniques mentioned above incorporate a refinement step that involves using a GRU to enhance optical flow by utilizing context features from frame 1. Although this technique improves accuracy, it also leads to an increase in the number of parameters and a reduction in the speed of inference Fig. 16. Moreover, the problem of local matching in CNN has been a challenge in earlier studies. To address this issue, GMFlow employs a transformer to enhance the features following a weight-sharing convolution network. These enhanced features are input into the global matching block to establish correspondences. Instead of employing iterative refinement, GMFlow utilizes a Transformer-based technique to generate optical flow at higher resolution, as illustrated in Fig. 10. Detailed information on each stage of the transformer is provided below:

  1. **Feature Enhancement using Transformer:** A weight-sharing modified ResNet-14 network is employed to extract features from both frames. In order to incorporate global context, these downsampled dense features are input into a Swin Transformer network [71], which consists of six layers of self-attention, cross-attention, and feed-forward network (FFN) with 2D sine and cosine positional encoding.
  2. **Global Matching with Transformer Inputs:** A correlation layer and a softmax layer are used to assess the similarity of features for each pixel in the feature maps of frame 1 in relation to all pixels in the feature maps of frame 2 through a straightforward matrix multiplication calculation.
  3. **Flow Propagation:** The softmax-based technique for flow estimation operates under the assumption that the pixels being compared are present in both images, a condition that does not hold true for obscured or off-image pixels. To address this limitation, the technique leverages high-fidelity flow estimations from corresponding pixels to inform unmatched ones by evaluating feature self-similarity through basic self-attention. Additionally, the model incorporates a refinement stage that applies GMFlow to higher resolution features to forecast a residual flow at a reduced 1/4 resolution, thus enhancing the overall precision of the predictions.

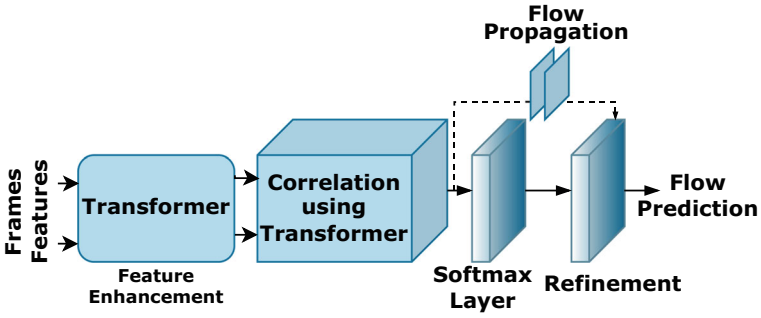


Fig. 10 GMFlow [87] architecture: illustrates the incorporation of the transformer in contrast to the phases of the baseline

- **FlowFormer: A Transformer Architecture for Optical Flow** [88] As noted previously, the Transformer has been integrated into CNN to improve optical flow by substitution or integration. However, cost-volume processing remains a fundamental element of the base technique. FlowFormer takes advantage of the transformer for encoding and decoding cost-volume data. In Fig. 11, three stages are depicted, each with a transformer. The features extracted from a pair of frames are refined using a transformer and then used to create a 4D correlation volume. Subsequently, this cost volume is fed through Alternate Group Transformer (AGT) layers to produce 2D cost characteristics. These 2D motion features are utilized to generate flow with the support of a cost-memory encoder and decoder. The following is a succinct overview of the elements.

  1. **Building the 4D Cost Volume:** The Siamese encoder on the Swin backbone takes the feature maps of size $H/8 * W/8$, where H and W represent the height and width of the input RGB image and generate the more global enhanced features of the same size. These feature maps are then used to construct a 4D cost volume of dimensions $(H \times W \times H \times W)$. The calculation involves finding the dot product between each pair of pixels from the source and target feature maps.
  2. **Encoding of Cost-Volume:** This block is used to transform the cost volume into the cost memory using a transformer. It consists of three stages: cost map patchification, cost patch token embedding, and cost-memory encoding. The initial step, cost map patchification, involves dividing the cost map of each source pixel into patches using strided convolutions to generate a series of embeddings for cost patches. Subsequently, in the cost patch token embedding phase, the features of each source pixel's patches are summarized by employing the dot-product attention mechanism with K latent code words. Lastly, the cost-memory encoding phase converts the cost volume into a cost memory by utilizing AGT layers that apply self-attention within and across the cost tokens in an interleaving manner to propagate information globally while keeping computation efficient.
  3. **Transformer Based decoding:** The cost-memory decoder employed a recurrent Transformer-based decoder to predict optical flows depending on the encoded cost memory. It uses cost queries to retrieve cost features from the cost memory and iteratively refines flow predictions with a recurrent attention decoder layer. The cost-memory aggregation step generates a sequence of cost queries, each of which estimates the flow of a single source pixel via co-attention to the cost memory. The recurrent flow prediction step iteratively regresses the flow residuals to refine the flow of each source pixel.

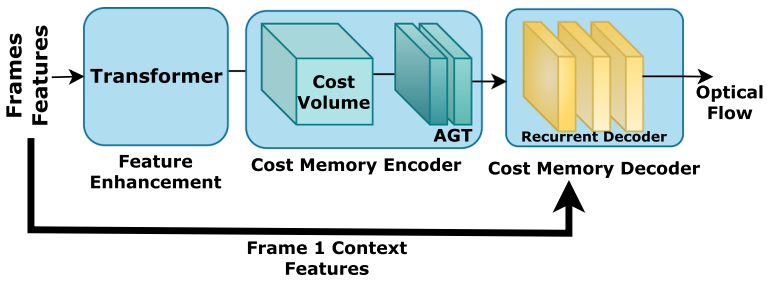


Fig. 11 Block representation of FlowFormer architecture: demonstrates the integration of a transformer in every phase of the standard model

- **FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation** [89] Xiaoyu Shi et al. proposed FlowFormer++ [89] to improve the performance of the FlowFormer [88] pipeline network. FlowFormer++ introduces a novel method for optical flow estimation by utilizing a Transformer-based architecture to successfully pre-train the cost-volume encoder. This method incorporates Masked Cost-Volume Autoencoding (MCVA) inspired by [94] within the transformer framework, with the goal of improving the cost-volume encoder performance. At its core, MCVA involves masking a portion of the cost volume data and training the network to reconstruct the masked tokens from the visible ones. However, directly applying this technique to FlowFormer's architecture faces challenges due to the cost volume's inherent redundancy and architectural incompatibility with masks. FlowFormer++ addresses these challenges by incorporating three essential components, as demonstrated in Fig. 12.

  1. **Masking technique:** FlowFormer architecture is modified to accommodate the masking technique during pre-training. These modifications ensure that the model can effectively process masked-cost volumes, preventing any mixing of masked and unmasked features. The careful integration of masked autoencoding within the transformer framework is essential to fully exploit the potential of the transformer in optical flow estimation. The implementation of a block-sharing masking technique addresses the problem of information leakage by making it difficult to deduce masked information from nearby pixels. This technique divides the image into non-overlapping blocks, where pixels within the same block share a common mask. By doing so, the model is encouraged to consider a wider context, which improves its ability to handle complex motion patterns and occlusions.
  2. **Pre-text Reconstruction Task:** A new pre-text task is better proposed to align the pre-training process with the fine-tuning phase. This task involves reconstructing larger cost map patches from smaller, randomly cropped patches. This design mimics the decoding process utilized during fine-tuning and ensures consistency between the pre-training and fine-tuning stages. Effectively, it prompts the model to learn representations that facilitate the aggregation of long-range information, which is crucial for accurate flow estimation.

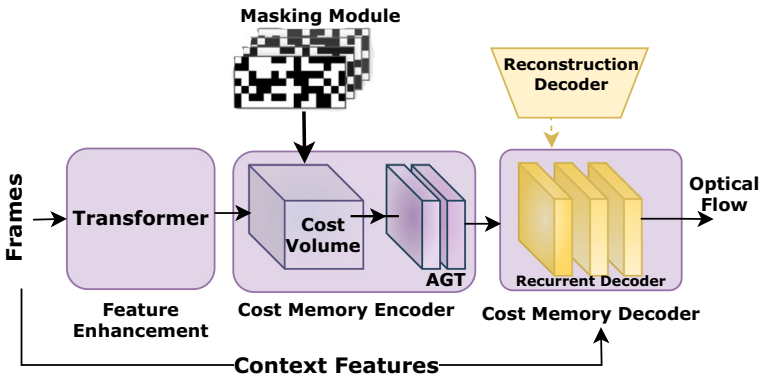


Fig. 12 Block representation of FlowFormer++ architecture: encapsulates the FlowFormer network with updation of masking-block and reconstruction decoder

FlowFormer++ illustrates the efficacy of integrating transformer with masked auto-encoding to pre-train models for optical flow estimation. By tackling the obstacles associated with applying MCVA to FlowFormer's structure and capitalizing on the advantages of transformer, FlowFormer++ attains superior performance in comparison to the conventional FlowFormer, especially in intricate real-world scenarios.

- **TransFlow: Transformer as Flow Learner** FlowFormer and FlowFormer++ both techniques have utilized a transformer with CNN to partially estimate the optical flow for two consecutive frames. Neither technique can solve the problem of tedious training, iterative refinement, and end-to-end transformer-based flow estimation baseline. Lu et al. [90] introduced TransFlow to estimate optical flow by taking multiple frames and using a pure transformer architecture. This technique addresses the limitations of traditional CNN-based techniques and achieves remarkable advances in accurate flow estimation. Figure 13 shows the function of the complex architecture using our baseline representation from Fig. 5. It is composed of three main components: a spatio-temporal encoder to extract the features for multiple frames and calculate the correlations among those frames. It also includes a flow decoder and a self-supervised pre-training module. This novel technique distinctly emphasizes several key aspects:

  1. **Spatial-Temporal Attention Mechanism:**
     (a) *Spatial Globalization:* Traditional CNN-based flow learners struggle with global spatial dependencies due to the local nature of convolution operations. It applies spatial attention mechanisms between consecutive frames to capture global spatial correlations, which are crucial for coherent region propagation and handling issues such as noise, motion blurs, and large displacements.
     (b) *Temporal Association:* Previous techniques lacked effectiveness in capturing long-range temporal associations. TransFlow addresses this by jointly modeling temporal association with spatial attention, allowing transformer layers to measure long-range interactions between input embeddings.
  2. **Advanced Flow Decoder:** Unlike traditional transformer decoders, the TransFlow decoder processes all frames simultaneously. This parallel processing eliminates the need for encoder-decoder cross-attention and beam search, making it more efficient. The decoder compares feature maps of consecutive frames to compute feature similarity. It uses a differentiable matching layer to establish correspondences

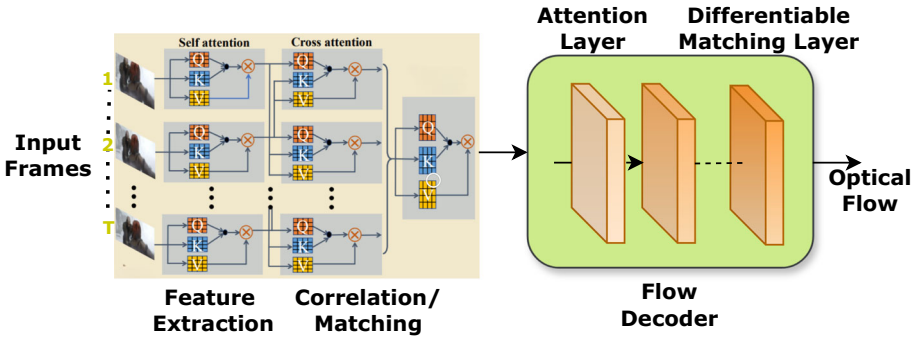


Fig. 13 Architecture of Transflow: this is the visual representation of TransFlow pipeline with the accordance of our baseline. It consists of two blocks; the first one works jointly as a feature extractor and similarity stage. The flow decoder shares the same architecture as the encoder with an extra differentiable matching layer

between frames, and occlusion detection is used to improve the accuracy of flow estimation.

3. **Self-Supervised Learning Technique:**
   (a) *Pre-training Strategy:* The model employs a self-supervised pre-training technique that masks out patch tokens and learns pixel representations by reconstructing signals from corrupted inputs. This technique differs from traditional, labor-intensive pre-training techniques.
   (b) *Strategic Masking and Score Map:* The model learns a score map to identify the most informative masking patches using a ranking algorithm and a soft-sort operation. This selective masking is crucial for effectively learning pixel representations and improving model performance.

- **NeuFlow: Real-time, High-accuracy Optical Flow Estimation on Robots Using Edge Devices** [91] As optical flow estimation is a critical task for real-time applications and requires a large number of parameters, Zhiyong et al. introduce an innovative method to address the issue. They focus on achieving high-accuracy real-time performance on edge devices such as the Jetson Orin Nano. The architecture shown in Fig. 14 utilizes a global-to-local processing scheme, combining CNNs and transformers to efficiently manage different image scales. Initially, a shallow CNN backbone extracts multiscale feature vectors, emphasizing essential low-level features. Transformer-based cross-attention layers perform global matching at 1/16 resolution, capturing extensive displacements. This initial optical flow is then refined to 1/8 resolution using lightweight CNN layers for enhanced accuracy. Finally, full-resolution optical flow is achieved via a convex upsampling module that integrates features from the 1/8-scale image. The transform components in NeuFlow significantly boost feature distinctiveness and reduce computational costs, leading to substantial performance improvements. NeuFlow operates at 30 FPS on the Jetson Orin Nano, offering a 10×-80× speedup over SoTA methods such as RAFT, GMFlow, and FlowFormer, while maintaining similar accuracy, making it ideal for real-time robotic and autonomous system applications.
- **MemFlow: Optical Flow Estimation and Prediction with Memory** [92] MemFlow introduces a memory module to optical flow estimation, significantly improving real-time

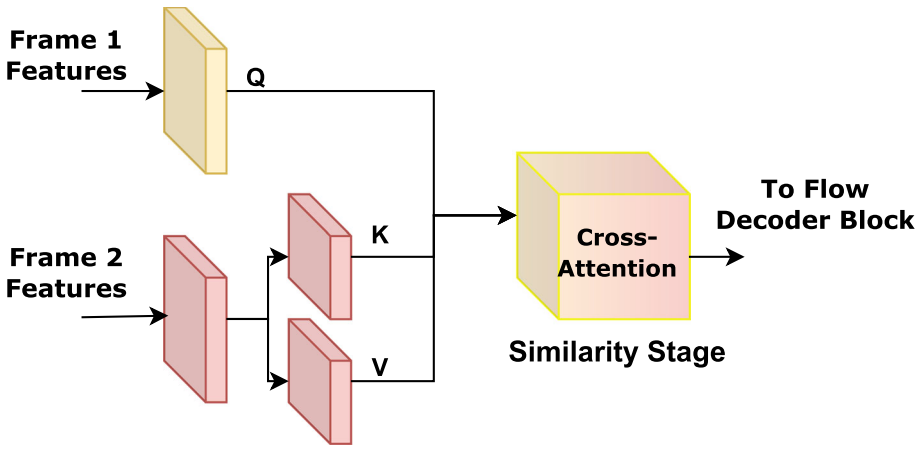


Fig. 14 NeuFlow architecture: cross-attention layers has been used to global matching to enhance the 2D motion features

performance and predictive accuracy. The architecture consists of three key components: feature extractors, a memory buffer, and update modules. Initially, the feature encoder extracts features from consecutive frames using the Swin transformer [71], constructing a 4D correlation volume. The motion encoder then processes these features to generate motion information. The memory buffer stores historical context and motion features, allowing the model to aggregate motion features through Flash Attention [95]. The GRU unit outputs residual flows, iteratively refining the optical flow. This architecture integrates a vision transformer for enhanced feature extraction, resulting in superior performance with minimal computational overhead. MemFlow demonstrates strong cross-dataset generalization and outperforms state-of-the-art methods in both efficiency and accuracy. Figure 15 highlights the integration of the transformer architecture.

4 Experimental analysis

This part of the research paper includes a thorough analysis of all the techniques. Each technique has a different setup for implementation. Implementing an experimental setup has some basic steps; these are metrics, training recipes, datasets, and hyperparameter tuning. Flow estimation has different metrices to close the gap between prediction and ground truth. The most popular metrices are the average EPE (End-Point-Error) [59] and the F1-all (outlier percentage averaged over all pixels). Due to the annotation complexity for optical flow datasets, most of the datasets are developed synthetically, and few are real environment-based. From the early days of research, there are a few popular datasets such as Middlebury [96], MPI Sintel [57], KITTI2012 [58], KITTI2015 [73], and others. An in-depth analysis of datasets, metrics, losses, performance, implementation, and computation is included. This gives the reader an understanding of how the transformer changed the performance in the flow estimation in relation to CNN.

4.1 Metrics and loss

All selected Transformer-based flow estimation techniques adopted the EPE as their primary metrics for the publication of results on the Sintel benchmark [57]. For the KITTI2015 benchmark [73], the authors adopted the average F1 metrics, and the loss follows the distance L_1. Further elaboration will be presented in the subsequent sections.

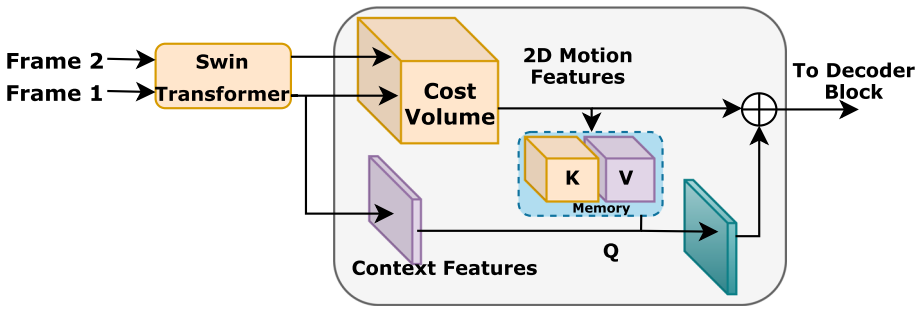


Fig. 15 MemFlow architecture: detailed depiction of the MemFlow pipeline, highlighting the integration of transformer components in memory buffer

4.1.1 End-point-error (EPE)

The EPE is a simple and intuitive metric that is easy to calculate. It is the average Euclidean distance between the predicted flow and the ground truth flow. The EPE follows the average distance L_2, which is explained in (8).

$$\text{EPE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(u_i^{\text{pred}} - u_i^{\text{gt}})^2 + (v_i^{\text{pred}} - v_i^{\text{gt}})^2} \tag{8}$$

where N is the number of pixels in Frame1 or Frame2, u_i^{pred} and v_i^{pred} are the predicted optical flow vectors in a pixel, u_i^{gt} and v_i^{gt} are the ground-truth optical flow vectors in the frame pixel.

Several techniques also considered various aspects of EPE calculation to evaluate the robust performance. Haofei Xu et al. calculated the EPE and also with a pixel of different magnitudes in the range of pixels from 0-10, 10-40 and 40 + pixels for the ground truth flow vector in GMFlow [87]. In the same way to prove the significance of large displacement, Xiuchao Sui et al. calculated the EPE on different magnitudes of flow vectors in CRAFT [85]. The large motion vectors include the range of pixels {(1 - 10), (10 - 20), (20 - 30), 30 +} in the frames.

4.1.2 F1-all

F1 is a valuable metric that is used to gauge the proportion of abnormal pixels. The simplest representation of F1-all is given in (9), where N is the number of abnormal pixels/outliers and H, W are height and width. These are defined as pixels with an EPE of 3 or greater and a relative EPE greater than 0.5. To calculate the relative EPE of a pixel, its EPE is divided by the square root of the sum of the EPE scores for all pixels in the same column as the pixel under examination.

$$F1 - all = \frac{N}{HW} * 100\% \tag{9}$$

4.1.3 Loss function

As mentioned in Table 6, all the techniques used the Multi-Iteration (L_1) loss, but with some adaptations. For example, GMFlowNet incorporated the matching loss, while TransFlow incorporated the occlusion detection loss. The multi-iteration loss L_1 is originally introduced by the authors of RAFT [34]. The specific formulation of the loss function can be found in (10).

$$L = \sum_{i=1}^{N} \gamma^{N-i} \left\| V_i - V_{gt} \right\| \tag{10}$$

N denotes the number of predicted optical flows, V_i is the predicted flow vectors, and V_{gt} is the ground-truth flow vectors. γ is the hyperparameter that increases exponentially and increases the weights to give higher weights for more refined predictions. Previous research has adjusted the parameter γ within the interval of $0 \rightarrow 9$.

In order to address the occlusion problem more efficiently and improve the convergence speed of the model, recent studies have incorporated the matching loss into the multi-iteration L1 loss. The updated loss functions are as follows:

- **Matching Loss:** GMFlowNet [86] uses the negative likelihood of P_c to compute the matching regions between the 4D cost volume and the coarse flow. M_c^{gt} represents the ground truth matching, P_c denotes matching confidence, and L_M is the definition of matching loss.

$$L_M = -\frac{1}{|M_c^{gt}|} \sum_{(\hat{i},\hat{j}) \in M_c^{gt}} \log P_c(\hat{i}, \hat{j}) \tag{11}$$

- **Occlusion Detection Loss** (M_occ): TransFlow [90] employed an extra occlusion detection loss to identify the discrepancy of pixels between frames by examining the forward consistency. In (12), the photometric distance is measured by f_D, while the estimated flow is represented by f.

$$M_{\text{occ}} = f_D(Frame1 - Frame2(x + f)) \tag{12}$$

4.2 Datasets

Numerous datasets have been developed to evaluate and compare optical flow estimation techniques. The Middlebury dataset [96], introduced at ICCV 2007, is an early benchmark to assess algorithm performance. The KITTI2012 [58] and KITTI2015 [73] datasets provide real-world scenarios for testing, including challenging driving scenes. The Sintel dataset [57] offers a diverse range of synthetic scenes derived from an animated movie, while the FlyingChairs dataset [59] and its successor, FlyingChairs2 [97], consist of synthetic images with random objects and motion of chairs. The FlyingThings3D dataset [98] introduces a large-scale dataset with ground truth disparity and optical flow, while HD1K [99] focuses on high-resolution images. VIPER [100] and ChairssdHom [101] offer additional synthetic datasets with challenging scenarios and various object motions. AutoFlow [102] provides a large-scale dataset for learning optical flow, which contains a wide range of synthetic scenes. MOVI-F is based on the dataset introduced in [103], which consists of intricate objects and a considerable amount of motion blur. These datasets collectively enable robust evaluation of optical flow estimation algorithms, ensuring their effectiveness and generalization across various conditions and applications. The techniques discussed in this research article make use of several datasets, such as FlyingChairs, FlyingThings, Sintel, KITTI2015, and HD1K. Xiuchao Sui et al. demonstrated a comparative generalization of pre-training with FlyingThings3D [98] in CRAFT [85] using the Autoflow dataset [102]. Table 3 provides a summary of the key datasets used in optical flow estimation, including their total size and the splits between trains and tests. A detailed discussion on datasets is also provided.

- **FlyingChairs (C)**: Dosovitskiy et al. proposed Flownet [59] to train this network and also proposed a dataset named FlyingChairs. It is a one-view optical flow dataset generated

Table 3 Summary of key datasets used in optical flow estimation, including their total size and train/test splits

Dataset	Total	Train	Test	Resolution	Remarks
FlyingChairs (C)	22,872	22,232	640	512×384	964 training scenes
FlyingThings3D (T)	26,066	21,818	4,248	960×540	2,247 distinct scenes
MPI-Sintel (S)	1,628	1,064	564	1024×436	25 distinct scenes, 3 versions
KITTI2015 (K)	400	200	200	1242×375	200 different scenes
HD1K (H)	1,040	840	200	2560×1080	50 real-world sequences

synthetically using the rendering of chairs over the images taken from the flicker. This dataset includes 22,872 training image pairs and 640 testing image pairs. This dataset contains 964 different types of training scenes, with images and a dense ground truth flow of resolution 512 × 384.

- **FlyingThings3D (T)**: To address the limitation of data samples and single view flow, Dosovitskiy et al. created synthetic datasets, namely FlyingThings3D, Monkaa, and Driving, as reported in [98]. The inclusion of 3D in Flyingthings3D implies 3D rendering of chairs in images. The dataset includes the samples for optical flow, disparity, and scene flow to train the CNN-based network. This dataset contains 25,000 pairs of stereo images with a resolution of 960 x 540 pixels, along with ground truth and images of 2,247 distinct scenes. The number of image pairs used for training and testing is 21,818 and 4,248, respectively.
- **MPI-Sintel or Sintel (S)**: Butler et al. derived this dataset in 2012 from an open source 3d animated movie called Sintel [57]. This collection of synthetic scenes is highly valued for its ability to simulate a variety of real-world scenarios that involve rapid movements, non-rigid deformations, and complex multilevel scenes. This dataset consists of 25 distinct scenes, with 1064 pairs of training images and 564 pairs of test images. The Sintel dataset has three versions with varying pixel counts, including albedo, clean pass, and final pass. Albedo is the most basic collection and does not include any extra features. The Clean pass provides a basic rendering of the scenes, while the Final pass adds more intricate details, such as motion blur, defocus blur, and atmospheric effects. The results are published for the clean pass and the final pass.
- **KITTI2015 (K)**: A realistic optical flow dataset is needed for fine-tuning, so Geige et al. created an early version of KITTI2012 [58]. This dataset consists of 389 frames of a car in motion. Menze and his team then extended this dataset, naming it KITTI2015 [73]. They added 3D CAD models to the dynamic scenes for all vehicles in motion. KITTI2015 has 400 frames of 200 different scenes, with 200 training frames and 200 test frames of 1242 x 375 resolution. The evaluation results are published for all EPE and F1 metrics in the KITTI2015 dataset.
- **HD1K (H)**: As we have seen, the earlier datasets used for pre-training and fine-tuning are not of high resolution. Therefore, the authors used the HD1K dataset [99] introduced by Kondermann et al. to fine-tune the optical flow estimation. It consists of 1,000 unique high-resolution frames with a resolution of 2560 x 1080 derived from 50 sequences of real-world dynamic scenes, which include fast motion, complex features, and non-rigid deformations, among others. The dataset is split into 800 training frames and 200 test frames.

The choice of benchmark datasets for quantitative analysis in Section 4.3 is based on several criteria:

- **Generalization:** The FlyingChairs dataset was chosen to illustrate the generalized behavior of the techniques. It offers a synthetic environment with random objects and motion, aiding in the assessment of how well the models generalize to unseen data.
- **Benchmarking:** The Sintel and KITTI2015 datasets are well-established benchmarks in the optical flow community. These datasets are frequently used by researchers to compare the performance of their techniques. The availability of test sets on their respective servers facilitates standardized evaluation and comparison of results across different studies.
- **Diversity and Challenge:** The selected datasets provide a diverse range of scenarios, from synthetic scenes in Sintel to real-world driving scenes in KITTI2015. This diversity

ensures that techniques are tested under various conditions, including different types of motion, occlusions, and lighting conditions.

- **Historical Significance:** These datasets have been extensively used in the literature, offering a historical perspective on the evolution of optical flow techniques. Using these well-established benchmarks, we can contextualize our findings within the broader scope of existing research.

By selecting these datasets, we aim to deliver a comprehensive evaluation of the techniques, ensuring that they are tested for both generalization and performance on widely accepted benchmarks.

4.3 Quantitative analysis

In this section, we assess the effectiveness of the techniques in the datasets we have already discussed. We are evaluating the performance of FlyingChairs (C) and Sintel (S), both versions of the Clean and Final pass and KITTI2015 (K) datasets. To measure performance, we use the endpoint error metric (EPE) for all datasets and the F1-all metric for the KITTI2015 dataset to show the percentage of invalid pixels. Training recipes and other details are discussed in a separate subsection. After evaluating the techniques, we also present the results of CNN-based techniques, including transformer techniques. The three versions of GMA, namely attention with position bias, joint position and content-wise attention, and aggregation-based attention, were reported by Shihao Jiang et al. [78]. However, in this paper, we present only the results for attention with position biasing. GMFlow [87] has two versions to assess performance based on inference speed, which are presented with and without refinement. The results of GMFlow with the refinement stage are shown in Table 4. In GMFlowNet, the authors discuss the results with two versions of the pipeline: one with multi-head axial attention, used for publishing the results on the Sintel benchmark, and the other without. FlowFormer [88] also has two versions in which the authors used the tensor with or without the PyTorch tile function. After thorough analysis and testing, we present only the most optimal results of these techniques. The authors of TransFlow [90] have not yet produced the code and checkpoints, so we cannot provide a complete analysis of TransFlow. NeuFlow

The experimental results in Table 4 clearly demonstrate the superiority of Transformer-based optical flow estimation architectures over traditional CNN-based techniques. This superiority is evident in consistently lower EPE and higher F1-all scores across various training datasets and benchmarks. For example, FlowFormer++ and MemFlow record an EPE of 0.90 on the Sintel Clean dataset, exceeding the top CNN-based model. Likewise, on the KITTI2015 dataset, Transformer-based models such as GMFlowNet achieve an F1-all score of 34.12 %, in contrast to 36.92 % for the CNN-based GMA model. The capability of transformers to capture long-range dependencies and intricate motion patterns makes them particularly well suited for optical flow estimation.

The choice of training data has a significant impact on the performance of both Transformer-based and CNN-based models. In Tables 4 and 5 the techniques marked with '*' are Transformer-based, and bold numbers indicate the highest result in each benchmark. C + T indicate training of the techniques in the FlyingChairs (C) dataset and fine-tuning in the FlyingThings3D (T) dataset. '+ S + K + H' denotes fine-tuning on a combination of Sintel (S), KITTI2015 (K), and HD1K (H). When trained on diverse datasets (C + T + S / K (+ H)), the models often excel in generalizability, performing well on the Sintel and KITTI2015 test sets. This finding emphasizes the importance of comprehensive training regimens in

Table 4 Experimental results of transformer and CNN-based techniques on FlyingChairs [59], Sintel [57] and KITTI2015 [73] datasets

Training data	Techniques	Chairs (train)	Sintel (train)		KITTI2015 (train)	
			Clean	Final	EPE	F1-all(%)
C	GMA [78]	0.79	2.32	4.11	10.33	36.92
	Flow1D* [84]	0.88	2.82	5.42	14.33	45.68
	CRAFT* [85]	0.76	**2.22**	**4.05**	10.67	**34.11**
	GMFlowNet* [86]	0.87	2.35	4.24	**9.13**	34.12
	GMFlow* [87]	0.92	3.22	4.43	17.82	56.14
	FlowFormer* [88]	**0.63**	2.69	4.41	10.15	34.45
	FlowFormer++* [89]	–	2.44	4.53	–	–
C+T	FlowNet 2.0 [29]	–	2.02	3.54	10.08	30.0
	PWCNet [30]	–	2.55	3.93	10.35	33.7
	LiteFlowNet [32]	–	2.48	4.04	10.39	28.5
	HD3 [104]	–	3.84	8.77	13.17	24.0
	LiteFlowNet2 [60]	–	2.24	3.78	8.97	25.9
	VCN [62]	–	2.21	3.68	8.36	25.1
	DICL-Flow [63]	–	1.94	3.77	8.7	23.6
	RAFT [34]	–	1.43	2.71	5.04	17.4
	ScopeFlow [64]	–	3.26	1.83	9.78	26.6
	MaskFlowNet [65]	–	2.25	3.61	–	23.1
	SCV [105]	–	1.29	2.95	6.80	19.3
	GMA* [78]	1.18	1.3	2.73	4.68	17.14
	Flow1D* [84]	1.27	1.98	3.27	6.69	22.95
	CRAFT* [85]	1.11	1.28	2.73	5.18	17.50
	GMFlowNet* [86]	**0.95**	1.14	2.71	4.42	16.69
	GMFlow* [87]	1.26	1.49	2.95	11.59	35.62
	FlowFormer* [88]	0.99	1	2.39	4.2	14.88
	FlowFormer++* [89]	–	**0.90**	**2.30**	**3.93**	**14.13**
	TransFlow* [90]	–	0.93	2.33	3.98	14.4
	NeuFlow* [91]	1.7	1.56	3.63	11.9	32.6
C+T+ S/K (+H)	FlowNet [59]	–	3.20	4.83	7.31	–
	SpyNet [31]	–	3.17	4.32	–	–
	FlowNet 2.0 [29]	–	1.45	2.01	2.30	6.8
	PWCNet [30]	–	1.71	2.34	1.50	5.3
	LiteFlowNet [32]	–	1.35	1.78	1.62	5.6
	HD3 [104]	–	1.87	1.17	1.31	4.1
	LiteFlowNet2 [60]	–	1.30	1.62	1.47	4.8
	VCN [62]	–	1.66	2.24	1.16	4.1
	IRR [33]	–	1.92	2.51	1.63	5.3
	DICL-Flow [63]	–	1.11	1.60	1.02	3.6
	RAFT [34]	–	0.77	1.27	1.61	2.86
	LiteFlowNet3 [61]	–	1.32	1.76	1.26	3.82
	SCV [105]	–	0.79	1.70	0.75	2.1

Table 4 continued

Training data	Techniques	Chairs (train)	Sintel (train)		KITTI2015 (train)	
			Clean	Final	EPE	F1-all(%)
	GMA* [78]	1.26	0.61	1.06	0.56	1.22
	Flow1D* [84]	–	0.84	1.25	–	1.6
	CRAFT* [85]	1.19	0.60	1.06	1.49	5.16
	GMFlowNet* [86]	**1.02**	0.59	0.91	0.63	4.85
	GMFlow* [87]	1.52	0.94	1.27	2.06	7.57
	FlowFormer* [88]	1.10	0.44	0.65	0.60	1.30
	FlowFormer++* [89]	–	**0.40**	**0.60**	0.57	1.16
	TransFlow* [90]	–	0.42	0.69	**0.49**	**1.05**
	NeuFlow* [91]	1.74	0.97	1.44	2.09	9.63
	MemFlow* [92]	1.1	**0.40**	0.62	0.78	1.92

Table 5 Experimental results: on test set of Sintel and KITTI2015

Training data	Techniques	Sintel (test)		KITTI2015 (test)
		Clean	Final	F1-all(%)
C+T+S/K (+H)	FlowNet	–	–	–
	SpyNet [31]	6.64	8.36	35.07
	FlowNet 2.0 [29]	4.16	5.74	11.48
	PWCNet [30]	3.45	4.6	7.72
	LiteFlowNet [32]	4.54	5.38	9.38
	HD3 [104]	4.79	4.67	6.55
	LiteFlowNet2 [60]	3.48	4.69	7.74
	VCN [62]	2.81	4.40	6.30
	IRR [33]	3.84	4.58	7.65
	DICL-Flow [63]	2.12	3.44	6.31
	RAFT [34]	1.94	3.18	5.10
	LiteFlowNet3 [61]	2.99	4.45	7.34
	MaskFlowNet [65]	2.52	4.17	6.10
	SCV [105]	–	–	–
	GMA* [78]	1.39	2.47	5.15
	Flow1D* [84]	2.24	3.81	6.27
	CRAFT* [85]	1.45	2.42	4.79
	GMFlowNet* [86]	1.39	2.65	4.89
	GMFlow* [87]	–	–	–
	FlowFormer* [88]	1.16	2.09	4.68
	FlowFormer++* [89]	1.07	1.94	4.52
	TransFlow* [90]	**1.06**	2.08	4.32
	MemFlow* [92]	1.08	**1.84**	**3.88**

the development of robust optical flow models. The following important observations are mentioned below:

- **MemFlow** [92]: MemFlow outperforms all other methods on the Sintel final and KITTI2015 datasets, highlighting the superior performance of transformer-based flow estimation pipelines compared to CNN-based ones.
- **FlowFormer/FlowFormer++** [88, 89]: These architectures exhibit consistently strong performance, particularly in the Sintel (S) dataset. The utilization of the transformer in the creation of cost volumes demonstrates the effectiveness of attention mechanisms in aligning visual characteristics.
- **GMFlowNet** [86]: GMFlowNet performs well when trained in FlyingChairs (C) and generalizes effectively to datasets such as KITTI2015 (K). Integration of the transformer with its geometric matching module probably improves its flexibility.
- **TransFlow** [90]: This technique's strength lies in direct flow regression. It performs consistently on the datasets, achieving a notably low EPE on the KITTI2015 (K) benchmark. This result suggests promise in its technique for direct flow prediction.

In summary, the choice of dataset greatly affects the comparison of techniques. FlowFormer/FlowFormer++ consistently performs better than other techniques when dealing with challenging motions and occlusions in the Sintel dataset (both clean and final versions). However, MemFlow, TransFlow and GMFlowNet show promising results in the KITTI2015 (K) dataset, suggesting their potential value for automotive-related applications. The evaluation analysis is conducted on training samples from the FlyingChairs, Sintel, and KITTI2015 datasets. Furthermore, the test results of the Sintel and KITTI2015 datasets are also included.

4.4 Experimental setup

Estimation of optical flow using either CNN or a transformer requires laborious training. Annotating the data for the optical flow is a complex task, as mentioned on page Section 4.2. The training samples available are mainly synthetic. Therefore, to publish benchmark results, many researchers have chosen to pre-train the models on C and T datasets, and then fine-tune them on the target benchmark data. The FLow1D, GMA, GMFlow, NeuFlow, and MemFlow methods all adhere to the RAFT training technique. CRAFT initially used AutoFlow [102] to compare pre-training in this dataset instead of C and T. FlowFormer is the first technique to introduce the use of a pre-trained Swin-large transformer on the ImageNet dataset to extract features. In FlowFormer++ , MCVA is used to enhance the robustness of the cost-volume encoder through pre-training on the YOUTube-VOS dataset. TransFlow eliminates the need for this laborious training.

The discussed techniques employ a one-cycle learning rate strategy, where the learning rate is initially set to its highest value during the pre-training stage and gradually decreases until the fine-tuning stage. A detailed comparison can be found in Table 6.

Transformer-based techniques have incorporated the multi-iteration L_1 loss originally introduced in RAFT [34]. Some modifications have been made by GMFlowNet [86] and TransFlow [90], which have also incorporated the matching loss (11) and the occlusion detection loss (12) together with the multi-iteration loss L_1.

Table 6 outlines the experimental setup for each technique, summarizing the datasets before training and fine-tuning, batch sizes, iterations, learning rates, and the type of loss functions used. The table lists the dataset used, the number of iterations, the batch size, the learning rate, the training loss, and the GPUs used for each technique. P-T and F-T refer

Table 6 Experimental and implementation details: The table provides a comprehensive overview of experimental setups and hyperparameter configurations of selected techniques

Techniques	Datasets		Iterations		Batch size		Learning rate	Training loss	GPUs used
	P-T	F-T	P-T	F-T	P-T	F-T			
GMA [78]	C, T	T, S, K, H	120k, 120k	120k for S, 50k for K	8, 6	6	2.5e-4 for C then 1.25e-4 for rest	Multi-iteration L_1 loss	Two 2080Ti
Flow1D [84]	C, T	T, S, K, H	100k, 100k	100k for S, 50k for K	6	6	4e-4 for C, 1.2e-4 for T and S+T+K+H, and 1e−4 for K	Multi-iteration L_1 loss	32G V100 GPU
CRAFT [85]	C, T	T, S, K, H	120k, 120k	120k for S, 50k for K	8, 6	6	4e-4 for C, 1.2e-4 for S and 1e-4 for K	Multi-iteration L_1 loss	–
GMFlowNet [86]	C, T	T, S, K, H	120k, 160k	120k for S, 50k for K	10, 6	–	–	Matching loss + Multi-iteration L_1 loss	–
GMFlow [87]	C, T	T, S, K, H	100k, 200k	800k for T, 120k for S and 50k for K	16	8	4e-4 for C then 2e-4 for rest	Multi-iteration L_1 loss	Four A100 (40GB)
FlowFormer [88]	C, T	T, S, K, H	120k, 120k	120k for S, 50k for K	8, 6	6	2.5e-4 for C then 1.25e-4 for rest	Multi-iteration L1 loss	–
FlowFormer++ [89]	YouTube-VOS [106], C, T	T, S, K, H	50k, 120k	120k for S, 50k for K	24,8,6	6	5e-4 for YouTube-VOS, 2.5e-4 for FlyingChairs and 1.2e-4 for others	MSE Loss	–

Table 6 continued

Techniques	Datasets		Iterations		Batch size		Learning rate	Training loss	GPUs used
	P-T	F-T	P-T	F-T	P-T	F-T			
TransFlow [90]	S,K	S,K	–	140k	–	6	1e-4 for pre-training, then 12.5e-5 for rest	Occlusion detection loss + Multi-iteration L_1 loss	–
NeuFlow [91]	C,T	T, S,K, H	100k	120k	42	56	2e-4 for pre-training, then 1e-4 for rest	Multi-iteration L_1 loss	RTX 2080
MemFlow [92]	C, T	T, S, K , H	600k	600k	8	8	2.1e-4 for pre-training, then 1.75e-4 for rest	Multi-iteration L_1 loss	Two NVIDIA A100

to pre-training and fine-tuning, respectively, with C, T, S, K, and H representing different datasets. Here are some notable points:

- GMA [78] and FlowFormer [88] share similarities in their setup, including learning rate and training loss, but diverge in performance, highlighting the impact of architectural differences.
- GMFlow [87] employs the largest batch sizes during pre-training, and is also the only technique using Four A100 (40GB) GPUs, suggesting a computationally intensive model.
- CRAFT [85], although it has similar training configurations to others, manages to outperform in key metrics, proving the efficacy of its unique components, such as the Semantic Smoothing Transformer.
- MemFlow [92] utilizes two NVIDIA A100 GPUs and strikes a balance between computational load and accuracy by employing a memory buffer to augment motion features.

4.5 Computational efficiency

This section emphasizes the crucial aspects of the techniques, including run-time, parameters, and FLOPs. The bar graph for computation shown in Fig. 16 provides a useful comparison of these Transformer-based techniques, which are crucial for scenarios where real-time deployment is necessary. The computation analysis is performed on two RTX-3090 GPUs, each with 24 GiB of VRAM, and the testing is conducted in an environment with Pytorch 1.12 and CUDA 11.1. The complexity and learning capacity of the models is reflected by the 'Number of Parameters' in the first chart. Larger models have a greater capacity to capture complex patterns, but also carry the risk of overfitting and require more computational resources. The FLOPs chart provides information on the computational load of the models, indicating how many floating point operations they can perform per second. Models with lower FLOPs are more efficient, making them ideal for environments with limited resources. The chart labelled 'Inference Time' provides information about the speed of the models by displaying the duration it takes for each model to make predictions on new data. This measure is particularly important for applications that require rapid processing, as shorter inference times offer advantages in such cases. TransFlow [90], SCV [105] computation is not analysed due to the unavailability of the code and checkpoints. Table 7 provides a detailed summary of the number of parameters, FLOPs, and inference time for each technique reviewed, complementing the visual comparison in Fig. 16.

Among the Transformer-based techniques, the following observations can be made:

- **GMA** [78] demonstrates a moderate parameter count of 5.88M with 2540.32G FLOPs, resulting in an inference time of 436.70 ms. This suggests a trade-off between computational load and performance.
- **Flow1D** [84] and **GMFlow** [87] show lower computational footprints with 5.73M and 4.71M parameters, respectively. However, their FLOPs (3000.00G for both) and inference times (500.00 ms for both) are higher than those of GMA, indicating lower computational efficiency.
- **CRAFT** [85] stands out with a relatively high parameter count of 6.30M and the highest FLOPs at 4204.72G, culminating in the longest inference time of 1379.29 ms, raising concerns about its suitability for time-sensitive applications.
- **GMFlowNet** [86], with 8.68M parameters, displays a balanced profile with its FLOPs and inference time positioned between the aforementioned models.
- **FlowFormer** [88] and **FlowFormer++** [89] present the largest models in terms of parameters, with 16.16M and 16.15M, respectively. Their FLOPs (4000.00G for both) and

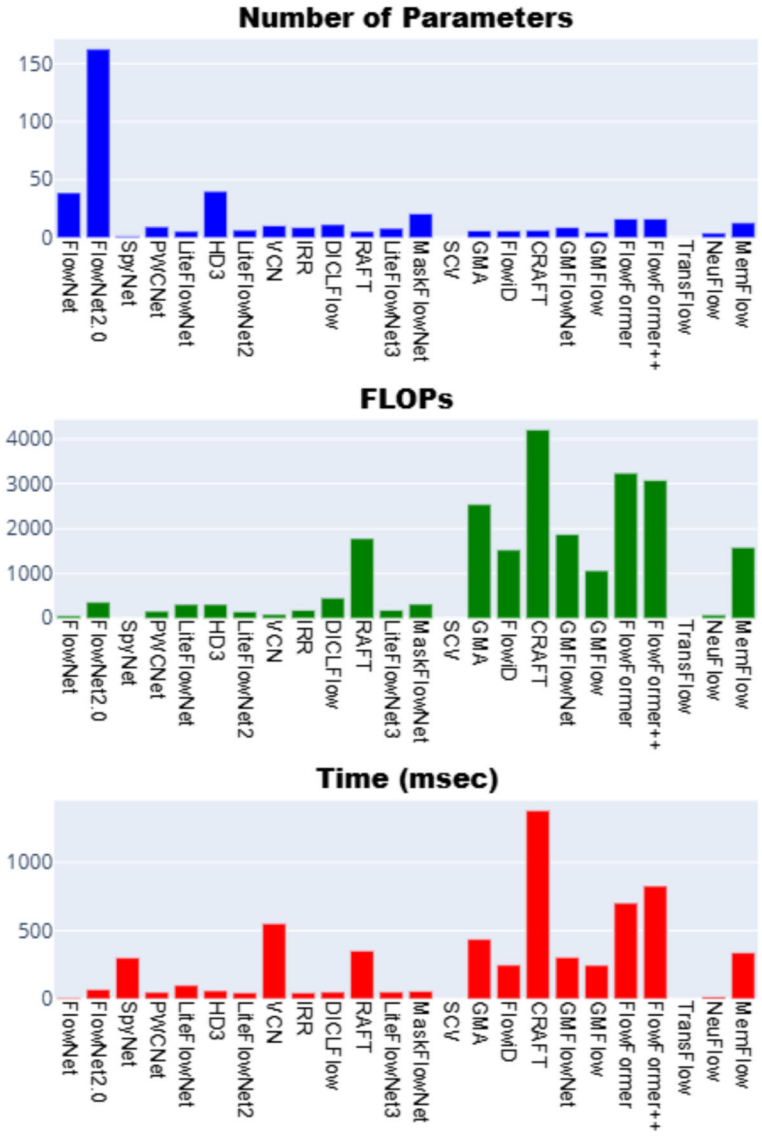


Fig. 16 This visual representation illustrates a comparative analysis of Transformer-based optical flow techniques, showcasing their respective parameters, FLOPs, and inference time. The chart consists of bars divided into three sections, each representing the number of parameters utilized, the computational cost (FLOPs), and the duration required to obtain the results. The length of each bar corresponds to the magnitude of the value it represents, allowing a clear distinction between resource-intensive techniques and those that offer faster performance

inference times (1000.00 ms for both) are higher, indicating a considerable demand for computational resources.

- **NeuFlow** [91] exhibits the lowest computational footprint among the reviewed techniques, with only 3.85M parameters, 76.21G FLOPs, and an inference time of 76.21 ms.

Table 7 Architectural details and computational metrics: this table provides a comprehensive overview of the number of parameters, FLOPs, and inference time for each reviewed technique

Technique	Parameters (M)	FLOPs (G)	Inference time (ms)	Architecture
GMA [78]	5.88	2540.32	436.70	Attention within GRU module
Flow1D [84]	5.73	3000.00	500.00	1D attention transformer
CRAFT [85]	6.30	4204.72	1379.29	Semantic smoothing transformer
GMFlowNet [86]	8.68	3500.00	600.00	Modified swin transformer
GMFlow [87]	4.71	3000.00	500.00	ViT
FlowFormer [88]	16.16	4000.00	1000.00	Siamese encoder transformer
FlowFormer++ [89]	16.15	4000.00	1000.00	Siamese encoder transformer
TransFlow [90]	–	–	–	ViT
NeuFlow [91]	3.85	76.21	76.21	Cross-attention transformer
MemFlow [92]	12.70	1581.62	337.99	Memory buffer block using flash attention [95]

However, its performance on the Sintel and KITTI datasets is suboptimal compared to other techniques.

- **MemFlow** [92] presents a balanced profile with 12.70M parameters, 1581.62G FLOPs, and an inference time of 337.99 ms. It achieves competitive results on the Sintel and KITTI benchmarks while maintaining a relatively efficient computational footprint.

In summary, although Transformer-based techniques like GMFlow [87] demonstrate impressive performance in terms of model size and speed, models such as FlowFormer [88] and FlowFormer++ [89] impose a significant computational burden, which can hinder their deployment in resource-constrained environments. NeuFlow [91] offers the most computationally efficient solution, but at the cost of performance, while MemFlow [92] strikes a balance between computational efficiency and competitive results. This analysis highlights the importance of finding a careful trade-off between the complexity of the model and the efficiency of inference, as this balance is crucial for the practical applicability of optical flow estimation in real-world scenarios.

5 Visual comparison and discussion

This section illustrates the visual analysis of Transformer-based optical flow techniques using a color wheel to underline the advantages and limitations of flow maps. The color-encoded wheel is described in the following subsection.

5.1 Colour-coded visualization for optical flow:

Visualizing optical flow presents a challenge due to the 2D fields for the u and v vectors it contains. Baker et al. [96] proposed a colour-encoded wheel to visualize the dense flow. Figure 17 uses a colour wheel to encode the direction of motion vectors: hues represent directions, with

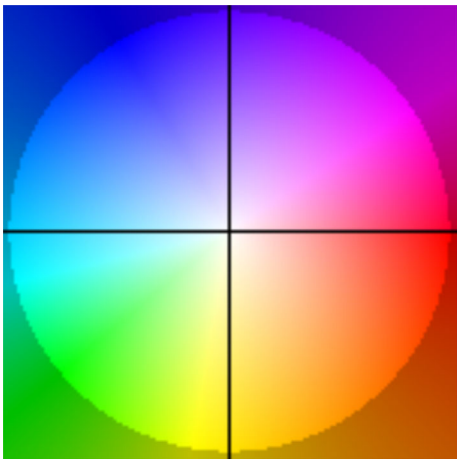


Fig. 17 Colour wheel for optical flow visualization: the visualization displays both the direction and magnitude of motion. The direction is represented by the hue of the color, while the length represents the magnitude

specific colours assigned to specific angles (e.g., red indicating rightward movement, green downward, etc.), ensuring that a full 360° spectrum of motion can be depicted. The magnitude of these vectors, reflecting the speed of motion, is conveyed through the colour's intensity or saturation, with brighter and more saturated colours denoting faster movements and paler shades indicating slower ones. This technique allows for an immediate and intuitive grasp of the directionality and velocity of motion across the image, facilitating a direct comparison between different optical flow estimation techniques through their visual outputs.

5.2 Visual analysis

Figures 18 and 19 show the two consecutive frames, ground truth, and predicted flow of the techniques MemFlow, NeuFlow, FlowFormer ++, FlowFormer, GMFlow, GMFlowNet, CRAFT, Flow1D and GMA. The training samples from the clean and final versions of the Sintel and KITTI2015 datasets are located in the first, second, and third columns. However, testing samples from the Sintel and KITTI2015 datasets can be found in the fourth, fifth, and sixth columns. Specifically, we present the visual results of the last stage of training, which incorporates the techniques C + T + S/K (+H). The first, second and third rows of Fig. 18 correspond to Frame 1, Frame 2, and Ground Truth, respectively, from the Sintel [57] and KITTI2015 [73] datasets, and the rest are the mentioned techniques. Our analysis is based on a visual inspection of representative images, with a particular focus on specific areas highlighted by colored boxes. These boxes indicate notable features or discrepancies in optical flow compared to ground truth. Here are the main findings we have made.

- **FlowFormer++** [89]: The technique demonstrates a strong ability to track flow in larger homogeneous regions, as shown by the green boxes. However, it deviates noticeably from the ground truth in areas with complex motion patterns or near object boundaries, as indicated by the red boxes. The presence of blue boxes suggests that the model may struggle to capture fine-grained details in densely textured regions.
- **FlowFormer** [88]: Similar to FlowFormer++, this model performs well in open and less cluttered spaces but has limitations in capturing complex movements, especially in areas with sudden changes in flow direction or speed. Blue boxes once again highlight areas where subtle flow nuances may not be fully captured.

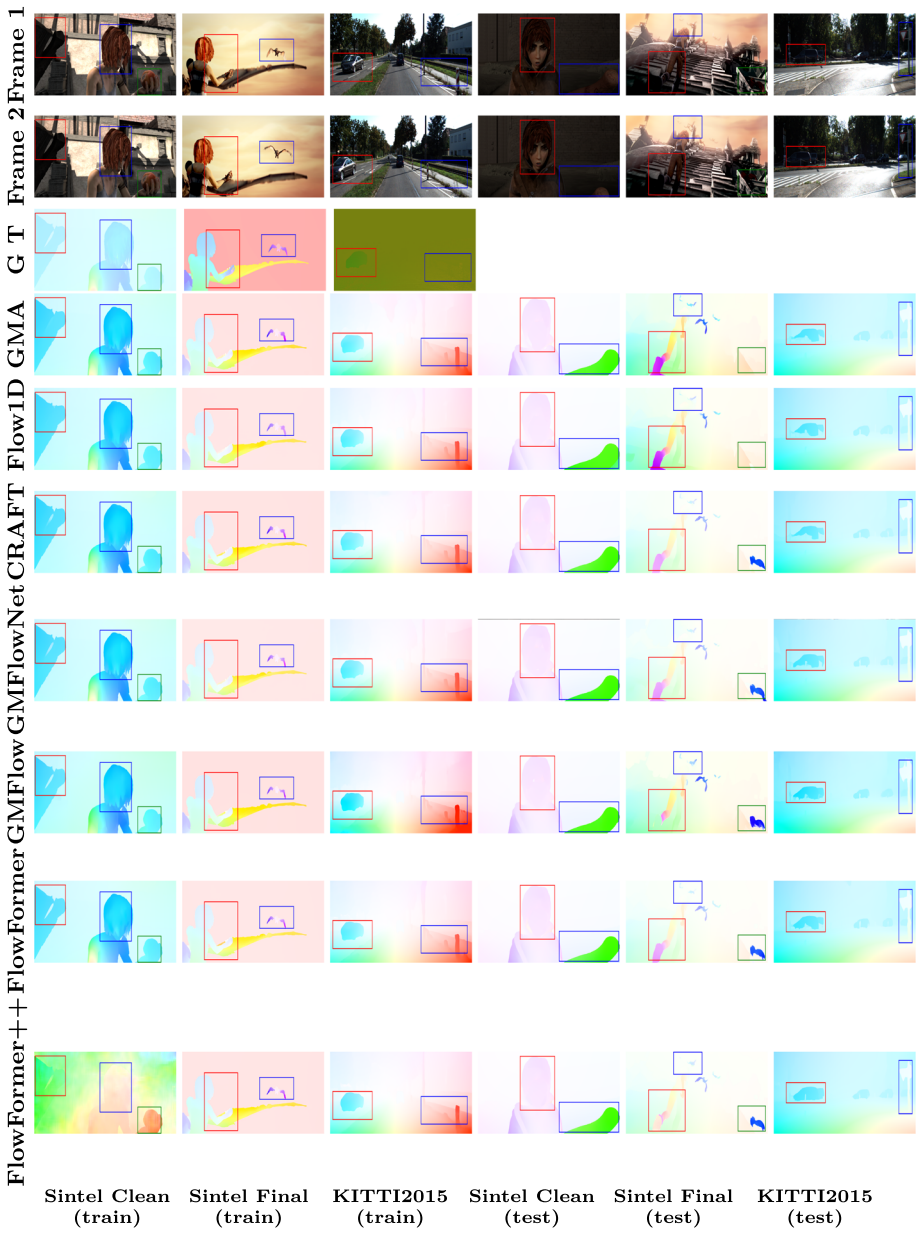


Fig. 18 Visual comparison of training and testing samples from the Sintel [57] and KITTI2015 [73] datasets. G T refers to the Ground Truth

- **GMFlow** [87]: This method showcases a balanced performance, effectively capturing the overall structure of the true data in various aspects. However, in complex situations, such as those indicated by the moderate number of red boxes, difficulties may occur. However, the detection of blue boxes indicates that it effectively handles intricate motion details.

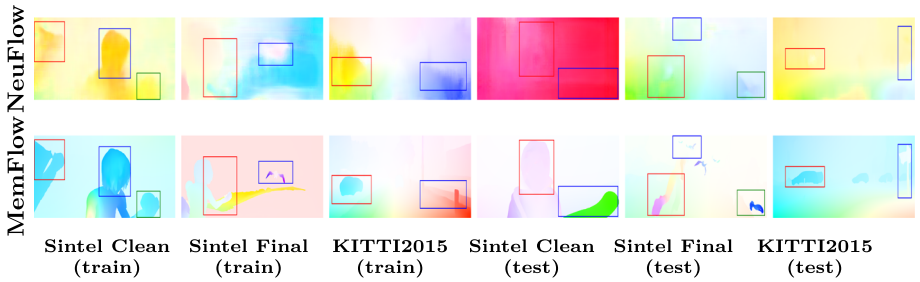


Fig. 19 Comparison of train and test results obtained from NeuFlow [91] and MemFlow [92]

- **GMFlowNet** [86]: GMFlowNet exhibits a performance closely aligned with GMFlow, indicating similar strengths and weaknesses. It effectively handles large, uniform flow fields but occasionally struggles with complex motion interfaces and can be observed in blue boxes.
- **CRAFT** [85]: CRAFT novel cross-attention module is capable of identifying the intricate and inflexible movement exhibited by the red box in the KITTI2015 test sample. Significant displacement in the Sintel clean and final samples indicates an inability to detect subtle motion.
- **Flow1D** [84]: The ability of Flow1D to identify localized characteristics in one dimension enables it to capture the complexities of flow across different textures effectively. Sintel's (final) test version indicates that the detection of fast-moving objects is inaccurate, as shown by the green box.
- **GMA** [78]: GMA's ability to detect occlusion using aggregated motion features allows it to accurately identify large displacement flows. When using samples from the Sintel dataset for training and testing, GMA successfully detects flying objects indicated by the blue box. However, it struggles to detect fast-moving objects, as observed in the green box in the Sintel (final) test sample.
- **NeuFlow** [91]: NeuFlow's visual results reveal significant discrepancies compared to the ground truth, particularly in regions with complex motion patterns and object boundaries, as indicated by the numerous red boxes. The model struggles to capture fine-grained details and subtle motion nuances, as evidenced by the presence of blue boxes across various samples.
- **MemFlow** [92]: MemFlow demonstrates a strong ability to capture the overall structure of the flow fields, effectively handling both large, uniform regions and areas with intricate motion details. The model shows a balanced performance across different datasets and scenarios, with only occasional difficulties in highly complex situations, as indicated by the relatively few red boxes.

In summary, FlowFormer++ and FlowFormer are proficient in capturing flow in extensive, uniform areas; however, they face challenges with complex motion patterns and detailed features. GMFlow and GMFlowNet offer balanced performance, adeptly capturing the overall structure of flow fields while sometimes struggling with detailed motion interfaces. CRAFT and Flow1D display specific strengths in identifying distinct motion traits but may overlook subtle movements or have problems with fast-moving objects. GMA is effective in detecting significant displacement flows and occlusions, though it can struggle with fast-moving objects. The visual results of NeuFlow show marked discrepancies and difficulty in capturing

detailed features, whereas MemFlow displays strong, balanced performance across various scenarios, effectively managing both uniform areas and intricate motion details.

5.3 Analysis, limitations and discussion

The findings presented in Table 4 clearly illustrate the advantage of Transformer-based architectures for optical flow estimation over traditional CNN-based methods. This advantage is reflected in consistently lower EPE and higher F1-all scores across different training datasets and benchmarks. Transformer-based models deliver superior accuracy in optical flow estimation compared to conventional CNN-based models due to several critical factors:

1. **Attention Mechanisms**: Transformers leverage self-attention mechanisms that allow them to capture long-range dependencies and contextual information across the entire image. This capability is particularly beneficial for optical flow estimation, where understanding the global context and relationships between distant features is crucial. In contrast, CNNs are limited to processing local contiguous sections of the input space, restricting their ability to capture such long-range dependencies. For instance, our results show that the FlowFormer++ model achieves an EPE of 0.90 on the Sintel Clean dataset, outperforming the best CNN-based model.
2. **Sequence Processing**: Originally designed for natural language processing, Transformers excel at handling sequential data. In the context of optical flow, this translates to effectively managing sequences of images and capturing the temporal dynamics between frames. This ability to model temporal relationships enhances the accuracy of motion estimation. For example, GMFlowNet achieves an F1-all score of 34.12
3. **Hierarchical Feature Representation**: Transformers can build hierarchical representations of features through their multi-head attention layers. This hierarchical representation allows for a more nuanced understanding of motion patterns and object interactions within the scene, leading to more accurate optical flow predictions.
4. **Flexibility and Generalization**: Transformer-based models have shown remarkable flexibility and generalizability. Techniques such as MemFlow, FlowFormer and GMFlowNet have demonstrated strong performance across diverse datasets, including synthetic and real-world scenarios. This generalization is attributed to the robust feature extraction and attention mechanisms inherent in Transformer architectures.

Despite their advantages, Transformer-based models also face several inherent limitations and challenges:

1. **Computational Complexity**: Transformers are computationally intensive, requiring significant computational resources for training and inference. The self-attention mechanism, in particular, has quadratic complexity with respect to the input size, leading to high memory and computational demands. This complexity can hinder the deployment of Transformer-based models in resource-constrained environments, such as mobile devices or embedded systems. For example, MemFlow and FlowFormer++ require four A100 (40GB) GPUs for training, highlighting its computational demands.
2. **Training Data Requirements**: Transformer-based models often require large amounts of training data to achieve optimal performance. The need for extensive pre-training on large datasets, followed by fine-tuning on specific tasks, can be a barrier to their practical application, especially in scenarios where labeled data is scarce.

3. **Interpretability**: The inner workings of Transformer models are less interpretable compared to CNNs. Understanding the mechanisms of attention and the interactions between different layers can be challenging, making it difficult to diagnose and address issues related to model performance and behavior.
4. **Optimization and Fine-tuning**: Fine-tuning Transformer-based models for specific tasks requires careful optimization of hyperparameters, such as learning rates, batch sizes, and training schedules. This process can be time-consuming and requires expertise in model training and optimization.

To address the computational demands and ensure the practical applicability of transformer-based optical flow techniques, several strategies can be employed.

1. **Model Compression and Pruning**: Techniques such as model compression, pruning, and quantization can be used to reduce the size and complexity of Transformer models, making them more suitable for deployment on resource-constrained devices.
2. **Specialized Hardware**: The development of specialized hardware accelerators, such as GPUs and TPUs optimized for transformer computations, can significantly alleviate the computational burden. These hardware solutions can enable real-time performance and energy-efficient implementation.
3. **Hybrid Architectures**: Combining the strengths of Transformers and CNNs in hybrid architectures can help balance computational efficiency and representational power. Using CNNs for local feature extraction and Transformers for global attention, hybrid models can achieve a harmonious blend of performance and efficiency.
4. **Efficient Training Techniques**: Employing efficient training techniques, such as transfer learning, knowledge distillation, and one-cycle learning rate strategies, can reduce training time and computational resources required for Transformer-based models.

In summary, while Transformer-based models offer superior accuracy in optical flow estimation due to their advanced attention mechanisms and sequence processing capabilities, they also present challenges related to computational complexity, interpretability, and training requirements. Addressing these challenges through model optimization, specialized hardware, hybrid architectures, and efficient training techniques can pave the way for the practical and widespread adoption of Transformer-based optical flow techniques in various real-world applications.

6 Conclusion and future work

In this survey, we have presented a systematic overview of recent CNN- and Transformer-based techniques. We discussed a baseline network for learning-based techniques and then used this to show the evolution of the transformer in the flow estimation pipeline. In addition to this, we also provide associated loss functions, metrics, and datasets. We conducted a thorough quantitative and visual examination and a detailed analysis of the state-of-the-art techniques. Moreover, we conducted an in-depth examination of their limitations and outlined the ongoing difficulties encountered in utilizing transformers in optical flow.

Future work in this area could focus on several promising directions. One avenue is the development of attention mechanisms and architectural designs specifically tailored for visual data, with the aim of achieving higher efficiency. Another potential direction

is the investigation of hybrid architectures that combine the strengths of Transformers and CNNs, which could overcome the limitations of each individual approach. Furthermore, the development of specialized hardware and accelerators optimized for efficient transformer computations could significantly alleviate the computational burden associated with these models. Addressing these challenges and exploring these future directions could pave the way for a more practical and widespread adoption of Transformer-based models in various real-world applications.

These results underscore the importance of comprehending Transformer-based methods for optical flow estimation among readers, stimulating further research to explore the capabilities of the Transformer model and improve its real-world implementations.

Author Contributions Nihal Kumar contributed to conceptualization, review of the literature, writing the original draft, and revisions. Anil Singh Parihar (corresponding author) and Om Prakash Verma contributed to the investigation, visualization, validation, and review/editing of the manuscript.

Funding The author acknowledges that no specific funding was received for the conduct of this research.

Data Availability Data sharing is not relevant to this article since no datasets are created in this research.

Declarations

Competing Interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wang M, Xu J, Ke F, Liao L (2023) A encoder-decoder deblurring network combined with high-frequency a priori. Multimed Tools Appl. https://api.semanticscholar.org/CorpusID:266104025
2. Ding X, Huang Y, Li Y, He J (2020) Forgery detection of motion compensation interpolated frames based on discontinuity of optical flow. Multimed Tools Appl 79:28729–28754. https://api.semanticscholar.org/CorpusID:221018490
3. Liang Z, Li Y, Yu R, Zhang K (2023) Aircraft type recognition in 3d-view optical image with contour segmentation. Multimed Tools Appl. https://api.semanticscholar.org/CorpusID:266068440
4. Obeso AM, Benois-Pineau J, García-Vázquez MS, Ramírez-Acosta AA (2021) Visual vs internal attention mechanisms in deep neural networks for image classification and object detection. Pattern Recognit 123:108411. https://api.semanticscholar.org/CorpusID:243472146
5. Singh LK, Pooja HG, Khanna M (2022) Performance evaluation of various deep learning based models for effective glaucoma evaluation using optical coherence tomography images. Multimed Tools Appl 81: 27737–27781. https://api.semanticscholar.org/CorpusID:247817622
6. Chen L, Yang X, Jeon G, Anisetti M, Liu K (2020) A trusted medical image super-resolution method based on feedback adaptive weighted dense network. Artif Intell Med 106:101857. https://api.semanticscholar.org/CorpusID:219438199
7. Furht B, Furht B, Greenblatt J (1996) Motion estimation algorithms for video compression. In: The Springer international series in engineering and computer science. https://api.semanticscholar.org/CorpusID:60683479
8. Azimjonov J, Özmen A, Varan M (2023) A vision-based real-time traffic flow monitoring system for road intersections. Multimedia Tools Appl 1 – 20. https://api.semanticscholar.org/CorpusID:256763598
9. Ali AA, El-Hafeez TA, Mohany YK (2019) An accurate system for face detection and recognition. J Adv Math Comput. https://api.semanticscholar.org/CorpusID:201133526
10. Saabia AAB, El-Hafeez T, Zaki AM (2018) Face recognition based on grey wolf optimization for feature selection. In: International conference on advanced intelligent system and informatics. https://api.semanticscholar.org/CorpusID:52134200

11. Ali AA, El-Hafeez TA, Mohany YK (2019) A robust and efficient system to detect human faces based on facial features. Asian J Res Comput Sci. https://api.semanticscholar.org/CorpusID:187398560
12. Eman M, Mahmoud TM, Ibrahim MM, El-Hafeez TA (2023) Innovative hybrid approach for masked face recognition using pretrained mask detection and segmentation, robust pca, and knn classifier. Sensors (Basel, Switzerland), 23. https://api.semanticscholar.org/CorpusID:260299938
13. Taha M, Mostafa T, El-Rahman TA (2023) A novel hybrid approach to masked face recognition using robust pca and goa optimizer. Sci J Damietta Fac Sci. https://api.semanticscholar.org/CorpusID:266201423
14. Xiao S, Wang Y, Wang Y (2024) Automatic video colorization based on contrastive learning and optical flow. Multimed Tools Appl 1–17. https://api.semanticscholar.org/CorpusID:266727535
15. Abd El-Hafeez T (2010) A new system for extracting and detecting skin color regions from pdf documents. International Journal on Computer Science and Engineering (IJCSE) 9(2):2838–2846
16. Girgis MR, Mahmoud TM, Abd-El-Hafeez T (2007) An approach to image extraction and accurate skin detection from web pages. Int J Comput Eng 1(3):697–705
17. Girgis MR, Mahmoud TM, Abd-El-Hafeez T (2010) A new effective system for filtering pornography images from web pages and pdf files. Int J Web Appl 2:1–13. https://api.semanticscholar.org/CorpusID:4975813
18. El Koshiry AM, Eliwa EHI, Abd El-Hafeez T, Khairy M (2024) Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique. Peer J Comput Sci 10. https://api.semanticscholar.org/CorpusID:268764971
19. Park SK, Chung J-H, Kang T-K, Lim MT (2021) Binary dense sift flow based two stream cnn for human action recognition. Multimed Tools Appl 80:35697 – 35720. https://api.semanticscholar.org/CorpusID:236283523
20. Pandey AK, Parihar AS (2023) A comparative analysis of deep learning based human action recognition algorithms. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp 1–7. https://doi.org/10.1109/ICCCNT56998.2023.10308200
21. Devanne M, Wannous H, Berretti S, Pala P, Daoudi M, Bimbo A (2015) 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. IEEE Trans Cybern 45:1340–1352. https://api.semanticscholar.org/CorpusID:6933411
22. Horn BKP, Schunck BG (1981) Determining optical flow. Artif Intell 17(1):185–203. ISSN 0004-3702. https://doi.org/10.1016/0004-3702(81)90024-2
23. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: International joint conference on artificial intelligence. https://api.semanticscholar.org/CorpusID:2121536
24. Chen Z, Zhang C, Li M (2013) Coarse-to-fine optical flow estimation with image structure tensor. In: 2013 6th International Congress on Image and Signal Processing (CISP), vol 2, pp 741–746. https://doi.org/10.1109/CISP.2013.6745263
25. Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow estimation based on a theory for warping. In: European conference on computer vision. https://api.semanticscholar.org/CorpusID:76390
26. Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on image analysis. https://api.semanticscholar.org/CorpusID:15601477
27. Hiraiwa A, Fuse K, Komatsu N, Komiya K, Ikeda H (1999) Accurate estimation of optical flow for fully automated tracking of moving-objects within video streams. In: 1999 IEEE International Symposium on Circuits and Systems (ISCAS), volume 4, pp 515–519 vol 4. https://doi.org/10.1109/ISCAS.1999.780055
28. Bruhn A, Weickert J, Schnörr C (2005) Lucas/Kanade Meets Horn/Schunck: combining local and global optic flow methods. Int J Comput Vis 61(3):211–231. ISSN 1573-1405. https://doi.org/10.1023/B:VISI.0000045324.43199.43
29. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) FlowNet 2.0: Evolution of optical flow estimation with deep networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1647–1655. https://doi.org/10.1109/CVPR.2017.179
30. Sun D, Yang X, Liu M-Y, Kautz J (2018) PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 8934–8943. https://doi.org/10.1109/CVPR.2018.00931
31. Ranjan A, Black MJ (2017) Optical flow estimation using a spatial pyramid network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2720–2729. https://doi.org/10.1109/CVPR.2017.291

32. Hui T-W, Tang X, Loy CC (2018) LiteFlowNet: a lightweight convolutional neural network for optical flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8981–8989
33. Hur J, Roth S (2019) Iterative residual refinement for joint optical flow and occlusion estimation. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 5747–5756. https://doi.org/10.1109/CVPR.2019.00590
34. Teed Z, Deng J (2020) Raft: Recurrent all-pairs field transforms for optical flow
35. Beauchemin SS, Barron JL (1995) The computation of optical flow. ACM Comput Surv 27:433–467. https://api.semanticscholar.org/CorpusID:1334552
36. Fortun D, Bouthemy P, Kervrann C (2015) Optical flow modeling and computation: a survey. Comput Vis Image Underst 134:1–21. https://api.semanticscholar.org/CorpusID:5945559
37. Tu Z, Xie W, Zhang D, Poppe R, Veltkamp RC, Li B, Yuan J (2019) A survey of variational and cnn-based optical flow techniques. Signal Process Image Commun 72:9–24. https://api.semanticscholar.org/CorpusID:57980549
38. Savian S, Elahi M, Tillo T (2020) Optical flow estimation with deep learning, a survey on recent advances. In: Deep Biometrics. https://api.semanticscholar.org/CorpusID:214149658
39. Hur J, Roth S (2020) Optical flow estimation in the deep learning age. arXiv:2004.02853. https://api.semanticscholar.org/CorpusID:214802969
40. Zhai M, Xiang X, Lv N, Kong X (2021) Optical flow and scene flow estimation: a survey. Pattern Recognit 114:107861. https://api.semanticscholar.org/CorpusID:232328500
41. Dobrički T, Zhuang X, Won KJ, Hong B-W (2022) Survey on unsupervised learning methods for optical flow estimation. 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), pp 591–594. https://api.semanticscholar.org/CorpusID:253881748
42. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Neural information processing systems. https://api.semanticscholar.org/CorpusID:13756489
43. Nagel H-H (1989) On a constraint equation for the estimation of displacement rates in image sequences. IEEE Trans Pattern Anal Mach Intell 11(1):13–30. https://doi.org/10.1109/34.23110
44. Anandan P (1989) A computational framework and an algorithm for the measurement of visual motion. Int J Comput Vis 2(3):283–310
45. Uras S, Girosi F, Verri A, Torre V (1988) A computational approach to motion perception. Biol Cybern 60:79–87
46. Hildreth EC, Ullman S (1982) The measurement of visual motion. ACM Computing Surveys (CSUR)
47. Adiv G (1985) Determining three-dimensional motion and structure from optical flow generated by several moving objects. IEEE Trans Pattern Anal Mach Intell 7(4):384–401
48. Black MJ, Anandan P (1996) The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comp Vision Image Underst 63(1):75–104
49. Thompson WB (1998) Exploiting discontinuities in optical flow. Int J Comput Vis 30:163–173
50. Weiss Y, Fleet DJ (2002) Velocity likelihoods in biological and machine vision. Probabilistic models of the brain: perception and neural function, pp 81–100
51. Roth S, Black MJ (2007) On the spatial statistics of optical flow. Int J Comput Vis 74:33–50
52. Sun D, Roth S, Lewis JP, Black MJ (2008) Learning optical flow. In: European conference on computer vision. https://api.semanticscholar.org/CorpusID:969406
53. Li X, Jia J, Matsushita Y (2011) Motion detail preserving optical flow estimation. IEEE Trans Pattern Anal Mach Intell 34(9):1744–1757
54. Weinzaepfel P, Revaud J, Harchaoui Z, Schmid C (2013) Deepflow: Large displacement optical flow with deep matching. In: 2013 IEEE International conference on computer vision, pp 1385–1392. https://doi.org/10.1109/ICCV.2013.175
55. Bailer C, Taetz B, Stricker D (2015) Flow fields: dense correspondence fields for highly accurate large displacement optical flow estimation. In: Proceedings of the IEEE international conference on computer vision, pp 4015–4023
56. Revaud J, Weinzaepfel P, Harchaoui Z, Schmid C (2015) EpicFlow: edge-preserving interpolation of correspondences for optical flow. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1164–1172. https://doi.org/10.1109/CVPR.2015.7298720
57. Butler DJ, Wulff J, Stanley GB, Black MJ (2012) A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. https://api.semanticscholar.org/CorpusID:4637111
58. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. 2012 IEEE Conference on computer vision and pattern recognition, pp 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

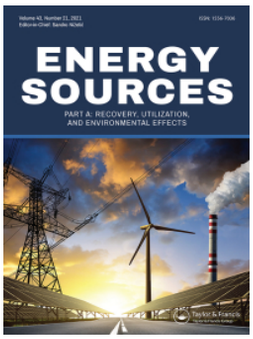
59. Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, van der Smagt P, Cremers D, Brox T (2015) FlowNet: learning optical flow with convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 2758–2766. https://doi.org/10.1109/ICCV.2015.316
60. Hui T-W, Tang X, Loy CC (2019) A lightweight optical flow cnn - revisiting data fidelity and regularization. In: arXiv:1903.07414
61. Hui T-W, Tang X, Loy CC (2021) A lightweight optical flow CNN –revisiting data fidelity and regularization. IEEE Trans Pattern Anal Mach Intell 43(8):2555–2569 . ISSN 1939-3539. https://doi.org/10.1109/TPAMI.2020.2976928
62. Yang G, Ramanan D (2019) Volumetric correspondence networks for optical flow. In: Neural information processing systems. https://api.semanticscholar.org/CorpusID:202763868
63. Wang J, Zhong Y, Dai Y, Zhang K, Ji P, Li H (2020) Displacement-invariant matching cost learning for accurate optical flow estimation. In: arXiv:2010.14851
64. Bar-Haim A, Wolf L (2020) ScopeFlow: Dynamic scene scoping for optical flow. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7995–8004. https://doi.org/10.1109/CVPR42600.2020.00802
65. Zhao S, Sheng Y, Dong Y, Chang EI-C, Xu Y (2020) MaskFlownet: asymmetric feature matching with learnable occlusion mask. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6277–6286. https://doi.org/10.1109/CVPR42600.2020.00631
66. Luo A, Yang F, Li X, Liu S (2022) Learning optical flow with kernel patch attention. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8896–8905. https://api.semanticscholar.org/CorpusID:250164362
67. Luo A, Yang F, Luo K, Li X, Fan H, Liu S (2022) Learning optical flow with adaptive graph reasoning. In: Proceedings of the AAAI conference on artificial intelligence 36:1890–1898
68. 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
69. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. arXiv:2005.12872. https://api.semanticscholar.org/CorpusID:218889832
70. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J-J, Ni LMS, Shum HY (2022) Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605. https://api.semanticscholar.org/CorpusID:247292561
71. 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. arXiv:2103.14030. https://arxiv.org/abs/2103.14030
72. Han K, Xiao A, Enhua W, Guo J, Chunjing X, Wang Yunhe (2021) Transformer in transformer. Adv Neural Inf Process Syst 34:15908–15919
73. Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3061–3070. https://doi.org/10.1109/CVPR.2015.7298925
74. Jaegle A, Borgeaud S, Alayrac J-B, Doersch C, Ionescu C, Ding D, Koppula S, Zoran D, Brock A, Shelhamer E et al (2021) Perceiver io: a general architecture for structured inputs & outputs. arXiv:2107.14795
75. Jiang W, Trulls E, Hosang J, Tagliasacchi A, Yi KM (2021) Cotr: correspondence transformer for matching across images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6207–6217
76. Liu H, Lu T, Xu Y, Liu J, Li W, Chen L (2022) Camliflow: bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5791–5801
77. Luo A, Yang F, Li X, Nie L, Lin C, Fan H, Liu S (2023) Gaflow: incorporating gaussian attention into optical flow. In: Proceedings of the IEEE/CVF International conference on computer vision, pp 9642–9651
78. Jiang S, Campbell D, Lu Y, Li H, Hartley R (2021) Learning to estimate hidden motions with global motion aggregation. In: 2021 IEEE/CVF international conference on computer vision (ICCV), pp 9752–9761. https://doi.org/10.1109/ICCV48922.2021.00963
79. Chen Y, Zhu D, Shi W, Zhang G, Zhang T, Zhang X, Li J(2023) Mfcflow: a motion feature compensated multi-frame recurrent network for optical flow estimation. In: Proceedings of the IEEE/CVF Winter conference on applications of computer vision, pp 5068–5077
80. Fang G, Chen J, Liang D, Asim M, Reeth FV, Claesen L, Yang Z, Liu W (2023) Feature correlation transformer for estimating ambiguous optical flow. Neural Process Lett 1–17
81. Dong Q, Cao C, Fu Y (2023) Rethinking optical flow from geometric matching consistent perspective. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 1337–1347

82. Weinzaepfel P, Lucas T, Leroy V, Cabon Y, Arora V, Brégier R, Csurka G, Antsfeld L, Chidlovskii B, Revaud J (2023) Croco v2: improved cross-view completion pre-training for stereo matching and optical flow. In: Proceedings of the IEEE/CVF International conference on computer vision, pp 17969–17980
83. Shi X, Huang Z, Bian W, Li D, Zhang M, Cheung KC, See S, Qin H, Dai J, Li H (2023) Videoflow: exploiting temporal cues for multi-frame optical flow estimation. arXiv:2303.08340
84. Xu H, Yang J, Cai J, Zhang J, Tong X (2021) High-resolution optical flow from 1D attention and correlation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 10478–10487. https://doi.org/10.1109/ICCV48922.2021.01033
85. Sui X, Li S, Geng X, Wu Y, Xu X, Liu Y, Goh R, Zhu H (2022) CRAFT: cross-attentional flow transformer for robust optical flow. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17581–17590. https://doi.org/10.1109/CVPR52688.2022.01708
86. Zhao S, Zhao L, Zhang Z, Zhou E, Metaxas D (2022) Global matching with overlapping attention for optical flow estimation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17571–17580. https://doi.org/10.1109/CVPR52688.2022.01707
87. Xu H, Zhang J, Cai J, Rezatofighi H, Tao D (2022) GMFlow: learning optical flow via global matching. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8111–8120. https://doi.org/10.1109/CVPR52688.2022.00795
88. Huang Z, Shi X, Zhang C, Wang Q, Cheung KC, Qin H, Dai J, Li H (2022) Flowformer: a transformer architecture for optical flow. arXiv:2203.16194. https://api.semanticscholar.org/CorpusID:247792986
89. Shi X, Huang Z, Li D, Zhang M, Cheung KC, See S, Qin H, Dai J, Li H (2023) Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1599–1610
90. Lu Y, Wang Q, Ma S, Geng T, Chen YV, Chen H, Liu D (2023) Transflow: transformer as flow learner. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 18063–18073
91. Zhang Z, Jiang H, Singh H (2024) Neuflow: real-time, high-accuracy optical flow estimation on robots using edge devices. arXiv:2403.10425. https://api.semanticscholar.org/CorpusID:268510480
92. Dong Q, Fu Y (2024) Memflow: optical flow estimation and prediction with memory. arXiv:2404.04808. https://api.semanticscholar.org/CorpusID:269005178
93. Shaw P, Uszkoreit J, Vaswani A (2018) Self-attention with relative position representations. arXiv:1803.02155
94. He K, Chen X, Xie S, Li Y, Doll'ar P, Girshick RB (2021) Masked autoencoders are scalable vision learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 15979–15988. https://api.semanticscholar.org/CorpusID:243985980
95. Dao T, Fu DY, Ermon S, Rudra A, R'e C (2022) Flashattention: fast and memory-efficient exact attention with io-awareness. arXiv:2205.14135. https://api.semanticscholar.org/CorpusID:249151871
96. Baker S, Scharstein D, Lewis JP, Roth S, Black MJ, Szeliski R (2011) A database and evaluation methodology for optical flow. Int J Comput Vis 92(1):1–31. ISSN 0920-5691, 1573-1405. https://doi.org/10.1007/s11263-010-0390-2
97. Ilg E, Saikia T, Keuper M, Brox T (2018) Occlusions, motion and depth boundaries with a generic network for disparity. Optical flow or scene flow estimation
98. Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A, Brox T (2016) A large dataset to train convolutional networks for disparity. Optical flow, and scene flow estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4040–4048. https://doi.org/10.1109/CVPR.2016.438
99. Kondermann D, Nair R, Honauer K, Krispin K, Andrulis J, Brock A, Güssefeld B, Rahimimoghaddam M, Hofmann S, Brenner C, Jähne B (2016) The HCI benchmark suite: stereo and flow ground truth with uncertainties for urban autonomous driving. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 19–28. https://doi.org/10.1109/CVPRW.2016.10
100. Richter SR, Hayder Z, Koltun V (2017) Playing for Benchmarks. 2017 IEEE International Conference on Computer Vision (ICCV), pp 2232–2241. https://doi.org/10.1109/ICCV.2017.243
101. Richter SR, Hayder Z, Koltun V (2017) Playing for benchmarks. In: Proceedings of the IEEE International conference on computer vision, pp 2213–2222
102. Sun D, Vlasic D, Herrmann C, Jampani V, Krainin M, Chang H, Zabih R, Freeman WT, Liu C (2021) AutoFlow: Learning a better training set for optical flow. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10088–10097. https://doi.org/10.1109/CVPR46437.2021.00996
103. Greff K, Belletti F, Beyer L, Doersch C, Du Y, Duckworth D, Fleet DJ, Gnanapragasam D, Golemo F, Herrmann C et al (2022) Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3749–3761

104. Yin Z, Darrell T, Yu F (2019) Hierarchical discrete distribution decomposition for match density estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6037–6046. https://doi.org/10.1109/CVPR.2019.00620
105. Jiang S, Lu Y, Li H, Hartley R (2021) Learning optical flow from a few matches. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 16587–16595. https://doi.org/10.1109/CVPR46437.2021.01632
106. Xu N, Yang L, Fan Y, Yue D, Liang Y, Yang J, Huang T (2018) Youtube-vos: a large-scale video object segmentation benchmark. arXiv:1809.03327
107. Black MJ, Jepson AD (1996) Estimating optical flow in segmented images using variable order parametric models with local deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(10):972–986. https://doi.org/10.1109/34.541407

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.



Taylor & Francis
Taylor & Francis Group

Energy Sources, Part A: Recovery, Utilization, and Environmental Effects

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ueso20

Experimental review on solid oxide fuel cell-based hybrid power generation system: Performance, economic, and environmental analysis

Anil Kumar Yadav, Anil Kumar, Shailendra Sinha, Samarjit Singh & S. K. Yadav

To cite this article: Anil Kumar Yadav, Anil Kumar, Shailendra Sinha, Samarjit Singh & S. K. Yadav (2024) Experimental review on solid oxide fuel cell-based hybrid power generation system: Performance, economic, and environmental analysis, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 46:1, 12815-12828, DOI: 10.1080/15567036.2024.2400305

To link to this article: https://doi.org/10.1080/15567036.2024.2400305

Published online: 11 Sep 2024.

Submit your article to this journal

Article views: 51

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ueso20

ENERGY SOURCES, PART A: RECOVERY, UTILIZATION, AND ENVIRONMENTAL EFFECTS
2024, VOL. 46, NO. 1, 12815–12828
https://doi.org/10.1080/15567036.2024.2400305

Taylor & Francis
Taylor & Francis Group

Check for updates

Experimental review on solid oxide fuel cell-based hybrid power generation system: Performance, economic, and environmental analysis

Anil Kumar Yadav[a], Anil Kumar[b], Shailendra Sinha[a], Samarjit Singh[c], and S. K. Yadav[d]

[a]Department of Mechanical Engineering, Institute of Engineering and Technology, Lucknow, India; [b]Department of Mechanical Engineering, Delhi Technological University, Delhi, India; [c]Department of Mechanical Engineering, GGU, Bilaspur, India; [d]Department of Mechanical Engineering, Lucknow University, Lucknow, India

ABSTRACT

Solid oxide fuel cells (SOFCs) have shown great promise as a means of producing clean and efficient electricity. This review focuses on doing a comprehensive analysis of SOFC-based experimental hybrid power systems, focusing on their thermodynamic (energy and exergy), economic, and environmental aspects. Performance of a standalone SOFC system can be improved by combining it with other power technologies, i.e. gas turbines (GTs), internal combustion engines (ICEs), or some other waste heat recovery cycles, which provide greater system flexibility and load-following capabilities. Maximum deviation between simulation results and experimental results have been reported within 10%. SOFC-GT hybrid system offers electrical efficiency 50–60%, exergy efficiency up to 52%, emission of CO_2 376.9 kg/MWh, and total cost of the system as 25 $/h. Most of the literatures have shown CO_2 emission of SOFC-based hybrid system in the range of 250–350 kg/MWh. Economic and environmental studies of the hybrid systems is essential for widespread adoption and future development.

ARTICLE HISTORY
Received 18 March 2024
Revised 23 August 2024
Accepted 23 August 2024

KEYWORDS
Economic; environmental analysis; experimental investigation; solid oxide fuel cell; thermal analysis

Introduction

The solid oxide fuel cell (SOFC) is a type of electrochemical device that uses a chemical interaction between fuel and air to produce electricity and heat (Yadav, Sinha, and Kumar 2024). Various researches have been exploring the performance of different SOFC-based hybrid configurations (Damo et al. 2019). They have mainly focussed on waste heat recovery cycle's integration with SOFC and their performance was mostly evaluated with simulation results. They have also not covered the control strategy of the configurations (Wang et al. 2021). Few researchers have demonstrated the performance of standalone-SOFC system without combination of hybrid power system (Zhang et al. 2022). Currently, several distinct SOFC topologies are being studied, including planar, tubular (Skabelund and Milcarek 2022). One important step in developing more effective concepts for combined heat and power (CHP) systems that use solid oxide fuel cells (SOFCs) is to conduct experiments on commercially viable SOFCs. These experiments should focus on key operating characteristics, like the ideal composition of fuel gas and cell temperature (Wu et al. 2009). The versatility of solid oxide fuel cell technology makes it a good fit for power cycles that use fuels other than pure hydrogen (Li et al. 2023). In Japan and Korea, where renewable energy is scarce and hydrogen is produced only from fossil fuels, the cogeneration plant of hydrogen, electricity, and heat could be put to practical use by allowing the realization of renewable refueling stations. When considering safety, storage, transport and cost, these nations appear to favor importing ammonia as

CONTACT Shailendra Sinha shailendra.sinha@ietlucknow.ac.in Department of Mechanical Engineering, Institute of Engineering and Technology, Lucknow 226021, India

© 2024 Taylor & Francis Group, LLC

a hydrogen transporter over pure hydrogen (Perna et al. 2018). SOFC has the capability to be effectively incorporated into methanol-based power generation systems with thermodynamic recuperation of waste heat (Abdin et al. 2023). At maximum power, performance drops 4.6% due to a discrepancy of 0.05 W cm^2 between the cell power densities recorded with H_2 and reformate (Tapiero et al. 2023).

An investigation was conducted into the fuel cell's performance at a working temperature of 650°C with three distinct fuel flow rates: 175 ml/min., 250 ml/min., and 325 ml/min. Fuel utilization factor (U_f) increases with decrease of fuel flow rates. It is confirmed that the maximum working temperature is better suitable for maximum electrical voltage generation, since the cell produced the maximum electrical voltage of around 1.08 V at 650°C (Ullah et al. 2015). The use of an ejector or blower to recycle off-gas from the anode is becoming increasingly popular as a means to increase the system electrical efficiency of SOFCs powered by various fuel (Yen et al. 2010). Researchers found that SOFCs powered by biogas might have their cell stack temperature gradient reduced with the addition of anode recycling (Du et al. 2023). Better system efficiencies are achievable with smaller cell U_f when the recirculation rate is greater than 80%. For instance, when the recirculation rate is larger than 85%, the system efficiency is determined to be higher for $U_f = 0.6$ compared to $U_f = 0.8$ (Barelli et al. 2022). Since exhaust of SOFC system is at high temperature, which can further be utilized in hybrid power cycles to improve the overall system performances (Mottaghizadeh et al. 2022). The majority of research has concentrated on fuel cell/GT hybrid system dynamics, but there has been very little comparison of experimental data with results from dynamic simulations (Yang et al. 2022). An experimental study has been conducted on SOFC-GT. Maximum deviation was less than 10% with simulation results (Zhu et al. 2018). An analysis has conducted on the sensitivity of hybrid SOFC-GT to ambient disturbances using experimental and dynamic simulation data. The component temperatures, pressures, voltage and system power were predicted with an error of 5°C, 2 kPa, 2 mV and 0.5% correspondingly (McLarty et al. 2012). Hydrogen rich fuel from solar energy and utilizing SOFCs with micro gas turbine (mGT) results in system energy efficiency (80.74%) and exergy efficiency (59.76%). SOFC-GT system achieves a net solar to power efficiency of 40.93% (Zheng et al. 2021). The SOFC-ICE systems offers higher efficiency and cost effectiveness of ICEs compared to GTs for the current power generating capacity of SOFCs, specially for capacities below several megawatts (Choi et al. 2019). For a 100-kW system, the SOFC-ICE hybrid system outperformed the SOFC-GT system in terms of energy efficiency (0.9% points higher) and economic feasibility (6.6% points lower) (Park, Lee, and Ahn 2014). A study on SOFC-GT-ORC has been investigated and it gives electrical efficiency (55.67%), exergy efficiency (53.55%), total cost rate (36.09 $/h) and CO_2 emission of 355.8 kg/MWh (Agrawal, Singh, and Sinha 2024). Table 1 summarizes the methodologies, findings, and conclusions of different studies, showcasing the diverse approaches to SOFC hybrid systems and their potential applications.

The objective of present review is to systematically analyze and evaluate the thermo-economic and environmental performance of SOFC-based experimental hybrid power systems. This article also provides a comprehensive understanding of their potential as a clean and efficient energy source, as well as the challenges and possibilities that prevent their widespread implementation.

Solid oxide fuel cell system

A standard SOFC stack consists of a thick layer of solid electrolyte that conducts oxygen sandwiched between two thin, porous electrodes known as the anode and cathode, as shown in Figure 1. The end of the cell called the cathode gets air, while the side called the anode gets hydrogen fuel, which is then mixed with oxygen ions from the air. When fuel with hydrogen ignites at the anode, the amount of oxygen at the cathode drops by a lot. When the oxidized fuel reacts with the oxygen ions that have passed through the crystalline grid of the ceramic electrolyte, free electrons are made and then run from the anode to the cathode through an outside circuit, which makes DC current. This DC current is converted into AC current with the help of an

Table 1. Methodologies, result and conclusions of SOFC-based hybrid systems.

Author	Methodology	Result	Conclusion
Roushenas, Zarei, and Torabi (2021)	Economic analysis and system design for 25 MW plant	System efficiency over 65% with a modular design costing under $400/kW	Highlights SOFC-GT system scalability and economic viability
Jouhara et al. (2018)	Experimental setup with thermionic and thermoelectric generators	Converted surplus heat into electricity, enhancing system efficiency	Integrating thermionic and thermoelectric technologies with SOFC may boost energy output
Deangelis (2004)	Comprehensive literature review	Discussed GT-integrated hybrid cycles and LT-cycles	Promotes novel hybrid designs to boost efficiency and energy recovery
Singh and Singh (2018)	Comparative analysis of various fuel cell technologies	Identified SOFC as one of the most efficient technologies available	Emphasizes the relevance of application-specific FC technology
Mekhilef, Saidur, and Safari (2012)	Economic and technical feasibility study of SOFC systems	Demonstrated economic viability for CCHP applications	SOFC systems may be a sustainable solution for multi-functional energy needs
Facci et al. (2017)	Thermodynamic modelling of SOFC-GT-ORC systems	Identified optimal operating conditions for maximum efficiency	integrated systems can significantly enhance energy output and efficiency
Yan et al. (2013)	Exergy analysis of SOFC-Stirling engine systems	Efficiency and enhancement opportunities in energy conversion processes	Suggests that exergy optimization can lead to more sustainable energy systems
Guo et al. (2020)	Thermodynamic modelling and optimization of SOFC-GT systems	Achieved significant gains in efficiency through optimization of operational parameters	Recommends continuous optimization for enhanced system performance

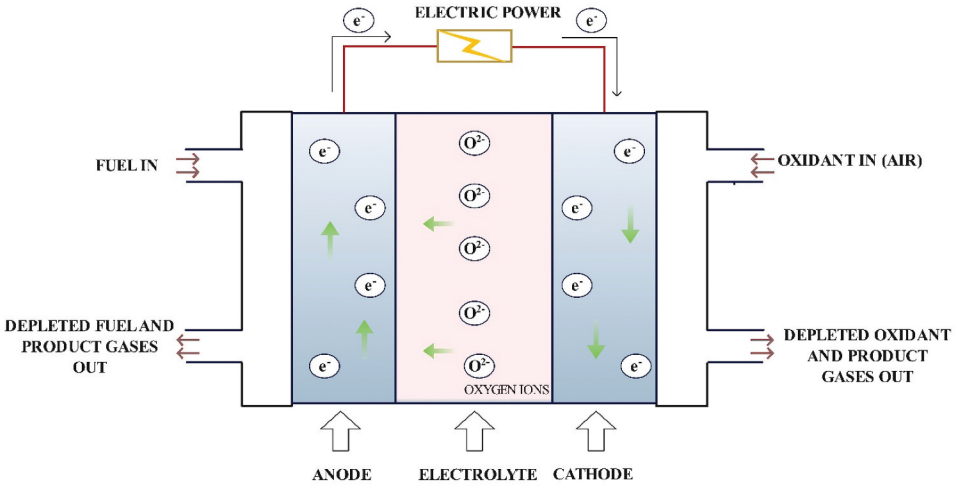


Figure 1. SOFC working principle.

inverter (Paulus 2024; Zarabi Golkhatmi, Asghar, and Lund 2022). SOFC reactions at cathode (Eq. 1) and anode (Eq. 2) are:

$$^1/_2O_2 + 2e^- \rightarrow O^{2-} \tag{1}$$

$$H_2 + O^{2-} \rightarrow H_2O + 2e^- \tag{2}$$

Due to its adaptability to a wide variety of fuels, its high efficiency, the low maintenance required, and its little impact on the environment, the SOFC has emerged as one of the most promising energy production techniques nowadays (Zhou et al. 2022). Even with improvements, SOFC continues to

Table 2. Strengths and weaknesses of various SOFC systems.

Author	SOFC system type	Strengths	Weaknesses
Wu et al. (2022)	Integrated SOFC system	• High efficiency under multiple working conditions; detailed dynamic response analysis	• Complexity in modeling; potential for performance degradation
Choi et al. (2019)	High temperature SOFC	• High efficiency (up to 60%) • Good thermal stability • Capable of using a variety of fuels	• High operating temperatures lead to material degradation • Requires complex thermal management
Yadav, Sinha, and Kumar (2024)	IT-SOFC	• Improve material longevity • Suitable for portable applications	• Lower efficiency compared to HT-SOFC • Limited fuel flexibility
Hjalmarsson and Mogensen (2011)	Anode supported SOFC	• Simple design and fabrication • High power density	• Sensitive to fuel impurities • Limited thermal cycling capability
Ni et al. (2020)	Cathode supported SOFC	• Better thermal stability • Good performance under varying loads	• More complex manufacturing • Heavier than anode-supported designs
Eisavi et al. (2018)	SOFC-GT hybrid systems	• High overall efficiency (up to 85%) • Effective for large-scale power generation	• Complex system integration • High capital costs
Palomba et al. (2018)	SOFC-adsorption cooling	• Integrates cooling and power generation • Reduces overall energy consumption	• Complexity in system design • Requires careful thermal management
Sinha and Mondal (2018)	Portable SOFC systems	• Lightweight and compact • Suitable for remote applications	• Limited power output • High cost per watt

produce exhaust gas at a high temperature, and its direct emissions result in a high rate of energy loss (Feng et al. 2023). Table 2 depicts strength and weakness of various SOFC systems.

SOFC-based hybrid power generation systems

Combined heat and power (CHP) systems have seen rapid growth in popularity over the past few years. In terms of micro-CHP systems, Stirling engines, and more specifically dynamic Stirling engines are promising options because of their high overall thermal efficiency, little noise and vibration, fuel flexibility, and little emissions (Zhu et al. 2018). Figure 2 depicts a conceptual diagram of a hybrid power production system based on SOFCs. The SOFC stack takes in fuel and air for operation. The SOFC system uses an electro-chemical reaction, with heat being a by-product, to generate electricity. The SOFC stack's by-product heat is harnessed to drive a thermodynamic cycle that generates electricity. This power cycle boosts the hybrid's overall efficiency.

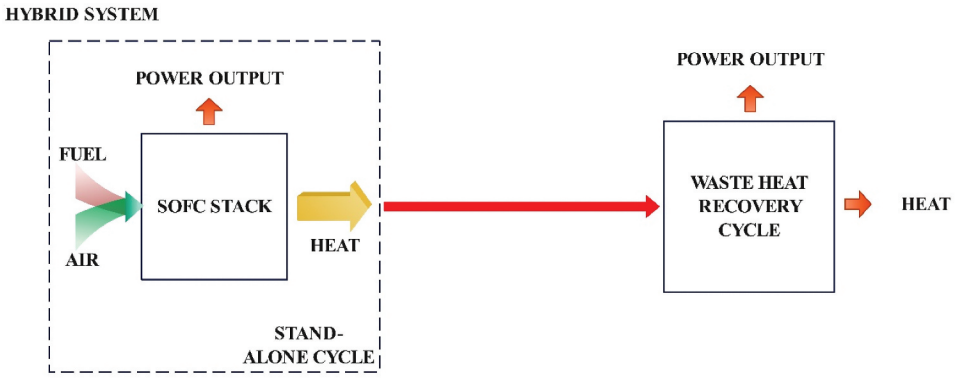


Figure 2. General concept of SOFC-based hybrid system (Kumar Yadav, Sinha, and Kumar 2023). [permission taken].

Achieving decarbonization and climate neutrality through the use of SOFC is a major goal of several long-term energy strategies (Subotić and Hochenauer 2022). To improve the system's energy conversion efficiency, a waste heat recovery (WHR) method can be implemented (Chehrmonavari et al. 2023). Researchers use both experiments and computer simulations to look into what an ammonia-fueled SOFC is like, and a well-built thermo-electrochemical model predicts how the proposed CHP plant will work and how it should be designed. Renewable refueling stations can be provided by the proposed hydrogen, power, and heat co-generation plant in renewable energy-poor countries like Japan and Korea, where hydrogen can only be produced from fossil fuels. Safety, storage, transport, and cost make ammonia a better hydrogen carrier in these countries than pure hydrogen (Perna et al. 2018). However, it is still difficult to integrate commercial GTs into a SOFC-GT hybrid because they are often designed for the direct use of natural gas (Rosner and Samuelsen 2022). The techno-economic viability of SOFC-CHP systems in commercial buildings has been performed by Marocco et al. (Marocco, Gandiglio, and Santarelli 2022). For the study, they drew heavily on the information gathered and knowledge gained during the European Comsos and DEMOSOFC initiatives. For a precise evaluation of the SOFC's performance, the optimization framework was updated to include real SOFC efficiency curves. Table 3 describes Experimental review of SOFC-based integrated system. Figure 3 (Zoupalis et al. 2023) depicts experimental test rig of SOFC system.

Performance evaluation

An experimental review of an integrated energy generation system based on SOFCs would normally involve analyzing a number of important parameters and metrics to ascertain the system's effectiveness and efficiency. There may be some variation in the evaluation criteria used based on the aims, scope, and focus of the experiment being reviewed. For this reason, it is suggested to refer to the review's methodology and research objectives in order to modify the performance evaluation properly.

Table 3. Experimental review of SOFC-based integrated system.

Plant type	Fuel	Important outcome	Ref.
100 kWe-SOFC	Natural gas	Using a heating rate of 1 °C/min, the system can start up 20% faster.	Leone and Lanzini (2013)
1 kW-SOFC-GT	Natural gas	At 750°C, the reformate gas had a combined hydrogen and carbon monoxide composition of about 70%.	Hong et al. (2016)
1.5 kWe-SOFC-CHP	Natural gas	Electrical efficiency was 57% (manufacturer's target was 60%).	Paulus and Lemort (2023)
Two SOFC in serial connection	Biogas	Compared to two in-parallel connections, the electrical efficiency of two stacks of SOFCs connected in series was 3% to 7% greater.	Kupecki et al. (2017)
SOFC-CHP	Biogas	The gas chromatography data and the simulation findings are in good agreement, with an error range of less than 5% for the matrix gases and less than 10% for water.	Tamburrano et al. (2024)
SOFC-HCCI engine	Methane	The plumbing and other plant components were able to recover 75.1% of the heat that was transported; the remaining 25%, or 638 W, was lost.	Koo et al. (2021)
kWe-SOFC	Methanol	The reforming reactor's predicted volume power was 303 W/L and weight power was 172 W/kg, using a 1125 W SOFC.	Hu et al. (2023)
5 kW-SOFC-Engine	Methane	Compared to operating with only FC (47.8% efficiency), the engine improves it by 5.3%.	Kim, Lee, and Ahn (2020)
NextCell-HP SOFC button cell	Carbonaceous fuels	A peak power density (PPD) of over 1000 mW cm^{-2} was achieved using pure H2, however it decreases to around 100 mW cm^{-2} when CO and H_2 are used as simulated syngas.	Narayana Sarma, Shivapuji, and Srinivasaiah (2022)
Planar-SOFC	Ammonia	System efficiency and LCOE were 60% and 0.221 $/$kWh^{-1}$ respectively.	Barelli (2020)
1 kWe diesel reformer-SOFC	Diesel	At an average cell voltage of 0.8V, the output power is around 110 kWe.	Kang et al. (2010)
Circular-planar type SOFC	Hydrogen	Open circuit voltage appears to be slightly yet significantly affected by fuel mass flow.	Bedogni et al. (2007)

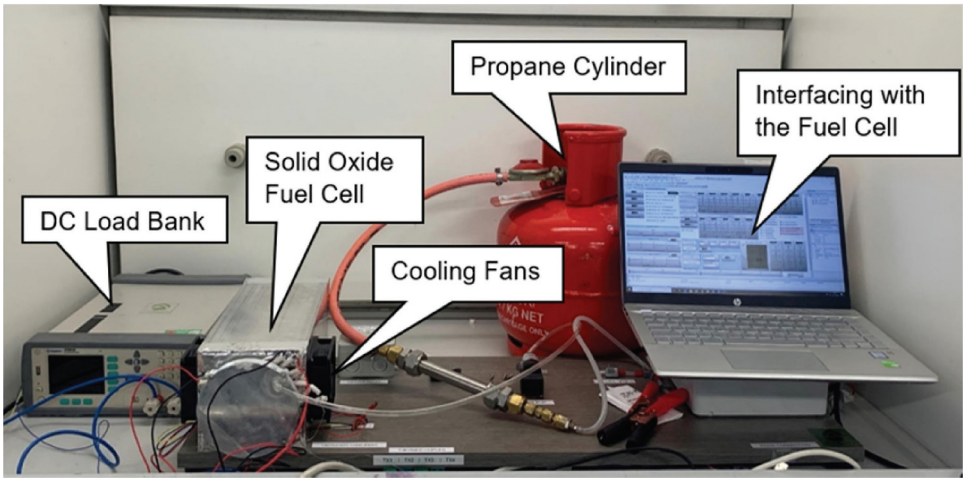


Figure 3. Thermal test rig of SOFC system (Zoupalis et al. 2023). [permission taken].

Energy and exergy methods should be used to evaluate the designed energy. Each part of the cycle has an associated mass-energy-entropy-exergy balance equation (Karapekmez and Dincer 2022; Ma et al. 2022). The mass balance equation (Eq. 3) is described as:

$$\sum \dot{m}_{in} = \sum \dot{m}_{out} \tag{3}$$

The first law of thermodynamics for hybrid systems (Eq. 4) can be described as:

$$\sum \dot{m}_{in} h_{in} + \dot{Q}_{in} + \dot{W}_{in} = \sum \dot{m}_{out} h_{out} + \dot{Q}_{out} + \dot{W}_{out} \tag{4}$$

The equation for the exergy balance (Eq. 5) can be stated as:

$$\sum \dot{m}_{in} ex_{in} + \dot{E}x^{Q_{in}} + \dot{W}_{in} = \sum \dot{m}_{out} ex_{out} + \dot{E}x^{Q_{out}} + \dot{W}_{out} + \dot{E}x_{<}/sup> \tag{5}$$

Where $\dot{E}x^{Q}$ refers to the thermal exergy rate. $\dot{E}x^{d}$ denoted exergy destruction rate. Energy efficiency is one of the most important ways to make sure one can have enough energy and protect the earth. In fact, the International Energy Agency (IEA) says that it can't reach 70% of our goals for reducing emissions if it does not improve the energy economy. Even though this is clear, improving energy efficiency is a big task for process operations. The objective is that the technique used to save energy put the running of the process at odds with safety requirements (Sharifzadeh, Meghdari, and Rashtchian 2017). Simulation was used to figure out how each control parameter affected the system, and the heat loss was used to figure out the ranges in which the system could work. At its best point, when heat loss was taken into account, the system works at a gross efficiency of 66.9%. Using the ICE in three distinct capacities (recirculation blower, reformer, and power generator), Author developed a high-efficiency SOFC-ICE hybrid system (Cho, Kim, and Ho Song 2022). An innovative integrated system has been studied, analyzing its energy, exergy, and thermodynamics in CHP process (Peng et al. 2020). A new multi-generation energy system has been created and assessed from an energy and exergy perspective to better use SOFC waste heat. A SOFC, MGT, s-CO_2 Brayton cycle, and lithium bromide absorption refrigeration are included. Under design, the multi-generation energy system generates 696.9 kW of electricity, 24.98 kW of heating, and 88.81 kW of cooling. Its energetic round-trip and electrical efficiency are 70.49% and 60.59% (Ran et al. 2023). A 5-kW SOFC-Engine hybrid proof-of-concept test yielded 4.4 kW and 731 W from hydrogen-based SOFC and engine, achieving 70% fuel utilization. The engine was 7.9% more efficient than fuel-cell

Table 4. Thermal, economic, and environmental analysis of SOFC-based hybrid system.

Sl. No.	Plant Type	Thermal Analysis	Economic Analysis	Environmental Analysis	Ref.
1	SOFC-GT	η=52.54% Ψ=50.54%	Total cost=24.99 $/h	CO_2 emission=376.9 kg/MWh	Eisavi et al. (2018)
2	SOFC-GT-ORC	Ψ=51.6%	LCOE=21.6 $/MWh	CO_2 emission=285 kg/h	Emadi et al. (2020)
3	SOFC-ORC-RO desalination	Ψ=46.29%	Product unit cost=28.59 $/GJ	Pollution emission factor=0.2587 kg/kWh	Alirahmi and Ebrahimi-Moghadam (2022)
4	SOFC-sCO_2-ORC-RO	Ψ=48.18%	Product cost=57.47 $/GJ	CO_2 emission=0.25 kg/kWh	Alirahmi, Gundersen, and (Yu 2023)
6	SOFC-Biomass Gasification (BG)-Geothermal	Ψ=64.49%	Cost rate=4.94 $/GJ	CO_2 emission=0.124 t/MWh	Hou et al. (2023)
7	SOFC-BG-GT-AR-ORC	η=86.70% Ψ=50.45%	Cost rate=19.6 $/h	CO_2 emission=0.4722 kg/kWh	Wang et al. (2023)
8	SOFC-OFC-desalination	η=53.4% Ψ=33.2%	Cost rate=8.89 $/day	Environmental impact rate=0.15 mpts/s	Wang et al. (2023)
9	SOFC-GT-Solar PV	Ψ=24.85%	LCOE=0.0911 $/kWh	CO_2 emission=0.257 kg/kWh	Cao et al. (2022)
10	SOFC-ORC	Ψ=41.27%	Cost rate=11.73 $/h	CO_2 emission=0.3 kg/kWh	Cao et al. (2022)

operation (47.8%). Over 200 hours were spent on the amended gas test. SOFC generated 3.3 kW and engine 1168 W at 50.6% fuel use. The engine generated 26.0% power. At fuel utilizations of 61.2%, 65.4%, and 70.1%, the engine's power share dropped to 17.3%, 13.1%, and 7.6%. Engine efficiency rose 5.3% at 70.1% fuel use (Kim, Lee, and Ahn 2020). Table 4 describe thermal, economic, and environmental investigation of SOFC-based integrated power generation system.

Economic analysis

Economic metrics, such as the system's maintenance cost and profit, are crucial for planning and establishing the system's infrastructure. By figuring out how much it will cost to run the system, how much it will help, and how long it will take to break even, an economic evaluation may be made (Rosner and Samuelsen 2022). The size of the global market for solid oxide fuel cells was assessed to be 1.9 billion US dollars in 2022, and it is anticipated to reach 20.02 billion US dollars by 2032. This represents a growth rate that is projected to be 26.6% of the compound annual growth rate for the period of forecasting from 2023 to 2032 (Power 2023). Calculating the plant's lifetime cost of products and payback period can help establish whether an integrated energy system is financially viable. Add expected carrying costs and expenses to get the total. Instead, then focusing on annual expenditure swings, levelized expenses may be more efficient. The objective of economics and thermodynamics (often exergy analysis) are brought together in thermo-economic analysis (Erdogan et al. 2022).

Pre-defined cost functions are commonly used to estimate capital and maintenance costs in economic modeling and optimization of SOFC integrated systems (Eq. 6). The cost functions are used to calculate an estimated capital investment (Z_k) for each part of the system. After accounting for time and maintenance for each process equipment, the full cost (Z_k) can be calculated (Dehghan, Fanaei, and Panahi 2022):

$$\dot{Z}_k = \frac{\emptyset \times CRF \times Z_k}{AOH \times 3600} \tag{6}$$

The following equation (Eq. 7) relates the interest rate (i) and the lifespan of the system (y) to the annual operating hours (AOH), maintenance factor (∅), and capital recovery factor (CRF).

$$CRF = \frac{i(1+i)^y}{(1+i)^y - 1} \tag{7}$$

By using the equation (Eq. 8), it is possible to calculate the system's overall cost rate, which includes the capital cost rate, the carbon dioxide emission penalty (Eq. 9), and the cost rate of natural gas.

$$\dot{C}_f = \left(C_f \times \dot{m}_f \times LHV_f\right) \times 3600 \tag{8}$$

$$\dot{C}_{env} = 3600 \times \dot{m}_{CO_2} \times C_{pen} \tag{9}$$

where the LHV of CH_4 is 50 MJ/kg, C_f is 0.003 $/MJ, and C_{pen} is 0.024 $/kg (Ni et al. 2022). The thermo-economic evaluations estimate that generated power costs USD 11.6 cents/kWh and USD 18.5 cents/kWh, respectively, based on Lazareto's model and the total revenue requirement model. The SOFC-GT hybrid system cost $1692/kW, twice as costly as a GT unit, including purchase, installation, and start-up (Pirkandi, Mahmoodi, and Ommian 2017).

Environmental evaluation

Because of environmental concerns, scientists and engineers are constantly on the lookout for alternative energy sources. The following relation (Eq. 10) has been applied to calculate the price tag for fixing the harm caused by CO_2 emissions (Hosseinpour et al. 2018).

$$\dot{C}_{CO_2} = \dot{m}_{CO_2} C_{CO_2} \tag{10}$$

The unit damage cost is denoted byC_{CO_2}. When comparing the environmental performance of different systems, the emission index is taken into account in addition to the environmental costs of CO_2 emissions. The definition of this aspect (Eq. 11) is (Cao et al. 2022):

$$CO_2\ emission = \frac{\dot{m}_{CO_2}}{\dot{W}_{net}} \tag{11}$$

In order to reduce CO_2 emissions from ocean-going vessels, researchers examine the viability of using SOFC hybrid generation. A four-stroke engine, SOFC system, steam boiler, were all integrated into the proposed SOFC integrated system. The SOFC integrated system's CO_2 emissions were measured and compared to those of a traditional system using a four-stroke generator engine. According to the preliminary findings, using a 2.5 MW and a 5.0 MW SOFC in a hybrid system may result in CO_2 emissions reductions of 5.4% and 8.8%, respectively (Lee, Yoon, and Cha 2023). Hybridising wind turbines with the biomass-driven SOFC to generate pure hydrogen through a PEM electrolyzer helps reduce biomass use. Wind turbines increased exergy efficiency by 7.3% and reduced CO_2 emissions by 13.0%. The wind turbine and water electrolyzer capital costs increased unit power costs by 6.4% to achieve these technological and environmental performance gains (Hai et al. 2023). Based on the experimental findings, the anode off-gas could be employed as a fuel in spark-ignition combustion, and a conventional engine could supplement the energy output of a hybrid SOFC-engine setup. The net indicated efficiency of combustion with the anode off-gas was comparable to that of CNG at a compression ratio of 13:1, but NO_x emissions were low and total hydrocarbon emissions were zero. However, due to the off-gas's high dilution levels, combustion with the off-gas resulted in worse volumetric efficiency and a lower load than CNG. Table 5 depicts the critical technological challenges that hinder the development and adaptation of the SOFC systems.

Table 5. Critical technological challenges and future research directions of SOFC systems.

Author	Critical technological challenges	Future research directions
Talukdar et al. (2024)	High manufacturing costs, durability concerns, and efficiency limitations	Research into cost-effective manufacturing techniques and long-lasting materials
Zarabi Golkhatmi, Asghar, and Lund (2022)	Nano-structural challenges affecting ion conductivity and mechanical stability	Investigation of nano-materials to enhance ionic conductivity and reduce mechanical failures
Lenser et al. (2022)	Mechanical instability and degradation under operational conditions	Enhanced modelling techniques to predict behavior under different conditions and materials development for stability
Ovtar et al. (2019)	Performance drop at high current densities and issues with catalyst stability	Research into novel catalyst materials and configurations to enhance performance
Navasa et al. (2018)	Complexity of modelling interactions within SOFC systems and their impact on performance	Development of simplified models for quicker assessments and predictions of performance
Chatzichristodoulou et al. (2015)	Understanding vacancy formation and its impact on ionic conductivity and overall performance	Further exploration of vacancy engineering to optimize ionic conduction in SOFC materials
Buonomano et al. (2015)	Balancing efficiency with emissions control in hybrid systems	Focus on emissions reduction technologies and their integration into hybrid systems
Bagherian and Mehranzamir (2020)	Complexity in system integration and heat recovery efficiency	Investigate advanced heat recovery systems and their integration in hybrid setups
Yuan et al. (2024)	Difficulty in diagnosing faults and maintaining system health	Develop advanced diagnostic tools and maintenance strategies for SOFC systems

Conclusions

The recent development of solid oxide fuel cells has piqued interest as a potential next-generation power source. These cells use electrochemical reactions to convert gaseous fuel and an oxidant into electrical power. However, several critical technological challenges hinder their future development and widespread adoption. Some challenges are high operating temperature, material challenges, manufacturing and stability, fuel flexibility and infrastructure, system integration and durability, and interconnect materials. Practical issues cause the electrochemical utilization of fuels to vary between 75% and 90% for present fuel cell technology. To maximize the utilization of heat generated from the high-temperature exhaust gas of the typical SOFCs, numerous studies have been focused on atmospheric SOFC-CHP systems. There is still lack on experimental work on SOFC-based CHP integration systems. Experimental studies have demonstrated that the electrical efficiency of hybrid systems based on SOFC is 30–60% and total thermal efficiency can reach 85–90%. Most of the studies have done experimentally on BioG-SOFC, SOFC-GT, and SOFC-ICE integration systems. Maximum deviation between simulation results and experimental results have been reported within 10%. SOFC-GT hybrid system offers electrical efficiency 50–60%, exergy efficiency up to 52%, emission of CO_2 376.9 kg/MWh, and total cost of the system as 25 $/h. Most of the literatures have shown CO_2 emission of SOFC-based hybrid system in the range of 250–350 kg/MWh. To achieve broad deployment of SOFCs, it is crucial to extent their lifespan, often referred to as longevity and lifetime, to ensure their reliability and economic viability.

Solid oxide fuel cell has better application prospects and may be a feasible energy utilization form in near future.

- Industry, rural locations without grid connections, and small to large-scale applications in cogeneration are all possible with tubular SOFC technology and a microgrid electrical power system.
- By using simulation technique, SOFC-CHP system has a staggering 90% overall efficiency. In future, more studies should be focused on experimental investigation of SOFC-GT based waste heat recovery cycles to improve overall system performance.
- More studies should be conducted on renewable-based hydrogen production techniques, which may be integrated with SOFC-CHP systems.

- Economic and environmental studies of the proposed literature should be analyzed for widespread adoption and future development.

Acknowledgements

Anil Kumar Yadav thankful to Dr. A.P.J. Abdul Kalam Technical University Lucknow, Uttar Pradesh, India for providing fellowship under the Homi Bhabha Teaching cum Research Fellowship. Authors acknowledge the encouragement provided by Department of Mechanical Engineering, Institute of Engineering and Technology Lucknow (India) and Centre for Energy and Environment, Delhi Technological University, Delhi (India). This paper was previously presented at ICARGET2023.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Anil Kumar Yadav is a full time Research Scholar in the Department of Mechanical Engineering at Institute of Engineering and Technology Lucknow, India. He received Bachelor's Degree in Mechanical Engineering from AKTU Lucknow, India, and Master's Degree in Mechanical Engineering from IET Lucknow, India. His current research interests include Fuel Cell Technology, Composite Materials, Cogeneration Systems, and Life Cycle Assessment.

Dr. Anil Kumar is a professor in the Department of Mechanical Engineering at Delhi Technological University Delhi, India. He is additional coordinator of Nodal Centre of Excellence in Energy Transition (NCEET), DTU Delhi, India. Dr. Kumar has been recognized for making significant contributions to the field of Renewable Energy, reflected by his outstanding publication track record and peer recognition. He has been granted several prestigious awards, including PLATINUM Cumulative Citation award, Early Research Impact and Influence award, Premier Research award, and Commendable Research award at DTU Delhi, India.

Dr. Shailendra Sinha is a Professor in the Department of Mechanical Engineering at IET Lucknow. Presently he is the Head of Department of Mechanical Engineering and Dean Post Graduate Studies and Research at Institute of Engineering and Technology Lucknow. He obtained his Ph.D. from Indian Institute of Technology Kanpur, India. His areas of interest include Biodiesel, Internal Combustion Engine, Thermodynamics, and Solid Oxide Fuel Cell based Hybrid Systems.

Dr. Samarjit Singh is an Assistant Professor in the Department of Mechanical Engineering at Guru Ghasidas Vishwavidyalaya Bilaspur, India. His research interests are Polymer Composites, Nanomaterials, and Genetic Algorithm.

Dr. S. K. Yadav is currently working as Assistant Professor in the Department of Mechanical Engineering, Lucknow University, Uttar Pradesh, India. He received his Ph.D. degree from IIT BHU, Varanasi, India. His Research interests are Aerosal, Nuclear Energy, and Thermodynamics.

ORCID

Anil Kumar Yadav http://orcid.org/0009-0000-8193-1015
Shailendra Sinha http://orcid.org/0000-0002-9284-6833

References

Abdin, Z., N. Al Khafaf, B. McGrath, K. Catchpole, and E. Gray. 2023. A review of renewable hydrogen hybrid energy systems towards a sustainable energy value chain. *Sustainable Energy and Fuels* 7 (9):2042–62. doi: 10.1039/d3se00099k.

Agrawal, B. N., C. Singh, and S. Sinha. 2024. An Estimation of Various Performance Parameters of CI Engine Fueled with Diesel and Biodiesel Blends. *Advances in Heat Transfer and Fluid Dynamics*, Singapore, 2022, 289–98. doi: 10.1007/978-981-99-7213-5_23.

Alirahmi, S. M., and A. Ebrahimi-Moghadam. 2022. Comparative study, working fluid selection, and optimal design of three systems for electricity and freshwater based on solid oxide fuel cell mover cycle. *Applied Energy* 323 (April):119545. doi: 10.1016/j.apenergy.2022.119545.

Alirahmi, S. M., T. Gundersen, and H. Yu. 2023. A comprehensive study and tri-objective optimization for an efficient waste heat recovery from solid oxide fuel cell. *International Journal of Hydrogen Energy* Xxxx:1–18. doi: 10.1016/j.ijhydene.2023.01.229.

Bagherian, M. A., and K. Mehranzamir. 2020. A comprehensive review on renewable energy integration for combined heat and power production. *Energy Conversion and Management* 224 (September):113454. doi: 10.1016/j.enconman.2020.113454.

Barelli, L. 2020. Operation of a solid oxide fuel cell based power system with ammonia as a fuel: Experimental test and system design. *Energies* 13 (23):6173. doi: 10.3390/en13236173.

Barelli, L., G. Bidini, G. Cinti, and P. A. Ottaviano. 2022. Solid oxide fuel cell systems in hydrogen-based energy storage applications: Performance assessment in case of anode recirculation. *Journal of Energy Storage* 54 (June):105257. doi: 10.1016/j.est.2022.105257.

Bedogni, S., S. Campanari, P. Iora, L. Montelatici, and P. Silva. 2007. Experimental analysis and modeling for a circular-planar type IT-SOFC. *Journal of Power Sources* 171 (2):617–25. doi: 10.1016/j.jpowsour.2007.07.014.

Buonomano, A., F. Calise, M. D. d'Accadia, A. Palombo, and M. Vicidomini. 2015. Hybrid solid oxide fuel cells-gas turbine systems for combined heat and power: A review. *Applied Energy* 156:32–85. doi: 10.1016/j.apenergy.2015.06.027.

Cao, Y., S. Alsharif, E. A. Attia, M. A. Shamseldin, and B. F. Ibrahim. 2022. A conceptual process design towards CO2 emission reduction by integration of solar-based hydrogen production and injection into biomass-derived solid oxide fuel cell. *Process Safety and Environmental Protection* 164 (June):164–76. doi: 10.1016/j.psep.2022.05.050.

Cao, Y., H. A. Dhahad, K. Sharma, A. S. El-Shafay, A. N. Ahmed, M. A. Shamseldin, S. F. Almojil, A. I. Almohana, A. F. Alali, and B. Farhang. 2022. Techno-economic evaluation and parametric study of generating green hydrogen from waste heat recovery of efficient solid oxide fuel cell. *International Journal of Hydrogen Energy* 47 (62):26632–45. doi: 10.1016/j.ijhydene.2022.02.160.

Chatzichristodoulou, C., P. Norby, P. V. Hendriksen, and M. B. Mogensen. 2015. Size of oxide vacancies in fluorite and perovskite structured oxides. *Journal of Electroceramics* 34 (1):100–07. doi: 10.1007/s10832-014-9916-2.

Chehrmonavari, H., A. Kakaee, S. E. Hosseini, U. Desideri, G. Tsatsaronis, G. Floerchinger, R. Braun, and A. Paykani. 2023. Hybridizing solid oxide fuel cells with internal combustion engines for power and propulsion systems: A review. *Renewable and Sustainable Energy Reviews* 171 (September 2022):112982. doi: 10.1016/j.rser.2022.112982.

Cho, M., Y. Kim, and H. Ho Song. 2022. Solid oxide fuel cell–internal combustion engine hybrid system utilizing an internal combustion engine for anode off-gas recirculation, external reforming, and additional power generation. *Applied Energy* 328 (August):120146. doi: 10.1016/j.apenergy.2022.120146.

Choi, W., J. Kim, Y. Kim, and H. H. Song. 2019. Solid oxide fuel cell operation in a solid oxide fuel cell–internal combustion engine hybrid system and the design point performance of the hybrid system. *Applied Energy* 254 (June):113681. doi: 10.1016/j.apenergy.2019.113681.

Damo, U. M., M. L. Ferrari, A. Turan, and A. F. Massardo. 2019. Solid oxide fuel cell hybrid system: A detailed review of an environmentally clean and efficient source of energy. *Energy* 168 (December):235–46. doi: 10.1016/j.energy.2018.11.091.

Deangelis, D. 2004. *Solid Oxide Fuel Cell Hybrid System for Distributed Power Generation SOFC Scaleup for Hybrid and Fuel Cell Systems*. April.

Dehghan, A. R., M. A. Fanaei, and M. Panahi. 2022. Economic plantwide control of a hybrid solid oxide fuel cell - gas turbine system. *Applied Energy* 328 (November):120232. doi: 10.1016/j.apenergy.2022.120232.

Du, Y., X. Gao, Q. Mu, S. Chen, J. Lou, J. Wang, and P. Zhao. 2023. Novel self-pressurized ammonia-powered solid oxide fuel cell hybrid system using anode recirculation with water removal for improving high-altitude performance: Feasibility exploration. *Energy Conversion and Management* 295 (September):117635. doi: 10.1016/j.enconman.2023.117635.

Eisavi, B., A. Chitsaz, J. Hosseinpour, and F. Ranjbar. 2018. Thermo-environmental and economic comparison of three different arrangements of solid oxide fuel cell-gas turbine (SOFC-GT) hybrid systems. *Energy Conversion and Management* 168 (January):343–56. doi: 10.1016/j.enconman.2018.04.088.

Emadi, M. A., N. Chitgar, O. A. Oyewunmi, and C. N. Markides. 2020. Working-fluid selection and thermoeconomic optimisation of a combined cycle cogeneration dual-loop organic rankine cycle (ORC) system for solid oxide fuel cell (SOFC) waste-heat recovery. *Applied Energy* 261 (February):114384. doi: 10.1016/j.apenergy.2019.114384.

Erdogan, A., B. Dursun, C. O. Colpan, and A. Ayol. 2022. A review on performance, economic, and environmental analyses of integrated solid oxide fuel cell and biomass gasification systems. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects* 44 (4):8403–26. doi: 10.1080/15567036.2022.2121451.

Facci, A. L., V. Cigolotti, E. Jannelli, and S. Ubertini. 2017. Technical and economic assessment of a SOFC-based energy system for combined cooling, heating and power. *Applied Energy* 192 (2017):563–74. doi: 10.1016/j.apenergy.2016.06.105.

Feng, Y., J. Qu, Y. Zhu, B. Wu, Y. Wu, Z. Xiao, and J. Liu. 2023. Progress and prospect of the novel integrated SOFC-ICE hybrid power system: System design, mass and heat integration, system optimization and techno-economic analysis. *Energy Conversion and Management: X* 18 (September 2022):100350. doi: 10.1016/j.ecmx.2023.100350.

Guo, Y., Z. Yu, G. Li, and H. Zhao. 2020. Performance assessment and optimization of an integrated solid oxide fuel cell-gas turbine cogeneration system. *International Journal of Hydrogen Energy* 45 (35):17702–16. doi: 10.1016/j.ijhydene.2020.04.210.

Hai, T., A. S. El-Shafay, A. Alizadeh, H. A. Dhahad, B. S. Chauhan, S. F. Almojil, A. I. Almohana, and A. F. Alali. 2023. Comparison analysis of hydrogen addition into both anode and afterburner of fuel cell incorporated with hybrid renewable energy driven SOFC: An application of techno-environmental horizon and multi-objective optimization. *International Journal of Hydrogen Energy* xxxx:1195–207. doi: 10.1016/j.ijhydene.2023.02.016.

Hjalmarsson, P., and M. Mogensen. 2011. La0.99Co0.4Ni0.6O3-δ- Ce0.8Gd0.2O1.95 as composite cathode for solid oxide fuel cells. *Journal of Power Sources* 196 (17):7237–44. doi: 10.1016/j.jpowsour.2010.11.065.

Hong, W. T., Y. L. Wu, T. H. Yen, C. N. Huang, H. I. Tan, and C. H. Lin. 2016. Experimental investigation into performance of integrated hotbox in 1-kW solid oxide fuel cell system. *International Journal of Green Energy* 13 (7):730–35. doi: 10.1080/15435075.2015.1087852.

Hosseinpour, J., A. Chitsaz, B. Eisavi, and M. Yari. 2018. Investigation on performance of an integrated SOFC-Goswami system using wood gasification. *Energy* 148:614–28. doi: 10.1016/j.energy.2018.01.162.

Hou, R., N. Zhang, W. Gao, K. Chen, and Y. Liu. 2023. Thermodynamic, environmental, and exergoeconomic feasibility analyses and optimization of biomass gasifier-solid oxide fuel cell boosting a doable-flash binary geothermal cycle; a novel trigeneration plant. *Energy* 265 (June 2022):126316. doi: 10.1016/j.energy.2022.126316.

Hu, Y., C. Han, W. Li, Q. Hu, H. Wu, and Z. Luo. 2023. Experimental evaluation of methanol steam reforming reactor heated by catalyst combustion for kW-class SOFC. *International Journal of Hydrogen Energy* 48 (12):4649–64. doi: 10.1016/j.ijhydene.2022.10.274.

Jouhara, H., N. Khordehgah, S. Almahmoud, B. Delpech, A. Chauhan, and S. A. Tassou. 2018. Waste heat recovery technologies and applications. *Thermal Science and Engineering Progress* 6 (April):268–89. doi: 10.1016/j.tsep.2018.04.017.

Kang, I., S. Yoon, G. Bae, J. Kim, S. Baek, and J. Bae. 2010. The tests of 1 kW e diesel reformer and solid oxide fuel cell system. *Journal of Fuel Cell Science and Technology* 7 (3):0310121–25. doi: 10.1115/1.3207875.

Karapekmez, A., and I. Dincer. 2022. Development of a new solar, gasification and fuel cell based integrated plant. *International Journal of Hydrogen Energy* 47 (6):4196–210. doi: 10.1016/j.ijhydene.2021.11.011.

Kim, Y. S., Y. D. Lee, and K. Y. Ahn. 2020. System integration and proof-of-concept test results of SOFC–engine hybrid power generation system. *Applied Energy* 277 (December 2019):115542. doi: 10.1016/j.apenergy.2020.115542.

Koo, T., Y. S. Kim, Y. D. Lee, S. Yu, D. K. Lee, and K. Y. Ahn. 2021. Exergetic evaluation of operation results of 5-kW-class SOFC-HCCI engine hybrid power generation system. *Applied Energy* 295 (December 2020):117037. doi: 10.1016/j.apenergy.2021.117037.

Kumar Yadav, A., S. Sinha, and A. Kumar. 2023. Comprehensive review on performance assessment of solid oxide fuel cell-based hybrid power generation system. *Thermal Science and Engineering Progress* 46 (October):102226. doi: 10.1016/j.tsep.2023.102226.

Kupecki, J., M. Skrzypkiewicz, M. Wierzbicki, and M. Stepien. 2017. Experimental and numerical analysis of a serial connection of two SOFC stacks in a micro-chp system fed by biogas. *International Journal of Hydrogen Energy* 42 (5):3487–97. doi: 10.1016/j.ijhydene.2016.07.222.

Lee, J. I., B. Y. Yoon, and S. W. Cha. 2023. Analysis of solid oxide fuel cell hybrid power system in marine application for CO2 reduction. *Energy Reports* 9:3072–81. doi: 10.1016/j.egyr.2023.01.123.

Lenser, C., J. Zhang, N. Russner, A. Weber, O. Guillon, and N. H. Menzler. 2022. Electro-chemo-mechanical analysis of a solid oxide cell based on doped ceria. *Journal of Power Sources* 541 (April):231505. doi: 10.1016/j.jpowsour.2022.231505.

Leone, P., and A. Lanzini. 2013. Experimental modeling of transients in large SOFC systems. *Journal of Fuel Cell Science and Technology* 10 (1):1–11. doi: 10.1115/1.4023217.

Li, S., Z. Zhang, G. Li, and S. Bai. 2023. Influence of off gas recirculation on the intermediate temperature SOFC with partial oxidation reformer. *Journal of Electrochemical Energy Conversion and Storage* 20 (3):1–10. doi: 10.1115/1.4055393.

Ma, Z., J. Wang, F. Dong, Z. Han, L. Tian, R. Yan, and Z. Liang. 2022. Thermodynamic analysis of fuel-cell-based combined cooling, heating, and power system integrated solar energy and chemical looping hydrogen generation. *Energy* 238:121974. doi: 10.1016/j.energy.2021.121974.

Marocco, P., M. Gandiglio, and M. Santarelli. 2022. When SOFC-based cogeneration systems become convenient? A cost-optimal analysis. *Energy Reports* 8 (November):8709–21. doi: 10.1016/j.egyr.2022.06.015.

McLarty, D., Y. Kuniba, J. Brouwer, and S. Samuelsen. 2012. Experimental and theoretical evidence for control requirements in solid oxide fuel cell gas turbine hybrid systems. *Journal of Power Sources* 209:195–203. doi: 10.1016/j.jpowsour.2012.02.102.

Mekhilef, S., R. Saidur, and A. Safari. 2012. Comparative study of different fuel cell technologies. *Renewable and Sustainable Energy Reviews* 16 (1):981–89. doi: 10.1016/j.rser.2011.09.020.

Mottaghizadeh, P., M. Fardadi, F. Jabbari, and J. Brouwer. 2022. Thermodynamic and dynamic analysis of a wind-powered off-grid industrial building integrated with solid oxide fuel cell and electrolyzer for energy management and storage. *Journal of Electrochemical Energy Conversion and Storage* 19 (3). doi: 10.1115/1.4052856.

Narayana Sarma, R., A. M. Shivapuji, and D. Srinivasaiah. 2022. Solid oxide fuel cells fueled by carbonaceous fuels: A thermodynamics-based approach for safe operation and experimental validation. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects* 44 (2):3509–31. doi: 10.1080/15567036.2022.2066738.

Navasa, M., C. Graves, C. Chatzichristodoulou, T. Løye Skafte, B. Sundén, and H. Lund Frandsen. 2018. A three dimensional multiphysics model of a solid oxide electrochemical cell: A tool for understanding degradation. *International Journal of Hydrogen Energy* 43 (27):11913–31. doi: 10.1016/j.ijhydene.2018.04.164.

Ni, T., J. Si, X. Gong, K. Zhang, and M. Pan. 2022. Thermodynamic and economic analysis of a novel cascade waste heat recovery system for solid oxide fuel cell. *Energy Conversion and Management* 259 (January):115562. doi: 10.1016/j.enconman.2022.115562.

Ni, W., T. Zhu, X. Chen, Q. Zhong, and W. Ma. 2020. Stable, efficient and cost-competitive Ni-substituted Sr(Ti,Fe)O3 cathode for solid oxide fuel cell: Effect of A-site deficiency. *Journal of Power Sources* 451 (October 2019):227762. doi: 10.1016/j.jpowsour.2020.227762.

Ovtar, S., X. Tong, J. J. Bentzen, K. T. S. Thydén, S. B. Simonsen, and M. Chen. 2019. Boosting the performance and durability of Ni/YSZ cathode for hydrogen production at high current densities: Via decoration with nano-sized electrocatalysts. *Nanoscale* 11 (10):4394–406. doi: 10.1039/c8nr07678b.

Palomba, V., M. Ferraro, A. Frazzica, S. Vasta, F. Sergi, and V. Antonucci. 2018. Experimental and numerical analysis of a SOFC-CHP system with adsorption and hybrid chillers for telecommunication applications. *Applied Energy* 216 (February):620–33. doi: 10.1016/j.apenergy.2018.02.063.

Park, S. H., Y. D. Lee, and K. Y. Ahn. 2014. Performance analysis of an SOFC/HCCI engine hybrid system: System simulation and thermo-economic comparison. *International Journal of Hydrogen Energy* 39 (4):1799–810. doi: 10.1016/j.ijhydene.2013.10.171.

Paulus, N. 2024. Experimental assessment of pollutant emissions from residential fuel cells and comparative benchmark analysis. *Journal of Environmental Management* In Press (6):121017. doi: 10.1016/j.jenvman.2024.121017.

Paulus, N., and V. Lemort. 2023. Experimental investigation of a solid oxide fuel cell (SOFC) used in residential cogeneration applications. *36th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems, ECOS 2023*, 613–24. doi: 10.52202/069564-0056.

Peng, M. Y. P., C. Chen, X. Peng, and M. Marefati. 2020. Energy and exergy analysis of a new combined concentrating solar collector, solid oxide fuel cell, and steam turbine CCHP system. *Sustainable Energy Technologies and Assessments* 39 (April):100713. doi: 10.1016/j.seta.2020.100713.

Perna, A., M. Minutillo, E. Jannelli, V. Cigolotti, S. W. Nam, and J. Han. 2018. Design and performance assessment of a combined heat, hydrogen and power (CHHP) system based on ammonia-fueled SOFC. *Applied Energy* 231 (September):1216–29. doi: 10.1016/j.apenergy.2018.09.138.

Pirkandi, J., M. Mahmoodi, and M. Ommian. 2017. An optimal configuration for a solid oxide fuel cell-gas turbine (SOFC-GT) hybrid system based on thermo-economic modelling. *Journal of Cleaner Production* 144:375–86. doi: 10.1016/j.jclepro.2017.01.019.

Power, E. A. 2023. *Solid oxide fuel cell market (by type: Linear, Tubular; by application: Portable, stationary, Transport; by end user, commercial & industrial, data centers, military & defense, residentials) - global industry analysis, size, share, growth, trends, Regiona*. https://www.precedenceresearch.com/solid-oxide-fuel-cell-market.

Ran, P., X. Zhou, Y. Wang, Q. Fan, D. Xin, and Z. Li. 2023. Thermodynamic and exergetic analysis of a novel multi-generation system based on SOFC, micro-gas turbine, s-CO2 and lithium bromide absorption refrigerator. *Applied Thermal Engineering* 219 (PB):119585. doi: 10.1016/j.applthermaleng.2022.119585.

Rosner, F., and S. Samuelsen. 2022. Thermo-economic analysis of a solid oxide fuel cell-gas turbine hybrid with commercial off-the-shelf gas turbine. *Applied Energy* 324 (June):119745. doi: 10.1016/j.apenergy.2022.119745.

Roushenas, R., E. Zarei, and M. Torabi. 2021. A novel trigeneration system based on solid oxide fuel cell-gas turbine integrated with compressed air and thermal energy storage concepts: Energy, exergy, and life cycle approaches. *Sustainable Cities and Society* 66 (December 2020):102667. doi: 10.1016/j.scs.2020.102667.

Sharifzadeh, M., M. Meghdari, and D. Rashtchian. 2017. Multi-objective design and operation of solid oxide fuel cell (SOFC) triple combined-cycle power generation systems: Integrating energy efficiency and operational safety. *Applied Energy* 185:345–61. doi: 10.1016/j.apenergy.2016.11.010.

Singh, R., and O. Singh. 2018. Comparative study of combined solid oxide fuel cell-gas turbine-organic rankine cycle for different working fluid in bottoming cycle. *Energy Conversion and Management* 171 (April):659–70. doi: 10.1016/j.enconman.2018.06.009.

Sinha, V., and S. Mondal. 2018. Recent development on performance modelling and fault diagnosis of fuel cell systems. *International Journal of Dynamics and Control* 6 (2):511–28. doi: 10.1007/s40435-017-0328-4.

Skabelund, B. B., and R. J. Milcarek. 2022. Review of thermal partial oxidation reforming with integrated solid oxide fuel cell power generation. *Renewable and Sustainable Energy Reviews* 168 (August):112852. doi: 10.1016/j.rser.2022.112852.

Subotić, V., and C. Hochenauer. 2022. Analysis of solid oxide fuel and electrolysis cells operated in a real-system environment: State-of-the-health diagnostic, failure modes, degradation mitigation and performance regeneration. *Progress in Energy and Combustion Science* 93(April). doi: 10.1016/j.pecs.2022.101011.

Talukdar, A., A. Chakrovorty, P. Sarmah, P. Paramasivam, V. Kumar, S. K. Yadav, S. Manickkam, and M. Ahmed. 2024. A review on solid oxide fuel cell technology: An efficient energy conversion system. *International Journal of Energy Research* 2024:1–20. doi: 10.1155/2024/6443247.

Tamburrano, G., D. Pumiglia, A. Monforti Ferrario, F. Santoni, and D. Borello. 2024. Analysis of the performances of a solid oxide fuel cell fed by biogas in different plant configurations: An integrated experimental and simulative approach. *International Journal of Hydrogen Energy* 52:745–60. doi: 10.1016/j.ijhydene.2023.07.324.

Tapiero, D., D. Diskin, N. Tzabar, and L. Tartakovsky. 2023. Performance of solid-oxide fuel cells fed with pure hydrogen and methanol steam reforming products – experimental study. *International Journal of Hydrogen Energy* xxxx:1–14. doi: 10.1016/j.ijhydene.2023.07.180.

Ullah, K. R., R. K. Akikur, H. W. Ping, R. Saidur, S. A. Hajimolana, and M. A. Hussain. 2015. An experimental investigation on a single tubular SOFC for renewable energy based cogeneration system. *Energy Conversion and Management* 94:139–49. doi: 10.1016/j.enconman.2015.01.055.

Wang, J., Z. Cui, W. Yao, and S. Huo. 2023. Regulation strategies and thermodynamic analysis of combined cooling, heating, and power system integrated with biomass gasification and solid oxide fuel cell. *Energy* 266 (June 2022):126430. doi: 10.1016/j.energy.2022.126430.

Wang, K., B. An, Z. Chu, Y. Wang, and Z. Liu. 2021. Evaluation of gas channel structures on cell performance and thermal stress in a solid oxide fuel cell. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects* 1–16. doi: 10.1080/15567036.2021.1916655.

Wang, S., X. Liu, X. Gu, X. Huang, and Y. Li. 2023. Analysis and multi-objective optimization of integrating a syngas-fed solid oxide fuel cell improved by a two-stage expander-organic flash cycle using an ejector and a desalination cycle. *Energy* 272 (July 2022):127095. doi: 10.1016/j.energy.2023.127095.

Wu, X. J., X. J. Zhu, G. Y. Cao, H. Y. Tu, and W. Q. Hu. 2009. A hybrid experimental model of a solid oxide fuel cell stack. *Journal of Fuel Cell Science and Technology* 6 (1):0110131–34. doi: 10.1115/1.2971125.

Wu, X. L., H. Zhang, H. Liu, Y. W. Xu, J. Peng, Z. Xia, and Y. Wang. 2022. Modeling analysis of SOFC system oriented to working condition identification. *Energies* 15 (5):1804–19. doi: 10.3390/en15051804.

Yadav, A. K., S. Sinha, and A. Kumar. 2024. Advancements in composite cathodes for intermediate-temperature solid oxide fuel cells: A comprehensive review. *International Journal of Hydrogen Energy* 59 (February):1080–93. doi: 10.1016/j.ijhydene.2024.02.124.

Yan, Z., P. Zhao, J. Wang, and Y. Dai. 2013. Thermodynamic analysis of an SOFC-GT-ORC integrated power system with liquefied natural gas as heat sink. *International Journal of Hydrogen Energy* 38 (8):3352–63. doi: 10.1016/j.ijhydene.2012.12.101.

Yang, Y., F. Chen, J. Yu, Y. Song, and Z. Guo. 2022. Design and experiment study of a micro radial-flow turbine for a SOFC-MGT turbine hybrid system. *Energy Conversion and Management* 266 (June):115861. doi: 10.1016/j.enconman.2022.115861.

Yen, T. H., W. T. Hong, W. P. Huang, Y. C. Tsai, H. Y. Wang, C. N. Huang, and C. H. Lee. 2010. Experimental investigation of 1 kW solid oxide fuel cell system with a natural gas reformer and an exhaust gas burner. *Journal of Power Sources* 195 (5):1454–62. doi: 10.1016/j.jpowsour.2009.09.021.

Yuan, Q., X. Li, S. Han, S. Wang, M. Wang, R. Chen, S. Kudashev, T. Wei, and D. Chen. 2024. Performance analysis and optimization of SOFC/GT hybrid systems: A review. *Energies* 17 (5):1–22. doi: 10.3390/en17051265.

Zarabi Golkhatmi, S., M. I. Asghar, and P. D. Lund. 2022. A review on solid oxide fuel cell durability: Latest progress, mechanisms, and study tools. *Renewable and Sustainable Energy Reviews* 161 (February):112339. doi: 10.1016/j.rser.2022.112339.

Zhang, L., W. Tang, F. Wang, C. Xie, W. Zhou, and H. Xie. 2022. Optimization and control for solid oxide fuel cell system hybrid DC microgrids from the perspective of high efficiency, thermal safety, and transient response. *Frontiers in Energy Research* 10 (July):1–11. doi: 10.3389/fenrg.2022.953082.

Zheng, Z., T. Liu, Q. Liu, J. Lei, and J. Fang. 2021. A distributed energy system integrating SOFC-MGT with mid-and-low temperature solar thermochemical hydrogen fuel production. *International Journal of Hydrogen Energy* 46 (38):19846–60. doi: 10.1016/j.ijhydene.2021.03.137.

Zhou, Y., Q. Hua, P. Liu, and L. Sun. 2022. Multi-objective optimal droop control of solid oxide fuel cell based integrated energy system. *International Journal of Hydrogen Energy* 48 (30):11382–89. doi: 10.1016/j.ijhydene.2022.04.071.

Zhu, S., G. Yu, O. Jongmin, T. Xu, Z. Wu, W. Dai, and E. Luo. 2018. Modeling and experimental investigation of a free-piston stirling engine-based micro-combined heat and power system. *Applied Energy* 226 (May):522–33. doi: 10.1016/j.apenergy.2018.05.122.

Zoupalis, K., A. Amiri, K. Sugden, M. Kendall, and K. Kendall. 2023. Micro solid oxide fuel cell thermal dynamics: Incorporation of experimental measurements and model-based estimations for a multidimensional thermal analysis. *Energy Conversion and Management* 277 (July 2022):116650. doi: 10.1016/j.enconman.2022.116650.

Journal of Electronic Materials
https://doi.org/10.1007/s11664-024-11463-5

ORIGINAL RESEARCH ARTICLE

Fabrication of Lead-Free PVDF/KNNLTS/MWCNT Piezoelectric Nanogenerator: Role of MWCNT in the Piezoelectric Performance of Nanogenerator for Energy-Harvesting Application

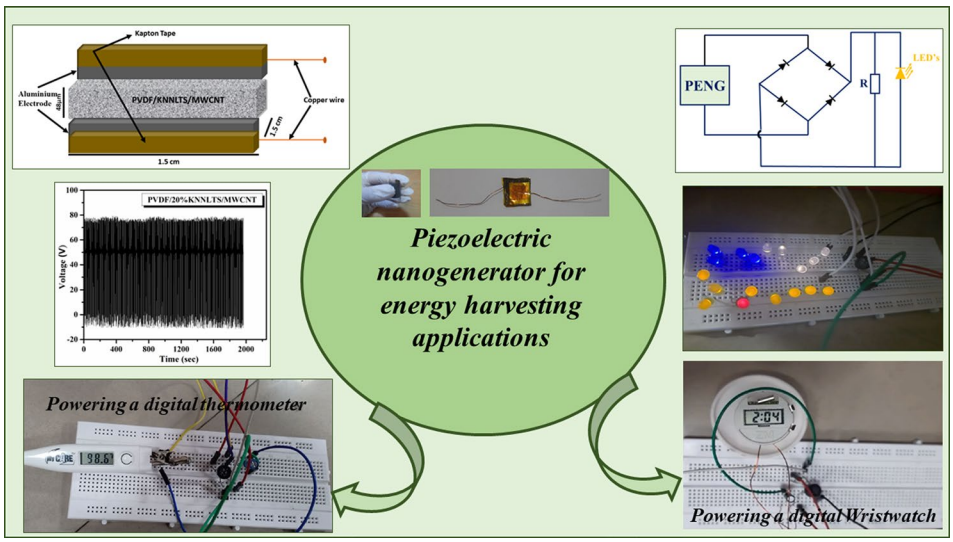
Komal Verma[1] · **Aman Kumar**[1] · **Richa Sharma**[1]

Received: 4 June 2024 / Accepted: 9 September 2024
© The Minerals, Metals & Materials Society 2024

Abstract

Lead-free polyvinylidene fluoride/potassium sodium niobate doped with lithium, tantalum, and antimony/multiwall carbon nanotube (PVDF/KNNLTS/MWCNT)-based piezoelectric nanocomposite film has been synthesized by drop-casting, in which KNNLTS nanoparticles were synthesized by using a high-energy ball mill and then incorporated into the PVDF matrix together with MWCNT. The nanocomposite films offer an eco-friendly solution to energy-harvesting applications. The objective of this present work demonstrates the role of MWCNT on KNNLTS (0–25 wt%)-based piezoelectric nanogenerators (PENGs). It was observed that the maximum piezoelectric output voltage for PVDF/KNNNLTS/MWCNT with 20wt% KNNLTS was 87.2 V, while the maximum current was 20.3 μA and power density of 145.48 μW/cm^2. Further, the PENG was utilized to check the potential performance of human body motion. The output voltages for finger tapping, thumb pressing, fist beating, elbow beating, elbow bending, and wrist stretching were 11.4 V, 27.2 V, 24.8 V, 34 V, 37 V, and 55.2 V, respectively. In addition, the converse piezoelectric coefficient was also measured, with average and maximum values of $d_{33}{}^{*}$ ~ 28.67 pm/V and 59.92 pm/V, respectively. Moreover, to test the ability to harvest energy, the nanocomposite film was integrated with conventional electronic components like a full-wave bridge rectifier and resistances. The fabricated PENG demonstrated practical applicability by powering a digital thermometer and digital wrist watch, and successfully lit 18 LEDs. In this study, we have analyzed lead-free piezoelectric nanogenerators which might be considered a good option for self-powered wearable and portable electronic gadgets that need a very small power source.

Graphical Abstract



Extended author information available on the last page of the article

Published online: 25 September 2024

Keywords PVDF · piezoelectric · KNNLTS · nanogenerator · nanocomposite · MWCNT

Introduction

Energy deficiency has become a global problem in the twenty-first century.[1–3] Over the past few decades, significant efforts have been undertaken to find alternative energy sources due to the challenges posed by the fluctuation of coal and petroleum fuel costs, the threat of climate change, and the anticipated use of non-sustainable fuel sources.[4–6] The energy crisis has become the inspiration for the fabrication of beneficial materials to harvest energy from the surrounding environment.[7,8] The rapid development of nanotechnology has garnered significant interest as it involves ongoing research and the creation of materials with superior properties, which has led to encouraging discoveries.[9] In 1880, Jacques and Pierre Curie, two Nobel laureates, discovered the piezoelectric phenomenon.[10] In 2006, Wang et al. made a nanogenerator based on the piezoelectric effect bringing scientific advancements closer to the objective of harnessing and collecting an abundance of mechanical energy.[11] Later, Wang and his colleagues also presented the concept of a nanogenerator that made use of triboelectric phenomena. They called the device a piezoelectric nanogenerator (PENG) as they offered the possibility of low cost, light weight, good flexibility, high power density, and highly efficient nanogenerators.[12–14] Over the past 10 years, research on collecting ambient mechanical energy has been driven by piezoelectric nanogenerators.[15] Many technologies, including self-powered devices, sustainable wearable devices, sensors, and actuators, have been made possible by the ability to harvest energy from human body motions.[16–18] So, nanogenerators have paved the way for the development of environmentally friendly power sources.

Selecting the appropriate materials is crucial for the piezoelectric effect.[19,20] Polyvinylidene fluoride (PVDF) is one of the most adaptable piezoelectric polymers with its low density, high flexibility, low resistance, and excellent piezoelectric response.[21,22] The crystalline polymorphs of PVDF include α, β, γ, δ, and ε, which vary according to the chain conformation.[23–25] Among these, the β and γ phases are the most important crystalline phases due to their spontaneous polarization and great piezoelectric sensitivity. There are several techniques for inducing the β phase in PVDF.[26] Poling, annealing, and stretching processes are utilized to convert PVDF from the α phase to the β phase.[27–29] The use of carbon nanotubes (CNTs) as fillers in a PVDF matrix results in a significant increase in the β phase, as demonstrated by Yu et al. and Kim et al.[30,31] Additionally, by forming a three-dimensional network of conductive (multi-walled carbon nanotubes) MWCNTs inside the polymer matrix, MWCNT increase the conductivity of composites by facilitating the electron transport during piezoelectric operations.[8]

To fabricate flexible piezoelectric energy harvesters, a variety of inorganic materials, including lead zirconium titanate (PZT), barium titanate ($BaTiO_3$), zinc oxide (ZnO), potassium sodium niobate ($KNaNbO_3$), zinc stannate ($ZnSnO_3$), gallium nitride (GaN), and molybdenum disulfide (MoS_2) nanoparticles[32–35] have been investigated for use in PVDF-based composites. PZT has shown the higher value of piezoelectric voltage and dielectric constants compared to other materials. However, PZT is lead-based and highly toxic, so harmful to the environment, and is probably not ideal for wearable applications.[36–38] Apart from these, alkali niobates, especially perovskite ABO_3 such as $KNaNbO_3$, have been considered promising lead-free piezoelectric ceramics.[39,40] Teka et al. described their work with KNN as a nano-piezoelectric filler because of its significant piezoelectric response, ferroelectric properties, high Curie temperature (> 400°C), low cost, and light weight requirements.[41] However, for practical and commercial uses, KNN-based nanogenerators still need to produce more energy. A piezoelectric nanogenerator based on PVDF as the polymer and KNNCNT as filler material was created by Bairagi et al. which showed a maximum output of 23.24 V and 18 μA; in the absence of the CNT, the outputs were ~ 17.5 V and ~ 0.522 μA.[42] To further improve the piezoelectric output performances of piezoelectric energy harvesters, doping of KNN ceramics has been carried out with appropriate dopants. Gao et al. synthesized KNN ceramics doped with Li, Ta, and Sb with piezoelectric coefficients of 413 pC/N.[43] Furthermore, there are many reports in the literature which demonstrate the superior piezoelectric properties of KNN ceramics doped with Li, Ta, and Sb elements compared to pure KNN ceramics.[44–46] To our knowledge, there has not been so much research work related to the KNNLTS ceramic-based piezoelectric generators. Wang and colleagues, for example, fabricated a composite-type piezoelectric nanogenerator by using Ag particles, CNT, and $(K_{0.4425}Na_{0.52}Li_{0.0375})(Nb_{0.86}Ta_{0.06}Sb_{0.08})O_3$ (KNNLTS) nanoparticles in a PDMS (polydimethylsiloxane) matrix. They reported a piezoelectric output voltage of 282 V and a short-circuit current of 32.2 mA.[47] Another work by Gupta et al.[48] used PDMS combined with $Na_{0.47}K_{0.47}Li_{0.06}NbO_3$ (KNNLi) to generate a piezoelectric energy-harvester device. By applying a vertical compressive force of ~ 2 kgf, the device produced a gigantic piezoelectric output voltage of 48 V. Based on the above literature, we have employed a

drop-casting approach in the current study to fabricate flexible PVDF/KNNLTS/MWCNT nanocomposite film.

In this current work, we have created lead-free and environmentally friendly nanocomposite films for energy-harvesting applications, where mechanical energy is transformed into electrical energy through piezoelectric mechanisms. To do this, we have synthesized KNNLTS using a high-energy ball mill process and then added synthesized KNNLTS and MWCNT to the PVDF. Scanning electron microscopy (SEM) and X-ray diffraction (XRD) spectroscopy were used to examine the synthesized KNNLTS particles. The nanocomposite film was attached with aluminum electrodes transformed into a PENG, and 4-Hz frequency was used for testing. In addition, the nanocomposite film was studied by human body motions (finger tapping, elbow bending, wrist stretching, fist beating, foot tapping). The influence of piezoelectricity was investigated by building a PENG. Additionally, different kinds of nanocomposite films (pure PVDF, PVDF/KNNLTS, PVDF/10%KNNLTS/MWCNT, PVDF/15%KNNLTS/MWCNT, PVDF/20%KNNLTS/MWCNT, and PVDF/25%KNNLTS/MWCNT) have been synthesized and tested. The PENG has shown a significantly improved piezoelectric performance with an output open-circuit voltage (V_{OC}) ~ 87.2 V and output short-circuit current (I_{SC}) ~ 20.3 μA for PVDF/20% KNNLTS/MWCNT) compared to the others. The existence of MWCNT with the PVDF/KNNLTS solution indicated a β-nucleating effect incorporated with the PVDF matrix, as proven by the beta fraction F(β) value of the materials (58.7% for PVDF/20% KNNLTS/MWCNT). In addition to forming a conducting path within the piezo device, MWCNT performs an additional key role by contributing conductivity to the PVDF/KNNLTS/MWCNT solution.

To further test the PENG's compatibility with traditional electrical components, resistors, LEDs, and a full wave bridge rectifier were combined with it. Overall, the material is a very strong contender for developing a nanogenerator for useful applications due to the combined effect of adding MWCNT into PVDF/KNNLTS.

Experimental

Synthesis and Materials

Materials

PVDF powder (Mw = 263,000 g mol^{-1}) was purchased from Alfa-Aesar. PVDF serves as a binder and dispersing agent in the ceramic synthesis process, which can impart some flexibility and mechanical stability to the composite material. N,N dimethylformamide (DMF; 99% purity) was procured from Thermo Fischer. DMF is used as a solvent in the synthesis process. It dissolves PVDF and helps in the dispersion of other materials, creating a uniform slurry. Sodium carbonate (Na_2CO_3, 99.8%), potassium carbonate (K_2CO_3, 99%), niobium oxide (Nb_2O_5, 99%), lithium carbonate (Li_2CO_3, 99%), tantalum oxide (Ta_2O_5, 99.95%), and antimony oxide (Sb_2O_5, 99%) raw ingredients were acquired from Alfa Aesar for the synthesis of $(K_{0.45}Na_{0.55})_{0.98}Li_{0.02}(Nb_{0.77}Ta_{0.18}Sb_{0.05})$ O_3 (KNNLTS) ceramic powder. In the KNNLTS composition, sodium (Na) and potassium (K) ions help to form the perovskite structure by substituting into the A-site of the ABO_3 perovskite lattice. On the other hand, niobium (Nb) ions occupy the B-site in the perovskite structure. Both these ions play a key role in the electrical properties of the ceramic, particularly in influencing its dielectric and piezoelectric responses. Lithium (Li) ions substitute into the KNN matrix, modifying the crystal structure and enhancing the piezoelectric properties. Tantalum (Ta) ions also occupy the B-site in the perovskite structure alongside niobium. Tantalum can modify the electrical properties of the ceramic, while antimony (Sb) ions can affect the grain growth, dielectric properties, and can improve the material's sinterability, which is essential for achieving a high-density and uniform microstructure. MWCNTs (> 98%) were procured from Sigma Aldrich. They can enhance the composite material's overall conductivity and mechanical properties, providing a synergistic effect when combined with the ceramic matrix. All the materials listed above were used as received without any further purification. Each component plays a specific role in the synthesis of the final composites.

Synthesis of KNNLTS Nanoparticles

KNNLTS nanoparticles were synthesized by using a solid-state reaction. All the raw ingredients in a stoichiometric ratio were combined in a ball mill. The propanol as a medium was used in a plastic jar filled with zirconia balls and left for 10 h. The obtained slurry was dried for 24 h at 65°C in a hot-air oven. After drying, the powder was calcined in the furnace at 900°C for 4 h in order to obtain KNNLTS ceramic powder, which was further ground by using a Retsch PM_{100} planetary high-energy ball mill. A sun wheel and grinding jar with powder rotating in the opposite direction with a speed ratio of 1:2 were used. The weight ratio between the raw materials and milling balls was maintained at 1:5. The KNNLTS calcined powder was milled at speed of 100–400 rpm at a regular interval of 50 rpm with the milling time constant for 20 h. To avoid excessive heating of the milling system, the planetary ball mill was configured so that the sun wheel's rotational direction and vial reversed once every 6 min, with a 2-min break, as shown in Fig. 1.

Development of Nanocomposite Based on PVDF/KNNLTS/MWCNTs

To develop the flexible nanocomposite film based on PVDF/KNNLTS/MWCNTs, 1 g of PVDF powder was first dissolved in 10 ml DMF solution with continuous stirring at 50°C until a transparent clear solution was obtained. Simultaneously, a dispersion solution of KNNLTS/MWCNTs was separately prepared in DMF solution in which KNNLTS was taken as (0–25 wt%) of PVDF and MWCNT ~ 0.005 g, as listed in Table I and then added into the PVDF solution, stirring constantly for 10 h. After that, the mixture was ultrasonicated for 30 m at room temperature, and a uniform dispersion of PVDF/KNNLTS/MWCNTs was prepared. This dispersion was evenly distributed into a Petri dish and placed in a hot-air oven for drying at 80°C. After this process, flexible PVDF/KNNLTS/MWCNTs nanocomposite films were finally peeled off from the Petri dish. For comparison, films of pure PVDF and PVDF/KNNLTS were also prepared by using a similar method. The detailed synthesis process is shown in Fig. 2.

Fabrication of Piezoelectric Nanogenerator Using the PVDF/KNNLTS/MWCNTs Nanocomposites

The developed PVDF/KNNLTS/MWCNTs nanocomposite films were cut into samples of 1.5 × 1.5 cm^2. Conductive aluminum tape was connected on both ends of the nanocomposite film, which served as an electrode material. For connection, copper wires were used with top and bottom electrodes to form the complete device. As seen in Fig. 3, the entire assembly was then wrapped with Kapton tape to shield it from the environment.

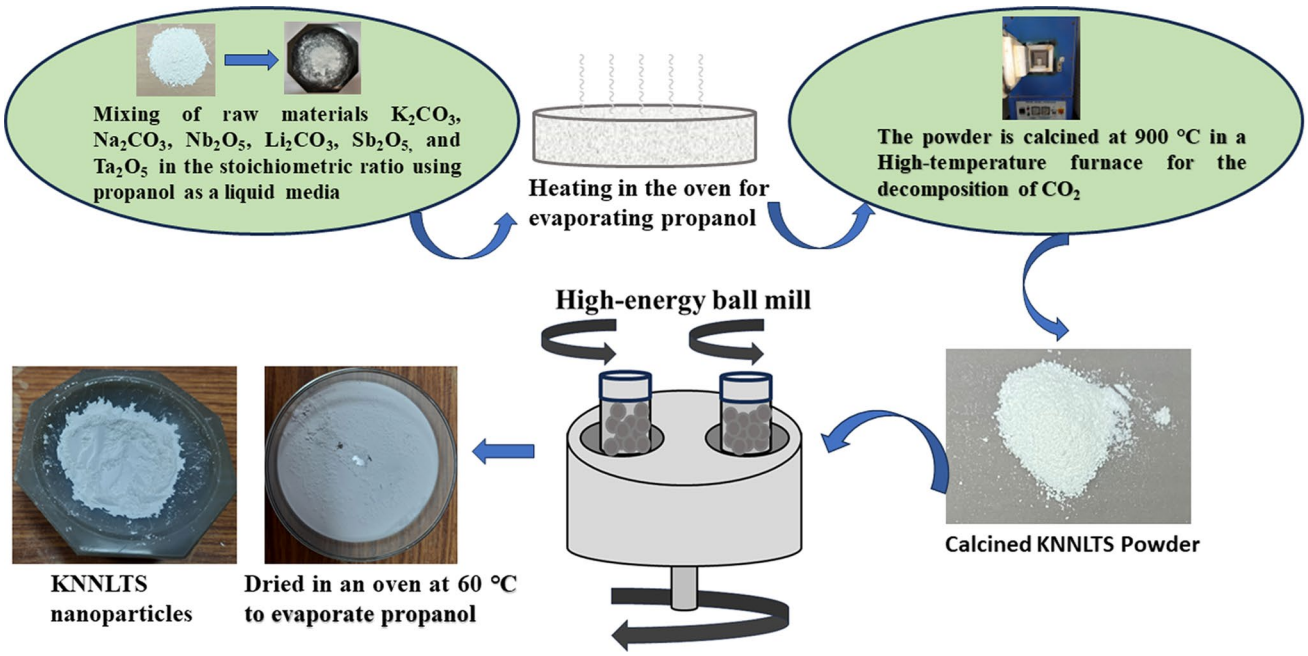


Fig. 1 Synthesis of KNNLTS nanoparticles by a high-energy ball mill via a solid-state reaction.

Table I Details of weight (g) of PVDF, KNNLTS, and MWCNT and volume (ml) of the DMF in the fabricated flexible nanocomposite films

No.	Polymer/ceramic	Amount (g)			Volume (ml)
		PVDF	KNNLTS	MWCNT	DMF
1	PVDF	1	0	0	10
2	PVDF/0%KNNLTS/MWCNT	1	0.00	0.005	10
3	PVDF/10%KNNLTS/MWCNT	0.9	0.10	0.005	10
4	PVDF/15%KNNLTS/MWCNT	0.85	0.15	0.005	10
5	PVDF/20%KNNLTS/MWCNT	0.80	0.20	0.005	10
6	PVDF/25%KNNLTS/MWCNT	0.75	0.25	0.005	10

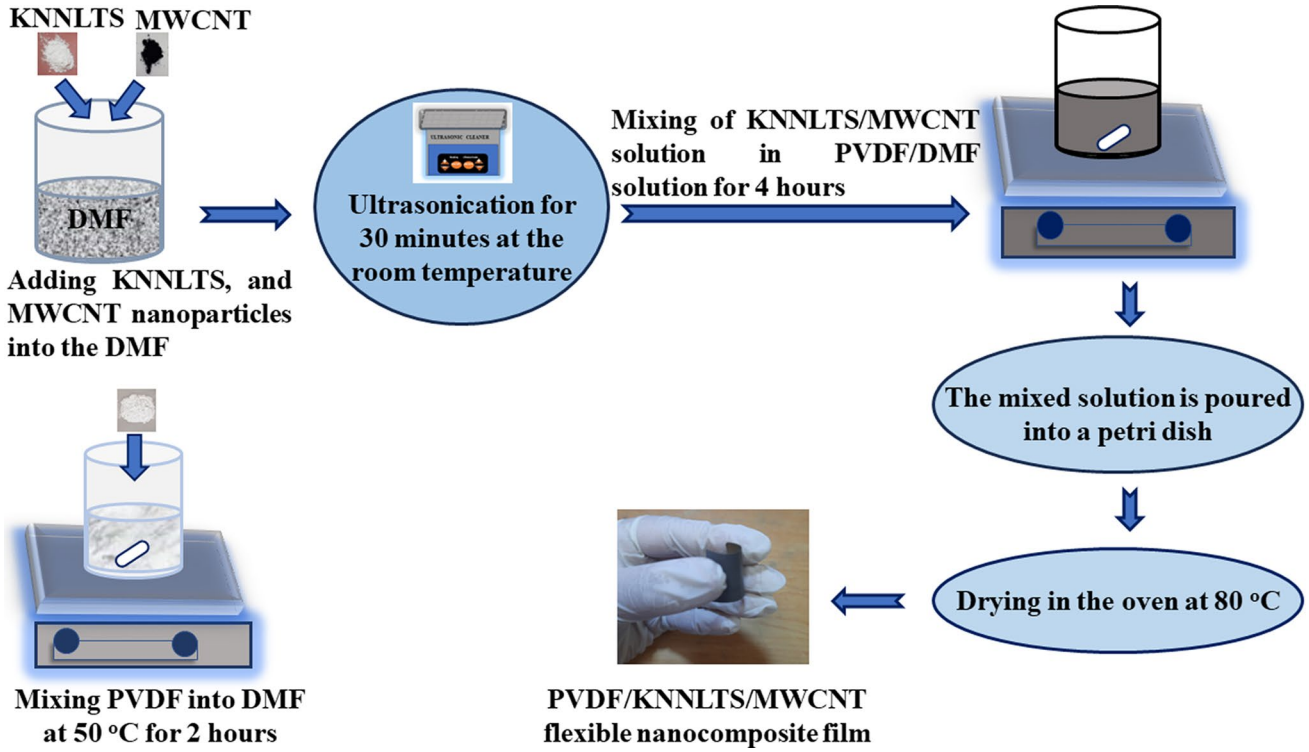


Fig. 2 Synthesis of PVDF/KNNLTS/MWCNTs-based flexible nanocomposite films.

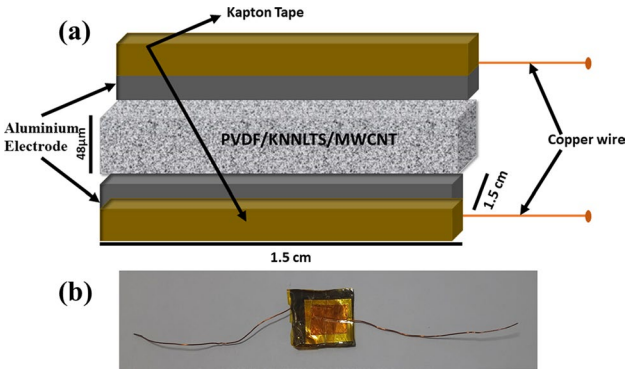


Fig. 3 (a) Schematic of the fabricated PENG device and (b) digital image of piezoelectric nanogenerator.

Characterization

The formation of α and β crystal phases in the nanocomposite films was evaluated by Fourier-transform infrared (FTIR; Perkin Elmer Spectrum-II) spectroscopy with a scanning range of 400–4000 cm^{-1}. The FTIR spectra showed significant peaks at 762 cm^{-1} denoted by the α phase. In addition, the peaks at 512 cm^{-1}, 840 cm^{-1}, and 1075 cm^{-1} strongly correspond to the presence of the β phase. The presence of the β phase in the film is very significant for the piezoelectric effect. To further corroborate the FTIR data, XRD (ULTIMA-IV; RIGAKU) patterns were employed with a Cu-Kα radiation source of 1.54 Å at a scanning rate of 2° min^{-1}. XRD analysis was performed to determine the crystalline structure of the material, and the pattern of the KNN-based ceramics, with peak splitting at the [002] and [200] crystalline planes around ~ 45° with the peak intensity ratio approaching 1:1, suggested the presence of coexistence of the orthorhombic and tetragonal phases. The diffraction peak corresponding to the nanocomposites around 18.6° can be assigned to the α phase of the PVDF which is non-polar, and the peak around 20.2° corresponds to the β phases of the PVDF which is polar. Again, the β phase is the most popular phase for its piezoelectric response. The morphology of the developed nanocomposite films was analyzed by SEM (USIC). ImageJ software was used to measure the average particle size of the KNNLTS powder. The thickness of the nanocomposite film was measured by a digital microscope. A digital storage oscilloscope (DSO; MDO500; Tektronix) and an electrometer/high-resistance meter (B2985B; Keysight) were used to measure the piezoelectric performance (output voltage) of the piezoelectric nanogenerator developed by the PVDF/KNNLTS/MWCNTs nanocomposites. The converse piezoelectric measurement was performed using a strain measurement system.

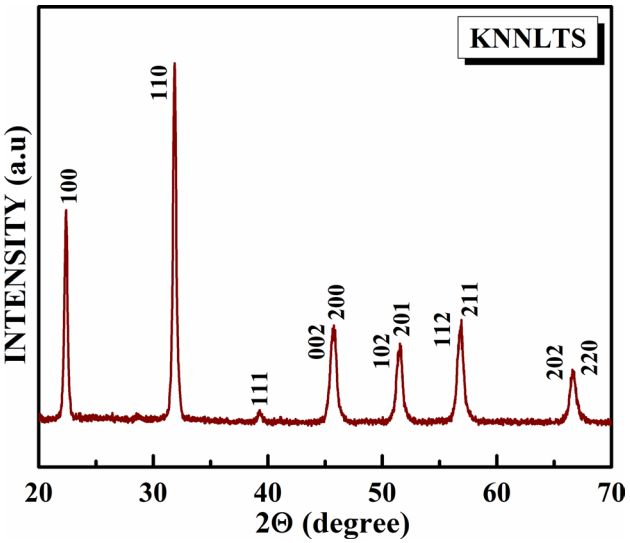


Fig. 4 XRD pattern of calcined KNNLTS nanoparticles.

Results and Discussion

Structural and Microstructural Analysis of KNNLTS Nanoparticles

XRD was used to analyze the crystal structural phases of the calcined KNNLTS powder and the corresponding plot is displayed in Fig. 4, in which it can be seen that the KNNLTS powder clearly exhibits a perovskite structure. The XRD plot shows no any extra peak, indicating that the KNNLTS powder has been successfully synthesized without any impurities. The structure of the KNNLTS powder was confirmed according to the JCPDS card numbers 71-2171 (orthorhombic) and 77-0037 (tetragonal), in the XRD patterns of the KNN-based ceramics by the peak splitting at the [002] and [200] crystalline planes around ~ 45° with apeak intensity ratio approaching 1:1, suggesting the presence of the coexistence of the orthorhombic and tetragonal phases.[49]The detail of this has been described in our previous paper.[50]

"The presence of different phases within a material plays a crucial role in determining its overall properties. The electrical properties (piezoelectric properties), mechanical properties (strength and hardness, ductility, and toughness), thermal properties (thermal stability, thermal conductivity), and optical properties (refractive index and transparency) of a material can be significantly affected by the presence of different phases. The piezoelectric property is highly dependent on the crystal structure of the phase composition, distribution, and transition. The presence and distribution of different phases within a material are fundamental to tailoring its properties for specific applications, from high-strength alloys to advanced composite materials".

Further, a SEM analysis was performed on the synthesized KNNLTS powder to analyze its morphology; the resulting picture is shown in Fig. 5. The scan revealed that the morphology of KNNLTS particles exhibit square or rectangular in form.

Structural Characterization Using XRD and FTIR Spectra of Nanocomposites

XRD was also used to analyze the structural phases of pure PVDF and PVDF/KNNLTS/MWCNT nanocomposite

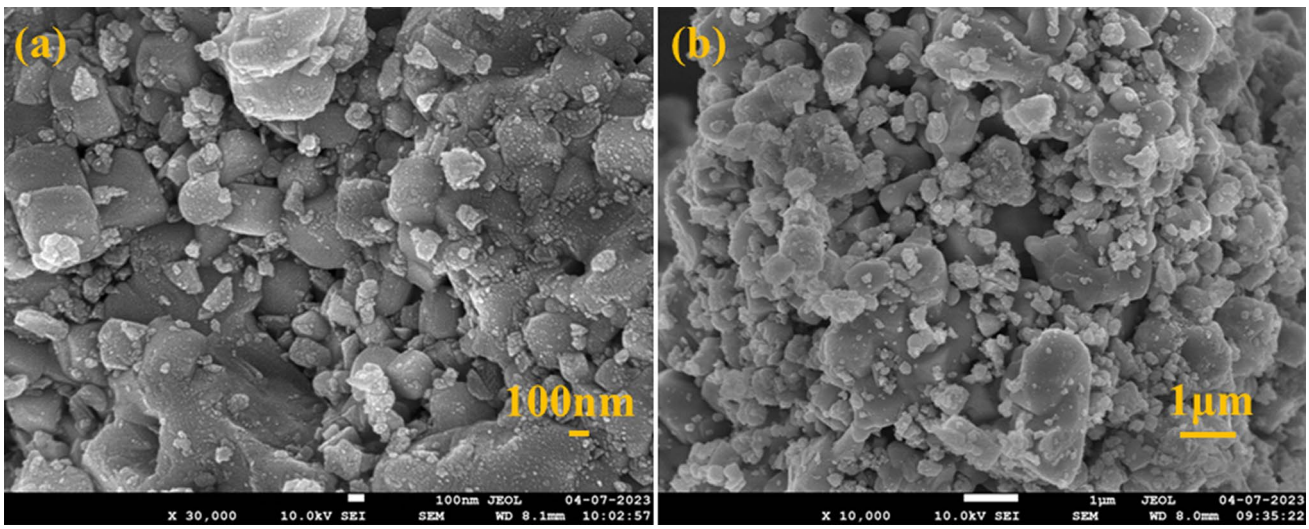


Fig. 5 SEM image of calcined KNNLTS nanoparticles (a) 100 nm scale and (b) 1 μm scale.

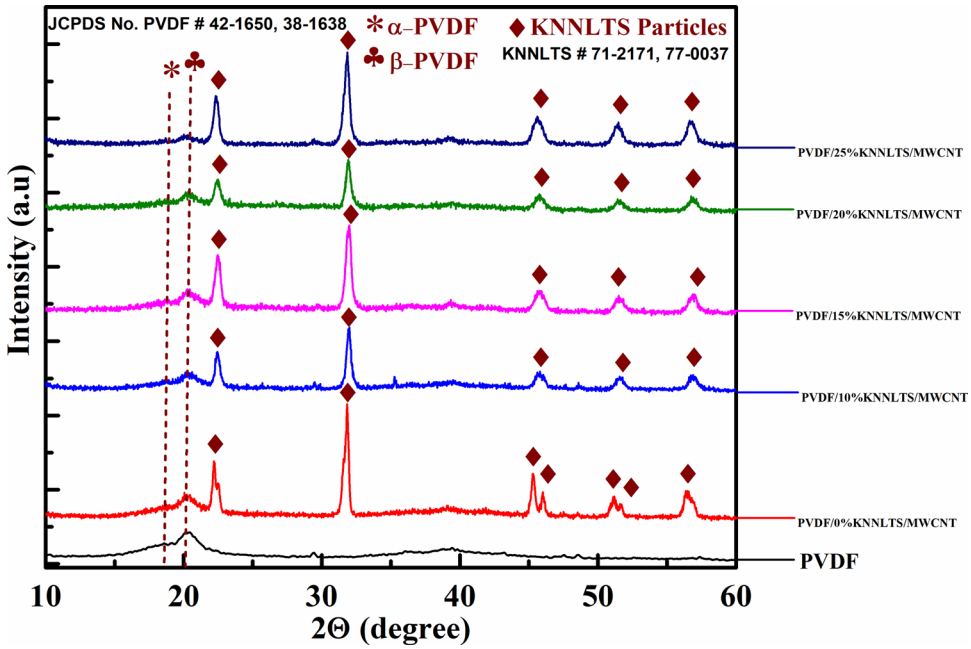


Fig. 6 X-ray diffraction patterns of the flexible nanocomposite films produced by pure PVDF and different contents of KNNLTS nanoparticles and MWCNT in the PVDF matrix.

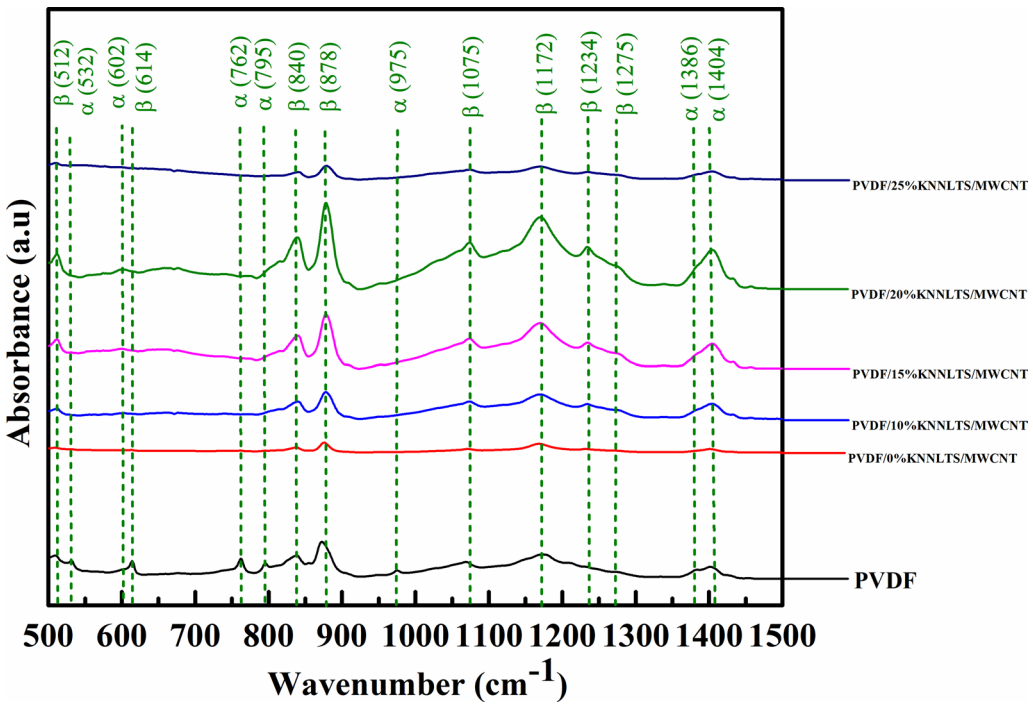


Fig. 7 FTIR spectra of the flexible nanocomposite films produced with pure PVDF and different contents of KNNLTS nanoparticles and MWCNT in the PVDF matrix.

films to obtain the crystallographic information of the developed films and related plots, as shown in Fig. 6. The diffraction peak around 18.6° can be assigned to the α phase of the PVDF which is non-polar, and peaks around 20.2° corresponding to the β phase of the PVDF which is polar[51,52] in all the prepared nanocomposite films. In the diffraction pattern of PVDF, the α phase and β phase correspond to the JCPDS card numbers 42-1650, and 38-1638,

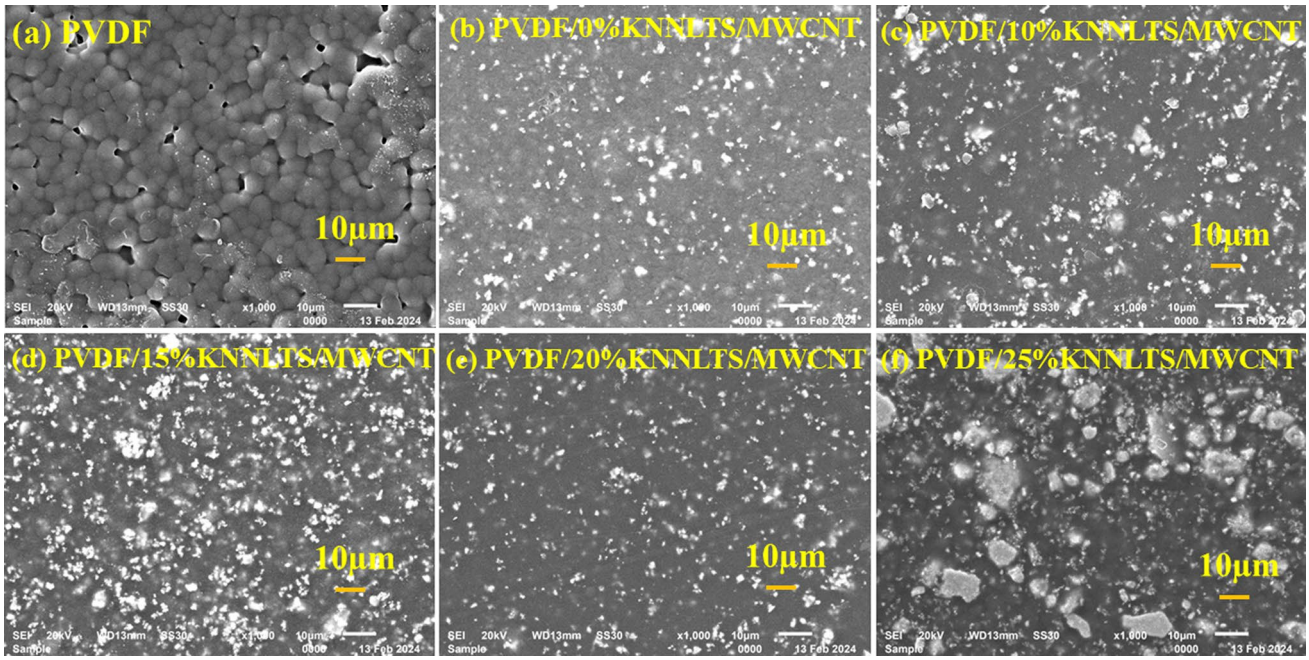


Fig. 8 SEM surface micrographs related to (a) pure PVDF, (b) PVDF/0%KNNLTS/MWCNT, (c) PVDF/10%KNNLTS/MWCNT, (d) PVDF/15%KNNLTS/MWCNT, (e) PVDF/20%KNNLTS/MWCNT, and (f) PVDF/25%KNNLTS/MWCNT.

respectively.[53] In PVDF, the β phase is the most popular phase for its piezoelectric response. The diffraction peak corresponding to the β phase has a stronger intensity in comparison to the α phase, indicating that the β phase of PVDF predominates over the α phase in all the nanocomposite films. Additionally, other peaks can be seen in the XRD plot assigned to the KNNLTS nanoparticles. The corresponding KNNLTS peaks can be confirmed by the JCPDS card numbers 71-2171 (orthorhombic) and 77-0037 (tetragonal).[49] The symbol "◆" is used to mark the presence of KNNLTS nanoparticles diffraction peaks. Furthermore, the XRD study of the PVDF/KNNLTS/MWCNT composites provides insights into the crystalline structure and phase composition of the material. In this specific context, the variation in peak intensity around 32° can be explained by considering the interaction between the PVDF matrix, KNNLTS ceramics, and MWCNT.

The peak around 32° is associated with the crystalline phases of the composite material, which includes contributions from the PVDF matrix and the KNNLTS ceramic. The specific phase of KNNLTS that give rise to this peak is sensitive to the concentration and distribution of the ceramic within the polymer matrix.

At a lower concentration of KNNLTS (10%), the ceramic particles may be insufficiently dispersed within the PVDF matrix, leading to poor interaction between the KNNLTS and the polymer, which could result in a reduction in the overall crystallinity or a decrease in the intensity of specific crystalline phases, leading to a lower peak intensity around 32°. As the concentration of KNNLTS increases to 15%, the ceramic particles are more effectively dispersed and interact

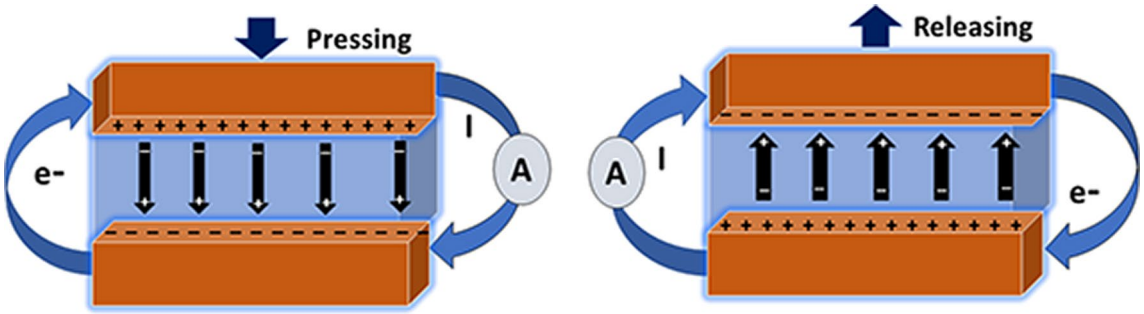


Fig. 9 Working mechanism of the PENG device.

more robustly with the PVDF matrix. This enhanced interaction can promote the crystallization of the PVDF or stabilize specific phases of the KNNLTS, resulting in a higher peak intensity. The better dispersion also ensures a more uniform phase distribution, contributing to the increase in intensity. Further increasing the KNNLTS content to 20% may lead to the formation of agglomerates, where the ceramic particles cluster together rather than dispersing evenly. These agglomerates can act as defects in the matrix, disrupting the crystallization process or altering the phase composition, leading to a decrease in the peak intensity. At even higher concentrations (25%), the system may reach a point where the KNNLTS particles start to percolate to form a continuous network within the PVDF matrix, which can enhance the crystallinity of the composite or stabilize a specific crystalline phase of the KNNLTS, resulting in an increase in peak intensity around 32°.

Conclusively, the observed variation in peak intensity is primarily due to the changes in dispersion, interaction, and crystallization behavior of KNNLTS within the PVDF matrix as its concentration is varied. At lower concentrations, poor dispersion leads to lower peak intensity, while at higher concentrations, better dispersion and phase interaction enhance the intensity, until agglomeration effects come into play at very high concentrations.

Figure 7 shows the FTIR analysis for the pure PVDF film and corresponding PVDF/KNNLTS/MWCNT nanocomposite films in the range of 500–1500 cm^{-1}. The sharp peaks observed at 878 cm^{-1} and 1404 cm^{-1} are due to the vibration of C–F stretching and C–H bending, respectively, of the PVDF.[54,55] The peak around 1172 cm^{-1} of the PVDF composite film indicates the presence of the β crystalline phase with high intensity.[56] The peak at 762 cm^{-1} is denoted by the α phase. In addition, the peaks at 512 cm^{-1}, 840 cm^{-1}, and 1075 cm^{-1} strongly correspond to the presence of the β phase.[27,57–61] It can be observed that the α and β phases of the PVDF occur in all the nanocomposite films. The presence of the β phase in the film is very significant for the piezoelectric effect.[27] The content of the β phase Fβ) in all the nanocomposite films can be calculated by using Beer–Lambert Law[62]:

$$F(\beta) = \frac{A_\beta}{\left(\frac{K_\beta}{K_\alpha}\right)A_\alpha + A_\beta} \tag{1}$$

where K_α and K_β denote the absorption coefficients at 840 cm^{-1} and 762 cm^{-1} whose values are $K_\alpha = 6.1 \times 10^4$ cm^2 mol^{-1}, and $K_\beta = 7.7$ x 10^4 cm^2 mol^{-1} K_α, respectively, and A_β and A_α are the absorbance intensities at 762 cm^{-1} and 840 cm^{-1}, respectively.[63] By introducing KNNLTS and MWCNT, the F(β) (β phase fraction) in the PVDF/20%KNNLTS/MWCNT nanocomposite film is more in comparison to other films. The results obtained are consistent with the XRD result, indicating that the prepared flexible nanocomposite films confirm the presence of the β phase. Simultaneously, the addition of KNNLTS and MWCNT has improved the crystallinity of the PVDF matrix, and, as a result, the fabrication of the device formed by PVDF/KNNLTS/MWCNT nanocomposite film may have

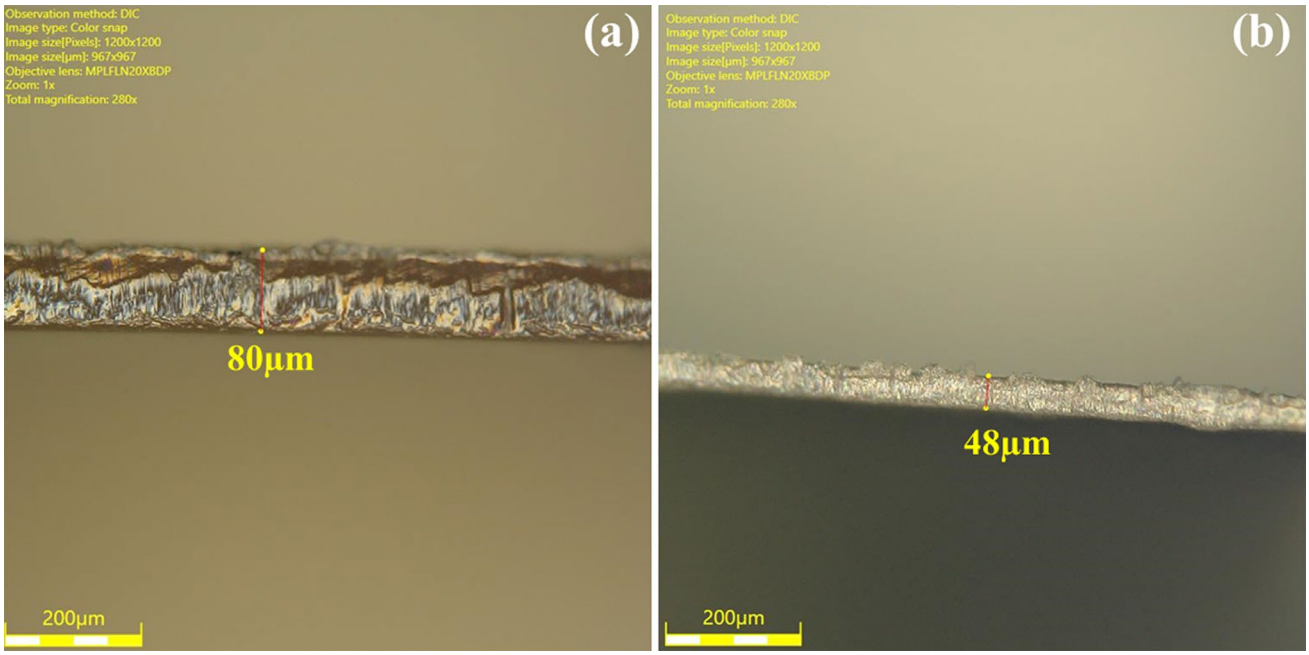


Fig. 10 Optical micrographs of fabricated composite films: (a) PVDF, and (b) KNNLTS/MWCNT in the PVDF matrix.

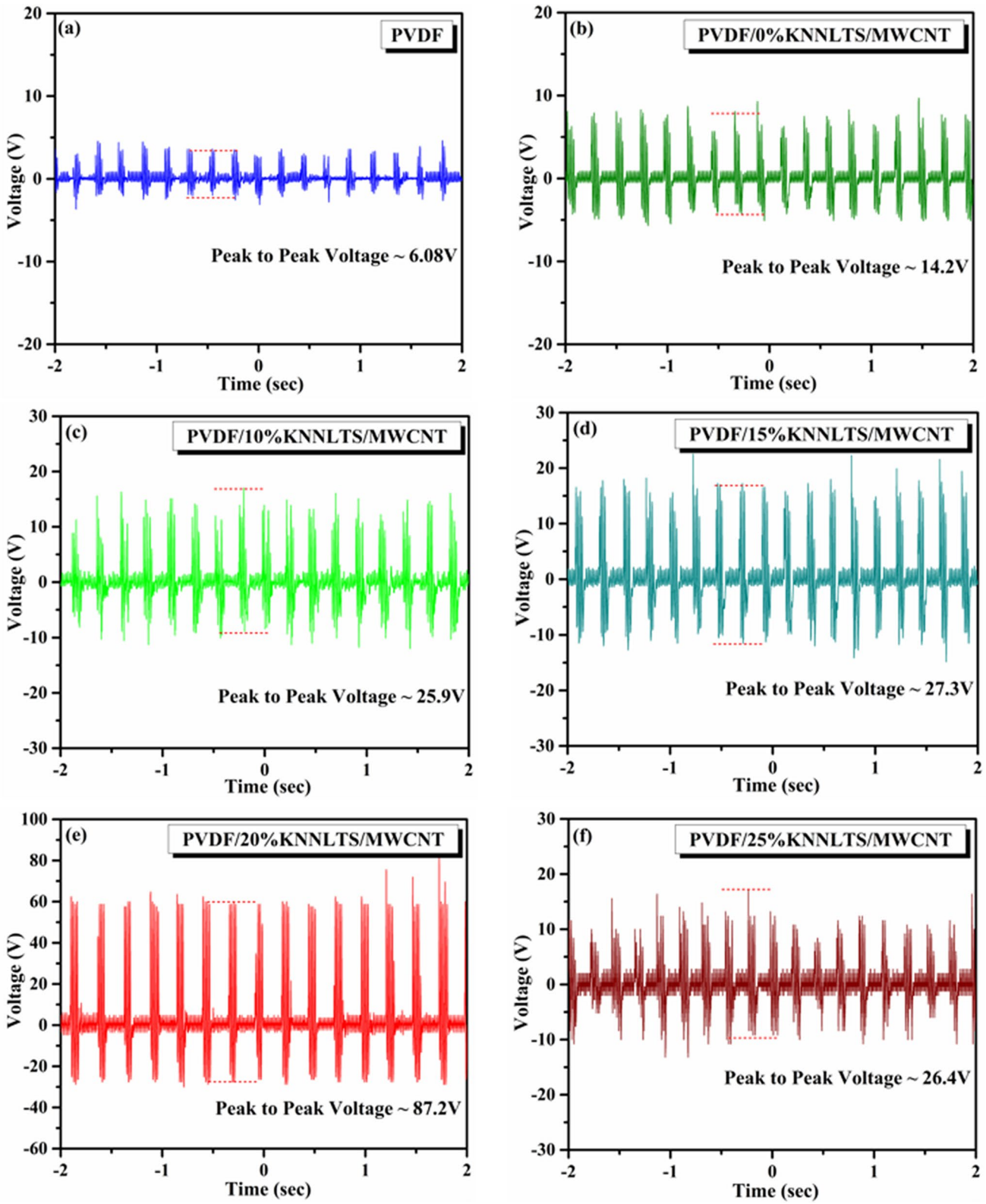


Fig. 11 Open-circuit voltage produced using (a) pure PVDF, (b) PVDF/0%KNNLTS/MWCNT, (c) PVDF/10%KNNLTS/MWCNT, (d) PVDF/15%KNNLTS/MWCNT, (e) PVDF/20%KNNLTS/MWCNT, and (f) PVDF/25%KNNLTS/MWCNT PENG devices.

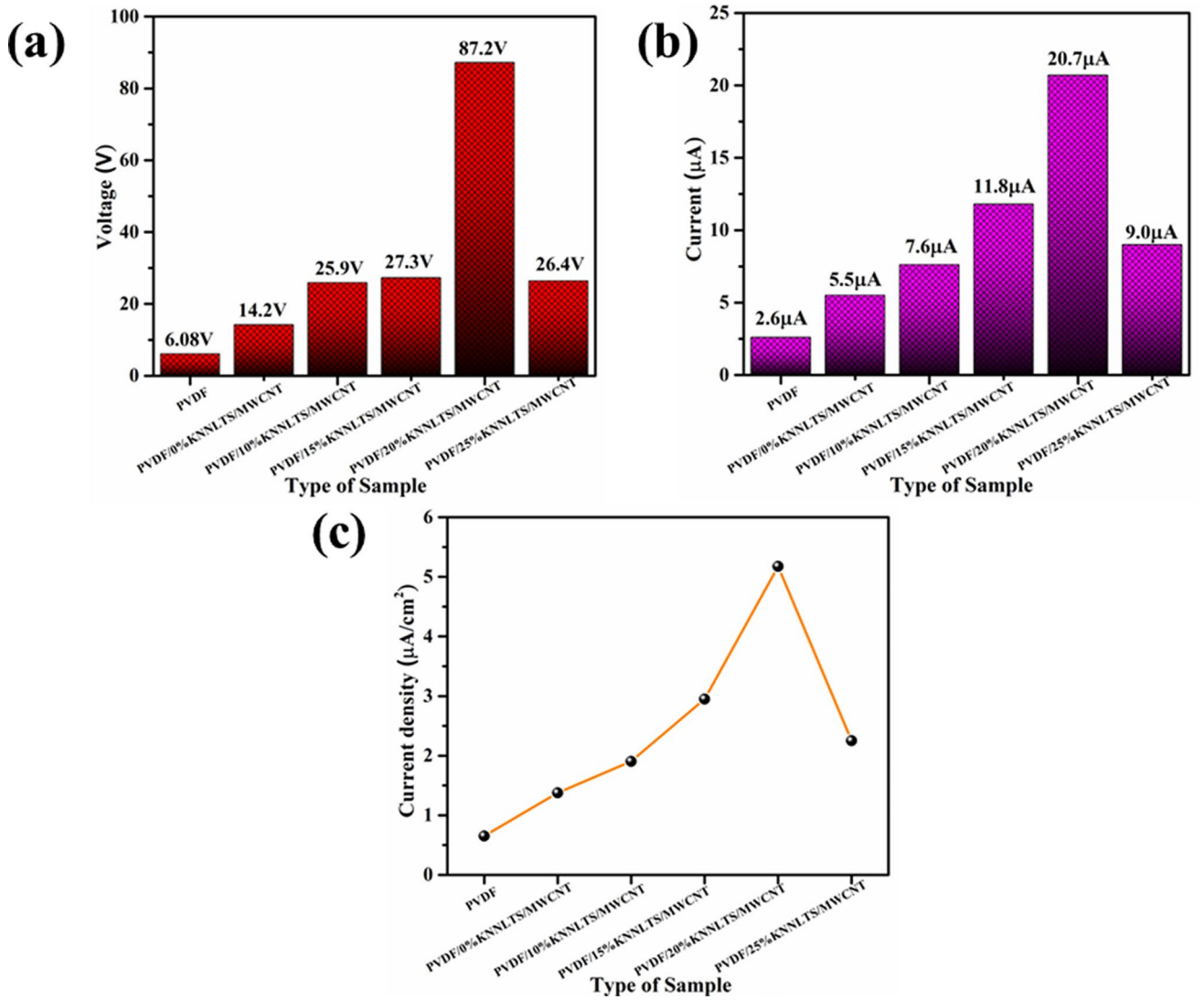


Fig. 12 (a) Output voltage, (b) current with the type of sample, and (c) plot of current density with type of sample.

a significant impact on the energy-harvesting performances of the prepared nanocomposite films.

Morphological Analysis of Nanocomposite Films

The morphology of the as-synthesized flexible nanocomposite films was analyzed by SEM to examine the surface and internal structure of the PVDF/KNNLTS/MWCNT nanocomposite films, and representative images are shown in Fig. 8a and f revealing a uniform distribution of KNNLTS nanoparticles and MWCNTs within the polymer matrix. The nanoparticles appear well-dispersed without any significant agglomeration, which is crucial for maintaining the structural integrity and performance of the nanocomposites. A smooth and dense surface is visible in the pure PVDF nanocomposite film, and it can be seen that KNNLTS nanoparticles were observed on the surface of the film, as shown in Fig. 8b. Further, uniform dispersion occurs in all the PVDF/KNNLTS/MWCNT nanocomposite films, and Fig. 8 clearly shows that KNNLTS particles are present in the PVDF matrix. The morphology of the PVDF/KNNLTS/MWCNT nanocomposite films plays a critical role in determining their overall performance.

MWCNT are known for their excellent electrical conductivity. The arrangement and connectivity of MWCNTs within the PVDF matrix influences the electrical conductivity of the nanocomposite films. A well-dispersed network of MWCNTs can create conductive pathways, improving the overall electrical performance of the films, which is beneficial for applications in sensors, actuators, and electronic devices. Further, KNNLTS is known for its piezoelectric properties. The distribution and alignment of KNNLTS

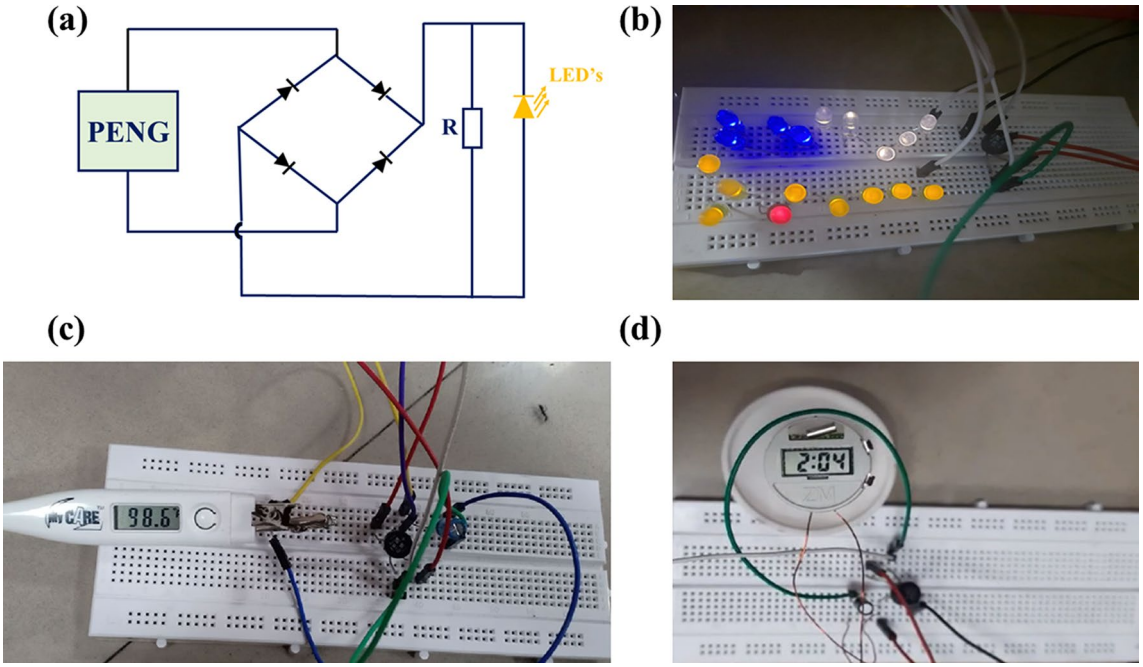


Fig. 13 (a) Electrical circuit use powering LEDs, (b) lighting LEDS, (c) digital thermometer, and (d) digital wrist watch derived by PVDF/20%KNNLTS/MWCNT-based PENG.

particles within the PVDF matrix affect the piezoelectric performance of the nanocomposite films. A well-dispersed and aligned structure can enhance the piezoelectric response, making these films suitable for applications in energy-harvesting and sensors. Overall, the morphological analysis of the KNNLTS/MWCNT/PVDF nanocomposite films provide valuable information on how the structural properties impact their functional properties.

Piezoelectric Performance of PENG Device

The fabricated PENG's operating mechanism is shown in Fig. 9. When a perpendicular force acts on the device, electric dipoles are oriented in a single direction generating potential differences between the two electrodes. As a result, the current flows via an outer circuit from the top to the bottom surface, as shown in Fig. 9a. Furthermore, as depicted in Fig. 9b, electric dipoles are oriented in the opposite direction of current flow from the bottom to the top electrodes when the device's perpendicular force is removed. The PENG device's output signals have dimensions of 1.5 cm × 1.5 cm. The thickness of the prepared nanocomposite films was determined by digital microscopy, as shown in Fig. 10. The thicknesses of the pure PVDF film, and KNNLTS/MWCNT embedded in the PVDF film were found to be ~ 80 μm, and 48 μm, respectively, by hitting the surface of the device with a vibrator shaker that applied a force of ~ 2 N at a frequency of 4 Hz. The potential difference across the generator develops when pressure is applied to the surface of the device; the output voltage generated may then monitored with the help of DSO. Figure 11 displays the open-circuit voltage plots generated by the fabricated devices that are constructed using different concentrations of KNNLTS ceramic and MWCNT as fixed in the PVDF matrix; the figure shows that the highest peak-to-peak open-circuit voltage ($V_{P\text{-}P}$) of the PVDF/20%KNNLTS/MWCNT-based PENG is 87.2 V. In the cases of pure PVDF-, PVDF/0%KNNLTS/MWCNT-, PVDF/10%KNNLTS/MWCNT-, PVDF/15%KNNLTS/MWCNT-, PVDF/20%KNNLTS/MWCNT-, and PVDF/25%KNNLTS/MWCNT-based PENGs the generated output voltages reach 6.08 V, 14.2 V, 25.9 V, 27.3 V, and 26.4 V, respectively. It can be seen in Fig. 12a that the output response for the device PVDF/20%KNNLTS/MWCNT is enhanced by almost 14.3 times for open-circuit voltage in comparison to pure PVDF and 6.1 times in comparison to the device containing only MWCNT in the PVDF matrix. Further, the piezoelectric performance of PVDF/KNNLTS/MWCNT was investigated and compared with existing literature. As shown in Table II, the output voltage and current values obtained in this study are higher than the reported work by Abdullah et al., who investigated PVDF/KNN/MWCNT nanocomposites by the same approach (drop-casting), whereas, Pratihar et al. performed the same kind of work using PVDF/ZnO/MWCNT

Table II Comparison of the piezoelectric performance between PVDF/KNNLTS/MWCNT PENG nanogenerator fabricated in the present work and various nanogeneratorz reported in the literature

Sample	Filler material in the polymer matrix	Polymer used	Film synthesis technique	Open circuit voltage (V)	Short circuit Current (μA)	Power density (μW/cm^2)	Piezoelectric coefficient ($d_{33}{}^*$) (pm/V)	References
1	ZnO nanoparticles	PVDF	Drop-casting	1.81	0.57	0.21	1.71	65
2	PZT particles	PVDF	Electro-spinning	0.184	–	30.69	–	66
3	KNN ceramic particles	PVDF	Finger tapping	11.2	0.3	–	16	36
4	$BaTiO_3$ particles	PVDF	Solvent-casting method	7.2	0.038	0.8	–	67
5	KNNLTS ceramic particles	PVDF	Drop-casting	36.58	–	–	24	50
6	KNN nanorods and CNT	PVDF	Electro-spinning	23.24	9	52.29	–	42
7	KNN particles and MWCNT	PVDF	Drop-casting	35.3	15.8	–	–	68
8	$BaTiO_3$ nanoparticles and CNT	PVDF	Drop-casting	4.4	0.66	–	–	69
9	ZnO nanoparticles and MWCNT	PVDF	Drop-casting	22	–	21.41	–	70
10.	Pt nanoparticles	PVDF	Electro-spinning	30	6	22	44	64
11	CNT	PVDF	Electro-spinning	1.89	0.011	–	31.3	71
12	TiO_2-Fe_3O_4-MWCNT	PVDF	Electro-spinning	0.68	–	–	51.42	72
13	KNNLTS nanoparticles and MWCNT	PVDF	Drop-casting	87.22	20.3	145.48	59.92	Present work

nanocomposites via drop-casting and reported an output voltage of 22 V, and power density of 21.41 μW/cm^2.[69] Moreover, the obtained result was considerably higher in comparison to the other fabricated PENG mainly attributed to its higher β phase content. This generated output voltage using PVDF/20%KNNLTS/MWCNT PENG was enough to light up 18 light-emitting diodes (LEDs) and power a digital thermometer as well as a digital wrist watch, as shown in Fig. 13b, c, and d, and Supplementary Videos 1, 2, and 3, respectively. Overall, the results suggest that PVDF/KNNLTS/MWCNT offers higher values of output voltage, current, and power density, making it a promising candidate for various applications in sensors, actuators, and energy-harvesting devices.

Further, an electrometer (B2985B) was used to measure the short-circuit current of the fabricated PENG device in vibrator shaker tapping mode. The related plots are presented in Fig. 14. The values of short-circuit current were 2.6 μA, 5.5 μA, 7.6 μA, 11.8 μA, 20.7 μA, and 8.96 μA for the samples of pure PVDF, PVDF/0%KNNLTS/MWCNT, PVDF/10%KNNLTS/MWCNT, PVDF/15%KNNLTS/MWCNT, and PVDF/25%KNNLTS/MWCNT PENGs, respectively, as shown in Fig. 12b with the frequency ~ 4 Hz.

Durability is an important parameter for energy harvesters in PENG devices. For testing the durability and stability of the fabricated PVDF/20%KNNLTS/MWCNT PENG device, we conducted an experiment on the device for a specific duration of time. A continuous tapping was applied on the device, and the measured output voltage was very stable during 2000 tappings, indicating that there was no degradation of peaks during the cycle and that μPENG can persist for a long period, as shown in Fig. 15. The current density was also calculated and the corresponding plot is shown in Fig. 12c, from which it can be seen that the highest current density was ~ 5.1 μA/cm^2 for the PVDF/20%KNNLTS/MWCNT PENG device.

With the help of external resistance, the PENG was further characterized, rising from 10^5 Ω to 10^8 Ω. Figure 16a shows the output voltage and current for varying external loads. Ohm's law ($V = IR$), states that, as the resistance rises, the voltage rises, and the current falls. The point where the voltage and current meet, at 2.2 M Ω, is known as the optimum point. It has been found that the power is expected to be maximum at this optimum resistance. The measured output power density for variable external load is shown in Fig. 16b. At an external resistance of 2.2 MΩ

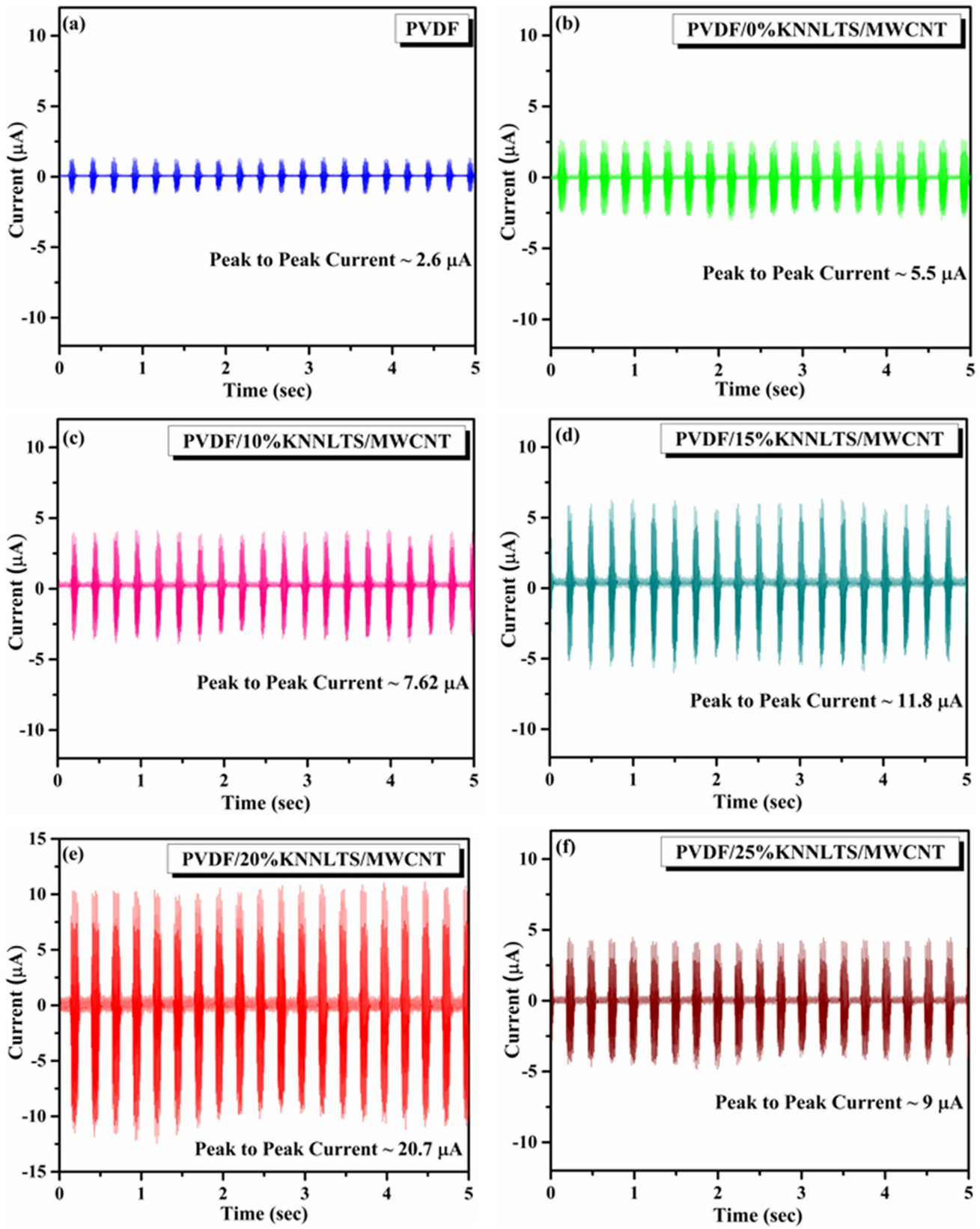


Fig. 14 Current generated using (a) pure PVDF, (b) PVDF/0%KNNLTS/MWCNT, (c) PVDF/10%KNNLTS/MWCNT, (d) PVDF/15%KNNLTS/MWCNT, (e) PVDF/20%KNNLTS/MWCNT, and (f) PVDF/25%KNNLTS/MWCNT PENG devices.

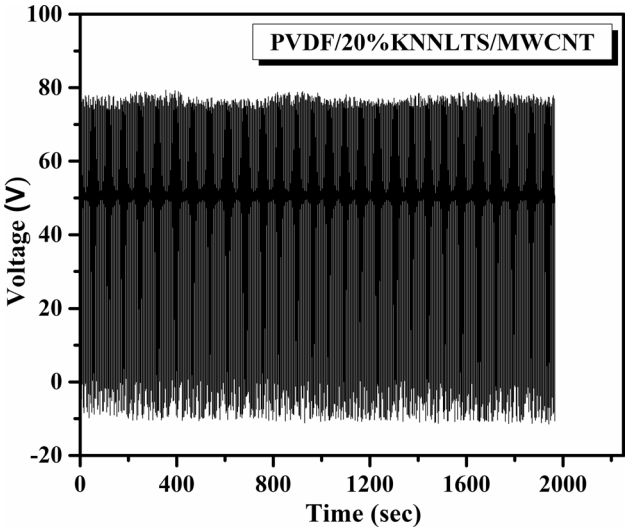


Fig. 15 Stability and durability tests of PVDF/20%KNNLTS/MWCNT film-based PENG under 2000 cycles.

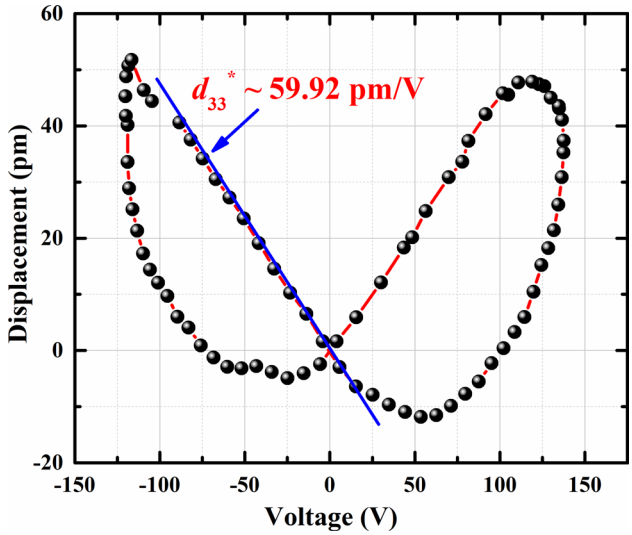


Fig. 17 Displacement vs. voltage (D-V) butterfly curve for PVDF/20%KNNLTS/MWCNT film-based PENG.

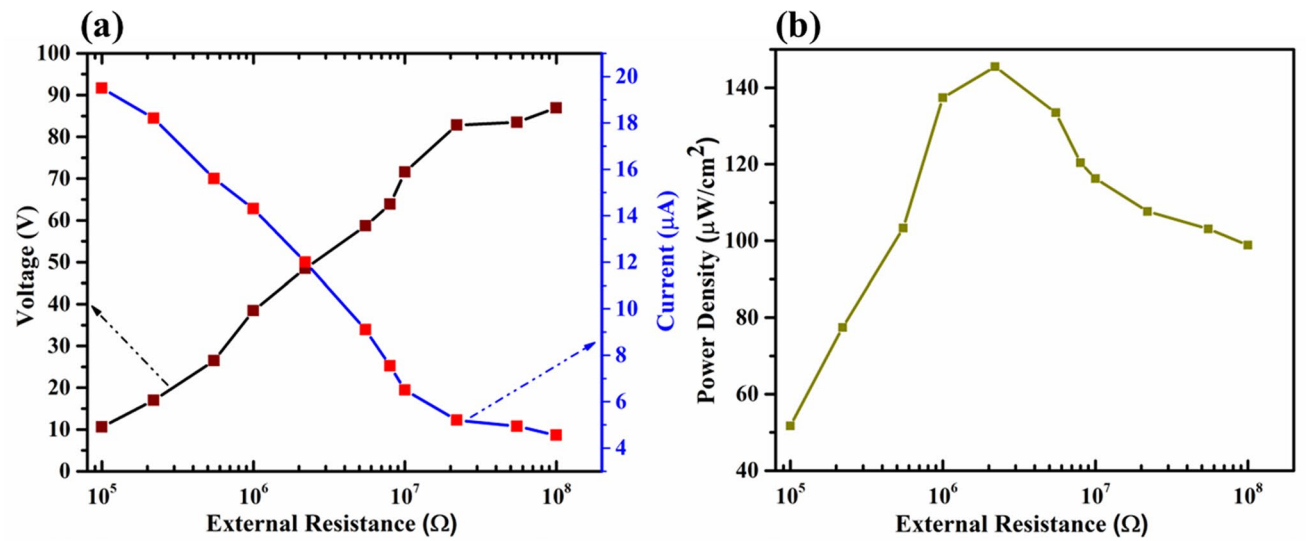


Fig. 16 (a) Output voltage and current measured with external resistance, (b) power density measured with external resistance by PVDF/20%KNNLTS/MWCNT film-based PENG.

(optimum point), the maximum power density measured was ~ 145.48 μW/cm^2.

In addition, by measuring the converse piezoelectric coefficient (d_{33}^{*}), the prepared PVDF/20%KNNLTS/MWCNT PENG device's piezoelectricity was examined. In the current study, the room-temperature displacement (D)–voltage (V) hysteresis loop was employed to determine the converse piezoelectric coefficient, and the related plot is displayed in Fig. 17. Based on the slope of the butterfly curve, the maximum and the average values of d_{33}^{*} were ~ 59.92 pm/V and 28.67 pm/V, respectively. These values are greater than those previously reported.[36,50] Ghosh and Mandal reported d_{33} values for PVDF composites enhanced with Pt nanoparticles, showing values ranging from 44 pC/N.[64] Singh et al. reported d_{33} values for PVDF composites with hybrid nanofillers using ZnO nanoparticles, and showed that the inclusion of hybrid fillers increased the d_{33} value to around 1.71 pC/N,[65] but still the d_{33} values obtained in this study are higher than those of the reported work. Overall, the results suggest that KNNLTS/MWCNT doped in the PVDF matrix offers

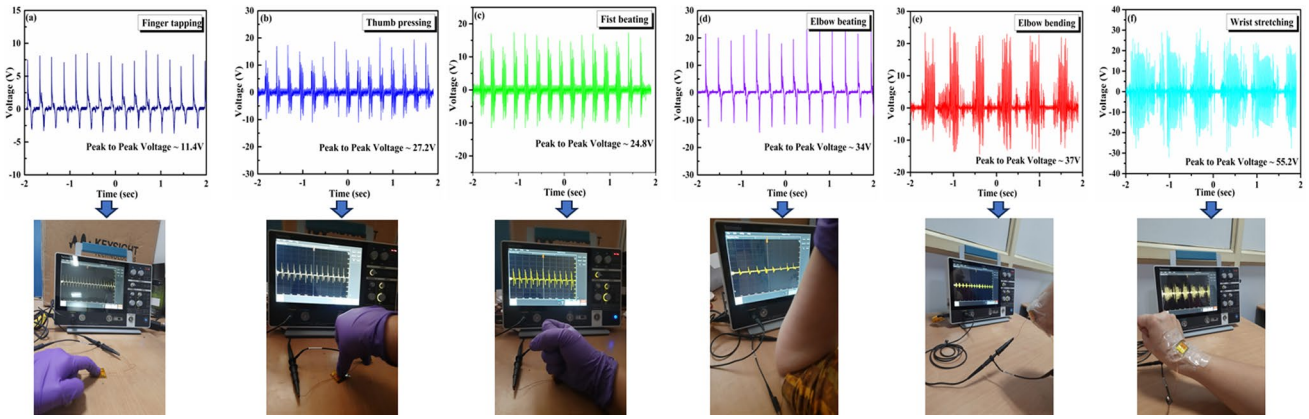


Fig. 18 The output voltages generated by PVDF/20%KNNLTS/MWCNT film-based PENG by (a) finger tapping, (b) thumb pressing, (c) fist beating, (d) elbow beating, (e) elbow bending, and (f) wrist stretching.

the higher values of piezoelectric coefficients, which make it a promising candidature for energy-harvesting applications (Fig. 17).

Furthermore, as mechanical energy is so abundant in our surroundings, it gets wasted in our daily lives. A PENG device can be utilized in human body motion especially by hand motion such as finger tapping, thumb pressing, fist beating, elbow beating, elbow bending, and wrist stretching. The generated open-circuit voltages for these applications were ~ 11.4 V, 27.2 V, 24.8 V, 34 V, 37 V, and 55.2 V, respectively, as shown in Fig. 18. This shows that the fabricated PENG device proves the utilization of mechanical energy in our daily life.

Conclusions

Environmentally friendly piezoelectric ceramic KNNLTS were synthesized by utilizing a high-energy ball mill technique to fabricate a piezoelectric nanogenerator with PVDF and MWCNT. It can be concluded that the piezoelectric performance of the PVDF/KNNLTS nanocomposite films is greatly influenced by a conductive filler (MWCNT). A PVDF/KNNLTS/MWCNT PENG wasn created by using nanocomposite films with dimensions of 1.5×1.5 cm^2 and having a thickness of ~ 48 μm. A significant dependency was found in the nanocomposites' piezoelectric performance, in terms of the output voltage, output current, and power density of the nanogenerator. It has been shown that the PENG based on PVDF/20%KNNLTS/MWCNT provided the maximum voltage and current of ~ 87.2 V and 20.3 μA. In fact, the MWCNT filler plays an intriguing role in the PVDF/KNNLTS combination. In one sense, the MWCNT act as a β-nucleating agent, but they also improve the β-crystalline orientation by providing conductivity. Furthermore, the PENG was successfully integrated with a full wave bridge rectifier, resistances, and LEDs. The optimum power density for the PENG was measured to be 145.48 μW/cm^2 with a 2.2-MΩ load resistance. Additionally, 18 LEDs were lit up, and a digital thermometer and a digital wrist watch powered. The outstanding piezoelectric performance in the PENG-based PVDF/KNNLTS/MWCNT has a considerable promise for the development of mechanical energy-harvesting applications, as well as the emerging use of smart materials.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11664-024-11463-5.

Acknowledgments The authors wish to thank the DTU Project Research Grant (F.NO.DTU/IRD/619/2019/2113) for providing the research grant to perform the research work. Komal Verma express her gratitude to Delhi Technological University (DTU), India for providing the DTU fellowship (2K19/PHDAP/06).

Author Contributions Komal Verma: Conceptualization, Writing-original Draft, methodology, investigation, Formal analysis. Aman Kumar: Data collection, investigation, methodology. Richa Sharma: Conceptualization, writing- review & editing, supervision of the manuscript at different stages, data curation.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Conflict of interest The authors declare that they have no known conflict of interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. D.K. Bharti, S. Veeralingam, and S. Badhulika, An ultra high performance, lead-free Bi_2WO_6: P(VDF-TrFE)-based triboelectric nanogenerator for self-powered sensors and smart electronic applications. *Mater. Horiz.Horiz.* 9, 663–674 (2022).
2. S.P. Muduli, S. Veeralingam, and S. Badhulika, Interface induced high-performance piezoelectric nanogenerator based on a

electrospun three-phase composite nanofiber for wearable applications. *ACS Appl. Energy Mater.* 4(11), 12593–12603 (2021).
3. I. Hussain, H.P. Tran, J. Moore, N. Islam, and M.J. Uddin, Functional materials, device architecture, and flexibility of perovskite solar cell. *Emergent Mater.* 1, 133–154 (2018).
4. F. Tasnim, S.A. Iqbal, and A.R. Chowdhury, Biogas production from anaerobic co-digestion of cow manure with kitchen waste and Water Hyacinth. *Renew. Energy* 109, 434–439 (2017).
5. S.M.R. Islam, D. Kwak, M.H. Kabir, M. Hossain, and K.-S. Kwak, The internet of things for health care: a comprehensive survey. *IEEE Access* 3, 678–708 (2015).
6. A.M. Abdullah, A.R. Chowdhury, Y. Yang, H. Vasquez, H.J. Moore, J.G. Parsons, K. Lozano, J.J. Gutierrez, K.S. Martirosyan, and M.J. Uddin, Tailoring the viscosity of water and ethylene glycol based TiO_2 nanofluids. *J. Mol. Liq.* 297, 111982 (2020).
7. X. Chen, X. Li, J. Shao, N. An, H. Tian, C. Wang, T. Han, L. Wang, and B. Lu, High performance piezoelectric nanogenerators with imprinted P(VDF-TrFE)/$BaTiO_3$ nanocomposite micropillars for self-powered flexible sensors. *Small* 13(23), 1604245 (2017).
8. A.R. Chowdhury, A.M. Abdullah, I. Hussain, J. Lopez, D. Cantu, S.K. Gupta, Y. Mao, S. Danti, and M.J. Uddin, Lithium doped zinc oxide based flexible piezoelectric-triboelectric hybrid nanogenerator. *Nano Energy* 61, 327–336 (2019).
9. S.H. Shin, Y.H. Kim, H.L. Min, J.Y. Jung, J.H. Seol, and J. Nah, Lithium-doped zinc oxide nanowires–polymer composite for high performance flexible piezoelectric nanogenerator. *ACS Nano* 8(10), 10844–10850 (2014).
10. P.K. Panda, Review: environmental friendly lead-free piezoelectric materials. *J. Mater. Sci.* 44, 5049–5062 (2009).
11. Z.L. Wang and J. Song, Piezoelectric nanogenerators based on zinc oxide nanowire arrays. *Science* 312, 242–246 (2006).
12. M. Stoppa and A. Chiolerio, Wearable electronics and smart textiles: a critical review. *Sensors* 14, 11957–11992 (2014).
13. D. Hua, M. Yao, Y. Fan, C. Ma, M. Fan, and M. Liu, Strategies to achieve high performance piezoelectric nanogenerators. *Nano Energy* 55, 288–304 (2019).
14. L. Li, J. Xu, J. Liu, and F. Gao, Recent Progress on piezoelectric energy harvesting: structures and materials. *Adv. Compos. Hybrid Mater.* 1, 478–505 (2018).
15. G.-T. Hwang, H. Park, J.-H. Lee, S. Oh, K.-I. Park, M. Byun, H. Park, G. Ahn, C.K. Jeong, K. No, H. Kwon, S.-G. Lee, B. Joung, and K.J. Lee, Self-powered cardiac pacemaker enabled by flexible single crystalline PMN-PT piezoelectric energy harvester. *Adv. Mater.* 26, 4880–4887 (2014).
16. Z. Lou, L. Li, L. Wang, and G. Shen, Recent progress of self-powered sensing systems for wearable electronics. *Small* 13, 1701791 (2017).
17. P. Bai, G. Zhu, Z.-H. Lin, Q. Jing, J. Chen, G. Zhang, J. Ma, and Z.L. Wang, Integrated multilayered triboelectric nanogenerator for harvesting biomechanical energy from human motions. *ACS Nano* 7, 3713–3719 (2013).
18. L. Jin, J. Tao, R. Bao, L. Sun, and C. Pan, Self-powered real-time movement monitoring sensor using triboelectric nanogenerator technology. *Sci. Rep.* 7, 1–6 (2017).
19. A.M. Abdullah, A. Flores, A. Chowdhury, J. Li, Y. Mao, and M.J. Uddin, Corrigendum to "Synthesis and fabrication of self-sustainable triboelectric energy case for powering smart electronic devices." *Nano Energy* 76, 104996 (2020).
20. A.R. Chowdhury, J. Jaksik, I. Hussain, R. Longoria, O. Faruque, F. Cesano, D. Scarano, J. Parsons, and M.J. Uddin, Multicomponent nanostructured materials and interfaces for efficient piezoelectricity. *Nano Struct. Nano Objects* 17, 148–184 (2019).
21. A.R. Chowdhury, Organic-Inorganic Hybrid Materials for Piezoelectric/Triboelectric Nanogenerator, M.S., The University of Texas Rio Grande Valley, Thesis (2019).
22. H.B. Kang, C.S. Han, J.C. Pyun, W.H. Ryu, C.-Y. Kang, and Y.S. Cho, (Na, K)NbO_3 nanoparticle-embedded piezoelectric nanofiber composites for flexible nanogenerators. *Compos. Sci. Technol.* 111, 1–8 (2015).
23. A.J. Lovinger, Annealing of poly(vinylidene fluoride) and formation of a fifth phase. *Macromolecules* 15, 40–44 (1982).
24. Y. Koseki, K. Aimi, and S. Ando, Crystalline structure and molecular mobility of PVDF chains in PVDF/PMMA blend films analyzed by solid-state 19 F MAS NMR spectroscopy. *Polym. J.* 44, 757–763 (2012).
25. R.G. Kepler and R.A. Anderson, Piezoelectricity and pyroelectricity in polyvinylidene fluoride. *J. Appl. Phys.* 49, 4490–4494 (1978).
26. S. Cha, S.M. Kim, H. Kim, J. Ku, J.I. Sohn, Y.J. Park, B.G. Song, M.H. Jung, E.K. Lee, B.L. Choi, J.J. Park, Z.L. Wang, J.M. Kim, and K. Kim, Porous PVDF as effective sonic wave driven nanogenerators. *Nano Lett.* 11, 5142–5147 (2011).
27. L. Ruan, X. Yao, Y. Chang, L. Zhou, G. Qin, and X. Zhang, Properties and applications of the β phase poly(vinylidene fluoride). *Polymers* 10, 228 (2018).
28. S. Bairagi and S.W. Ali, A unique piezoelectric nanogenerator composed of melt-spun PVDF/KNN nanorod-based nanocomposite fibre. *Eur. Polym. J.* 116, 554–561 (2019).
29. J. Yang, Y. Zhang, Y. Li, Z. Wang, W. Wang, Q. An, and W. Tong, Piezoelectric nanogenerators based on graphene oxide/PVDF electrospun nanofiber with enhanced performances by in-situ reduction. *Mater. Today Commun.* 26, 101629 (2021).
30. S. Yu, W. Zheng, W. Yu, Y. Zhang, Q. Jiang, and Z. Zhao, Formation mechanism of β-phase in PVDF/CNT composite prepared by the sonication method. *Macromolecules* 42, 8870–8874 (2009).
31. G.H. Kim, S.M. Hong, and Y. Seo, Piezoelectric properties of poly(vinylidene fluoride) and carbon 648 nanotube blends: beta-phase development. *Phys. Chem. Chem. Phys.* 11, 10506–10512 (2009).
32. A. Boochani, G. Rasoolian, F. Kafi, A. Aminian, and S.M. Elahi, Electronic, optical, vibrational and EFG properties of tetragonal $BaTiO_3$ under pressure: by first-principles study. *Chin. J. Phys.* 59, 357–371 (2019).
33. H. Takahashi, Y. Numamoto, J. Tani, and S. Tsurekawa, Piezoelectric properties of $BaTiO_3$ ceramics with high performance fabricated by microwave sintering. *Jpn. J. Appl. Phys.* 45, 7405–7408 (2006).
34. L. Liu, H. Fan, L. Fang, X. Chen, H. Dammak, and M.P. Thi, The Effect of Na/K evaporation on electrical properties and intrinsic defects in $Na_{0.5}K_{0.5}NbO_3$ ceramics. *Mater. Chem. Phys.* 117, 138–141 (2009).
35. K. Verma, S. Goel, and R. Sharma, Influence of calcination and sintering temperature on the microstructure, dielectric, ferroelectric and piezoelectric properties of the lead-free KNN ceramics. *J. Mater. Sci. Mater. Electron.* 33, 26067–26085 (2022).
36. K. Verma and R. Sharma, A flexible piezoelectric generator based on KNN/PVDF composite films: role of KNN concentration on the piezoelectric performance of generator. *Chin. J. Phys.* 84, 198–215 (2023).
37. L. Paralı, M. Koç, and E. Akça, Fabrication and characterization of high performance PVDF-based flexible piezoelectric nanogenerators using PMN-xPT (x:30, 32.5, and 35) particles. *Ceram. Int.* 49, 18388–18396 (2023).
38. M.-G. Kang, W.-S. Jung, C.-Y. Kang, and S.-J. Yoon, Recent progress on PZT based piezoelectric energy harvesting technologies. *Actuators* 5(1), 5 (2016).
39. T.-Y. Ke, H.-A. Chen, H.-S. Sheu, J.-W. Yeh, H.-N. Lin, C.-Y. Lee, and H.-T. Chiu, Sodium niobate nanowire and its piezoelectricity. *J. Phys. Chem. C* 112, 8827–8831 (2008).

40. Y. Shiratori, A. Magrez, W. Fischer, C. Pithan, and R. Waser, Temperature-induced phase transitions in micro-, submicro-, and nanocrystalline $NaNbO_3$. *J. Phys. Chem. C* 111, 18493–18502 (2007).
41. A. Teka, S. Bairagi, M. Shahadat, M. Joshi, S.Z. Ahammad, and S.W. Ali, Poly(vinylidene fluoride) (PVDF)/potassium sodium niobate (KNN)–based nanofibrous web: a unique nanogenerator for renewable energy harvesting and investigating the role of KNN nanostructures. *Polym. Adv. Technol.* 29, 2537–2544 (2018).
42. S. Bairagi and S.W. Ali, Investigating the role of carbon nanotubes (CNTs) in the piezoelectric performance of a PVDF/KNN-based electrospun nanogenerator. *Soft Matter* 16, 4876 (2020).
43. Y. Gao, J. Zhang, Y. Qing, Y. Tan, Z. Zhang, and X. Hao, Remarkably strong piezoelectricity of lead-free $(K_{0.45}\ Na_{0.55})_{0.98}\ Li_{0.02}\ (Nb_{0.77}\ Ta_{0.18}\ Sb_{0.05})\ O_3$ ceramic. *J. Am. Ceram. Soc.* 94, 2968 (2011).
44. B.-Q. Ming, J.-F. Wang, P. Qi, and G.-Z. Zang, Piezoelectric properties of (Li, Sb, Ta) modified (Na, K)NbO_3 lead-free ceramics. *J. Appl. Phys.* 101, 054103 (2007).
45. K. Wang and J.-F. Li, Low-Temperature sintering of Li-modified lead-free ceramics: sintering behavior, microstructure, and electrical properties. *J. Am. Ceram. Soc.* 93, 1101 (2010).
46. Y.G. Lv, C.L. Wang, J.L. Zhang, L. Wu, M.L. Zhao, and J.P. Xu, Tantalum influence on physical properties of $(K_{0.5}Na_{0.5})(Nb_{1-x}Ta_x)\ O_3$ ceramics. *Mater. Res. Bull.* 44, 284 (2009).
47. Y. Huan, T. Wei, Z. Wang, H. Shen, X. Lin, S. Huang, and X. Wang, Ultrahigh energy harvesting properties in Ag decorated potassium-sodium niobite particle-polymer composite. *J. Materiomics* 6, 355 (2020).
48. M.K. Gupta, S.-W. Kim, and B. Kumar, Flexible high-performance lead-free $Na_{0.47}\ K_{0.47}\ Li_{0.06}\ NbO_3$ microcube-structure-based piezoelectric energy harvester. *ACS Appl. Mater. Interfaces* 8, 1766 (2016).
49 P. Kumar and P. Palei, Effect of sintering temperature on ferroelectric properties of $0.94(K_{0.5}\ Na_{0.5})\ NbO_3$-$0.06LNbO_3$ system. *Ceramic International* 36, 1725–1729 (2010).
50. K. Verma and R. Sharma, Development of KNNLTS-PVDF-based flexible piezoelectric generator for energy-harvesting application. *Bull. Mater. Sci.* 47, 38 (2024).
51. A.K. Zak, W.C. Gan, W.A. Majid, M. Darroudi, and T.S. Velayutham, Experimental and theoretical dielectric studies of PVDF/PZT nanocomposite thin films. *Ceram. Int.* 37, 1653–1660 (2011).
52. S. Bairagi and S.W. Ali, Poly (vinylidene fluoride) (PVDF)/Potassium Sodium Niobate (KNN) nanorods based flexible nanocomposite film: influence of KNN concentration in the performance of nanogenerator. *Org. Electron.* 78, 105547 (2020).
53. A. Peles, O. Aleksic, V.P. Pavlovic, V. Dojcilovic, Z. Nikolic, F. Marinkovic, M. Mitric, V. Blagojevic, B. Vlahovic, and V.B. Pavlovic, Structural and electrical properties of ferroelectric poly(vinylidene fluoride) and mechanically activated ZnO nanoparticle composite films. *Phys. Scr.* 93, 105801 (2018).
54. S. Janakiraman, A. Surendran, S. Ghosh, S. Anandhan, and A. Venimadhav, Electroactive poly(vinylidene fluoride) fluoride separator for sodium ion battery with high coulombic efficiency. *Solid State Ionics* 292, 130–135 (2016).
55. Y.-J. Kim, C.H. Ahn, M.B. Lee, and M.-S. Choi, Characteristics of electrospun PVDF/SiO_2 composite nanofiber membranes as polymer electrolyte. *Mater. Chem. Phys.* 127, 137–142 (2011).
56. X. Cai, T. Lei, D. Sun, and L. Lin, A critical analysis of the α, β and γ phases in poly(vinylidene fluoride) using FTIR. *RSC Adv.* 7, 15382–15389 (2017).
57. W.A. Yee, M. Kotaki, Y. Liu, and X. Lu, Morphology, polymorphism behavior and molecular orientation of electrospun poly(vinylidene fluoride) fibers. *Polymer* 48, 512–521 (2007).
58. S. Badatya, D.K. Bharti, A.K. Srivastava, and M.K. Gupta, Solution processed high performance piezoelectric eggshell membrane—PVDF layer composite nanogenerator via tuning the interfacial polarization. *J. Alloys Compds.* 863, 158406 (2021).
59. T. Prabhakaran and J. Hemalatha, Ferroelectric and magnetic studies on unpoled Poly (vinylidene Fluoride)/Fe_3O_4 magnetoelectric nanocomposite structures. *Mater. Chem. Phys.* 137, 781–787 (2013).
60. D. Guo, K. Cai, P. Deng, G. Si, L. Sun, F. Chen, H. Ning, L. Jin, and J. Ma, Structure tailorable triple-phase and pure double-polar-phase flexible IF-WS_2@poly(vinylidene fluoride) nanocomposites with enhanced electrical and mechanical properties. *J. Materiomics* 6, 563–572 (2020).
61. N. Abzan, M. Kharaziha, and S. Labbaf, Development of three-dimensional piezoelectric polyvinylidene fluoride-graphene oxide scaffold by non-solvent induced phase separation method for nerve tissue engineering. *Mater. Des.* 167, 107636 (2019).
62. Y. Li, X. Su, K. Liang, C. Luo, P. Li, J. Hu, G. Li, H. Jiang, and K. Wang, Multi-layered BTO/PVDF nanogenerator with highly enhanced performance induced by interlaminar electric field. *Micro Electron. Eng.* 244, 111557 (2021).
63. R.K. Singh, S.W. Lye, and J. Miao, Holistic investigation of the electrospinning parameters for high percentage of β-phase in PVDF nanofibers. *Polymer* 214, 123366 (2021).
64. S.K. Ghosh and D. Mandal, Synergistically enhanced piezoelectric output in highly aligned 1D polymer nanofibers integrated all-fiber nanogenerator for wearable nano-tactile sensor. *Nnao Energy* 53, 245–257 (2018).
65. H.H. Singh, S. Singh, and N. Khare, Enhanced β-phase in PVDF polymer nanocomposite and its application for nanogenerator. *Polym. Adv. Technol.. Adv. Technol.* 29, 143–150 (2017).
66. N. Chamankar, R. Khajavi, A.A. Yousefi, A. Rashidi, and F. Golestanifard, A flexible piezoelectric pressure sensor based on PVDF nanocomposite fibers doped with PZT particles for energy harvesting applications. *Ceram. Int.* 46, 19669–19681 (2020).
67. M. Sahu, S. Hajra, K. Lee, P.L. Deepti, K. Mistewicz, and H.J. Kim, Piezoelectric nanogenerator based on lead-free flexible PVDF-barium titanate composite films for driving low power. *Electronics* 11, 2 (2021).
68. A.M. Abdullah, M.U.K. Sadaf, F. Tasnim, H. Vasquez, K. Lozano, and M.J. Uddin, KNN based piezo-triboelectric lead-free hybrid energy films. *Nano Energy* 86, 106133 (2021).
69. A.S.M.I. Uddin, D. Lee, C. Cho, and B. Kim, Impact of multi-walled CNT incorporation on dielectric properties of PVDF-$BaTiO_3$ nanocomposites and their energy harvesting possibilities. *Coatings* 12, 77 (2022).
70. S. Pratihar, A. Patra, A. Sasmal, S.K. Medda, and S. Sen, Enhanced dielectric, ferroelectric, energy storage and mechanical energy harvesting performance of ZnO-PVDF composites included by MWCNT as an additive third phase. *Soft Matter* 17, 8483–8495 (2021).
71. C.-M. Wu, M.-H. Chou, and W.-Y. Zeng, Piezoelectric response of aligned electrospun polyvinylidene fluoride/carbon nanotube nanofibrous membranes. *Nanomaterials* 8(6), 420 (2018).
72. A. Samadi and S.M. Hosseini, Influence of TiO_2-Fe_3O_4-MWCNT hybrid nanotubes on piezoelectric and electromagnetic wave absorption properties of electrospun PVDF nanocomposites. *Org. Electron.* 75, 105405 (2019).

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

Komal Verma[1] · Aman Kumar[1] · Richa Sharma[1]

✉ Richa Sharma
richasharma@dtu.ac.in

[1] Department of Applied Physics, Delhi Technological University, Delhi 110042, India

JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-19, No.-7, July (2024) pp 112-125 ISSN (Print) 0973-8975

FLOW THROUGH STEAM TURBINE CASCADE FOR ROUGHNESS ANALYSIS

Manjunath K.[1], Ajeet Singh Sikarwar[2], Naushad Ahmad Ansari[3]

[1]Mechanical Engineering Department, Delhi Technological University Delhi 110042, India

[2]Mechanical Engineering Department, Institute of Technology and Management, Gwalior 474001, India

[3]Mechanical Engineering Department, Delhi Technological University Delhi 110042, India

Email : [1]manjunath.k@dtu.ac.in, [2]ajeetsingh.sikarwar10@gmail.com, [3]naushad.nsr@gmail.com

Corresponding Author: **Naushad Ahmad Ansari**

https://doi.org/10.26782/jmcms.2024.07.00008

(Received: May 07, 2024; Revised: June 23, 2024; Accepted: July 05, 2024)

Abstract

The flow of liquids is essential to our understanding of the world. Traditionally, this is done by studying the flow of liquids using wind tunnels in the model. However, the field of computer fluid dynamics has been born over the past century. A program that can model fluid flow is CFD software. Gambit 2.4.6 created a three-dimensional geometry of four reaction blades with a square cascade and studied the secondary losses using FLUENT 6.2. The air is chosen as a working material. Air passes through the turbine cascade at a maximum input speed of 102 m/s. The cascade opens to the atmosphere when exiting. Firstly, the two surfaces of the blade cascade have been smoothed and the secondary losses analyzed. This total flow loss was compared with a roughness applied individually to the suction and pressure surfaces of 250 microns, 750 microns, and 1000 microns in thickness and examined the effect of the thickness on flow loss.

Keywords: Blade surface, Effect of roughness, End loss phenomena, Loss Coefficient, Turbine steam path.

I. Introduction

As energy savings are increasingly needed, sustainable efforts are being made worldwide to improve the efficiency of thermal power plants. Today, 80 percent of the world's electricity is produced by steam turbines. Historical divisions of loss into profile loss, end loss, and leakage loss are still widely used, but it is clear that loss

generation mechanisms are rare, but it is now widely recognized that they are independent. Final loss is the main component of internal turbine aerodynamic loss. End losses are sometimes referred to as secondary losses because they are partly generated by secondary flows generated by three-dimensional flows near the end walls. End-of-loss phenomena are mainly important for two reasons in turbo machines. First, it causes a drop in pressure in a stage and second, it makes the exit flow of a stage not uniform, which can cause a decrease in pressure losses in a downstream line. From the above-mentioned point of view, improving the efficiency of turbo machines has become a key field of research. Thus, research efforts on highly efficient turbines have improved the understanding of flow fields and the losses associated with turbine stages.

Using a k-ξ turbulence model, J. Y. Moon and K. S. Yong [X] examined how end wall fencing affected secondary flow reduction. They also justified the end wall fence's optimal placement in order to minimize losses due to secondary flow. This is because the end-wall fencing keeps the path vortex and the pressure-side horseshoe vortex from merging, which lowers overall pressure loss. Brear et al. [XI] looked into changing the leading edge geometry to lessen pressure surface separation. By expanding the force near the wall, they found that raising the sharp edge thickness at the strain surface decreases the strength of the optional stream.

Tests focus on rectangular response fountains and drive cutting edges with soft and hard surfaces by Samsher [XVII]. The 220, 100, and 50-grade emery paper roughness at the upper edge, middle harmony, and upper edge covers 33% of the surface. Attractions and resistance surfaces are gradually coated in these rough spots. Saha et al. [III] examined the violent flow through a three-layer non-axisymmetric cut tube. They found that end wall shaping brings down the pitch-wise tension slope close to the end wall, bringing down the probability of stream partition.

Singh and Kale [VIII] numerically simulated a rectilinear blade cascade with a surface roughness of up to 100 m at various turbine blade locations. It has been pointed out that the roughness affects the efficiency of the turbine blade more on the input surface than on the pressure side. When the roughness is placed on the subsequent edge of the suction surface instead of on the middle chord, the mass-average profile losses are higher and the losses are lower when the top edge is rough. Sonoda et al. [XIX] used the axis symmetrical contour method for the end wall. To minimize secondary losses in low-view high-pressure turbines. They investigated the effects of three different techniques of final wall contours. They found that the average mass loss was lower than the baseline.

Fixed edge overflow testing is as yet essential for the execution expectation of hub stream blowers and turbines. Zachos et al. [XV]. Finding the general strain misfortune and outlet stream point through the cutting edges is the point of the review. The information will be utilized to approve and adjust a mathematical solver, upgrading its ability to isolate streams. This exploration examines the three-layered direct blower outpouring's presentation at an extremely regrettable frequency point. Comparing a laminar boundary layer to an analogous turbulent layer, Goodhand [XIV] found that the latter has less skin friction. This feature has not yet been applied to the compressor

aerofoil. The leading edge geometry of the aerofoil could be carefully controlled to maintain sizable amounts of laminar flow over its suction surfaces.

Adeniyi et al. [I] used the k - SST cuteness model and the Enormous Swirl Reenactment with the help of ANSYS-CFX to concentrate on the wakes in the field of the stream of a direct flow cut end. This results in pressure differences in the flow of liquid due to the opening between the wall of the end and the tip of the knife. A large speed deficit was found to characterize both the edge wake and the tip leakage vortex. Singoria and Samsher [XX] examined the impact of the sharp edge surface hardness on the optional misfortune of the stream (end loss) and the misfortune of the profile in an air overflow with four reaction cutting edges. These results show that when a 1,000 m roughness is added to both the pressure surface and the suction surface, the mass loss is almost doubled.

The secondary loss is lowest in a cascade with smooth blade suction surfaces and largest in one with rough pressure surfaces. When both surfaces of a blade are rough, the impact of roughness on secondary loss is intermediate.

Azim et al. [XVI] used CFD analysis to study the boundary layer delay in NACA aerofoil on the suction side. At increasing angles of attack, transition flow over airfoils exhibits a great deal of flow instability, including turbulence, boundary layer transition, and localized separation areas. Experiments were performed on the NACA aerofoil cascade of compressor with inlet flow by Zhang et al. [XXII]. The progress zone extends from the pulling surface to the end wall as the unfavorable tension inclination increments, and the way of behaving of the end-wall limit layer decays with expanding entry compressing capacity.

Peyvan and Benisi [V] examined the design of a single-stage axial compressor and forecasted its performance outside the design with a 1D and 3D model. In a single-layer model, conditions are dealt with in the mean line at three main stations: the state output, the rotor bay, and the rotor output. Bai et al. [XVIII] studied the impact of surface hardness on the improvement of limits and misfortunes at the turbine edge with different Reynolds numbers. The result shows that the limit layer speed profile is a smooth edge with a bumpier surface than a smooth edge.

Tang et al. [XXI] studied the robustness effect on the development of 3D corner partitions in a rapid and low-perspective blower cascade. When the two end-side disconnection zones and the first-line dissection point push simultaneously in front of the throat position of the cutting surface, the newly proposed sharp edge attraction surface slows down. Sillitoe et al. [II] used a large-eddy simulation to analyze the border layer transition mechanism of two rectilinear compressor cascades. The study looks at how well different sub-grid scale models work and how unstable wakes affect transition. To create novel compressor designs, it is crucial to accurately forecast the transitional behavior in gas-turbine compressors, as this behavior can have a substantial impact on the compressor's aerodynamic performance.

Baumgartner et al. [VI] explored the impact of the obscure isentropic example on the stream field inside a turbine vane. The researchers led a few exploratory and computational tests that empowered annular fountain testing with different working

liquids and working circumstances, utilizing a newly changed transient air stream. The discoveries show that over the range of isentropic examples assessed, the misfortune could shift by 20% to 35% relying upon the vane leave Mach number. The exploration of the effect of essential mathematical factors and stream qualities on optional stream misfortunes in a turbine overflow was finished by Mesbah et al. [XIII], who distributed the consequences of their mathematical reproduction. The effects of the number of exits of Mach, the pitch-to-chord ratio, the angle of intake, and the angle ratio of the blade on the loss of the end wall and the generation of secondary flow were investigated.

Li et al. [IX] fostered a slowdown expectation model in light of Koch's two-layered diffuser relationship way to deal with evaluate the pivotal stream blower stage's greatest static tension ascent capacity. Dissimilar to the first Koch's model, which just records for the expansion in the channel dynamic strain factor brought about by a high stun point, this paper considers the impact of stream points on the dispersion length and determines the numerical articulation of the changed non-layered dissemination length. Majcher et al. [XII] offer a mathematical process of wind current lift coefficients, drag coefficients, and flux rotation points as part of a cutting-edge overflow configuration approach in a pivotal flow machine. Alongside a procedure and results segment, the paper has an exhaustive classification and catchphrases segment.

The flow of the axial turbine is very complex, 3D, and unstable, resulting in high losses, low efficiency, and low performance. Despite great progress in this field over the past 50 years, the mechanisms for the final flow of the walls are still not very clear. Even today, these flows are a challenging task for turbine designers and researchers. A better understanding of complex three-dimensional flows, their origins, and associated losses will certainly be useful in any attempt to improve turbine performance. Previous work in this field has focused on the two-dimensional geometry of blade profile modeling and simulation. Consequently, the present work is a step forward in comparison with the previous work, and we are trying to see if FLUENT 6.2 can capture complex three-dimensional secondary flow vortices near the end wall area.

The main objective of this study is to perform computational studies on the end-loss phenomena in the turbine steam routes. The analysis includes a three-dimensional computational investigation of the end-loss phenomena of the impulsive blade profile (profile 3525) in a stable turbine flow path. Gambit 2.4.6 is used as a preprocessor [IV], developed a three-dimensional cascade geometry model, and FLUENT 6.2 [VI] is used as a solution and postprocessor for flow simulation. Once the model is verified, more simulations will be performed to investigate the effect of final loss on flow rates of higher Mach values of 0.4 to 0.95, typical of modern utility steam turbines. Simulation results will investigate the impact of roughness on the average mass loss coefficient. It also studied the effects of roughness on secondary losses.

II. Analysis

The friction loss and the secondary loss occur when the liquid flows through the passage of the turbine blade. The roughness of the aerofoil surface has a great influence on these losses. In the current computational study, computational fluid

dynamics (CFD) determined the final loss of a 3D circular wheel turbine cascade by Samsher [XVII].

a) Description of 3D geometry modeling and computational methodology

The accuracy of the computational model depends on the accurate selection of the boundary conditions and the analysis hypotheses. The quality of networks, density, and real geometry representations are the second factors determining computer simulation accuracy. The quality of the grid depends on the shape, low aspect ratio, and low extension ratio of the mesh cell. If further improvements to the network do not change the result, i.e. if grid-dependent solutions are achieved, the grid density is sufficient. The accurate representation of geometry is important, as computer simulation only predicts the flow of ideal models. The accuracy of discrete planning and turbulent modeling is usually dependent on the code and cannot be influenced by the user and it is found by Singoria and Samsher [XX].

The current CFD simulation of the turbine cascade is performed with the commercially available FLUENT 6.2 CFD package as a solution and postprocessor. It is a finite-volume solution based on pressure correction. FLUENT 6.2 is particularly applicable to three-dimensional boundary flow, as the code allows adaptive grids based on gradients to achieve grid-dependent results.

b) Gambit to generate geometry

In CFD analysis, the most important part is the generation of geometry and engraving, which has a major impact on the solution. Therefore, great care should be taken to choose the right sizes of mesh and mesh Singoria and Samsher [XX]. The measurements of the cascade parameters and the flow rate for the test area are shown in Table 1. Gambit 2.4.6 is a preprocessor for the generation of geometry and meshes. First, using the blade profile x and y coordinates to create a two-dimensional model, Gambit 2.4.6 maintained the same dimensions from the input measurement plane to the exit of the tunnel as in the experimental configuration. Then, the 2D model is converted to 3D by dividing the edges of the 2D model by the height of the blade. The flow domains are modeled using four blades instead of six blades in experiments to optimize the power/processor capacity of computers. It is assumed that the flow is symmetric at a medium-range level. The 3D computing domain of the cascade of 3525 turbine blades is shown in Fig. 1.

Table 1: Dimensions of Cascade and other parameters, Samsher [XVII]

Parameters	**Blade-Profile 3525**
Cascade-type	Rectilinear
Inlet cross-section	95 x 93.2 mm^2
Type of test-blade	Impulse type
Blade chord (mm)	50
Blade rows pitch (mm)	29 mm
Blade height (mm)	95

Blade angle	80°
Inlet-angle	40°
Blade numbers	4
Channels numbers	3
Working-fluid	Air
Air inlet temperature	30°C
Velocity at the inlet plane	102m/s
Nature of flow inside cascade	Subsonic (compressible)
Reynolds number at exit	4.7×10^5

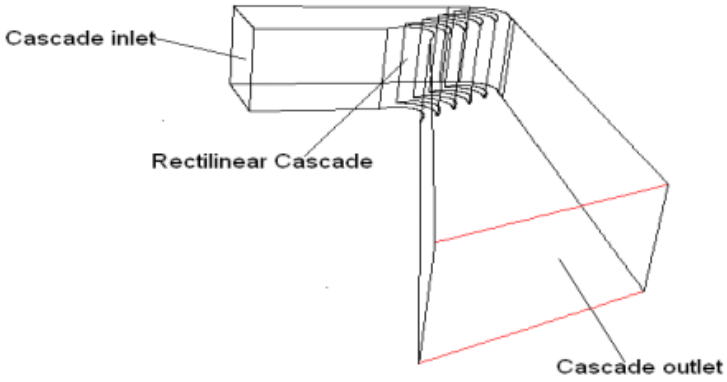


Fig. 1. 3-D computational domain of turbine blade 3525 cascade.

3-D computational domain along with turbine blade 3525 cascade created in Gambit 2.4.6 is shown in Fig. 2.

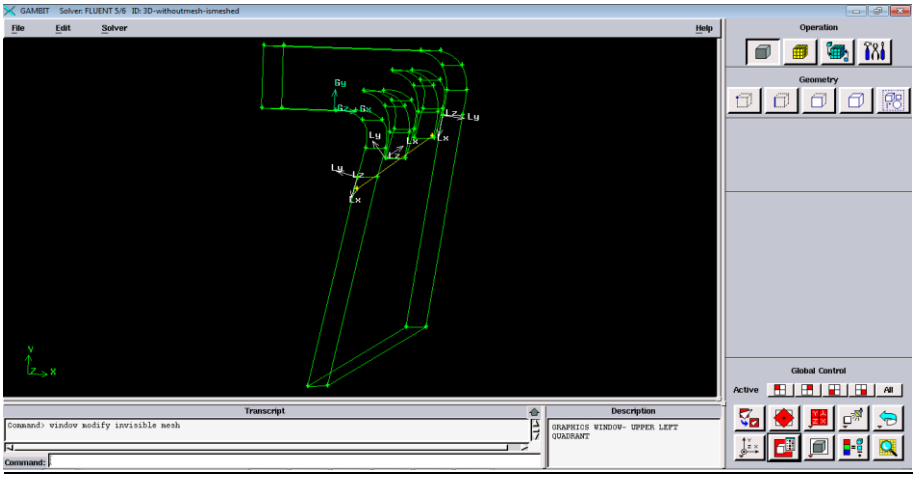


Fig. 2. 3-D computational domain along with turbine blade 3525 cascade

Table 2: Boundary conditions created as zones from various faces & boundary types

Boundary-zone	Inlet-faces	Outlet-faces	External-faces	Suction surfaces 1 to 5	Pressure-surfaces 1 to 5	Bottom end wall	Top end wall
Boundary-type	Velocity-inlet	Pressure-outlet	Wall	Wall	Wall	Wall	Wall

c) Grid generation & Meshing

The mixing of turbomachinery liquid areas involves the mixing of blades and blade passages. Meshes are largely divided into two types, namely structured and non-structured. Consequently, the structured mesh is easy to handle in terms of discretization, linearization, and solution, and is used in current cascade geometry. The typical structure of the blade's blade includes H-line, O-line, and C-line combinations. Each has its advantages and limitations. In general, in Turbomachine laser simulation, structured networks, particularly boundary layers, are preferred and more accurate than non-structured networks. The accuracy of the solution, the cost of computer hardware, and the calculation time needed depend on the quality of the grid. The quality of a good solution network depends on the density of the network, the relationship between the length and volume of adjacent cells, and the tilt. Therefore, every effort is being made to produce a good quality network without excessively increasing the computation cells. The calculation area is shown in Fig. 3.

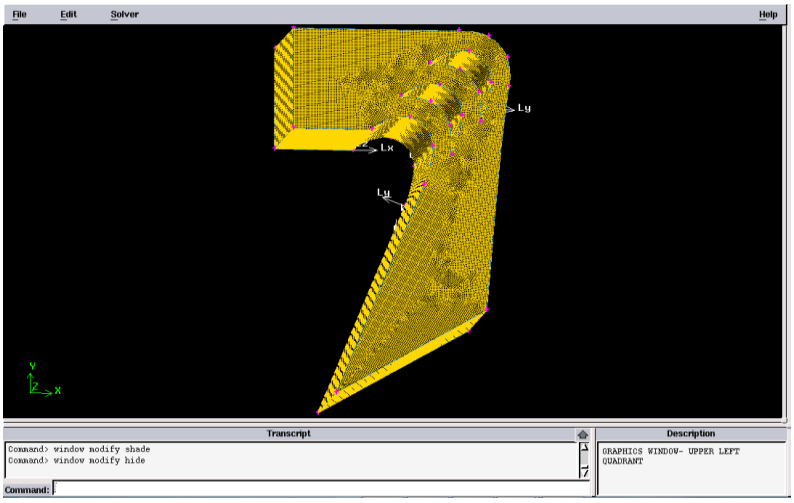


Fig. 3. Meshing of the computational domain

The surface of the 2D geometry profile is swept to generate the volume. The blade pressure and suction surfaces are divided into three parts. This was done to study the effects of localized roughness on secondary loss. The blade volume is removed from the outer shell volume, and the blade position is empty (due to lack of flow). Finally,

the size of the mesh is selected according to the computer capacity of the masked geometry. The areas are as follows. – Pressure input, velocity output, wall, pressure surface, suction surface, wall are also defined. As shown in Table 2 [XX], the boundary conditions created as different faces and boundary types are given names.

d) Validation of results

The present CFD results for the smooth blades were validated with the experimental result obtained by Samsher [XVII] to ensure that the k-epsilon model-RNG which was used in Fluent for simulation purposes is correct as presented in Fig. 4.

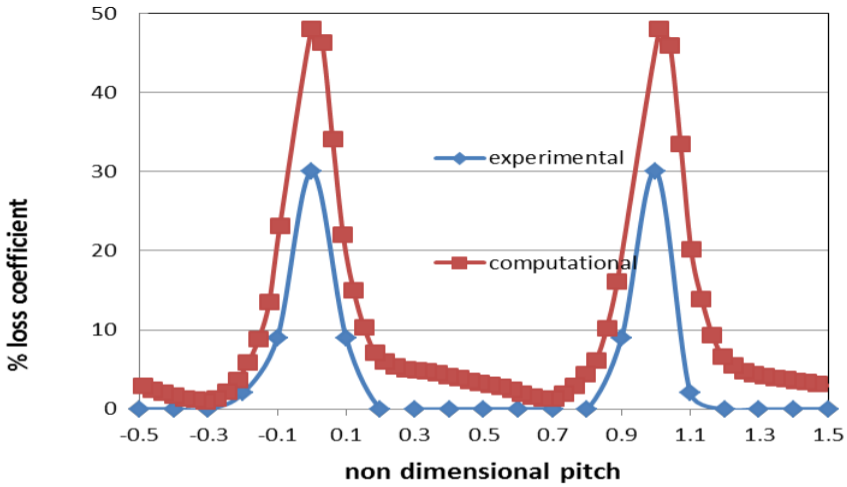


Fig. 4. Computational results validation with that of experimental results by Samsher [XVII]

The results were validated as both graphs appeared to be very close and the % loss coefficient was computed and found to be 6.03%. Two peaks were obtained because only one channel through the cascade was considered. The peaks indicate the presence of both profile-loss and secondary-loss; whereas in the middle portion, the profile of the blade offers no obstruction, and therefore only secondary losses are observed.

III. Results and Discussion

The generated mesh is exported to FLUENT 6.2. The equations of continuous, energy, and momentum are solved separately. A separate solver is used to solve the moment equation for the first time using the guess pressure and generate the pressure correction equation. Due to the "guess and correction" nature of linear systems, many repetitions are usually required. This requires a careful selection of the relaxation parameters of the variables.

After the completion of the model, the quality of the mesh was checked in FLUENT 6.2. The simulation was carried out by a three-dimension double-precision method. As the first step, losses were computed for roughness provided on all surfaces. The total losses for roughness heights of 250, 500, 750, and 1000 micrometers were compared with that for the smooth profile applied separately on all blades: (i) all-pressure-

surfaces (PSR) (ii) all-suction-surfaces (SSR) (iii) both-surfaces (resistance & pressure) of all blades (ASR).

The roughness effects on the suction surface are presented in Fig. 5. Different grades of the roughness of 250µm, 500µm,750µm, and 1000µm were applied and results were obtained. The total blade loss was found to increase with the increase in roughness. The increased losses due to roughness may be attributed to the increased turbulence and earlier boundary layer separation resulting in a wider wake.

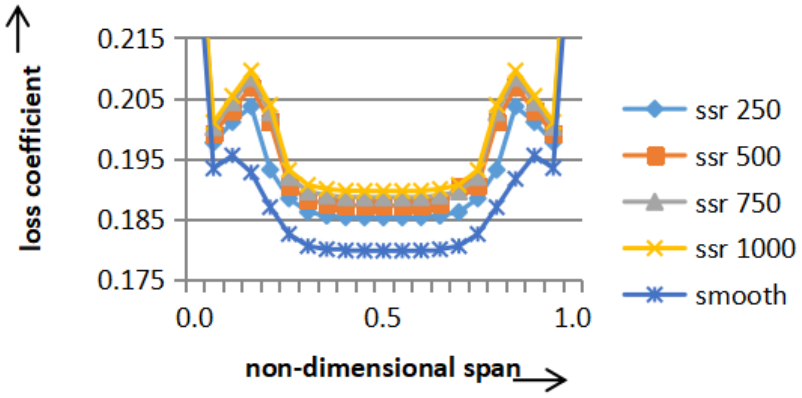


Fig. 5. The variation in total loss along the span with the roughness of the different levels at the suction surface.

The roughness effects when applied to the pressure surface are shown in Fig. 6. Different grades of the roughness of 250µm, 500µm, 750µm, and 1000µm were applied and results were obtained. The total losses were found to meagerly increase with increasing roughness. The reason for the meager increase may be attributed to the lesser effect of roughness on boundary layer separation from the pressure surface.

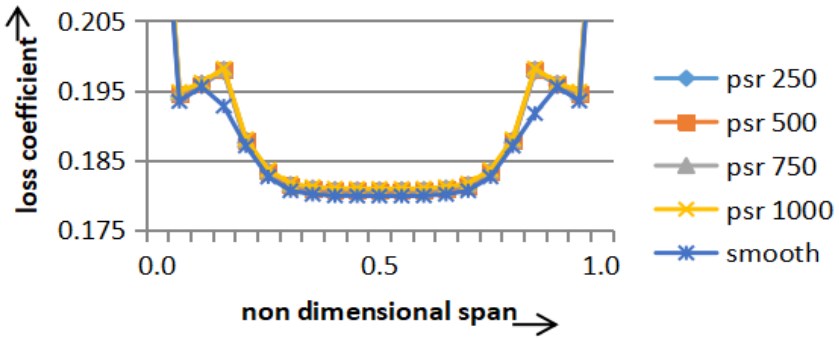


Fig. 6. Variation in total loss along span with roughness of different levels at pressure surface.

The effects of roughness when applied over the entire surface are shown in Fig. 7. Different grades of the roughness of 250µm, 500µm, 750µm, and 1000µm were applied and results were obtained. Here, too the total as well as the profile losses were found to increase with increasing roughness.

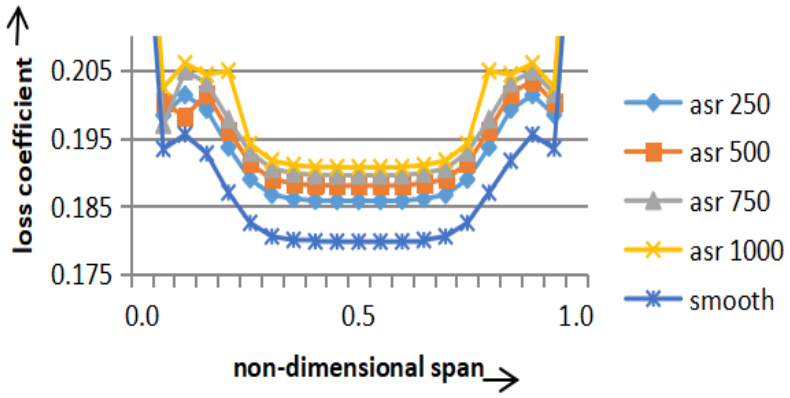


Fig. 7. Variation in total loss along span with the roughness of different levels at the entire surface

The total and profile losses were found to be the maximum in ASR, followed by SSR, followed by PSR, and the lowest in smooth cases, as presented in Fig. 8. The loss of profile in the case of SSR 500 was found to be 3.71% more than that of PSR 500 and 4.13% more in case of ASR 500 for turbine 3525.

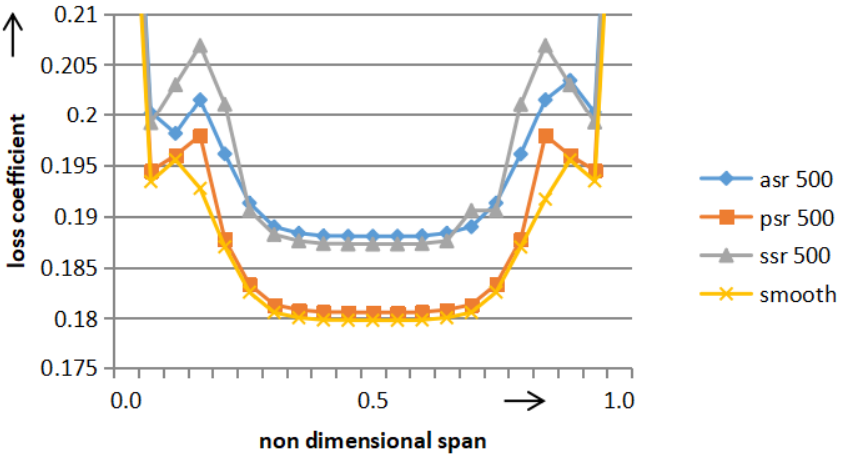


Fig. 8. Variation in total loss along span with roughness of different levels at all surfaces

a) Comparison of secondary losses for turbine profile 3525

The secondary losses or the end losses however show a different trend. In the case of ASR, the end losses initially(250µm) were found to be lower than that of a smooth profile and were found to decrease with increasing roughness up to a grade of 500µm, after which they start showing an increasing trend with increasing roughness. Variation in secondary loss (expressed as % of total loss) for different levels and locations of roughness is shown in Fig. 9. The end losses in the case of PSR were found to be always invariable and almost equal at all roughness grades. The end losses were however found to be greater than that of a smooth profile. The secondary losses in case of SSR was

initially found to increase with increasing roughness up to a roughness grade of 500µm, after which it started showing a decreasing trend.

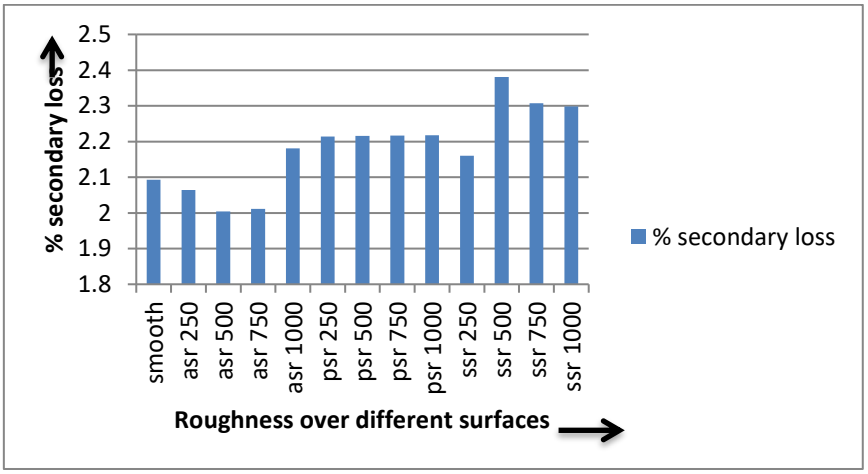


Fig. 9. Variation in secondary loss (expressed as % of total loss) for different levels and locations of roughness

b) Practical Implications

Corrosion is essentially the loss of a turbine material due to a variety of materials that contain contaminants such as steam, water, or gas injection. Erosion is a loss of material due to hard particles present in the stream. The degradation and failure of the turbine components depend on various parameters, such as the Mach number, Reynolds number, the size of the particles, and the material of the particles. Particles with a diameter of less than 10 microns are toxic because they do not have sufficient motion energy to cause erosion. These particles are usually attached to the surface of the blade and accumulate, and the particle diameter exceeds 10 microns, which is less stick efficiency. In essence, erosion is caused by solid particles of large diameter, and the likelihood of erosion is found mainly at the top and bottom of the blade. Despite progress in gas filtration systems, other problems include the formation of particles during combustion by low-grade fuels. Thus, we must analyze the flow behavior of the turbine under the injection of solid and liquid particles.

Using Fluent, you can analyze the length of the blade because, during the flow of fluid, heavy and large-diameter particles hit most of the edge of the blade's release surface, other particles hit the pressure surface of the blade, and almost all of the blades are generated by particles, as you can see in the blade. The shape of the blade is also damaged by the particles as its front and rear edges are completely affected. Lighter particles do not cause erosion but stick to the blade and cause corrosion on the blade. We can eliminate the damage effects by cleaning and polishing the blade, or we can say that some damage losses can be recovered, but erosion losses cannot be recovered. These losses reduce the strength of the blade and the overall efficiency of the turbine. The particle traces describe the path of the particle by the time of the particle's residence, and here we see that, with the increase in diameter, the particle reaches the

top edge, then parallels the length of the blade, and the effect increases in the direction of the suction surface. It can be said that under low speeds and minimal diameter particles, the probability of deposits is higher due to decreased inertia effects. It is recommended to increase the length of the blade by increasing the velocity and diameter of the injection particles on the side of the suction and pressure.

IV. Conclusions

For steam turbines, efficiency is arguably the most crucial performance factor. We can gain a better understanding of the loss mechanisms in steam turbines by researching the impact of end-loss phenomena on machine efficiency. Computer fluid dynamics is used to computerize the end loss phenomena of turbine steam paths. In this study, end-loss phenomena of the Impulse Blade Profile (profile 3525) in the turbine flow path under stable conditions are investigated using three-dimensional computing methods.

Gambit 2.4.6 was used as a preprocessor, FLUENT 6.2 as a solvent, and a postprocessor for flow simulation, creating a three-dimensional model with cascade geometry. After modeling, simulations are performed to examine the impact of end-loss on higher Mach flow in modern commercial steam turbines, with a range of 0.4 to 0.95. Based on simulation results, the impact of roughness on the mass average loss coefficient is investigated. In addition, the impact of roughness on secondary losses is investigated.

It is observed that for a given grade of roughness, the profile losses are found to be maximum for roughness on the suction surface of the mid-chord section and minimum for roughness on the pressure surface of the mid-chord section. The profile loss for roughness being applied on the entire mid-chord section was found to be less. This implies that any roughness on the pressure surface of the mid-chord section tends to reduce total losses. It has also been observed that any roughness applied on the entire leading edge is most detrimental followed by roughness being applied on the entire trailing edge and is least for the mid-chord section. Furthermore, any roughness imparted to a portion of the blade's suction surface is more harmful than the same roughness applied to the blade's pressure surface.

Conflict of Interest

There was no relevant conflict of interest regarding this paper.

References

I. A. A. Adeniyi, A. Mohammed and S. O. Emmanuel. : ‘CFD Modelling of Wakes on Cascade Compressor Blades’. *International Journal of Advances in Science and Technology*. Vol. 4(2), pp. 60-67, 2012.

II. A. D. Scillitoe, P. G. Tucker and P. Adami. : 'Large eddy simulation of boundary layer transition mechanisms in a gas-turbine compressor cascade'. *Journal of Turbomachinery*. Vol. 141(6), pp. 61-68, 2019. 10.1115/1.4042023

III. A. K. Saha and S. Acharya. : 'Computations of turbulent flow and heat transfer through a three-dimensional nonaxisymmetric blade passage'. *ASME Journal of Turbo-machinery*. Vol. 130(3), pp. 1008-1018, 2008. 10.1115/1.2776952

IV. ANSYS Fluent Meshing User's Guide. (2015).

V. A. Peyvan, and A. H. Benisi. : 'Axial-Flow Compressor Performance Prediction in Design and Off-Design Conditions through 1-D and 3-D Modeling and Experimental Study'. *Journal of Applied Fluid Mechanics*. Vol. 9(5), pp. 2149-2160, 2016. 10.18869/acadpub.jafm.68.236.25222

VI. D. Baumgärtner, J. J. Otter and A. P Wheeler. : 'The effect of isentropic exponent on transonic turbine performance'. *Journal of Turbomachinery*. Vol. 142(8), pp. 81-87, 2020. 10.1115/1.4046528

VII. FLUENT, 2005. FLUENT 6.2 Users Guide. Lebanon, USA.

VIII. H. R. Singh and S. R. Kale. : 'A numerical study of the effect of roughness on the turbine blade cascade performance'. *Progress in Computational Fluid Dynamics*. Vol. 8(7), pp. 439-46, 2008. 10.1504/PCFD.2008.021320

IX. J. Li, J. Teng, M. Ferlauto, M. Zhu and X. Qiang. : 'An improved stall prediction model for axial compressor stage based on diffuser analogy'. *Aerospace Science and Technology*. Vol. 127, 2022. 10.1016/j.ast.2022.107692

X. J. Y. Moon and K. S. Yong. : 'Counter-rotating stream-wise vortex formation in the turbine cascade with end wall fencing'. *Int. J. Computers and Fluids*. Vol. 30(4), pp. 473-490, 2001. 10.1016/S0045-7930(00)00026-8

XI. M. J. Brear, H. P. Hodson, P. Gonzalez and N. W. Harvey. : 'Pressure surface separations in low-pressure turbines—Part 2: Interactions with the secondary flow'. *Int. J. Turbomach*. Vol.124(3), pp. 402-409, 2002. 10.1115/1.1450765

XII. M. Majcher, M. Frant and R. Kieszek. : 'Preliminary Numerical Study of a Rectilinear Blade Cascade Flow for a Determination of Aerodynamic Characteristics. *International Review of Aerospace Engineering* (I.RE.AS.E), Vol. 16(4), pp. 143-151, 2023. 10.15866/irease.v16i4.24065

XIII. M. Mesbah., V. G. Gribin and K. Souri. : 'Investigation of the effects of main geometric parameters and flow characteristics on secondary flow losses in a turbine cascade'. *Journal of Physics: Conference Series*. Vol. 3, pp. 21-31, 2021. 10.1088/1742-6596/2131/3/032081

XIV. M. N. Goodhand. : 'Laminar flow compressor blades'. *9th Osborne Reynolds Colloquium and Research Student Award*, Department of Aeronautics Imperial College London, 2011.

XV. P. K. Zachos, N. Grech, B. Charnley, V. Pachidis and R. Singh. : 'Experimental and numerical investigation of a compressor cascade at highly negative incidence'. : '*Engineering Applications of Computational Fluid Mechanics*. Vol. 5(1), pp. 26-36, 2011. 10.1080/19942060.2011.11015350

XVI. R. Azim, M. M. Hasan and M. Ali. : 'Numerical investigation on the delay of boundary layer separation by suction for NACA 4412'. *Procedia Engineering*. Vol. 105, pp. 329-334, 2015. 10.1016/j.proeng.2015.05.013

XVII. Samsher. : 'Effects of localized roughness over reaction and impulse blades on loss coefficient'. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*. Vol. 221(1), pp. 21-32, 2007. 10.1243/09576509JPE214

XVIII. T. Bai, J. Liu, W. Zhang, and Z. Zou. : 'Effect of surface roughness on the aerodynamic performance of turbine blade cascade'. *Propulsion and Power Research*. Vol. 3(2), pp. 82-89, 2014. 10.1016/j.jppr.2014.05.001

XIX. T. Sonoda, M. Hasenjäger, T. Arima and B. Sendhoff. : 'Effect of end wall contouring on performance of ultra-low aspect ratio transonic turbine inlet guide vanes'. *ASME Journal of Turbomachinery*. Vol. 131(1), pp. 11020-11031, 2009. 10.1115/1.2813015

XX. V. K. Singoria and Samsher. : 'The Study of End Losses in a Three Dimensional Rectilinear Turbine Cascade'. *Int. J. of Emerging Technology and Advanced Engineering*. Vol. 3(8), pp. 782-797, 2013. https://www.scribd.com/document/389319869/IJETAE-0813-122

XXI. Y. Tang, Y. Liu and L. Lu. : 'Solidity effect on corner separation and its control in a high-speed low aspect ratio compressor cascade'. *Int. J. of Mechanical Sciences*. Vol. 142, pp. 304-325, 2018. 10.1016/j.ijmecsci.2018.04.048

XXII. Zhang, Y. Wu, Y. Li, and H. Lu. : 'Experimental investigation on a high subsonic compressor cascade flow'. *Chinese Journal of Aeronautics*. Vol. 28(4), pp. 1034-1043, 2015. 10.1016/j.cja.2015.06.019

Forecasting Electric Vehicle Sales in India using Singular Spectrum Analysis

Presented by

YASH AGGARWAL
MA Economics
Delhi Technological University
yash17aggarwal@gmail.com

Under the Supervision of

DR. ROMA DEBNATH
Associate Professor
Indian Institute of Public Administration
romamitra@iipa.education

JULY 2024

सत्यमेव जयते
Government of India
NCGG
National Centre for Good Governance
The Torch Bearer of Good Governance

National Centre for Good Governance (NCGG)
Department of Administrative Reforms and Public Grievances, Ministry of Personnel, Public Grievances and Pensions, Government of India

CERTIFICATE

This is to certify that **Yash Aggarwal,** a student of Delhi Technological University (DTU), has successfully concluded his research project titled **"Forecasting Electric Vehicle Sales in India using Singular Spectrum Analysis"** as part of the internship program at the National Centre for Good Governance (NCGG) under my mentorship.

I, Dr. Roma Debnath, hereby validate the successful completion of the internship report within the internship program at the National Centre for Good Governance (NCGG). The report submitted by Yash Aggarwal is an authentic work carried out by him under my supervision and guidance. I have reviewed and assessed the intern's performance throughout the internship period.

(Dr. Roma Mitra Debnath)
Associate Professor
Indian Institute of Public Administration

Date: 01.07.2024

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to Shri. V. Srinivas Sir, Secretary, DARPG and DG, NCGG for providing me with this opportunity. I am extremely grateful to my mentor Dr. Roma Debnath, Associate Professor, Indian Institute of Public Administration (IIPA). Despite her numerous responsibilities, she graciously took the time to guide us, ensuring we stayed on the right path throughout our research paper. I am thankful to Dr. Gazala Hasan, Professor, NCGG for her unwavering support throughout the period of the internship. I also extend my heartfelt thanks to Dr. Zaid Fakhar, Consultant, Ms. Megha Tomar and Mr. Akash Sikdar, Young Professionals, and Mr. Sachin, DEO for their invaluable assistance in addressing my queries related to my research paper.

I would also like to express my appreciation to my co-interns at NCGG for their help in broadening my perspectives and deepening my understanding of the research process. Their collaboration and insights have been immensely valuable. I am profoundly thankful to my parents for their encouragement, understanding, and patience throughout this research journey. Lastly, I offer my humble thanksgiving to God for his love, strength, and infinite blessings.



(Yash Aggarwal)
M.A. Economics
Delhi Technological University

ABSTRACT

The global transition to sustainable mobility is crucial for mitigating climate change, with the transportation sector adopting electric vehicles (EVs) to reduce greenhouse gas emissions. Countries like Norway, China, and the United States lead this shift, integrating EVs into their national strategies. In India, the push for green mobility is driven by severe air pollution, rapid urbanization, and heavy oil import reliance. Despite initiatives like the National Electric Mobility Mission Plan (NEMMP) 2020 and the Faster Adoption and Manufacturing of Hybrid and Electric Vehicles (FAME) schemes, challenges such as high costs, limited range, and inadequate infrastructure remain.

This paper addresses the scarcity of empirical research on EV demand in India by employing advanced analytical techniques to forecast EV demand, offering insights for policymakers and industry stakeholders. Using monthly sales data from June 2017 to March 2024 for two-wheelers, three-wheelers, and four-wheelers, this study applies Singular Spectrum Analysis (SSA) to decompose sales data into trends, periodic elements, and noise, adapting to the non-linearity and non-stationarity of EV sales data.

The study forecasts significant growth in the Indian EV market across all vehicle categories. SSA captures sales dynamics, indicating potential for accelerated EV adoption driven by technological advancements, regulatory support, and shifting consumer preferences. These findings inform strategic decisions for infrastructure development and policy formulation, emphasising the need for continued government support, innovation, and consumer awareness to achieve sustainability goals.

CONTENT

TITLE	Page no.
Certificate	
Acknowledgement	i.
Abstract	ii.
List of Figures	v.
List of Tables	v.
List of Abbreviations	vi.
CHAPTER 1: INTRODUCTION	
1.1. Background	1.
1.2. Global EV Outlook	3.
1.3. EV Adoption Scenario in India	5.
1.4. Government Initiatives to promote EV adoption	6.
1.5. Types of EVs in Indian Market	8.
CHAPTER 2: OVERVIEW OF LITERATURE	
2.1. Factors influencing EV Demand	9.
2.2. Forecasting Techniques and Models	12.
2.3. Emerging Trends and Future Projections	13.
2.4. Research Gap	14.
CHAPTER 3: METHODOLOGY	
3.1. Rationale of Using SSA	15.
3.2. Mathematical Formulation	16.
3.3. Data	17.
CHAPTER 4: Empirical Analysis	
4.1. Data Transformation	18.
4.2. SSA Decomposition	19.
4.3. Trend Extraction and Forecasting	21.
4.4. Forecasting Future Sales for Electric Vehicles	23.

CHAPTER 5: FINDINGS 25.

CHAPTER 6: POLICY SUGGESTION, CONCLUSION AND FUTURE SCOPE

6.1. Policy Support 25.

6.2. Conclusions and Way Forward 26.

6.3. Future Scope of Study 27.

Plagiarism Report viii.

References ix.

LIST OF FIGURES

Figure Number	Figure Name
1.	Raw Sales data of 2W, 3W and 4W
2.	Logarithmic Transformed Sales data of 2W, 3W and 4W
3.	Singular Spectrum Component: 2 Wheeler
4.	Singular Spectrum Component: 3 Wheeler
5.	Singular Spectrum Component: 4 Wheeler
6.	Trend Extraction: 2 Wheeler
7.	Trend Extraction: 3 Wheeler
8.	Trend Extraction: 4 Wheeler
9.	Forecasted Sales using SSA: 2 Wheeler
10.	Forecasted Sales using SSA: 3 Wheeler
11.	Forecasted Sales using SSA: 4 Wheeler

List of Tables

Table Number	Table Name
1.	Forecasted Values of 2W, 3W and 4W

List of Abbreviations

Abbreviation	Full Form
ACC	Advanced Chemistry Cell
BEVs	Battery Electric Vehicles
CEEW	Council on Energy, Environment and Water
COP26	26th UN Climate Change Conference of the Parties
EV	Electric Vehicle
FAME	Faster Adoption and Manufacturing of Hybrid and Electric Vehicles
FY	Fiscal Year
GDP	Gross Domestic Product
GST	Goods and Services Tax
HEVs	Hybrid Electric Vehicles
IEA	International Energy Agency
LEAP	Long-range Energy Alternatives Planning
MoRTH	Ministry of Road Transport & Highways

NEMMP	National Electric Mobility Mission Plan
PHEVs	Plug-in Hybrid Electric Vehicles
PLI	Production Linked Incentive
RBI	Reserve Bank of India
SSA	Singular Spectrum Analysis
SVD	Singular Value Decomposition
ICE	Internal Combustion Engine

1. INTRODUCTION

1.1 Background

As the world pivots towards sustainable solutions to address the pressing issues of climate change and environmental degradation, the mobility sector stands at the forefront of this transformation. As clearly pointed out by NITI Aayog, “If cities are the “Engines of Economic Growth” its mobility systems are the “wheels of that engine” (Global Mobility Summit, 2018). The urgency for this shift is compounded by the transportation sector's significant contribution to global greenhouse gas emissions, primarily driven by fossil fuels. The advent of electric vehicles (EVs) offers a promising avenue to mitigate these environmental impacts while also revolutionising the mobility landscape.

Globally, the transition towards electric mobility is gaining momentum, driven by a confluence of technological advancements, regulatory policies, and shifting consumer preferences. Nations such as Norway, China, and the United States have made substantial headway, fostering environments where EVs are increasingly becoming the norm. This global shift is not just about adopting new technologies but also about realigning economic and environmental strategies to pave the way for a sustainable future.

In the Indian context, the necessity for transitioning to green mobility is underscored by several critical factors. India, as one of the fastest-growing economies, faces unique challenges that stem from its vast population, rapid urbanisation, and escalating energy demands. India was declared as the third-most polluted country in 2023 in terms of air pollution, after Bangladesh and Pakistan (IQAir, 2024). The country is currently grappling with severe air pollution levels, particularly in urban centres, which poses health risks and economic costs. With India being home to 42 of the

world's 50 most polluted cities according to the World Air Quality Report 2023, the need for cleaner transportation options becomes even more critical.

Moreover, India's heavy reliance on oil imports, which accounted for a staggering USD 119.2 billion on crude oil imports in FY 2021-22, which is up from USD 62.2 billion spent in FY 2020-21, further accentuates the need for an indigenous and sustainable energy solution (Petroleum Planning & Analysis Cell, 2023). EVs offer a viable alternative that could reduce this dependency significantly, thus enhancing energy security and economic resilience. Further, India's commitment, as evidenced at the COP26 in Glasgow, to reach a Net Zero target by 2070, further amplifies the relevance of this transition (Ministry of Environment, Forest and Climate Change, 2022).

Historically, India has taken proactive steps towards fostering an electric mobility ecosystem. Initiatives like the National Electric Mobility Mission Plan (NEMMP) 2020 and the Faster Adoption and Manufacturing of Hybrid and Electric Vehicles (FAME) schemes illustrate the government's commitment to this cause. FAME II, for instance, significantly increased budgetary allocations to bolster EV adoption across various vehicle segments, focusing on public transport and high-usage commercial vehicles to maximise environmental and economic benefits.

The Indian EV market has witnessed a remarkable evolution over the past decade. Key milestones include the launch of Mahindra REVAi, India's first electric car, back in 2001. While initial adoption was slow, recent years have seen a surge in consumer interest and market activity driven by factors such as government incentives, declining battery costs, and a growing range of EV models from domestic and international manufacturers.

The current context presents a pivotal moment for India's EV sector. Government policies, technological advancements, infrastructure development, and shifting consumer preferences are

converging to create a conducive environment for accelerated EV adoption. The EV market in India includes two primary categories: battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs). BEVs operate solely on battery power, whereas PHEVs combine electricity and gasoline, offering a flexible approach to reducing oil dependency. Despite their benefits, widespread adoption of EVs in India faces hurdles such as high retail prices, limited battery range, and inadequate charging infrastructure. Therefore, a thorough examination of projected EV demand is essential to guide policymakers and industry participants.

Forecasting EV demand in India holds far-reaching implications. Understanding the trajectory of EV adoption will inform critical decisions regarding the allocation of resources towards charging infrastructure development, ensuring grid stability to meet increasing electricity demand, and formulating policies that encourage both EV production and consumer uptake. Furthermore, accurate forecasts empower stakeholders to proactively address potential challenges and capitalise on emerging opportunities within the rapidly evolving EV landscape.

By delving into the complexities of EV technology, macroeconomic dynamics, policy measures, and market trends, this research paper endeavours to shed light on the future of electric mobility in India. With a particular focus on forecasting EV demand, this research aims to contribute valuable insights to shape a sustainable, accessible, and efficient transportation system for the nation.

1.2 Global EV Outlook

The global landscape for EVs is undergoing a remarkable transformation, characterised by rapid growth and evolving market dynamics. In 2022, the electric car market witnessed exponential growth, with sales surpassing 10 million units, representing 14% of all new car sales globally, a significant increase from approximately 9% in 2021 and less than 5% in 2020 (IEA Report, 2023).

This surge in EV adoption is primarily concentrated in three major markets namely Chine, Europe and US.

China continues to lead global electric car sales, accounting for about 60% of the total market. The nation has not only surpassed its 2025 target for new energy vehicle sales but also now hosts more than half of the electric cars currently on the roads worldwide. In Europe, the second-largest market, there was a 15% increase in electric car sales in 2022, with electric vehicles constituting over 20% of new car sales. Meanwhile, in the United States, which ranks as the third largest market, electric car sales grew by 55% in 2022, achieving an 8% market share of total car sales (IEA Report, 2023).

The momentum of electric car sales is expected to continue robustly into 2024. Projections for the end of 2024 anticipate a total of 16.84 million electric car sales, representing a 21% year-on-year increase (ABI research, 2024). This surge is likely to result in electric cars making up 18% of total global car sales by the end of the year. Such growth is supported by national policies and incentives, and potentially further fuelled by high oil prices, which encourage consumers to transition to more economically efficient electric vehicles.

In regions outside the primary markets, the adoption of EV is rising from a smaller base but showing promising growth. In 2022, countries like India, Thailand, and Indonesia saw electric car sales more than triple compared to the previous year, reaching a total of 80,000 units. Government support in India, including a USD 3.2 billion incentive program, has catalysed significant investments totalling USD 8.3 billion, enhancing EV and component manufacturing (IEA report, 2023). Similarly, Thailand and Indonesia are bolstering their policy frameworks, potentially offering valuable lessons for other emerging economies looking to promote EV adoption.

Looking forward, under the International Energy Agency's Stated Policies Scenario (STEPS), the global outlook for electric car sales is optimistic. By 2030, the share of electric car sales is projected to reach 35%, up from an earlier forecast of less than 25%. China is expected to maintain its lead with 40% of the global total, while the United States could double its share to 20% by the decade's end, spurred by recent policy developments. Europe is anticipated to sustain its 25% share.

This shift towards EVs has significant implications for energy markets and climate objectives. According to STEPS projections, oil demand from road transport is expected to peak around 2025, with EVs displacing over 5 million barrels per day by 2030. Consequently, this transition is projected to avoid approximately 700 million tonnes of CO_2 equivalent emissions by 2030.

The expansion of battery manufacturing is being accelerated by the optimistic outlook for EVs. As of March 2023, announced expansions in battery manufacturing capacity are projected to not only meet the anticipated demand based on government pledges but also cover potential needs under the Net Zero Emissions by 2050 Scenario. This suggests that higher adoption rates for EVs than currently anticipated by government policies and targets are well within reach, further bolstering the global shift towards sustainable mobility.

1.3. EV adoption Scenario in India

The recent surge in crude oil prices, exacerbated by the Russia-Ukraine conflict, Israel-Palestine war and other international tensions, has significantly influenced India's strategic shift towards EVs as a mainstream mode of transportation (Deb et al., 2021). This shift is also supported by stringent environmental regulations, the government's scrapping policy, and increased public awareness of environmental issues (Singh et al., 2022; Jones et al., 2020; Pawar and Ahire, 2020;

Dixit and Singh, 2022). The Indian government's tax policies and incentives have further fuelled consumer interest towards EVs.

India's automotive industry, ranked as the third largest globally in 2023, is experiencing a robust transformation with major domestic manufacturers like Tata Motors and Mahindra & Mahindra Ltd. launching their EV models (Das and Bhat, 2022). The market is also welcoming international EV manufacturers such as Hyundai, Kia, MG, and BMW, enriching the consumer choices with a range of vehicles tailored to diverse needs and price points (Rohini and Asha, 2022). This influx has significantly contributed to the rapid expansion of the EV market in India within a remarkably short timeframe.

1.4. Government Initiatives to Promote EV Adoption

The Indian government has implemented several strategic measures to foster the adoption of EVs, reflecting its commitment to enhancing environmental sustainability and energy efficiency.

FAME India Scheme: Initiated by the Ministry of Heavy Industries in 2015, the Faster Adoption and Manufacturing of (Hybrid &) Electric Vehicles (FAME India) Scheme has been pivotal in promoting the adoption of electric and hybrid vehicles. Currently, Phase II of the FAME India Scheme is active, starting from April 1, 2019, with a duration of five years and a total budgetary support of ₹10,000 crores. This phase focuses on supporting the electrification of public and shared transportation and offers incentives for electric buses, passenger vehicles, three-wheelers, and two-wheelers.

Production Linked Incentive (PLI) for Advanced Chemistry Cell (ACC) Manufacturing: On May 12, 2021, the government approved the PLI scheme to promote the domestic manufacturing of Advanced Chemistry Cells, aiming to reduce the costs of batteries in India. A decrease in battery

costs is expected to directly result in lower prices for EV, making them more affordable for consumers.

PLI Scheme for Automobile and Auto Components: Approved on September 15, 2021, this scheme has a budgetary outlay of ₹25,938 crores over five years. It covers EVs and aims to enhance India's manufacturing capabilities by incentivising the production of various automobile components, including those for EVs.

GST Reductions: To make EVs and their related components more affordable, the government has reduced the Goods and Services Tax (GST) on EVs from 12% to 5%. Additionally, the GST on chargers and charging stations for EVs has been reduced from 18% to 5%.

Green License Plates and Permit Exemptions: The Ministry of Road Transport & Highways (MoRTH) has introduced green license plates for battery-operated vehicles to distinguish them from conventional vehicles. EVs with these plates are exempted from permit requirements, easing operational restrictions and encouraging their use.

Road Tax Waiver: MoRTH has also advised states to waive the road tax on EVs. This policy helps decrease the initial ownership costs of EVs, making them a more attractive option for potential buyers.

Amendment of Model Building Bylaws: In 2016, the Model Building Bylaws were amended to facilitate the establishment of charging stations in private and commercial buildings. This amendment supports the development of necessary EV charging infrastructure, addressing one of the significant barriers to EV adoption—charging facility availability.

Through these measures, the Government of India aims to accelerate the transition to a more sustainable transportation system by increasing the adoption of EVs. This not only helps in

reducing the environmental impact of the transportation sector but also positions India as a significant player in the global electric vehicle market.

1.5. Types of EVs in Indian Market

EVs can be classified into several types based on their energy sources and operational mechanisms. Each type offers distinct features and benefits, catering to various consumer needs and environmental objectives.

Battery Electric Vehicles (BEVs): Commonly known as All-Electric Vehicles (AEVs), BEVs operate exclusively on electricity stored in high-capacity battery packs. These batteries are charged from an external electrical power source and provide electricity to the motors that drive the vehicle. BEVs produce zero exhaust emissions, significantly reduce noise pollution compared to conventional vehicles, and are seen as crucial in the shift towards sustainable transportation. The environmental benefits of BEVs include the absence of tailpipe emissions and the potential to utilise renewable energy sources for charging.

Hybrid Electric Vehicles (HEVs): HEVs combine an internal combustion engine(ICE) with one or more electric motors that utilise energy stored in batteries. Unlike BEVs, HEVs are not plugged into charge, instead, their batteries are charged through regenerative braking systems and by the engine. This dual system allows HEVs to optimise fuel efficiency and lower emissions during operation. HEVs are particularly effective in reducing emissions when used in city driving, where frequent stopping provides opportunities for regenerative braking.

Plug-in Hybrid Electric Vehicles (PHEVs): PHEVs share similarities with HEVs but are equipped with larger battery packs that can be recharged by connecting to the electrical grid. This capability allows them to operate significant distances on electric power alone, thereby reducing

fuel consumption and emissions. Once the battery is depleted, the vehicle can switch to operate on the ICE, thereby providing greater flexibility and range than BEVs.

Fuel Cell Electric Vehicles (FCEVs): Often referred to as Zero-Emission Vehicles, FCEVs utilise hydrogen fuel cell technology to generate electricity that powers the vehicle's motors. Instead of storing energy in a battery, hydrogen gas is converted into electricity through a chemical reaction in the fuel cell, with water vapour as the only emission. FCEVs combine the benefits of quick refuelling and long range, characteristics similar to conventional ICE vehicles, while maintaining minimal environmental impact.

2. OVERVIEW OF LITERATURE

2.1. Factors Influencing EV Demand

The adoption of EVs is influenced by an intricate matrix of technical, economic, environmental, and social factors, with distinct variances observed across different global contexts. Key vehicle performance attributes such as driving range, speed, and safety, alongside critical economic considerations like purchase price, resale value, and operating costs, substantially shape consumer decisions in diverse markets (Thananusak et al., 2017; Wang et al., 2018). Notably, Thananusak et al. (2017) highlight that in Thailand, performance factors significantly outweigh concerns related to charging infrastructure and costs. Conversely, in denser urban environments such as Shanghai, Wang et al. (2018) found that environmental awareness and technological sophistication are dominant factors swaying consumer preferences towards EVs.

Complicating the landscape for EV adoption are persistent barriers such as limited driving ranges, a narrow range of available models, prolonged recharging times, and prevalent battery issues,

which collectively dampen consumer enthusiasm for fully EVs (Zhang et al., 2017; Cheron & Zins, 1997; Hu, X. et al., 2016; Hu, X., Martinez, & Yang, 2017; EPRI, 2010). Further exacerbating these challenges are the high perceived purchasing costs and the general convenience and benefits associated with conventional vehicles (Bockarjova & Steg, 2014). Zhang, Yu & Zou (2011) also identified the scarcity of charging stations and the extended duration required for charging as significant deterrents to EV penetration.

However, various incentives have been identified as effective in mitigating these challenges, including tax-based policies, designated lane access, complimentary parking, and enhanced insurance products, which collectively contribute to an improved value proposition for EVs (Zhang et al., 2017). Moreover, advancements in lithium-ion battery technology that prolong battery life up to 2000 charge cycles before significant degradation are also crucial in reducing the frequency and costs associated with battery replacement, thereby enhancing the economic attractiveness of EVs (Jin et al., 2022). A life cycle assessment further illustrates the long-term economic benefits of EVs compared to conventional internal combustion engine vehicles, presenting a compelling breakeven point that enhances consumer adoption (Verma et al., 2021).

The availability and enhancement of charging infrastructure play a pivotal role in alleviating range anxiety and augmenting the attractiveness of EVs. Studies conducted in diverse settings such as the U.S. by Carley et al. (2013) and in Malaysia by Sang & Bekhet (2015) demonstrate that infrastructure readiness significantly boosts EV adoption rates by addressing consumer concerns about charging availability and vehicle range. Extending this analysis to a broader international context, Yong & Park (2017) examined the factors influencing EV deployment across 24 countries, utilising a fuzzy-set qualitative comparative analysis methodology to derive insights from country-based aggregate data. Their findings underscore that robust policy support, environmental factors,

and reliable charging infrastructure are essential for the widespread adoption of EVs. This notion is further supported by research indicating that governmental interventions, including financial incentives and supportive policies, are critical in sculpting the EV market landscape (Daziano & Chiew, 2013; Yong & Park, 2017).

Adding a regional perspective, Verma et al. (2019) in their Bangalore study, identified environmental benefits and financial incentives as key motivators for EV adoption. Similarly, Kumar et al. (2020) emphasised the importance of the sharing economy and public utilities in facilitating EV purchases in India, particularly given the high costs and inadequate charging infrastructure. Additionally, a 2020 Castrol study projected the Indian EV market to reach $2 billion by 2025, pinpointing vehicle price, charging time, and driving range as primary adoption challenges. In their study, Dixit and Singh (2022) employed logistic regression to analyse survey data on 245 potential EV buyers in India, identifying age, income, vehicle cost, running cost, and driving range as significant predictors of EV purchases. Interestingly, their findings indicated that education level, employment status, and government subsidies were not significant factors influencing EV adoption in India.

Additionally, socio-cultural factors and consumer psychology significantly affect EV purchasing decisions. Helveston et al. (2015) in their comparative study of the U.S. and China, emphasised how cultural differences impact preferences for EV technologies—U.S. consumers are predominantly influenced by direct financial savings, whereas Chinese consumers are motivated by broader environmental concerns. Similarly, a study by Leabeau et al. (2013) in Belgium found that the ability to charge vehicles at home was deemed a crucial factor by consumers, highlighting the importance of understanding and integrating consumer lifestyle preferences and regional characteristics into EV marketing and infrastructure strategies.

2.2. Forecasting Techniques and Models

Forecasting the demand for EVs presents unique challenges, particularly due to the nascent stage of the market and the absence of extensive historical sales data. The complexity in predicting EV markets stems from several dynamic factors, including shifting government policies, fluctuating component costs such as batteries, developing charging infrastructure, and variable technology adoption rates across different regions (Toppr-guides, 2018; Struben & Sterman, 2008). Addressing these challenges necessitates sophisticated forecasting models that can navigate the nuanced interplay of these variables under uncertain market conditions.

Consumer choice models are pivotal in understanding how potential buyers might respond to EVs compared to traditional vehicles. For instance, Struben (2008) integrated discrete consumer choice theory with diffusion models to estimate the adoption rates of PHEVs. This approach described the consumer's predisposition towards EVs as the anticipated utility of the vehicle, taking into account the influence of social dynamics and the availability of infrastructure, using frameworks like the multinomial logit choice model. Similarly, Bolduc et al. (2008) employed a hybrid choice model that melds discrete choice theory with psychometric variables, providing insights into Canadian consumers' preferences for HEVs over traditional vehicles and capturing the psychological traits that may influence their decisions.

Moreover, various traditional and innovative statistical techniques have been utilised to forecast EV demand. Time-series analysis, for instance, has been widely applied with models like ARIMA used by Amini et al. (2016) to predict EV charging demands based on daily usage patterns—helping not only in forecasting demand but also in optimising charging infrastructure to minimise operational costs. Zhang et al. (2017) extended this approach by employing both univariate and multivariate time-series models to predict short-term and long-term EV sales in China. In addition

to these methods, Hamed and Eideh (2018) explored the use of an Information Index and Poisson Regression model based on survey data to forecast EV demand in Jordan.

Further expanding the array of forecasting tools, researchers have combined stated and revealed preference data to refine their predictions. This technique has been evident in works by Ida et al. (2014), Brownstone et al. (2000), and Axsen et al. (2009), who have utilised multinomial logit and mixed logit models to explore consumer preferences for alternative fuel vehicles. Additionally, Ziegler (2012) applied multinomial probit models to examine potential car purchasing choices, including EVs, while Helveston et al. (2015) addressed consumer preferences in the U.S. and China using both multinomial logit and mixed logit models.

The integration of diffusion models further enhances our understanding of how EVs might penetrate various market segments over time. Struben & Sterman (2008) discussed how these models could incorporate the effects of policy changes, technological advancements, and social trends to forecast long-term adoption rates, considering the dynamic interactions between consumers, manufacturers, and regulatory environments, ensuring that policy and infrastructure development align with the projected growth in EV adoption.

2.3. Emerging Trends and Future Projections

EV market is at a critical juncture, with potential growth paths significantly influenced by technological advancements, policy evolution, and shifting consumer attitudes. Global movements towards sustainability and the transition away from fossil fuels are driving substantial policy shifts, as seen in countries like India and UK, where ambitious targets for phasing out ICE vehicles are set (Das & Bhat, 2022; Zhang et al., 2017).

Emerging trends also indicate a significant decrease in the costs of critical components such as lithium-ion batteries, spurred by advancements in technology and scaling production (Jin et al.,

2022). As battery costs decline, EVs become more financially accessible to a broader segment of consumers, potentially increasing market penetration. Furthermore, the integration of EVs into smart grids and renewable energy systems presents an opportunity to enhance the environmental value proposition of electric mobility (Guo et al., 2021). With increasing emphasis on sustainable urban planning and the development of smart cities, EVs are poised to play a pivotal role in future transportation ecosystems, especially in densely populated countries like India where urban air pollution is a significant concern.

2.4. Research Gap

Despite the growing importance of EVs in the global automotive market, there is a notable absence of empirical research focusing on the demand for EVs in India. Existing literature predominantly addresses EV adoption and demand forecasting in developed markets such as the United States, Europe, and China, where infrastructure and policy frameworks are significantly different from those in India. These studies often utilise time series analytical techniques to predict trends and assess factors influencing EV adoption, providing valuable insights into consumer behaviour and market dynamics.

However, the unique socio-economic landscape, regulatory environment, and infrastructural challenges in India necessitate a localised analysis to accurately forecast EV demand. Current research in India primarily explores the barriers to EV adoption, policy implications, and potential environmental benefits, but lacks robust empirical models that predict future demand.

Therefore, this research paper aims to fill this critical gap by employing Singular Spectrum Analysis to forecast EV demand in India. This research will not only contribute to the academic discourse but also offer valuable insights for policymakers and industry stakeholders striving to promote sustainable transportation solutions in India.

3. METHODOLOGY

The theoretical foundation of this paper revolves around forecasting the future sales of EVs in India, segmented into three categories: two-wheelers (2W), three-wheelers (3W), and four-wheelers (4W). The dynamic nature of the EV market, influenced by technological advancements, regulatory changes, and shifting consumer preferences, necessitates a robust methodological approach. Singular Spectrum Analysis (SSA) is employed as the primary tool for univariate time series analysis to forecast EV demand in India.

3.1. Rationale of using SSA

SSA aims to decompose a given time series into distinct components, such as trend, periodicity, quasi-periodicity, and noise. This process involves several sequential steps: transformation, decomposition, grouping, and reconstruction (Afshar and Bigdeli, 2011). The objective is to transform the original series into a format where these individual components can be identified and analysed separately (Hassani et al., 2013). The rationale for using SSA is grounded in several theoretical considerations. Firstly, SSA is a non-parametric method that does not require the specification of a predefined model structure, making it adaptable to the inherent non-linearity and non-stationarity of EV sales data. Secondly, SSA excels in isolating trend and periodic components from noise, which is crucial for accurate forecasting in markets characterised by technological and economic fluctuations. Additionally, SSA can be effectively applied in scenarios with incomplete data and is robust against outliers, which are common in emerging market datasets. When compared to traditional time series methods like ARIMA or exponential smoothing, SSA provides a more nuanced analysis of underlying data dynamics. Unlike these methods, which often require data to be stationary or involve extensive pre-processing, SSA directly handles the original series, preserving all vital information.

3.2 Mathematical Formulation

Mathematically, SSA follows the following steps-

Step 1: Transformation to Trajectory Matrix

The SSA begins with the transformation of the time series into a trajectory matrix, which serves as the basis for extracting significant data structures.

Formation of the Trajectory Matrix: Given a time series x_t where $t = 1,2,\ldots,N$, and selecting a window length M (where $M < N$), the trajectory matrix X is constructed as follows-

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_{N-M+1} \\ x_2 & x_3 & \cdots & x_{N-M+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & \cdots & x_N \end{bmatrix}$$

Each column of X represents a lagged version of the time series, embedding it into a M × (N − M + 1) dimensional space.

Step 2: Singular Value Decomposition (SVD)

Decomposition of the trajectory matrix through SVD reveals the principal components of the data, indicating underlying patterns.

Compute the SVD: Decompose X into its singular values and vectors:

$$X = U\Sigma V^T$$

where U and V are orthogonal matrices containing the left and right singular vectors, respectively, and Σ is a diagonal matrix of singular values σ_i.

Lag-Covariance Matrix: Optionally, one may also compute the covariance matrix C of the trajectory matrix to understand the variance structure:

$$C = \frac{1}{N-M+1} XX^T$$

Step 3: Grouping of Singular Components

This step involves grouping the singular components derived from SVD to separate the signal components from noise.

Identification of Significant Components: Components associated with larger singular values represent more significant data structures (e.g., trends and cycles), whereas those associated with smaller singular values typically represent noise:

$$X_{singal}=U_{singal}\Sigma_{singal}V^{T}{}_{singal}$$

where U_{signal}, Σ_{signal}, and V_{signal} are the parts of U, Σ, and V corresponding to significant singular values.

Step 4: Reconstruction and Forecasting

The final step involves reconstructing the time series from significant components and using this cleaner, more informative series for forecasting.

Reconstruct the Series: The original time series can be reconstructed using only the significant components:

$$\hat{x}_t = \sum_{i=1}^{r} U_{ti} \cdot \sigma_i \cdot V_{ti}^T$$

where r is the number of significant components retained for reconstruction. Forecasting: Extend the decomposition to forecast future values. This involves estimating future singular vectors or extending the time series by applying the patterns found in the decomposition.

3.3 Data

The dataset used in this research comprises a range of variables critical to understanding and forecasting the sales of EVs in India. The data spans a monthly timeframe, ensuring sufficient granularity for detailed time series analysis. In this research, Singular Spectrum Analysis (SSA) is

utilised as a univariate time series model, focusing on sales data of electric Vehicles in India, spanning a monthly timeframe to ensure detailed and granular analysis. The key variables include the monthly sales figures for electric two-wheelers (Y1), electric three-wheelers (Y2), and electric four-wheelers (Y3). All data is sourced from the Society of Manufacturers of Electric Vehicles (SMEV), providing a reliable basis for the analysis. This comprehensive dataset allows for an in-depth examination of sales trends and dynamics in the Indian EV market.

4. EMPIRICAL ANALYSIS

4.1. Data Transformation

Initially, the raw sales data were converted into time series format and adjusted via logarithmic transformation to mitigate issues related to non-linearity and heteroscedasticity.

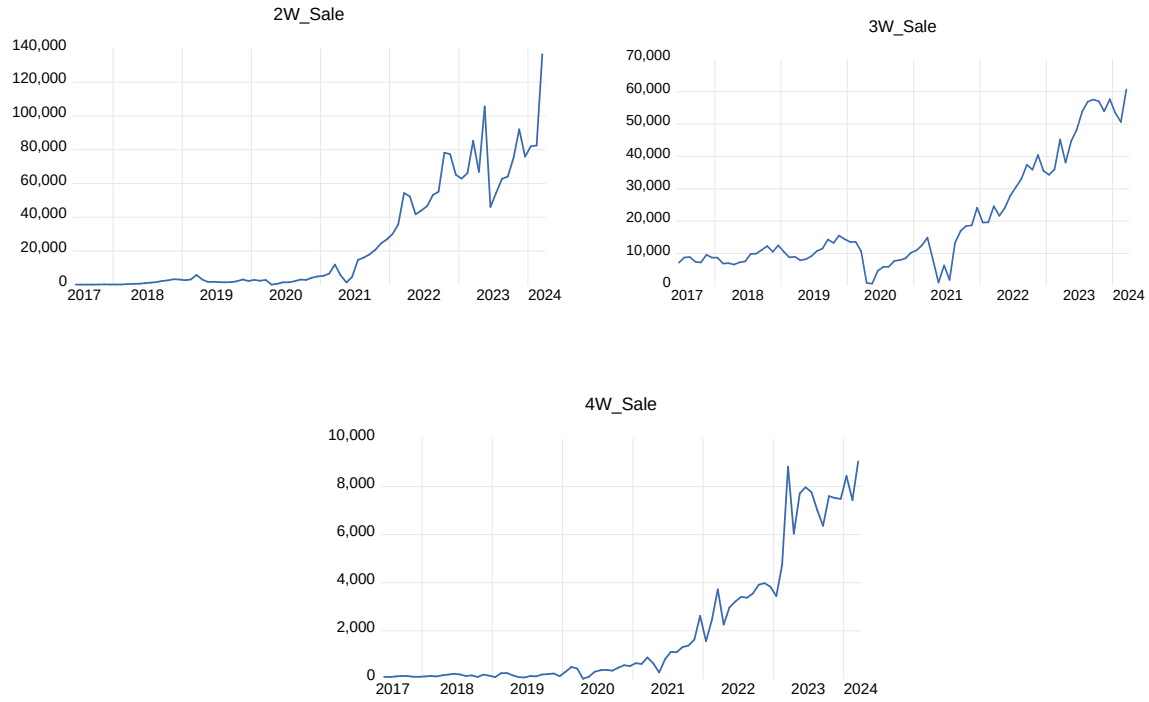


Figure 1: Raw Sales data of 2W, 3W and 4W

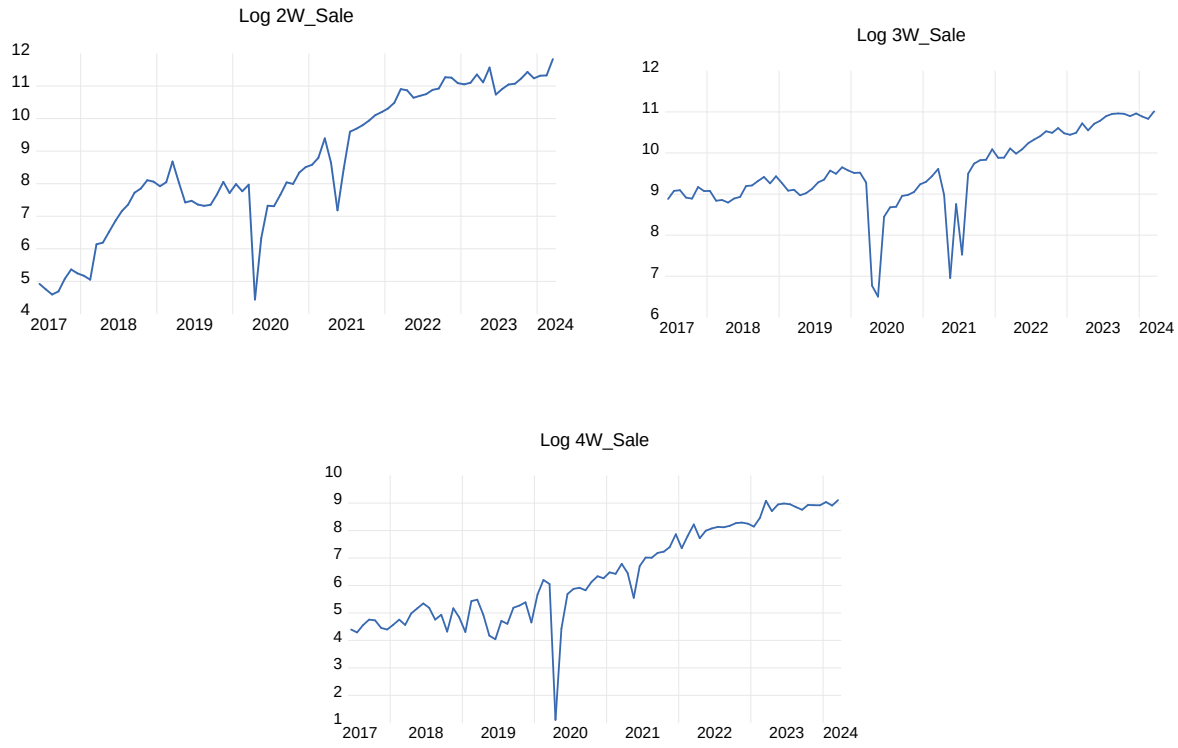


Figure 2: Logarithmic Transformed Sales data of 2W, 3W and 4W

4.2. SSA Decomposition

SSA was then applied to each transformed series with a window length of eight, aiming to decompose the series into various intrinsic components.

2-Wheelers EV Sales Analysis

For the 2W segment, the decomposition exhibits a steep decline in the eigenvalues from the very first component, suggesting a pronounced aggregation of variance within the initial components. This trend indicates that the predominant share of the sales dynamics can be effectively captured and explained by these leading components. Such a pattern is indicative of a strong, underlying trend within the data, with subsequent components likely reflecting lesser oscillations or seasonal effects, which are comparatively minor in their contribution to the overall sales dynamics.

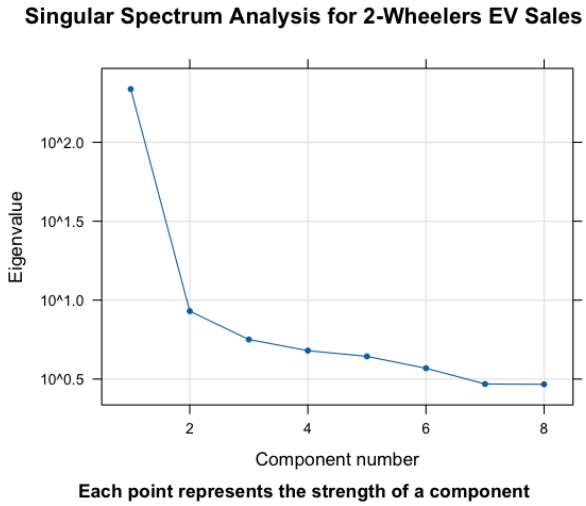


Figure 3 : Singular Spectrum Component : 2 Wheeler

3-Wheelers EV Sales Analysis

The eigenvalue plot for the 3W category mirrors the characteristics observed in the 2W segment, displaying a rapid decline in eigenvalue magnitude. This drop signifies that the major underlying trend and periodic components are predominantly captured by the initial few singular values. The swift decrease in eigenvalue size highlights the effectiveness of the leading components in summarising the principal sales trend, with minimal contribution from the noise or less significant oscillatory components.

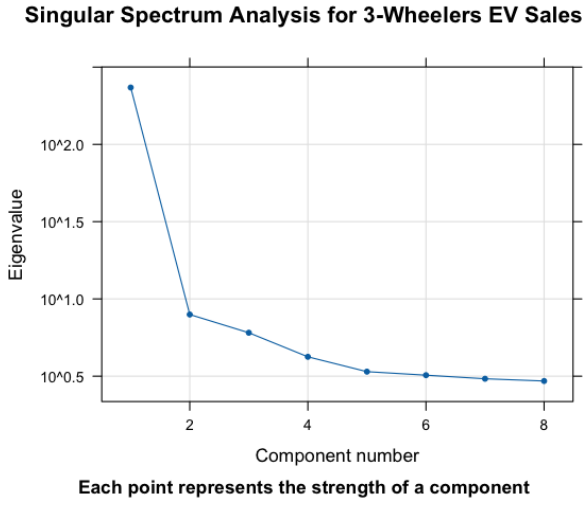


Figure 4 : Singular Spectrum Component : 3 Wheeler

4-Wheelers EV Sales Analysis

Similar to the previous categories, the 4W sales data analysis through SSA reveals a significant encapsulation of information within the first few components. The eigenvalue plot for this segment shows that these components capture the essential dynamics of the series, thus providing a clear depiction of the primary sales trends and cycles. The pattern observed in the eigenvalue decline is consistent with the 2W and 3W analyses, reinforcing the presence of strong, dominant sales trends that can be succinctly described by the initial singular values.

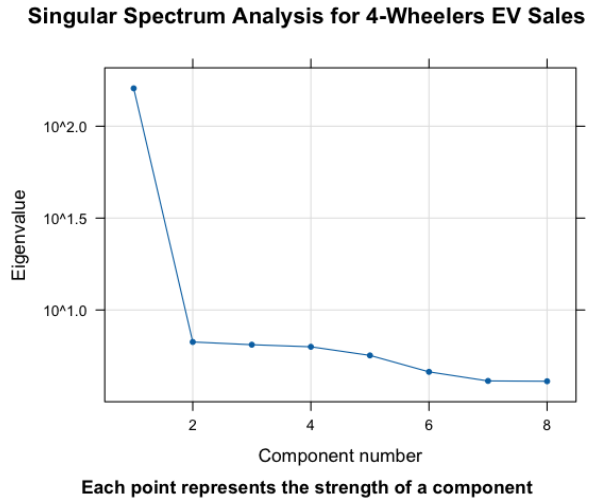


Figure 5 : Singular Spectrum Component : 4 Wheeler

4.3. Trend Extraction and Forecasting

Following the decomposition and analysis of singular values, the next crucial step in our empirical study involves extracting and examining the trend components from the time series of EV sales in India. This phase focuses on isolating the underlying trends and assessing the residuals to understand the systematic growth patterns and irregular fluctuations.

The trend and residual analysis for EV sales across different categories provides a comprehensive view of the market dynamics. For 2W, the analysis reveals a clear and steady increase in sales, depicted on a logarithmic scale to better illustrate exponential growth, with minor residuals

pointing to occasional market disturbances or seasonal variations. The 3W segment similarly shows significant growth, with a smooth trend line that underscores the effectiveness of SSA in capturing long-term market movements. Minimal residuals in this segment suggest deviations might be influenced by external factors or policy changes. The 4W segment displays a consistent and smooth upward trend, indicating stable growth, which is significant given the higher costs and investments associated with 4W compared to 2W and 3W.

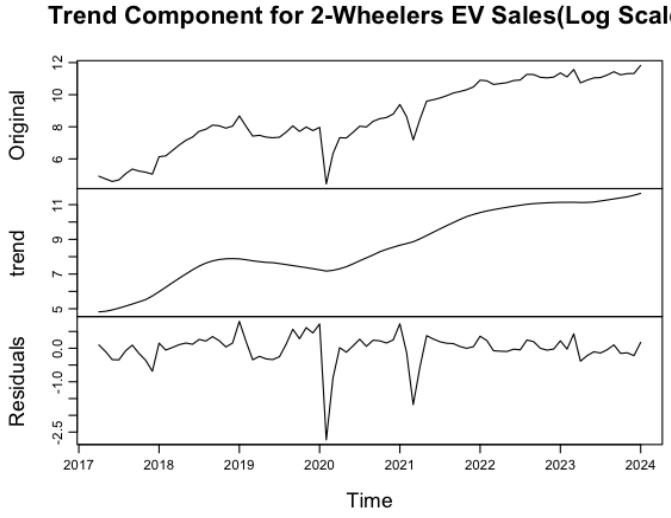


Figure 6 : Trend Extraction : 2 Wheeler

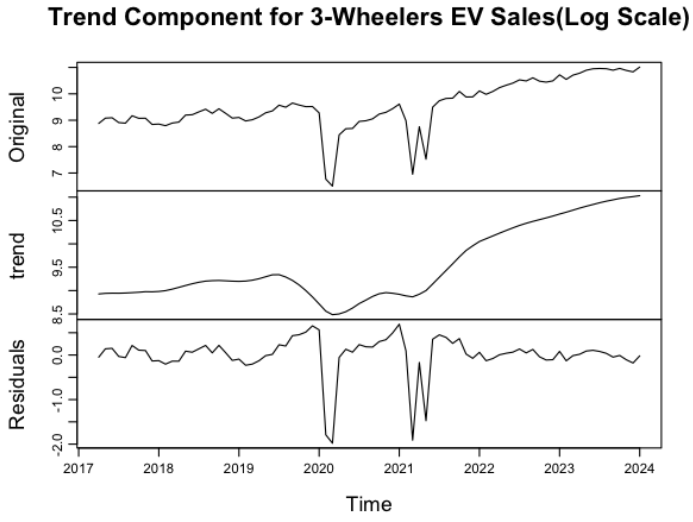


Figure 7 : Trend Extraction : 3 Wheeler

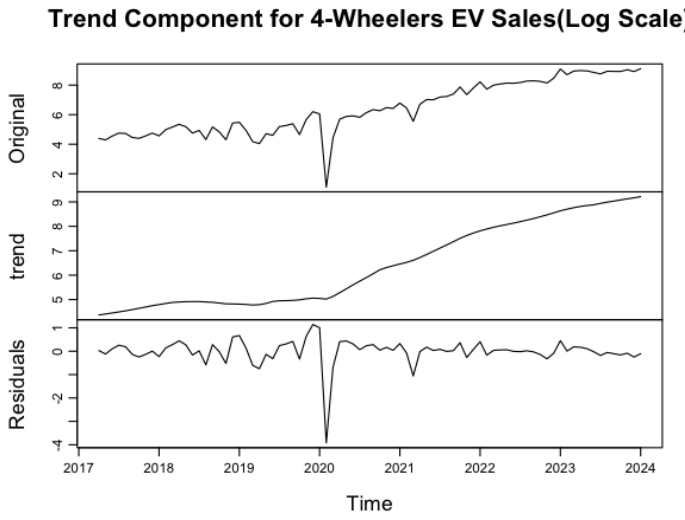


Figure 8 : Trend Extraction : 4 Wheeler

4.4. Forecasting Future Sales for Electric Vehicles

The following graphs illustrate the short-run forecast of 2W, 3W, and 4W EV sales, respectively, using SSA. Initially presented in a logarithmic scale, these forecasts highlight the relative changes and expected growth rates across each vehicle segment, showcasing a consistent upward trajectory. Subsequently, the forecasts were transformed from the logarithmic scale back to absolute sales figures through exponentiation. This step renders the projections in more tangible terms, offering direct insights into the expected sales volumes. The absolute number forecasts reveal substantial growth in the 2W segment from 1.5 Lakh to more than 3 Lakh by the end of 2024, which can be attributed to the increasing popularity of electric 2W as affordable and efficient urban transport solutions. Similarly, the 3W and 4W segments as shown in figure 8 and figure 9, demonstrating a significant growth potential, with the forecasts indicating robust increase in sales, driven by the broader acceptance of EVs in commercial transport and personal use due to technological advancements and growing environmental consciousness.

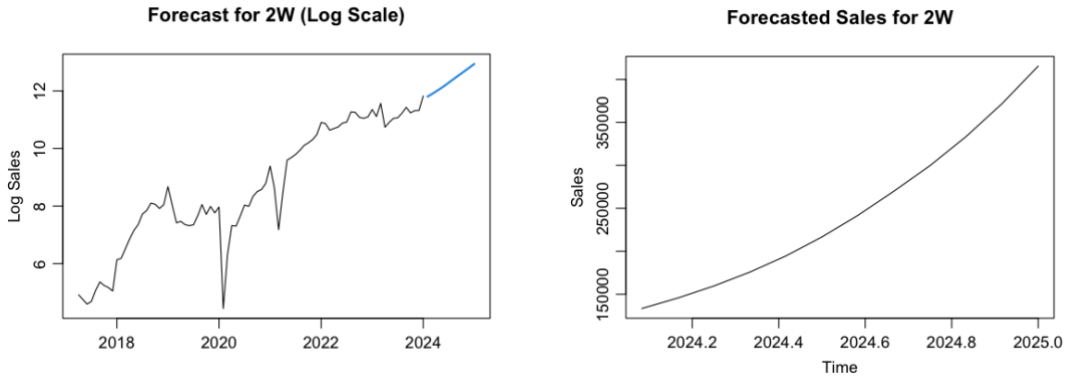


Figure 7 : Forecasted Sales using SSA : 2 Wheeler

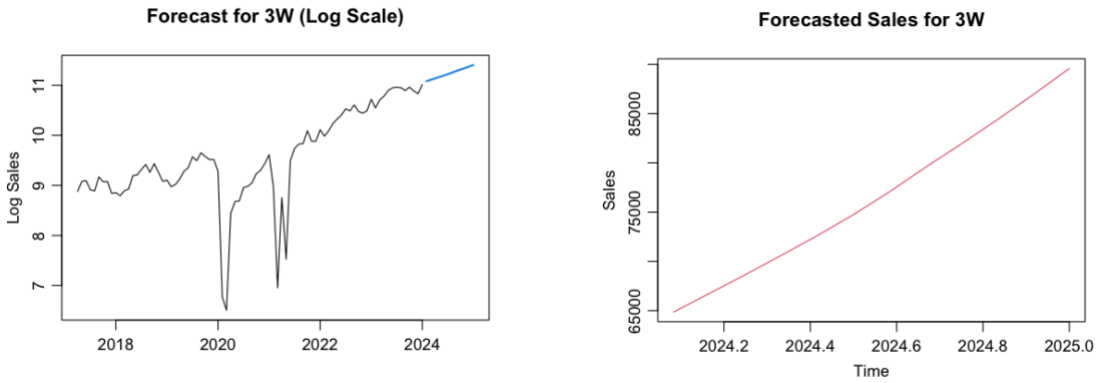


Figure 8 : Forecasted Sales using SSA : 3 Wheeler

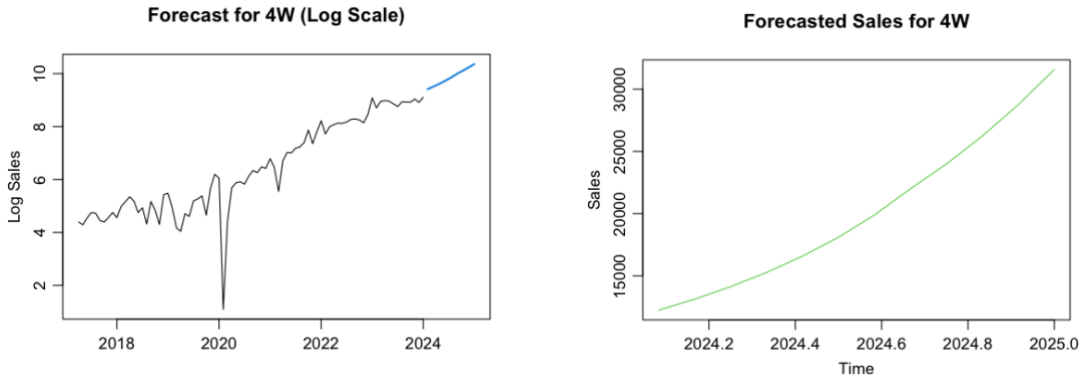


Figure 9 : Forecasted Sales using SSA : 4 Wheeler

5. FINDINGS

The SSA model provides forecasts for EV sales across three categories: 2W, 3W, and 4W. The forecasts span from March 2024 to December 2024.

The forecasted sales for 2W show a significant increase from 145,623 units in March 2024 to 321,915 units in December 2024, which is more than a two-fold increase. This substantial growth highlights the rapidly rising demand for electric two-wheelers. Similarly, 3W sales are predicted to rise from 66,759 units in March 2024 to 86,934 units in December 2024, indicating a 30% increase over the period. For 4W, the sales are expected to grow from 13,128 units in March 2024 to 28,755 units in December 2024, representing a significant increase of over two times.

Month-Year	**F_2W_SSA**	**F_3W_SSA**	**F_4W_SSA**
Mar-24	145623.1	66759.48	13128.74
Apr-24	159623.5	68669.90	14128.46
May-24	175781.9	70630.60	15272.85
Jun-24	194597.4	72628.46	16581.77
Jul-24	216680.4	74748.61	18101.05
Aug-24	241865.0	77059.44	19890.91
Sep-24	270039.3	79548.21	21991.82
Oct-24	289748.2	81927.75	23991.41
Nov-24	300525.8	84386.05	26235.24
Dec-24	3,21,915.3	86,934.94	28,755.44

Table 1 : Forecast values of 2W, 3W and 4W

6. Policy Suggestions and Conclusion

6.1. Policy Support

Given the current scenario and the projections for the electric vehicle market, it is clear that significant growth is anticipated across all categories of EVs. According to the Singular Spectrum

Analysis, there is a predicted 2.2-fold increase in two-wheeler sales, a 30% rise in three-wheeler sales, and a 2.2-fold increase in four-wheeler sales by December 2024. These forecasts consider the existing conditions without factoring in government interventions.

If the government enhances its focus on charging infrastructure and reducing battery costs, the potential for EV adoption could be much higher. This model serves as a baseline scenario under current conditions, suggesting that actual outcomes could exceed these expectations. For instance, a 1% increase in EV sales in a city can reduce local CO_2 emissions by 0.096% (Xiaolei Zhao et al., 2023). Additionally, achieving a 30% EV sales share in India could save approximately INR 1.1 lakh crore in crude oil imports (Abhinav Soman et al., 2020).

To capitalise on these opportunities, several policy measures are recommended. First, it is essential to enhance battery and motor technologies by focusing on improving the size, lifespan, and efficiency of batteries and electric motors while reducing material costs. Second, expanding the charging infrastructure is crucial to establish a widespread and reliable network of charging stations to assure potential buyers of constant charging availability. Third, increasing public awareness of the benefits and sustainability of EVs will drive consumer demand and foster greater adoption. Lastly, promoting diverse EV options by ensuring high-quality and varied choices across all vehicle types will cater to different market segments and enhance overall market penetration**.**

6.2. Conclusion and Way Forward

The anticipated growth in the EV market, as projected by SSA model, highlights the significant potential for the electric mobility sector in India. These models, while providing a conservative estimate, underscore the transformative impact that increased EV adoption could have on the environment and the economy.

To fully realise this potential, it is crucial to focus on several strategic areas. First, promoting new economic activities such as battery recycling (urban mining) and other services related to electric mobility can create job opportunities and stimulate economic growth. Second, implementing training and skilling initiatives to develop a workforce capable of meeting the demands of EV manufacturing and maintenance is essential. This will ensure that the industry has the necessary human resources to support its expansion. Third, developing pre-emptive fiscal strategies to mitigate the expected decline in petroleum tax revenues as EV adoption increases is vital. This

will ensure stable central and state government revenue collections while promoting electric mobility.

By adopting these measures, India can achieve significant environmental benefits and drive economic growth and job creation. The country has the potential to position itself as a global leader in the transition to sustainable mobility, provided that these strategies are implemented effectively and timely.

6.3. Future Scope of the Study

The future scope of this study is multifaceted, encompassing various dimensions of the EV market and its broader implications. Firstly, future research could integrate the impact of government policies and initiatives, such as subsidies, tax incentives, and infrastructural developments, to provide a more comprehensive analysis of their effects on EV adoption rates. Secondly, exploring advancements in battery technology, including improvements in energy density, charging speed, and cost reduction, will be crucial in understanding their potential to accelerate market growth. Additionally, the integration of smart grid technologies and renewable energy sources with EV charging infrastructure could be examined for their roles in enhancing system efficiency and sustainability. Consumer behaviour analysis is another vital area, where studies could delve into factors influencing purchasing decisions, such as environmental awareness, economic considerations, and technological preferences. Furthermore, detailed environmental impact assessments could quantify the benefits of EV adoption in terms of CO_2 emissions reduction and air quality improvements. Lastly, the economic impacts of EV transition, including job creation, effects on GDP, and shifts in the energy and automotive industries, warrant comprehensive investigation to inform policymakers and stakeholders about the broader economic benefits and challenges of fostering an EV ecosystem.

PLAGIARISM REPORT

turnitin

Similarity Report ID: oid:16608:61844145

PAPER NAME

NCGG_Research_Paper_Yash_Aggarwal_240621_150146.pdf

WORD COUNT	CHARACTER COUNT
8360 Words	**48745 Characters**
PAGE COUNT	FILE SIZE
40 Pages	**817.7KB**
SUBMISSION DATE	REPORT DATE
Jun 22, 2024 1:32 AM GMT+5:30	**Jun 22, 2024 1:32 AM GMT+5:30**

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database

Excluded from Similarity Report

- Submitted Works database
- Bibliographic material
- Cited material
- Small Matches (Less then 10 words)

REFERENCES

1. ABI Research. (2024, April 11). Global EV market analysis 2024: Trends and forecasts. Retrieved from https://www.abiresearch.com/blogs/2024/04/11/global-ev-market-analysis-2024-trends-and-forecasts/
2. Amini, M. H., Kargarian, A., & Karabasoglu, O. (2016). ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation. Electric Power Systems Research, 140, 378–390. https://www.sciencedirect.com/science/article/pii/S0378779616302115
3. Bockarjova, M., & Steg, L. (2014). Can Protection Motivation Theory predict pro-environmental behavior? Explaining the adoption of electric vehicles in the Netherlands.
4. Canalys. (2023). Global EV sales H1 2023. Retrieved from https://canalys.com/newsroom/global-ev-sales-h1-2023
5. Council on Energy, Environment and Water (CEEW). (n.d.). India's electric vehicle transition: Electric mobility support sustainable economic recovery. Retrieved from https://www.ceew.in/publications/indias-electric-vehicle-transition-electric-mobility-support-sustainable-economic-recovery
6. Das, P. K., & Bhat, M. Y. (2022). Global electric vehicle adoption: Implementation and policy implications for India. Environmental Science and Pollution Research, 29(27), 40612–40622.
7. Das, P. K., & Bhat, M. Y. (2022). Global electric vehicle adoption: Implementation and policy implications for India. Environmental Science and Pollution Research, 29(27), 40612–40622. https://doi.org/10.1007/s12626-022-00078-4
8. Dixit, S. K., & Singh, A. K. (2022). Predicting electric vehicle (EV) buyers in India: A machine learning approach. The Review of Socionetwork Strategies, 1–18. https://doi.org/10.1007/s12626-022-00078-4
9. Hamed, M., & Al-Eideh, B. (2018). Modelling the demand for electric vehicles. European Journal of Social Sciences, 56.
10. Helveston, J. P., Liu, Y., Feit, E. M., Fuchs, E., Klampfl, E., & Michalek, J. J. (2015). Will subsidies drive electric vehicle adoption? Measuring consumer preferences in the U.S. and China. Transportation Research Part A: Policy and Practice, 73, 96–112.

11. International Energy Agency. (2024). Global EV outlook 2024. Retrieved from https://iea.blob.core.windows.net/assets/a9e3544b-0b12-4e15-b407-65f5c8ce1b5f/GlobalEVOutlook2024.pdf
12. International Energy Agency. (n.d.). India: Oil. Retrieved from https://www.iea.org/countries/india/oil
13. Jin, S., Mu, D., Lu, Z., Li, R., Liu, Z., Wang, Y., Tian, S., & Dai, C. (2022). A comprehensive review on recycling of spent lithium-ion batteries: Urgent status and technology advances. Journal of Cleaner Production, 130535.
14. Jones, B., Elliott, R. J., & Nguyen-Tien, V. (2020). The EV revolution: The road ahead for critical raw materials demand. Applied Energy, 280, 115072. https://doi.org/10.1016/j.apenergy.2020.115072
15. MetalMiner. (2023). Car Demand Up, Sideways Steel Prices Helping Industry. https://agmetalminer.com/2023/08/08/automotive-mmi-consumer-demand-steel-prices/
16. NITI Aayog. (n.d.). State summary for electric mobility. Retrieved from https://e-amrit.niti.gov.in/assets/admin/dist/img/new-fronend-img/report-pdf/State-Summary-Final.pdf
17. Pawar, M. S. R., & Ahire, M. R. R. (2020). Potential of Electrical Vehicle (EV) in Indian market and a review of public policy outlook for 2020. Solid State Technology, 63(6), 2076–2087.
18. Petroleum Planning & Analysis Cell (PPAC). (n.d.). Import & export. Retrieved from https://ppac.gov.in/import-export
19. Press Information Bureau. (n.d.). PM launches Electric Vehicle Policy. Retrieved from https://pib.gov.in/PressReleasePage.aspx?PRID=1795071
20. PV Magazine India. (2024, April 3). Electric vehicle sales in India surged 42% YoY in FY2024. Retrieved from https://www.pv-magazine-india.com/2024/04/03/electric-vehicle-sales-in-india-surged-42-yoy-in-fy2024/
21. Reserve Bank of India (RBI). (n.d.). Business outlook. Retrieved from https://cimsdbie.rbi.org.in/BOE/OpenDocument/2311211338/OpenDocument/opendoc/openDocument.jsp?logonSuccessful=true&shareId=4
22. Sang, Y. N., & Bekhet, H. A. (2017). Exploring factors influencing electric vehicle usage intention: An empirical study in Malaysia.

23. ScienceDirect. (n.d.). Retrieved from https://www.sciencedirect.com/science/article/pii/S2405844023075047
24. Singh, P., Kaur, I., Kumar, A., & Sahay, A. (n.d.). Accelerating the E-Mobility Revolution: A mixed-method enquiry into EV adoption behaviour. Available at SSRN 4119543.
25. Society of Manufacturers of Electric Vehicles (SMEV). (n.d.). Statistics. Retrieved from https://www.smev.in/statistics
26. Struben, J., & Sterman, J. D. (2008). Transition challenges for alternative fuel vehicle and transportation systems. Environment and Planning B: Planning and Design, 35(6), 1070–1097.
27. Tang, H. (2021). How Does the Fluctuation of Steel Prices Impact Your Project Budgets? LinkedIn. https://www.linkedin.com/pulse/how-does-fluctuation-steel-prices-impact-your-project-heidi-tang-qd9xc
28. The Roundup. (n.d.). CO2 greenhouse gas emission statistics. Retrieved from https://theroundup.org/co2-greenhouse-gas-emission-statistics/
29. Times of India. (2023, January 1). India beats Japan to become world's third largest vehicle market. Retrieved from https://timesofindia.indiatimes.com/auto/news/india-beats-japan-to-become-worlds-third-largest-vehicle-market/articleshow/96874402.cms#google_vignette
30. Toppr-guides. (2018). Demand forecasting. Retrieved from https://www.toppr.com/guides/business-economics/theory-of-demand/demand-forecasting/
31. United Nations. (n.d.). The second United Nations Global Sustainable Transport Conference. Retrieved from https://www.un.org/en/conferences/transport2021/about#:~:text=The%20second%20United%20Nations%20Global,Paris%20Agreement%20on%20climate%20change
32. Verma, M., Verma, A., & Khan, M. (2020). Factors influencing the adoption of electric vehicles in Bengaluru. Transportation Developing Economies, 6(2), 17. https://doi.org/10.1007/s40890-020-0100-x
33. Wang, J., Li, Y., & Zhang, Y. (2022). Research on carbon emissions of road traffic in Chengdu city based on a LEAP model. Sustainability, 14(9), 5625.

34. World Resources Institute. (n.d.). Climate Watch - CAIT country greenhouse gas emissions data. Retrieved from https://www.wri.org/data/climate-watch-cait-country-greenhouse-gas-emissions-data
35. Zhang, S. (2016). Studying the factors affecting sales of new energy vehicles from supply side. Proceedings of 2016 2nd International Conference on Humanities and Social Science Research (ICHSSR 2016). https://doi.org/10.2991/ichssr-16.2016.71
36. Zhang, Y., Zhong, M., Geng, N., & Jiang, Y. (2017). Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China. PLOS ONE, 12(5), e0176729. https://doi.org/10.1371/journal.pone.0176729

Optical and Quantum Electronics (2024) 56:1643
https://doi.org/10.1007/s11082-024-07499-w

Ge–As–Se–Te chalcogenide based rib-waveguide for highly coherent on-chip mid-infrared supercontinuum generation: design and analysis

Deepak Garg[1] · **Akash Khamaru**[1] · **Ajeet Kumar**[1]

Received: 11 April 2024 / Accepted: 9 September 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
We report numerical analysis and design of a rib waveguide composed of a core of Ge–As–Se–Te chalcogenide glass with upper and lower cladding of ZnSe for on-chip mid-infrared supercontinuum generation. By implementing the Finite Element Method (FEM), we analyzed the impact of geometrical parameters of waveguide on optimizing them in order to minimize the dispersion effects and to obtain a flat dispersion curve. Our design possesses Zero Dispersion Wavelength at 5.76 µm and 10.50 µm. The dispersion value at pump wavelength of 9.8 µm is 1.353 ps/nm/km. Our simulation reports a supercontinuum broadening from 3 to 25 µm, when pumped at 9.8 µm wavelength with pulse width 50 fs and peak power 15 kW in only 10 mm long waveguide, making it a promising candidate for mid-infrared supercontinuum light sources for many non-linear applications.

Keywords Rib-waveguide · Supercontinuum generation · Coherent source · On-chip

1 Introduction

In the recent years, optical waveguides based supercontinuum generation sources has gained wide popularity for various applications in the field of spectroscopy (Ke et al. 2009), optical coherence tomography (Hartl et al. 2001), sensing (Bock et al. 2007), pulse compression (Husakou et al. 2001) etc. due to enhanced bandwidth and small size. On-chip supercontinuum generation sources have an advantage over photonic crystal fiber (PCF) sources as they are able to provide more robustness, better power efficiency and are more cost efficient. The electromagnetic spectrum window ranging from 1 to 25 µm is of utmost importance because it contributes to near and mid-infrared region which opens the door for the applications like mid-infrared molecular fingerprint spectroscopy which has the potential to provide better understanding of molecular structures of atoms and molecules (Saini et al. 2015).

✉ Ajeet Kumar
ajeetdph@dtu.ac.in

[1] Advanced Photonics Simulation Research Laboratory, Department of Applied Physics, Delhi Technological University, Bawana Road, Shahabad Daulatpur, New Delhi, Delhi 110042, India

Supercontinuum generation (SCG) is the phenomenon of spectral broadening of short and high intensity pulses when they are allowed to propagate through a non-linear medium. The process of supercontinuum generation is completed when a number of non-linear effects like self-phase modulation, cross-phase modulation, four wave mixing, stimulated Raman scattering, soliton dynamics and stimulated Brillouin scattering (Agrawal 2000; Boyd et al. 2008; Dudley et al. 2006) etc. comes into play and acts on the initial pulse. In previous years a lot of research have been carried out on supercontinuum sources based on waveguides and photonic crystal fibers. Silica has been widely used as background material for SCG but due to the multiphonon transmission edge phenomena and high absorption coefficient, the spectral broadening is restricted in mid-infrared region and is limited to 2.5 µm only (Chauhan et al. 2018). Other materials like ZBLAN (Jiang et al. 2016), tellurite (Klimczak et al. 2019), fluoride (Gauthier et al. 2018), and chalcogenide glasses offers high non-linearity, low losses and wider transparency windows in near and mid-infrared region making them a good candidate for the exploration in the field of supercontinuum sources.

Among all the materials, chalcogenide materials are best suited for the supercontinuum generation as they offer sufficient high kerr non-linearity, low material absorption, lower response time and some compositions of chalcogenide glasses offers transparency window upto 25 µm which covers both near and mid-infrared region (Shiryaev and Churbanov et al. 2013). The non-linear refractive indices of some compositions of chalcogenide glasses are tens or hundreds time greater than that of tellurite and fluoride glasses. Fabrication of chalcogenide based rib waveguides can be done through existing techniques with high control such as chemical vapour deposition, ion etching techniques, sputtering, thermal evaporation and photolithography (Cognolato 1995; Tan et al. 2010; Righini and Chiappini 2014; Han et al. 2010; Gai et al. 2010).

There are many research works that have been demonstrated both experimentally and theoretically for SCG in waveguides as well as in PCF geometries. A channel waveguide using air cladding and $Ge_{11.5}As_{24}Se_{64.5}$ as core material has been numerically reported for mid-IR SC covering 2–6 µm when pumped at 3.1 µm (Karim et al. 2015). A graphene surface plasmon waveguide has been reported which is capable of producing supercontinuum broadening from 6 to 13 µm with 0.5mW peak power in 150 nm long waveguide (Gorbach et al. 2015). Mid-Infrared spanning of 1.2–7.2 µm has been theoretically generated in 2.5 mm long asymmetrical waveguide consisting of As_2Se_3 core and MgF_2 as top cladding, when pumped at 2.8 µm with peak power 2.5 kW (Saini et al. 2018). Another asymmetrical waveguide is reported with As_2Se_3 as core material capable of generating an SC span ranging from 2 to 15 µm at 2.5 µm pump wavelength in 4 mm length of waveguide (Saini et al. 2017). In addition, a buried chalcogenide waveguide structure has been experimentally reported, generating an SC span from 1.2 to 5.2 µm with a very high peak power of 1.76 MW when pumped at 4.26 µm central wavelength in 9 mm long waveguide (McCarthy et al. 2013). Further, a $Ge_{11.5}As_{24}Sb_{64.5}$ based rib waveguide has been experimentally reported with spectrum generation from 1.8 to 7.5 µm at 4 µm pump wavelength and 20mW average power (Yu et al. 2014). Due to increase in absorption losses in As_2S_3 fibers and waveguides above 5 µm and As_2Se_3 being a toxic and harmful material, the utilization of these materials for broadband non-linear devices has been limited.

Ge–As–Se–Te (GAST) possesses high kerr non-linearity, low losses, high linear refractive index and wider transparency range upto 25 µm which covers atmospheric transparent window, molecular fingerprint region as well as far-infrared band. The numerical and experimental analysis of GAST has been done previously in fibers, PCFs and waveguide structures. Experimental demonstration of mid-infrared SCG in a step

index fiber using Ge-As-Se-Te as the core material has been reported, producing SC broadening from 1.5 to 14 μm with 4.5 μm pump wavelength in 23 cm long fiber (Zhao et al. 2016). Further, Experimental and theoretical demonstration of SCG in double-clad structure has been reported with SC broadening from 1.5 to 11 μm when pumped with 5 μm wavelength with high peak power of 100 kW (Xiao et al. 2020). Furthermore, a theoretical PCF model was also reported capable of covering the SC span from 3 to 25 μm with a pump wavelength of 6 μm and very high peak power of 80 kW (Gao et al. 2022). Also, mid-infrared spanning from 2 to 13 μm using 5.8 μm pump has also been demonstrated in waveguide structure with the GAST as core material (Xia et al. 2021).

In this study, we report theoretical analysis of waveguide structure by dispersion engineering on the waveguide parameters to optimize its dispersion profile. We analyzed the influence of geometrical parameters of waveguide to optimize them in order to minimize the dispersion effects and finally proposed a design and compared it with the previous waveguides designs. The simulated results shows that the proposed waveguide is able to produce a spectrum broadening from 3 to 25 μm when pumped at 9.8 μm laser source (Cheng et al. 2016) with 50 fs pulses and 15 kW peak power in only 10 mm long waveguide.

2 Proposed waveguide design

We have designed and theoretically investigated a Ge–As–Se–Te (GAST) chalcogenide glass based rib waveguide structure for broadband supercontinuum generation in mid-infrared region. The core of the proposed structure is made up of Ge–As–Se–Te with upper and lower cladding made up of ZnSe. The upper layer and lower layer are chosen to be made up of ZnSe because it possesses wide transmission window with lower refractive index than GAST glass which helps in better confinement of field in the core region. The width of the core region is denoted as a, core rib height is denoted as b and core maximum height as h. In this work, we have altered these parameters for dispersion engineering to obtain low dispersion values and a flat dispersion curve. After studying the influence of geometrical parameters, a waveguide design is finalized for the further analysis. The cross sectional view of the proposed rib-waveguide structure is shown in Fig. 1a and the confinement of fundamental mode in core region at 9.8 μm is shown Fig. 1b.

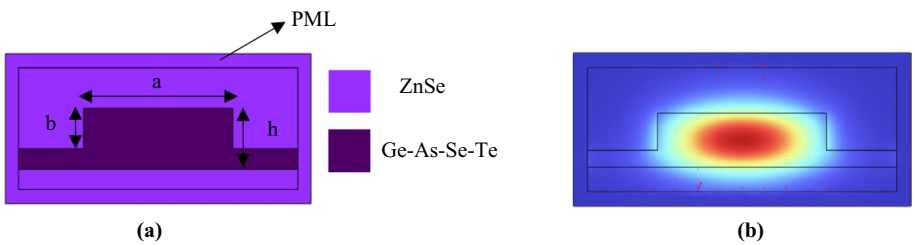


Fig. 1 **a** Cross-sectional representation of designed waveguide **b** modal profile fundamental mode at 9.8 μm

3 Linear and non-linear characteristics of SC generation

Since the linear refractive index n depends on the wavelength of light that is propagating through the material, its refractive index can be calculated with the help of an empirical equation known as Sellmeier equation, whose general form can be written as (Sharma et al. 2020):

$$n^2(\lambda) = A_0 + \sum_{n=1}^{3} \frac{A_n \lambda^2}{\lambda^2 - B_n^2} \tag{1}$$

where A_0, A_n and B_n are the Sellmeier fitting constants. The value of these constants for GAST and $ZnSe_2$ are given in Table 1.

For modal analysis of the proposed waveguide structure we employed Finite Element Method (FEM) based software "COMSOL MULTIPHYISICS" which gives us the information about the distribution of electric field and values of effective refractive index (n_{eff}) of different propagating modes in the waveguide.

Chromatic Dispersion and material non-linearity are the two important factors that determines the quality and efficiency of the supercontinuum broadening. The non-linearity can be calculated according to the following formula (Sharma et al. 2020):

$$\gamma = \frac{2\pi n_2}{\lambda A_{eff}(\lambda)} \tag{2}$$

where n_2 represents the non-linear coefficient of the material which in our case is 6.1757×10^{-18} $m^2.W^{-1}$ (Gao et al. 2022) and A_{eff} represents effective mode area which can be calculated using the following expression (Saini et al. 2018):

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

where E is the transverse component of propagating electric field.

The chromatic dispersion determines the extent to which the spectral components of the ultra-short input pulse spread at different pulse velocity while propagating in the non-linear

Table 1 Sellmeier fitting coefficients for the materials used

Sellmeier coefficient	Ge–As–Se–Te (Core)	ZnSe (Upper and lower cladding)
A_0	1	1
A_1	9.3179	4.9257
A_2	$4.7797e^{-4}$	$3.2477e^{-5}$
A_3	7.9541	22.2063
$\vec{B}_2^2$	0.29015	0.0513
$\vec{B}_1^2$	270.06	19.315
$\vec{B}_3^2$	8422.70	16,574

medium. Thus, it plays significant role in the SC generation. Chromatic dispersion depends on the variation of effective refractive index(n_{eff}) of material with the wavelength of incident light and can be computed with the help of following expression (Saini et al. 2018):

$$D(\lambda) = \frac{-\lambda}{c}\frac{d^2\text{Re}(n_{\text{eff}})}{d\lambda^2} \tag{4}$$

For SC generation a generalized non-linear Schrodinger equation which is given by the following expression is solved (Chauhan et al. 2018):

$$\frac{\partial A}{\partial Z} + \frac{\alpha}{2}A - \left(\sum_{n\geq 2}\frac{i^{n+1}}{n!}\beta_n\frac{\partial^n A}{\partial T^n}\right) = i\gamma\left(1 + \frac{i}{\omega_0}\frac{\partial}{\partial T}\right)\left(A(z,t)\int_{-\infty}^{\infty}R(T)\left|A\left(z,T-T'\right)\right|^2 dT + i\hbar_R(z,T)\right) \tag{5}$$

whereβ_n denotes the n^{th} order of dispersion parameters and α is the fiber transmission loss coefficient. Spatial co-ordinate along the fiber length is termed as z and t is the time variable. We are working under the assumption that $\hbar_R$=0. The non-linear response $R(t)$ which depends on instantaneous electronic and delayed Raman contributions can be expressed as (Sharma et al. 2020):

$$R(t) = \left(1 - f_R\right)\delta(t) + f_R h_R(t) \tag{6}$$

where the value of f_R for GAST is taken as $f_R = 0.115$ (Gao et al. 2022). The Raman response function $h_R(t)$ is stated and calculated as:

$$h_R(t) = \frac{\tau_1^2 + \tau_2^2}{\tau_1\tau_2^2}\exp\left[\frac{-t}{\tau_2}\right]\sin\left(\frac{t}{\tau_1}\right) \quad) \tag{7}$$

τ_1 is related to photon non-oscillation frequency and τ_2 represents damping time of vibrating atoms. We have chosen τ_1=23.1 fs, τ_2=195 fs (Gao et al. 2022). For the generation of pump pulses, a tunable hyperbolic secant pulse is used which can be expressed as (Chauhan et al. 2018):

$$A(z=0,t) = \sqrt{P_0}\text{sech}\left(\frac{t}{P_0}\right) \tag{8}$$

where P_0 denotes the peak power.

We have employed one photon per mode noise model which can help in investigating the coherent property of mid-infrared SCG. In this study, modulus of complex degree of first order coherence g_{12} is used to calculate the degree of coherence which varies from 0 to 1 (Khamis and Abbasi 2024).

$$\left|g_{12}^{(1)}(\lambda, t_1 - t_2)\right| = \left|\frac{< E_1^*(\lambda, t_1)E_2(\lambda, t_1) >}{\sqrt{< \left|E_1(\lambda, t_1)\right|^2 >< \left|E_2(\lambda, t_2)\right|^2 >}}\right| \tag{9}$$

4 Results and discussion

In the proposed structure, the dispersion characteristic is controlled by altering the geometrical parameters like core width (a), core-rib height (b) and core maximum height (h) in order to optimize these parameters to obtain low dispersion values at pump wavelength along with flat dispersion curve. The effect in dispersion characteristics by altering core width (a) and by keeping other parameters b and h to be constant is shown in Fig. 2. It can be realized from the figure that with the increase in the core width (a) there is no significant shift in the zero dispersion wavelengths i.e. the dispersion curves almost overlaps, but the dispersion values in anomalous region is slightly lesser for $a = 8.2$ µm. Thus, the optimized value chosen for core width (a) is 8.2 µm. Similarly, the effect of core rib height (b) is also considered and is shown in Fig. 3. It can be observed that with the increase in the core rib height the second ZDW shifts towards the right side i.e. shifts towards the longer wavelength side. There is no significant effect on the shift of first ZDW but the values of dispersion in the anomalous dispersion regime increases with the increase in the parameter. As b increases from 1.8 µm to 2.2 µm the values of dispersion changes from 1.8941 to 16.528 ps/nm/km at a pump wavelength of 9.8 µm. Thus, the chosen optimized value for core rib height (b) is 1.8 µm. Further, the effect of core maximum height on the dispersion characteristic is also studied which is shown in Fig. 4. It can be seen that with the increase in core maximum height there is a prominent shift of the second ZDW towards the right hand side or longer wavelength side. Also, as h increases the value of dispersion in the anomalous region also increases. Thus, the chosen optimized value for the core maximum height is 2.8 µm.

The optimized parameters of the finally proposed waveguide are tabulated in Table 2. From the above analysis it can be concluded that for the chosen optimized parameters $a = 8.2$ µm, $b = 1.8$ µm and $h = 2.8$ µm, the value of effective mode area and dispersion value is minimum at the pump wavelength of 9.8 µm. The dispersion characteristics offered by the waveguide structure for the optimized parameters is shown in Fig. 5. It

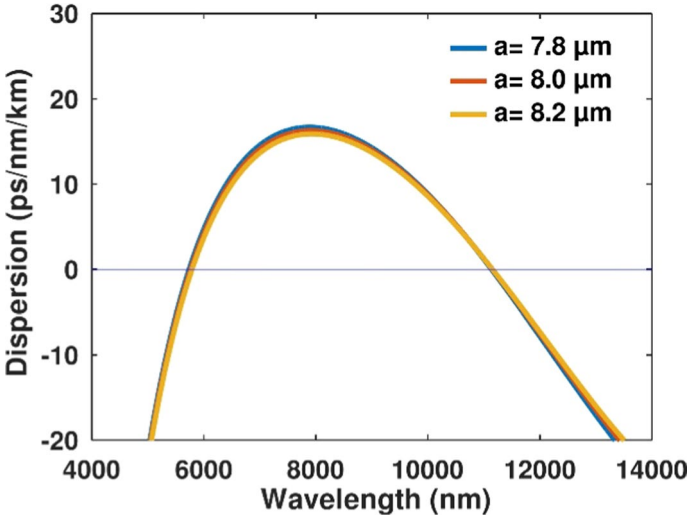


Fig. 2 Dispersion characteristics with variation in core width (a) by keeping other parameters constant

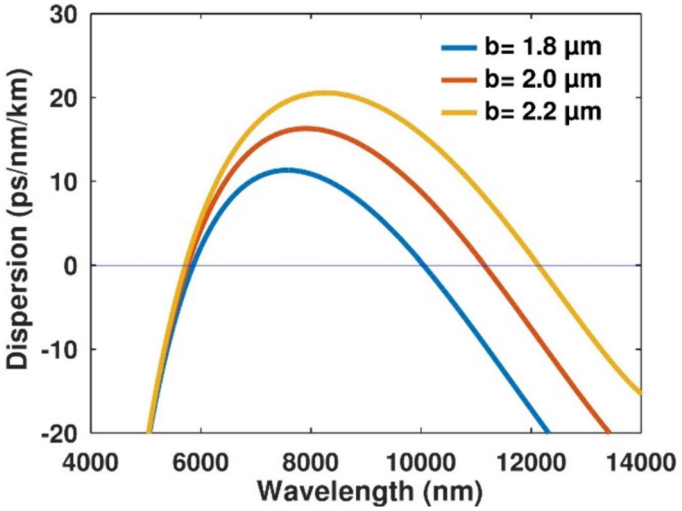


Fig. 3 Dispersion characteristics with variation in rib height (b)

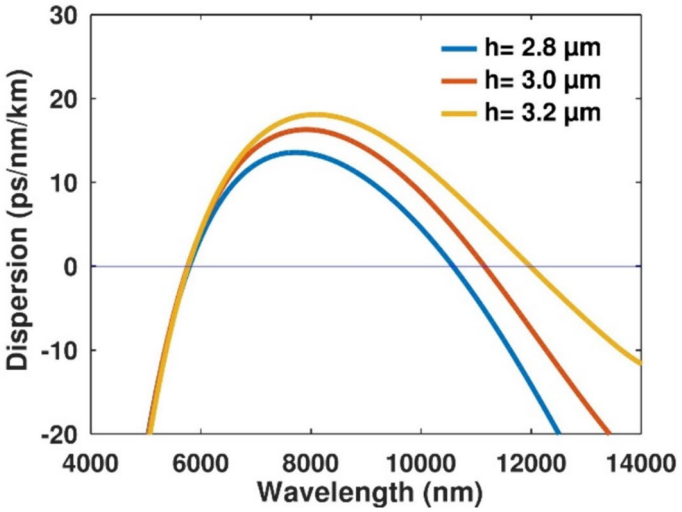


Fig. 4 Dispersion characteristics with variation in core maximum height (h)

can be seen that our designed waveguide exhibits two zero dispersion wavelengths at 5.76 μm and 10.50 μm. There is a linear decrease in effective mode index of the fundamental mode from 3.302 to 2.769 in the wavelength range of 2–14 μm. In the wavelength range of 4.5–12 μm, the dispersion curve is almost flat. At pump wavelength of 9.8 μm, the dispersion profile of proposed waveguide falls under the anomalous dispersion regime. The value of dispersion at pump wavelength is 1.3530 ps/nm/km. Periodic evolution of optical solitons is responsible for the SC broadening while pumping in the

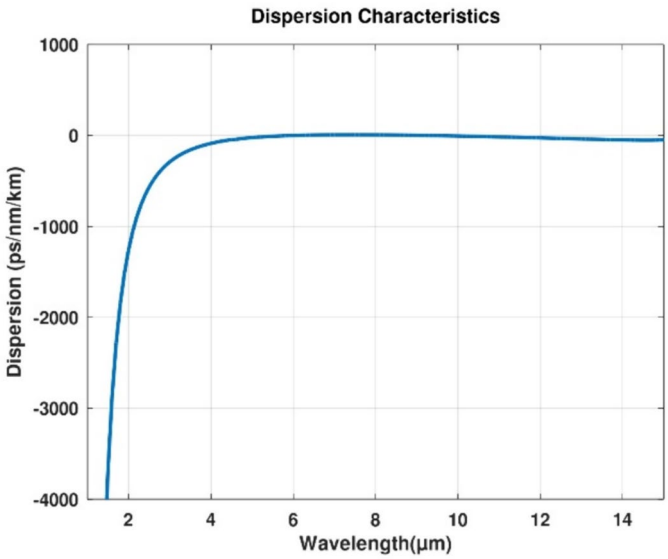


Fig. 5 Dispersion characteristics of proposed waveguide

Table 2 Optimized parameters of the designed waveguide

Geometrical parameters	Core width (a)	Core rib height (b)	Core maximum height (h)
Values	8.2 μm	1.8 μm	2.8 μm

anomalous region. Interplay of non-linear chirp influenced by the self-phase modulation and linear chirp by second order dispersion helps in stable spectrum broadening.

The variation of Effective mode area and non-linear coefficient for the optimized parameters with the whole wavelength range is shown in Fig. 6. The values of effective mode area and non-linear coefficient at pump wavelength of 9.8 μm are 20.57 μm^2 and 192.32 $W^{-1}km^{-1}$, respectively.

Supercontinuum generation is studied by solving Eq. 5. Left hand side of this equation contains β_n terms which represents order of dispersion coefficients. In our simulation, we have considered the coefficients upto 9th order because of the insignificant effect of coefficients greater than 9th order. SC spectrum is investigated under different laser operating conditions like pulse width of the input pulses and input peak power. In our simulations, we have varied the power from 10 kW to 20 kW in steps of 5 kW. The other operating condition, i.e. pulse width is varied from 50 fs to 150 fs in steps of 50 fs. In a 10 mm length of waveguide, the effect of variation of pulse width and peak power is shown in Figs. 7 and 8, respectively.

Pulse width of input pulses is an important parameter that affect the supercontinuum broadening. The variation in the SC generation with the change in pulse width is shown in Fig. 7. In our simulation we have used hyperbolic secant pulse and varied the pulse duration from 50 fs to 150 fs. It can be seen that with reducing the pulse width results in the increase in the spectral broadening in short propagation length. Hence, we can draw a

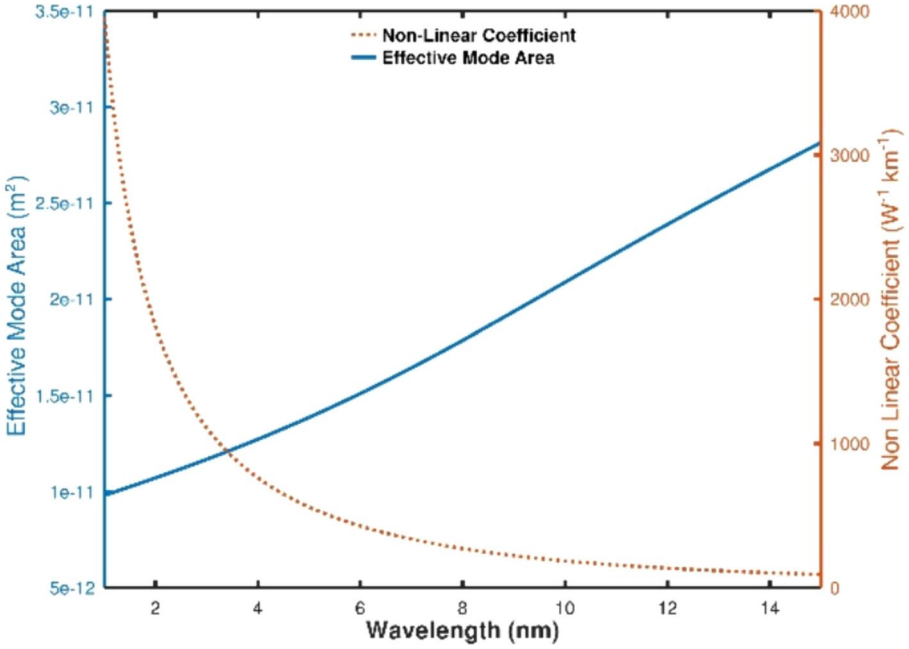


Fig. 6 Effective mode area and non-linearity with wavelength

simple conclusion that shorter pulse widths favors the spectral broadening if the propagation distance is short which in our case holds true. The comparison as can be seen from the Fig. 7 reveals that with the increase in pulse width from 50 fs, the broadening at 40 dB level decreases accompanied by decrease in the spectral evolution.

Furthermore, the effect of variation in peak power of the input pulse is also studied and is shown in Fig. 8. We have varied peak power from 10 kW to 20 kW. It can be observed that as peak power increases from 10 kW to 15 kW, the SC broadening increases and on further increasing the peak power SC spectrum at 40 dB level gets distorted. Hence, the proposed waveguide design is able to produce spectrum broadening from 3 μm to 25 μm when pumped with 9.8 μm wavelength in 10 mm length waveguide with 50 fs pulse width and 15 kW peak power.

Upon studying how the variation in laser operating conditions such as pulse width and peak power of input pulse affects the SC spectrum, we finally confirm our decision to propose the waveguide outlined in Sect. 2 for mid-infrared SC generation. Our simulation results in SC broadening of 3 to 25 μm, when pumped at 9.8 μm wavelength with 50 fs pulse width and 15 kW peak power in only 10 mm length of waveguide. Further, we have compared our design with previously reported designs of waveguides proposed by various researchers on different materials. We have compared our simulation results with the theoretical and experimental reporting's of GAST material used in PCF for SC generation. We observed that as compared to the previous PCF designs (Zhao et al. 2016; Xiao et al. 2020; Gao et al. 2022; Xia et al. 2021) using the same GAST material our proposed waveguide design is able to produce significant broadening covering mid-infrared region from 3 to 25 μm with relatively low peak power and high degree of coherence. Finally, the broadband spectrum from 3 to 25 μm along with the spectral

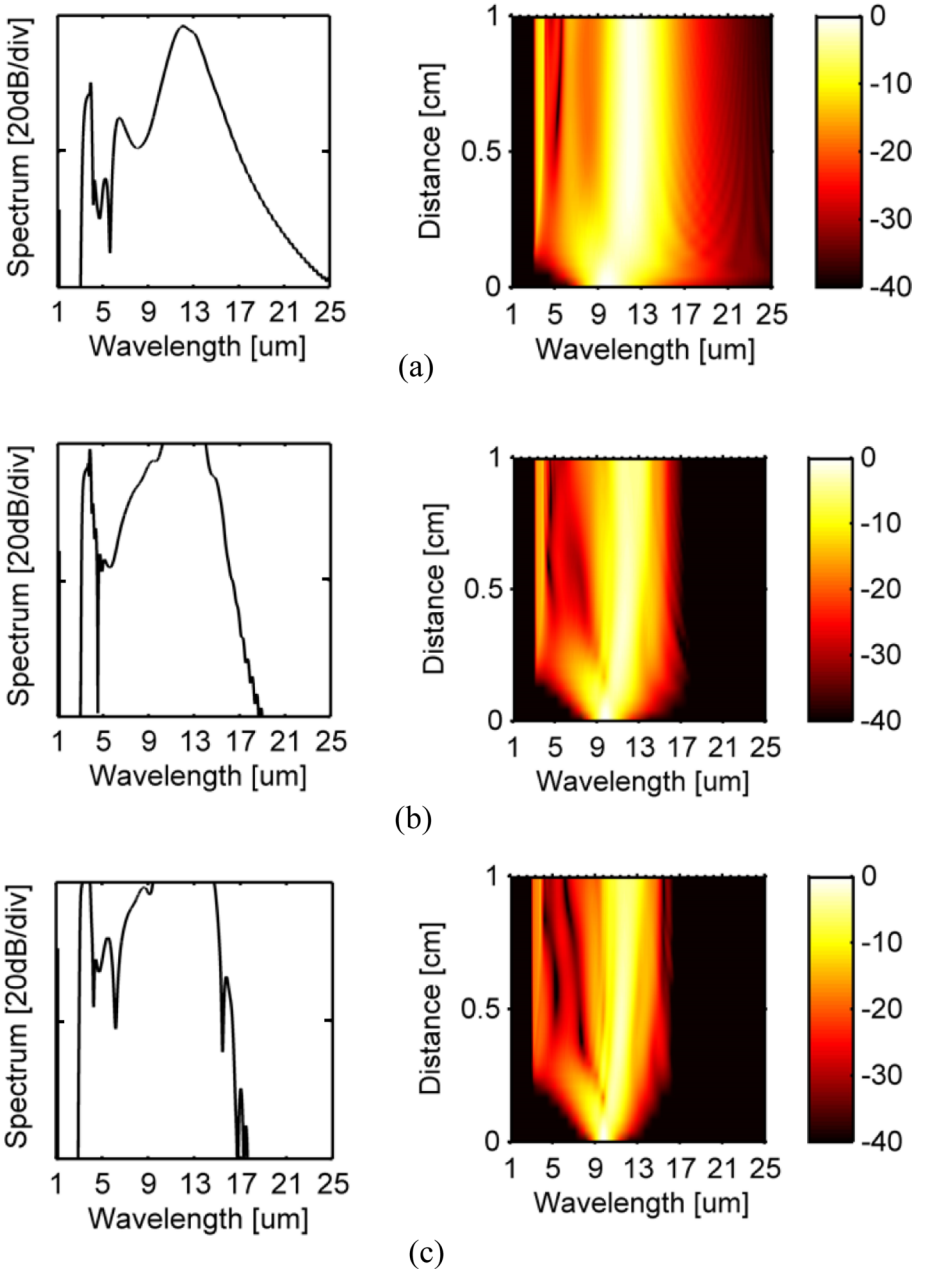


Fig. 7 SC broadening and spectral evolution along 10 mm length of fiber with pulse width of **a** 50 fs, **b** 100 fs and **c** 150 fs

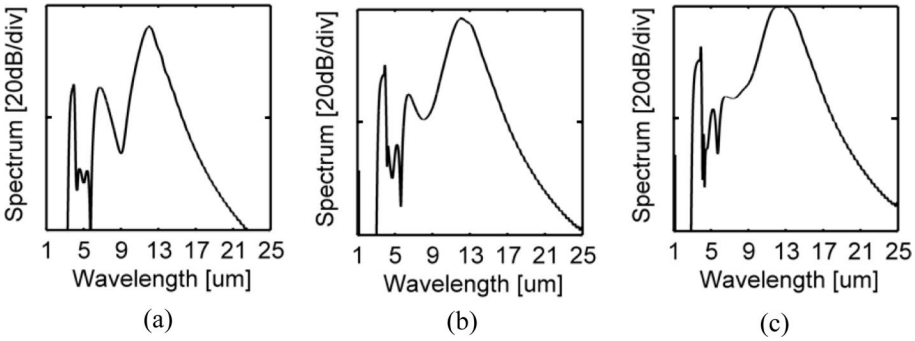


Fig. 8 SC broadening with variation in power, **a** 10 kW, **b** 15 kW and **c** 20 kW

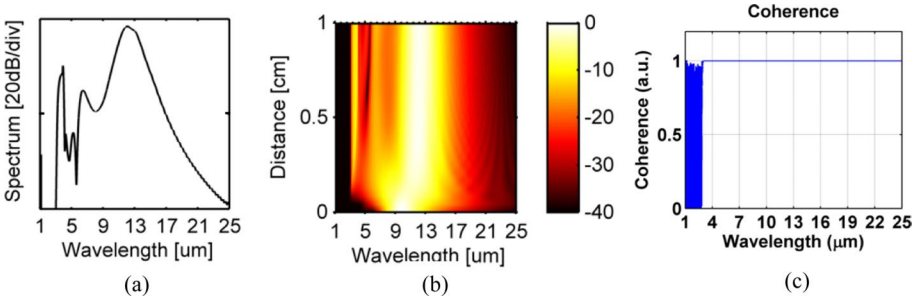


Fig. 9 **a** SC broadening, **b** spectral evolution and **c** degree of coherence in a 10 mm waveguide with 50 fs pulse width pumped at 9.8 μm wavelength with peak power 15 kW

Table 3 Comparison of the proposed waveguide with some previous works

Core material	Structure	SC Broadening (μm)	Pump wavelength(μm)	References
Buried chalcogenide	Waveguide	1.2–5.2	4.26	McCarthy et al. (2013)
$Ge_{11.5}As_{24}Sb_{64}$.	Waveguide	1.8–7.5	4	Yu et al. (2014)
As_2Se_3	Rib-Waveguide	1.2–7.2	2.8	Saini et al. (2018)
Ge–As–Se–Te	Double clad	1.5–11	5	Xiao et al. (2020)
Ge–As–Se–Te	Waveguide	2–13	5.8	Xia et al. (2021)
Ge–As–Se–Te	PCF	3–25	6	Gao et al. (2022)
Ge–As–Se–Te	Waveguide	3–25	9.8	This work

evolution of input pulses while propagating in a 10 mm waveguide and degree of coherence is shown in Fig. 9. The tabulated form of comparison is shown in Table 3. However, it is worth to report that since the size of the waveguide is smaller than pump wavelength, the laser coupling and transmission may present challenge. The coupling efficiency can be improved by using efficient fiber-to-waveguide couplers and by incorporating suitable methods (Soref 2006; Ballester et al. 2020).

5 Conclusion

In conclusion, we have numerically designed and analyzed a rib-waveguide structure featuring Ge–As–Se–Te (GAST) chalcogenide glass as core material and ZnSe as upper and lower cladding material for on-chip mid-infrared supercontinuum generation for non-linear applications. We analyzed the effect of geometrical parameters on dispersion characteristics and finally proposed an optimized design structure having very low dispersion values with two zero dispersion wavelengths at 5.76 μm and 10.50 μm. Analyzing the proposed design with varying laser operating conditions finally results in SC broadening which covers both atmospheric transparent window (3–15 μm and 8–13 μm) and molecular fingerprint region (2–15 μm) as well as far-infrared band. Such broadband mid-infrared SC spectrum is achieved by launching 50 fs pulse laser pulses with peak power of 15 kW in only 10 mm length of waveguide. The proposed rib-waveguide design can be helpful in many applications including sensing, spectroscopy, supercontinuum sources and optical coherence tomography.

Acknowledgements Deepak Garg acknowledges Delhi Technological University for providing an excellent research facility for conducting research during his Master's degree in Physics at the University.

Author contributions D.G. did the simulation and presented results and wrote the original manuscript, A. Khamaru helped in the explanation of results and helped in paper writing, A. K. reviewed paper and concept building.

Funding The authors have not disclosed any funding.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Agrawal, G.P.: Nonlinear fiber optics. In: Nonlinear Science at the Dawn of the 21st Century, pp. 195–211. Springer Berlin Heidelberg, Berlin, Heidelberg (2000)

Ballester, M.M., Lafforgue, C., Frigerio, J., Ballabio, A., Vakarin, V., Liu, Q., Ramirez, J.M., Le Roux, X., Bouville, D., Barzaghi, A., Ramos, C.A., et al.: On-chip mid-infrared supercontinuum generation from 3 to 13 μm wavelength. ACS Photonics, **7** (2020)

Bock, W.J., Gannot, I., Tanev, S. (eds.): Optical Waveguide Sensing and Imaging. Springer Science & Business Media (2007)

Boyd, R.W., Alexander, L.G., Enno, G.: Nonlinear optics. In: Springer Handbook of Atomic, Molecular, and Optical Physics, pp. 1097–1110. Springer International Publishing, Cham (2008)

Chauhan, P., Kumar, A., Kalra, Y.: Mid-infrared broadband supercontinuum generation in a highly nonlinear rectangular core chalcogenide photonic crystal fiber. Opt. Fiber. Technol. **46**, 174–178 (2018)
Cheng, T., Nagasaka, Tuan, T.H., Xue, X., Matsumoto, M., Tezuka, H., Suzuki, T., Ohishi, Y.: Mid-infrared supercontinuum generation spanning 2.0 to 15.1 μm in a chalcogenide step-index fiber. Opt. Lett. **41**, 2117–2120 (2016)
Cognolato, L.: Chemical vapour deposition for optical fibre technology. Le J. De Phys. IV. **5**, C5–975 (1995)
Dudley, J.M., Genty, G., Coen, S.: Supercontinuum generation in photonic crystal fiber. Rev. Mod. Phys. **78**, 1135 (2006)
Gai, X., Han, T., Prasad, A., Madden, S., Choi, D.Y., Wang, R., Bulla, D., Davies, B.L.: Progress in optical waveguides fabricated from chalcogenide glasses. Opt. Express. **18**, 26635–26646 (2010)
Gao, W., Dai, W., Zheng, Z., Chen, K., Chu, Y., Luo, Q., Wang, X., et al.: Simulation and analysis of supercontinuum generation in the waveband up to 25 μm. Appl. Opt. **61**, 6697–6703 (2022)
Gauthier, J.C., Robichaud, L.R., Fortin, V., Vallée, R., Bernier, M.: Mid-infrared supercontinuum generation in fluoride fiber amplifiers: Current status and future perspectives. Appl. Phys. B. **124**, 1–14 (2018)
Gorbach, A.V.: Graphene plasmonic waveguides for mid-infrared supercontinuum generation on a chip. Photonics. **2**, 825–837 (2015)
Han, T., Madden, S., Bulla, D., Davies, B.L.: Low loss chalcogenide glass waveguides by thermal nanoimprint lithography. Opt. Express. **18**, 19286–19291 (2010)
Hartl, I., Li, X.D., Chudoba, C., Ghanta, R.K., Ko, T.H., Fujimoto, J.G., Ranka, J.K., Windeler, R.S.: Ultrahigh-resolution optical coherence tomography using continuum generation in an air–silica microstructure optical fiber. Opt. Lett. **26**, 608–610 (2001)
Husakou, A.V., Kalosha, V.P., Herrmann, J.: Supercontinuum generation and pulse compression in hollow waveguides. Opt. Lett. **26**, 1022–1024 (2001)
Jiang, X., Joly, N.Y., Finger, M.A., Babic, F., Pang, M., Sopalla, R., Frosz, M.H., et al.: Supercontinuum generation in ZBLAN glass photonic crystal fiber with six nanobore cores. Opt. Lett. **41**, 4245–4248 (2016)
Karim, M.R., Rahman, B.M.A., Agrawal, G.P.: Mid-infrared supercontinuum generation using dispersion-engineered $Ge_{11.5}As_{24}Se_{64.5}$ chalcogenide channel waveguide. Opt. Express. **23**, 6903–6914 (2015)
Ke, K., Xia, C., Islam, M.N., Welsh, M.J., Freeman, M.J.: Mid-infrared absorption spectroscopy and differential damage in vitro between lipids and proteins by an all-fiber-integrated supercontinuum laser. Opt. Express. **17**, 12627–12640 (2009)
Khamis, M.A., Abbasi, Z.Q.: Design of ultra-broadband and highly coherent MIR supercontinuum generation using a photonic crystal fiber. Opt. Quant. Electron. **56**, 1–11 (2024)
Klimczak, M., Michalik, D., Stępniewski, G., Karpate, T., Cimek, J., Forestier, X., Pysz, D., Stępień, R., Buczyński, R.: Coherent supercontinuum generation in tellurite glass regular lattice photonic crystal fibers. JOSA B. **36**, A112–A124 (2019)
McCarthy, J., Bookey, H., Beecher, S., Lamb, R., Elder, I., Kar, A.K.: Spectrally tailored mid-infrared super-continuum generation in a buried waveguide spanning 1750 nm to 5000 nm for atmospheric transmission. Appl. Phys. Lett. **103**, 151103 (2013)
Righini, G.C., Chiappini, A.: Glass optical waveguides: a review of fabrication techniques. Opt. Eng. **53**, 071819–071819 (2014)
Saini, T.S., Kumar, A., Sinha, R.K.: 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)
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)
Saini, T.S., Phuoc, N., Hoa, T., Nagasaka, K., Luo, X., Tuan, T.H., Suzuki, T., Ohishi, T.: Coherent mid-infrared supercontinuum generation using a rib waveguide pumped with 200 fs laser pulses at 2.8 μm. Appl. Opt. **57**, 1689–1693 (2018)
Sharma, R., Kaur, S., Chauhan, P., Kumar, A.: Computational design & analysis of $GeSe_2$-As_2Se_3-PbSe based rib waveguide for mid-infrared supercontinuum generation. Optik, **220**, 165032 (2020)
Shiryaev, V.S., Churbanov, M.F.: Trends and prospects for development of chalcogenide fibers for mid-infrared transmission. J. Non-cryst. Solids. **377**, 225–230 (2013)
Soref, R.: The past, present, and future of silicon photonics. IEEE J. Sel. Top. Quantum Electron. **12**, 1678–1687 (2006)
Tan, W.C., Solmaz, M.E., Gardner, J., Atkins, R., Madsen, C.: Optical characterization of a-As2S3 thin films prepared by magnetron sputtering. J. Appl. Phys. **107**, 033524 (2010)

Xia, D., Huang, Y., Zhang, B., Yang, Z., Zeng, P., Shang, H., Cheng, H., et al.: On-chip broadband mid-infrared supercontinuum generation based on highly nonlinear chalcogenide glass waveguides. Front. Phys. **9**, 598091 (2021)

Xiao, J., Tian, Y., Zhao, Z., Yao, J., Wang, X., Chen, P., Xue, Z., et al.: Investigation of tellurium-based chalcogenide double-clad fiber for coherent mid-infrared supercontinuum generation. Opt. Fiber. Technol. **55**, 102144 (2020)

Yu, Y., Gai, X., Ma, P., Choi, D.Y., Yang, Z., Wang, R., Debbarma, S., Madden, S.J.: Davies. B.L: A broadband quasi-continuous, mid-infrared supercontinuum generated in a chalcogenide glass waveguide. Laser Photonics Rev. **8**, 792–798 (2014)

Zhao, Z., Wang, X., Dai, S., Pan, Z., Liu, S., Sun, L., Zhang, P., et al.: :1.5–14 μm mid-infrared supercontinuum generation in a low-loss Te-based chalcogenide step-index fiber. Opt. Lett. **41**, 5222–5225 (2016)

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.



Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Ge'ez Digit Recognition Model Based on Convolutional Neural Network

Ruchika Malhotra & Maru Tesfaye Addis

To cite this article: Ruchika Malhotra & Maru Tesfaye Addis (2024) Ge'ez Digit Recognition Model Based on Convolutional Neural Network, Applied Artificial Intelligence, 38:1, 2400641, DOI: 10.1080/08839514.2024.2400641

To link to this article: https://doi.org/10.1080/08839514.2024.2400641

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 20 Sep 2024.

Submit your article to this journal

Article views: 25

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uaai20

APPLIED ARTIFICIAL INTELLIGENCE
2024, VOL. 38, NO. 1, e2400641 (21 pages)
https://doi.org/10.1080/08839514.2024.2400641

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Ge'ez Digit Recognition Model Based on Convolutional Neural Network

Ruchika Malhotra[a] and Maru Tesfaye Addis[b,c]

[a]Department of Software Engineering, Delhi Technological University, New Delhi, India; [b]Department of Computer Science, Debre Tabor University, Debre Tabor, Ethiopia; [c]Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

ABSTRACT

Despite the historical significance of the Ge'ez script, there is a notable scarcity of studies on Ge'ez digit recognition, compounded by the challenges posed by the absence of publicly accessible datasets. The complex structure of Ge'ez digits further complicates the recognition task. In response to this gap, our study addresses the development of a digit recognition model based on deep learning (DL) specifically tailored for printed Ge'ez digits, accompanied by the creation of a comprehensive study dataset. The proposed model architecture seamlessly integrates one input layer, six convolutional layers, and three Max-Pooling layers. To assess the model's performance, we meticulously curated a Ge'ez digit dataset comprising 72,000 images, well-suited for DL applications using the ocropus-linegen model within OCRopus, a free document analysis tool. Leveraging cutting-edge DL algorithms, our proposed model demonstrates an impressive accuracy of 97.29%. This surpasses the performance of previous Ge'ez digit recognition models, marking a noteworthy advancement in this underexplored domain.

ARTICLE HISTORY
Received 15 December 2022
Revised 12 March 2024
Accepted 29 August 2024

Introduction

Ge'ez (ግዕዝ) language is a prehistoric Semitic language that dates back to the 5th century BC and has a writing system called Abugeda (አቡጊዳ). The language is used primarily in the Ethiopian Orthodox Church to compose several holy scriptures. Ethiopian Christians and the Jewish community of Beta Israel also use this language for liturgical purposes. Currently, this language is used in church activities and also serves as a teaching language in schools. Teaching this language in schools has expanded beyond Ethiopia to other countries, including Canada, Germany, and Italy. Additionally, several languages, such as Amharic, Awngi, and Guragegna, adopt the Ge'ez alphabet as the foundation for their writing systems within Ethiopia (Nur, Abebe, and Rajendran 2022).

CONTACT Maru Tesfaye Addis marutcomp@gmail.com Department of Computer Science, Debre Tabor University, Debre Tabor, Ethiopia

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Consequently, the Ge'ez writing system continues to be prolific, generating a substantial number of historical documents across various domains.

In the realm of optical character recognition (OCR), digit recognition remains an active and evolving area of exploration (Shaukat et al. 2020). Notably, languages that have historically received less attention, such as the Ge'ez script, present an ongoing focus in research efforts, driven by the aim to enhance recognition efficiency. While Latin script recognition research has reached maturity, complete datasets are publicly accessible. However, despite its antiquity and widespread applications, the Ge'ez script has not garnered sufficient attention in OCR research, posing a critical gap.

Given that a significant portion of Ethiopia's ancient history is documented in the Ge'ez script, the imperative to digitize these materials for historical preservation is evident (Beyene 2019). Ge'ez digits pose unique challenges for recognition due to their cursive symbols, presenting complexity for machine learning algorithms. This complexity is exacerbated by variations in writing materials, font sizes, types, and styles found in printed documents. Ge'ez numerals, commonly inscribed on animal skins called birhana (ብርሃና), further contribute to diverse writing styles and alignment within texts.

Compounding these challenges is the absence of a standardized dataset for Ge'ez digit recognition systems. Recognizing this gap, our proposed study aims to develop an accurate automatic Ge'ez digit recognition model using deep learning techniques. Our research includes the preparation of experimental datasets, addressing the current deficiency, and paving the way for future applications. In essence, this endeavor signifies a promising avenue for advancing the field and contributing to the preservation of Ge'ez script heritage.

Traditionally, digit recognition models have relied on k-nearest neighbor support, support vector machines, multilayer perceptrons, and random forests (Shaukat et al. 2020). These traditional methods can be time-consuming and error-prone, often requiring extensive manual effort in feature extraction (Malhotra and Addis 2023). Therefore, there is a need for automated and efficient techniques that can accurately recognize Ge'ez digits. However, with the ability to automatically extract features, deep learning techniques, particularly through CNNs, are now increasingly used for digit recognition. This method has gained widespread recognition across various scripts due to its ability to detect intricate patterns. CNN is the preferred choice in OCR technologies for text extraction from scanned images due to its learning ability.

We hypothesize that the development of an accurate automatic Ge'ez digit recognition model using DL techniques, particularly CNNs, will significantly improve recognition efficiency. This hypothesis is based on the premise that CNNs, with their ability to automatically extract features, are well-suited for recognizing intricate patterns in Ge'ez digits. The

proposed model is expected to outperform traditional methods, contributing to the advancement of digit recognition in the specific domain of the Ge'ez script and offering insights applicable to the broader field of pattern recognition.

This study significantly advances the field in three key areas:

(1) **Ge'ez Numerals Recognition Model**: Addressing a historical understudied aspect, our study introduces a Ge'ez numerals recognition model based on the Convolutional Neural Network (CNN). The model achieves superior accuracy rates, marking a substantial contribution to this unexplored domain.
(2) **Publicly Accessible Dataset**: We present a meticulously curated dataset for Ge'ez digit recognition, openly accessible for noncommercial and scientific purposes. This resource, hosted on Google Drive (link: https://drive.google.com/file/d/1SBHjTV4sO4klDleEP22_1f0dCq8waMfF/view?usp=sharing), serves as a valuable asset for the research community.
(3) **Experimental Evaluation and Promising Results**: Through extensive experiments and analyses, our study not only achieves high recognition accuracy in Ge'ez digit recognition but also advances the state of the art in pattern recognition. The findings contribute not only to the specific domain of Ge'ez script but also offer insights beneficial to the broader field of pattern recognition.

The Ge'ez Digits

Like the Latin script, the Ge'ez language adheres to a left-to-right reading and writing system. In this linguistic realm, a distinctive numeral scheme is employed to convey digital information, mirroring the structure of the regular alphabet. In Figure 1, we observe the Ge'ez digits, a set comprising 20 distinct symbols.

Notably, the Ge'ez numeral system diverges from its Latin counterpart. Intriguingly, there is no specific symbol designated for the number 0 as shown in Table 1. Instead, each digit from 1 to 9 possesses its unique representation. Additionally, numerical multiples of ten—10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 10,000—equally boast their individual symbols within the Ge'ez system.

The nuanced intricacies of the Ge'ez numeral system involve the utilization of the base symbols to construct representations for the remaining digits. Each digit is adorned with two horizontal lines, positioned both above and below, adding a distinctive visual element to the numerical characters. This unique

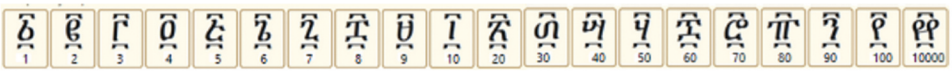


Figure 1. Ge'ez numbers representation (Beyene 2019).

Table 1. Ge'ez numbering and English equivalent numeric symbols.

Digits*	English symbol	Numbers*	English symbol
፩	1	፲	10
፪	2	፳	20
፫	3	፴	30
፬	4	፵	40
፭	5	፶	50
፮	6	፷	60
፯	7	፸	70
፰	8	፹	80
፱	9	፺	90
		፻	100
		፼	10,000

blend of symbolism and structure renders the Ge'ez numeral system a captivating facet of the language's written expression.

Digit recognition, a subset of the broader optical character recognition challenge, is integral to document analysis, especially in identifying numerical characters. Its relevance spans beyond academia, finding applications in check truncation systems, postal identification, vehicle license plate number recognition, and banking transactions (Ahlawat et al. 2020), underscoring the practical utility of digit recognition technologies.

Our proposed approach not only addresses academic aspects but also holds substantial industrial significance. Beyond traditional character recognition scenarios, it has the potential to enhance efficiency and accuracy in various industrial processes. For example, the developed model could optimize check truncation systems, improving accuracy and expediting financial transactions. Similarly, in postal and license plate identification systems, the proposed approach could offer tangible benefits in terms of speed and precision (Aggarwal and Singh 2016).

Moreover, recognizing the significance of digit recognition in the context of Ge'ez numerals, as elucidated by (Nur, Abebe, and Rajendran 2022), extends beyond cultural and historical contexts. It plays a vital role in practical applications, such as date recognition in Ethiopian calendars and text extraction in religious literature, showcasing the dual academic and industrial importance of our proposed approach.

The remainder of this paper is structured as follows: Section 2 discusses related work; Section 3 presents the methodology; Section 4 presents the results and discussion; and Section 5 draws the conclusions and recommendations of the paper.

Related Works

Numerous published studies have described various pattern recognition and feature extraction strategies for digit recognition. Most previous studies achieved excellent results when they used conventional feature extraction methods for their development. In this section, we have attempted to provide an overview of the studies that have been conducted in the area of digit recognition.

The authors, Abdulrazzaq and Saeed, compare three recognition algorithms (Abdulrazzaq and Saeed 2019). The three algorithms are Naive Bayes (NB), Multilayer Perceptron (MLP), and the K* algorithm. For their experiment, the handwritten dataset from the National Institute of Standards and Technology (NIST) was used. Each image is 128×128 pixels and was randomly selected from 48,060 samples. In addition, preprocessing was performed during the experiment and 16×16 feature images were discovered. The authors used the "correlated feature selection" method to select the features, and as a result, 37 features were obtained, which were used to train the algorithms. The main objective of the study is to determine the best classifier among the three algorithms that achieves reasonable accuracy with few features. The recognition result of K* algorithm is better than that of NB and MLP. A recognition rate of 82.36% is obtained. The resizing of the images in the pre-processing phase of the study and the random selection affect the recognition accuracy as they lead to an increase in the number of misrecognized digits. Since the dataset NIST contains numerous confusing images, it was difficult for the study to find the best classifier by randomly selecting samples.

Liu, Wei, and Meng (2020) compared the recognition time and recognition rate of neural networks with backpropagation (BP), support vector machines (SVM), k-nearest neighbor (KNN), and CNN. Liu et al. took the MNIST handwritten dataset as a sample. The MNIST handwritten dataset is a popular dataset that contains a total of 70,000 handwritten 28×28 images for training and testing samples. From the MNIST dataset, 5000 samples were randomly selected for training and 1000 samples for testing activities. To train and develop the classifier, the authors follow their procedures. Input, preprocessing, feature selection, classifier design and classification decision are the steps for the training phase. In the testing phase, they follow the input, preprocessing, feature selection, and classification decision steps to obtain recognition results. The advantages and disadvantages of the four algorithms are analyzed and reported. The comparison of the simulation test results shows that the CNN algorithm achieves the best results in terms of recognition accuracy.

Abu Ghosh and Maghari (2017) compared the three most popular neural network (NN) approaches. These are CNN, deep belief networks (DBN), and

deep neural networks (DNN). Their proposed system is divided into six main steps: pre-processing, segmentation, feature extraction, classification, training, and recognition. The authors used the MNIST handwritten dataset and a random dataset they created themselves. The random dataset is used to verify the efficiency and validity of the results obtained with the MNIST standard dataset. The random data set consists of 85 different handwritten digits obtained from different sources. The authors compared the three algorithms in terms of accuracy and performance, although these were not the only evaluation criteria. In addition to accuracy, they also compared the three algorithms based on execution time. DNN performed better than the other two algorithms with an accuracy of 98.08%.

Ahlawat et al. (2020) proposed an improved digit recognition using the MNIST dataset and a CNN recognition algorithm. They designed two CNN architectures with three layers (i.e., CNN_3L) and four layers (i.e., CNN_4L). The main objective of their work is to roughly investigate all the parameters of the CNN architecture using the MNIST handwritten dataset. Optimization algorithms are used to optimize neural networks for better performance and faster results. Stochastic gradient descent with momentum, Adam, adagrad and adadelta are the optimization parameters used in their study to achieve optimal performance. Among all optimization parameters, the Adam optimizer has a detection rate of 99.89% for the CNN model with three layers, which is better than the previously reported results in this area.

The authors (de Sousa 2018) proposed the recognition of Arabic handwritten characters using ensemble convolutional neural networks. They used adaptive gradient descent in the first epoch to facilitate convergence and regular stochastic descent in the last epoch. Their study tested K-fold and Monte Carlo cross-validation methods. Using the MADbase dataset, a test accuracy of 99.47% and a classification accuracy of 99.74% were achieved.

Ahlawat and Choudhary (2019) proposed a model for automatic recognition of handwritten digits using CNN for feature extraction and SVM for output prediction. In this model, the authors try to take advantage of CNN and SVM. They achieved a classification accuracy of 99.28% for the MNIST dataset of handwritten digits.

An effective and improved method for recognizing handwritten digits based on a convolutional neural network was developed by (Ali et al. 2019). In their study, a CNN with a DeepLearning4J (DL4J) framework with rectified linear unit (ReLU) activation is used to recognize handwritten digits. The model is trained and tested with the MNIST handwritten dataset. Three convolutional layers, one max-pooling layer and two fully connected layers form the CNN model in the study. The architecture consists of two main components (i.e., feature extractor and feature classifier). The parameters used determine how effective CNN models are. They use a dataset of 5,130 images for testing and 60,000

images for training their better model. The result: out of 5,130 images tested, 42 were misidentified and 5088 were correctly identified, and the overall accuracy was 99.21%. The computation time for training and testing was improved.

A study on the recognition of handwritten Ge'ez digits using Deep Learning was conducted by (Nur, Abebe, and Rajendran 2022). To achieve a high recognition rate, they proposed a CNN architecture with convolutional layers, max-pooling, dropout, flatten, fully connected, and SoftMax layers. 52400 handwritten Ge'ez digits were used for their study, and 524 different individuals with different writing styles contributed to the collection of the data. With a model loss of 0.2013 and an accuracy of 96.21%, the proposed approach outperformed the competition in Ge'ez digits. The lack of handwritten Ge'ez digits makes the study more challenging. They recommended multi-digit recognition as a future direction.

The Ge'ez digits are also not limited to only printed and handwritten studies. Some studies have also been conducted for visually impaired people. Tamiru et al. proposed a study on the recognition of Amharic alphabetic sign language (Tamiru, Tekeba, and Salau 2022). Their research benefits people with hearing loss. Their recognition model consists of four stages of development, such as pre-processing, regimentation, feature extraction, and recognition. The first three stages are for digital image processing and the last stage is for classification by machine learning algorithms. Digital image processing and machine learning algorithms were used to develop the model. The two most common machine learning algorithms are artificial neural networks and multiclass support vector machines. In their study, the recognition accuracy is achieved using SVM (98.06%) and ANN (80.86%).

Abeje et al. (2021) studied Ethiopian sign language using a deep convolutional neural network. Their study takes images of Amharic sign language and translates them into the desired Amharic texts. Their proposed system consists of three main stages: pre-processing, feature extraction, and classification. They also applied digital image processing techniques such as data acquisition, pre-processing, normalization, resizing, region of interest identification, noise removal, and feature extraction. For end-to-end learning, the authors used deep convolutional neural networks. The image dataset used was collected from hearing-impaired students.

The majority of researchers focused on the recognition of English numerals. The availability of a public dataset for Latin scripts motivated researchers to focus on the area. However, in the case of Ge'ez numerals, there is no publicly available dataset to study research in this area. This discourages researchers from making efforts to recognize Ge'ez numerals. In this study developing a recognition model with better accuracy and preparation of the experimental dataset advances the state of the art in Geez digit recognition.

Methodology

In this work, DL techniques are used to build a model for the recognition of Ge'ez digits. To realize this, the details of the dataset preparation and tools, the deep learning algorithms, and the experimental setup of the model are described in the following subsection.

Dataset Preparations and Tools

Given the absence of an existing Ge'ez numeral dataset in this study, we addressed this gap by crafting a syntactically generated dataset using OCRopus. Unlike datasets derived from real-world occurrences, this synthetic dataset was artificially created for our purposes. OCRopus, a free optical character recognition (OCR) and document analysis solution, facilitated this task with its command line interfaces and modular design. Developed by Thomas Breuel, OCRopus is accessible on GitHub (Breuel 2017).

To create our Ge'ez digit image dataset, we employed the ocropus-linegen model, utilizing text files containing Ge'ez digits in various font types as inputs. The methodology for dataset preparation, as outlined by (Belay, Habtegebrial, and Stricker 2018) and (Addis, Liu, and Ta 2018), was directly applied. This process yielded a primary dataset comprising 72,000 cropped images, each with dimensions of 32 × 32. Figure 2 illustrates examples of these images.

Illustrated in Figure 3 is a balanced distribution of sample sizes across classes, with each class exhibiting a nearly equal frequency. This uniform

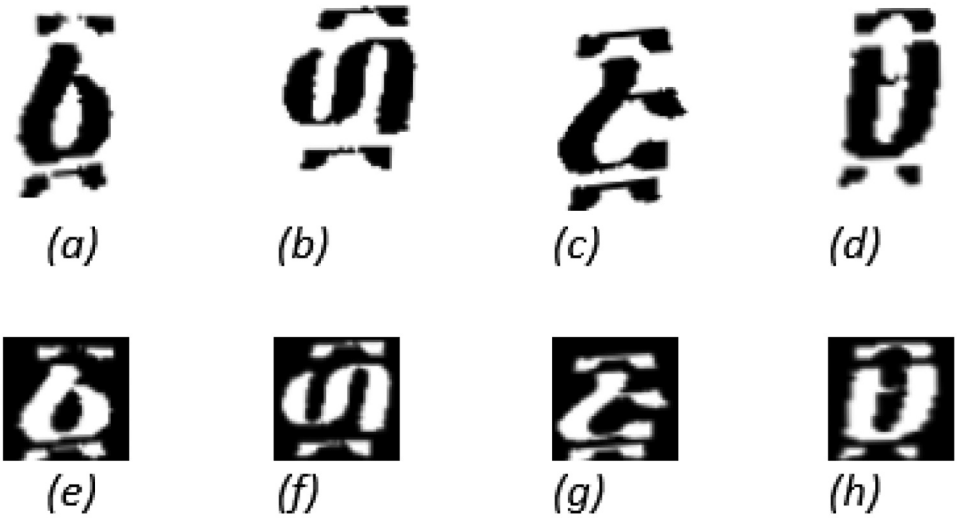


Figure 2. Sample images from the Ge'ez digits dataset.

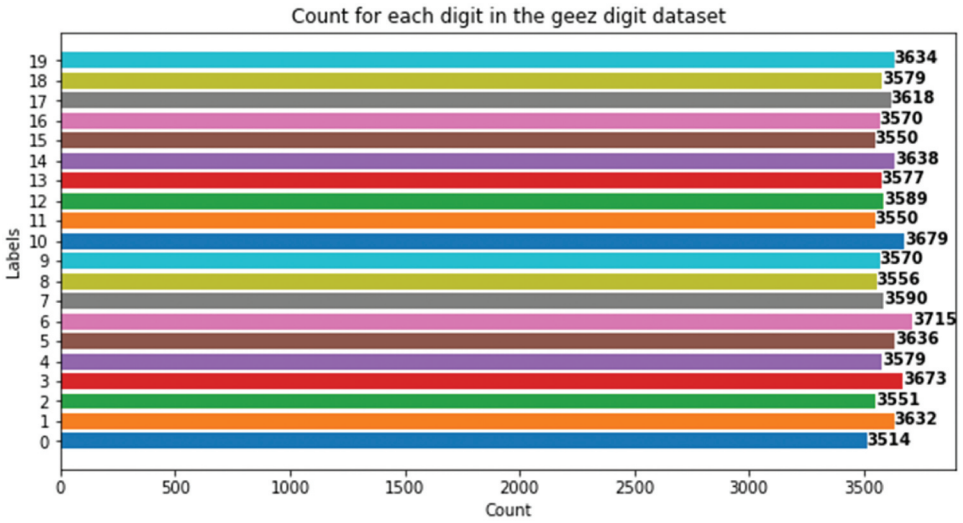


Figure 3. Numbers of samples in the Ge'ez-digit database.

distribution ensures that each character within the Ge'ez digit dataset has approximately the same number of samples across distinct class labels. Notably, the total number of classes in the Ge'ez digit dataset is 20, contributing to a comprehensive representation of the characters within the dataset. This balanced distribution enhances the robustness and fairness of the dataset, fostering equitable training and evaluation across all classes.

The implementation of the programming for training the networks was carried out using Python 3.7. To construct a model for recognizing printed Ge'ez digits, scripts were developed within Kaggle's notebook, leveraging the Keras framework with TensorFlow as the backend. Kaggle's notebook offers dedicated virtual hardware resources throughout code execution, featuring a robust configuration including a 73.1 GB hard disk and 13 GB of GPU memory. These hardware specifications ensure ample capacity for thorough model testing, demonstrating the suitability of the environment for the project's computational requirements.

Deep Learning Algorithms

Deep learning, a specialized domain within artificial intelligence, mimics neural networks akin to the human brain, as highlighted by research such as (Gerber and Chung 2016) and (Gurucharan 2020). This powerful subset of machine learning excels even when confronted with highly heterogeneous, unstructured, and interconnected data, enabling machines to adeptly address complex challenges in pattern recognition. At the heart of this capability are

artificial neural networks, playing a pivotal role in processing intricate information.

However, when dealing with diverse and extensive datasets, the scalability of conventional artificial neural networks may prove insufficient. This limitation is circumvented by the advantage of integrated feature extractors and classifiers offered by deep neural networks, as emphasized by the work of (Kim and Xie 2015). This adaptability enables deep learning models to navigate and derive meaningful insights from vast and varied datasets, showcasing their proficiency in handling the intricacies of real-world data.

The CNN, a multilayer hierarchical model, proves to be well-suited for a variety of pattern recognition challenges. One key advantage lies in the weight-sharing mechanism between CNN layers, as discussed by (Belay, Habtegebrial, and Stricker 2018). This approach not only diminishes the number of parameters, promoting computational efficiency but also enhances overall performance, making CNNs particularly effective in discerning complex patterns and features within diverse datasets.

CNNs or ConvNets stand out as a leading deep learning approach, exhibiting remarkable promise in various supervised learning tasks and pattern recognition endeavors, as noted by (Asha and Krishnappa 2018). CNNs, renowned for their adept handling of image and video data, excel in tasks such as pattern recognition, object detection, and facial recognition (Malhotra and Addis 2023; Mukhoti, Dutta, and Sarkar 2020).

The efficacy of CNNs has been notably harnessed in character and digit recognition tasks in recent years, as highlighted by (Ahamed et al. 2020). CNNs, irrespective of the script used, demonstrate a consistent improvement in digit recognition performance. To delve into the intricacies of this, the subsequent subsection aims to provide a simplified explanation of the fundamental layers constituting the CNN model.

Convolutional Layers

The convolution layer is responsible for the detection and extraction of visual features from spatially localized neighborhoods of the image (Mukhoti, Dutta, and Sarkar 2020)he initial stage in processing the input image involves the first convolutional layer, a pivotal component utilizing CNNs to extract essential features before transmitting the data to the recurrent layer. This convolutional layer is composed of key elements: convolution, activation, and maximum pooling, seamlessly integrated into the CNN architecture to perform targeted feature extraction. Within each CNN layer, a set of filters plays a crucial role in extracting significant features from the input data. Several parameters define the characteristics of each convolutional layer, including the size of the input, filter dimensions, depth of the feature map, padding, and stride. These

parameters collectively contribute to shaping the network's ability to discern and capture relevant patterns within the data.

Convolution

This initial layer of the CNN receives raw image data as input, which may be in either color or grayscale format. Represented as a three-dimensional matrix of pixel values, the input undergoes a crucial reshaping process to form a single row. To optimize learning speed and enhance model performance, pixel values are normalized by dividing each value by 255, effectively scaling them within the range of 0 to 1.

Central to the convolutional layers, convolution is a vital operation that enables CNNs to extract pertinent image features. By systematically traversing a kernel across different regions of the input image, the convolution process is computed (Jain, Singh, and Goel 2019). The multiplication of elements in the matrix and their summation results in values stored in a feature map, constructed by the filter (kernel). Subsequent layers gain access to this feature map, fostering the propagation of important information throughout the network.

The feature map F is calculated in "(1)," where I represent the input image, K denotes the kernel size or feature detector, and x and y signify the width and height indices of the result matrix, respectively.

$$F_{x,y} = I * K_{x,y} = \Sigma_{i,j} I_{x-i,y-j} \cdot K_{i,j} \tag{1}$$

In Figure 4, we visually depict the generation of a 4×4 feature map through the convolution process using a 3×3-pixel kernel on a 6×6-pixel input image. The size of the resulting feature map (n_{new}) is calculated in "(2)."

$$n_{new} = n - f + 1 \tag{2}$$

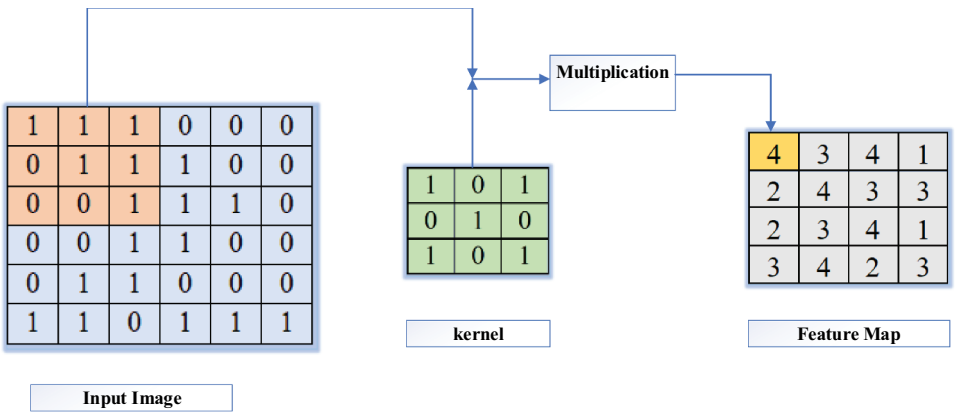


Figure 4. 3-by-3 convolution computation.

where n_{new} represents the feature map dimension, n is the size of the input image, and f is the size of the filter or kernel. In each convolutional layer throughout this study, we utilized a 3-by-3 kernel.

Activation/Non-Linearity

Efficiency in a neural network demands the inclusion of activation or non-linearity following the summation of weights and biases. ReLU (Rectified Linear Unit) emerges as the predominant activation function in CNNs. This function, represented by Equation "(3)," unleashes the full potential of convolutional results and bias values by setting all negative values to zero.

$$ReLU = max(0, (w \cdot I) + b) \tag{3}$$

where ReLU is the activation function, I is the input value, w signifies weight, and bias is denoted by b.

Pooling Layer

Connecting the convolutional layer and the fully connected layer, the pooling layer plays a crucial role in downsampling the output of the preceding layer, as discussed by (Jain, Singh, and Goel 2019). By selecting the most prominent features, the pooling layer effectively reduces the dimensionality of each convolutional layer, optimizing computational efficiency and mitigating overfitting risks. Max-pooling, a prevalent method, maintains the most significant traits by extracting the maximum value within "n x n" windows. In this study, we employed max pooling for its effectiveness in retaining key features, as illustrated in Figure 5.

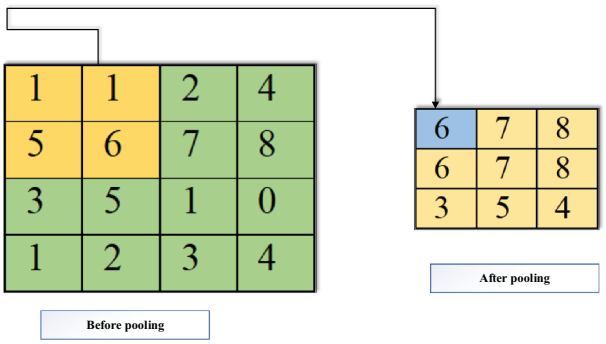


Figure 5. 4x4 input and 2x2 windows sliding with 1-step stride.

Table 2. Experimentation software tools, packages, and versions.

Tools and Packages	Version	Description
Python	3.7.	Scripting language to write code
TensorFlow	2.10.0	Symbolic math library
OpenCV	4.5.2	Supports image processing
Keras	2.10.0	Wrapper for TensorFlow
Pillow	Latest	Supports image file management
Kaggle Notebook	Latest	Back-end editor with Google compute engine

Table 3. The hyperparameters used in the model and their values.

Parameter type	Values
Optimizer	Adam
Loss function	Categorical cross-entropy
Learning rate	0.001
Padding	Same
Epochs/Iterations	100
Batch size	32
Accuracy	Percentage

Experimental Setup and Model Configuration

The setup for the experiment and the model are described in this section. It also covers software tools informations. As shown in Table 2, all algorithms are created and implemented using the required software tools, packages, and libraries Table 3 lists the hyper-parameters used in the model along with their associated values.

The architecture of the trained model is depicted in Figure 6, generated using the model plotting utility package in Keras. The model encompasses key layers essential for the CNN, including one input layer, six convolutional layers, and three Max-Pooling layers. Additionally, the model incorporates

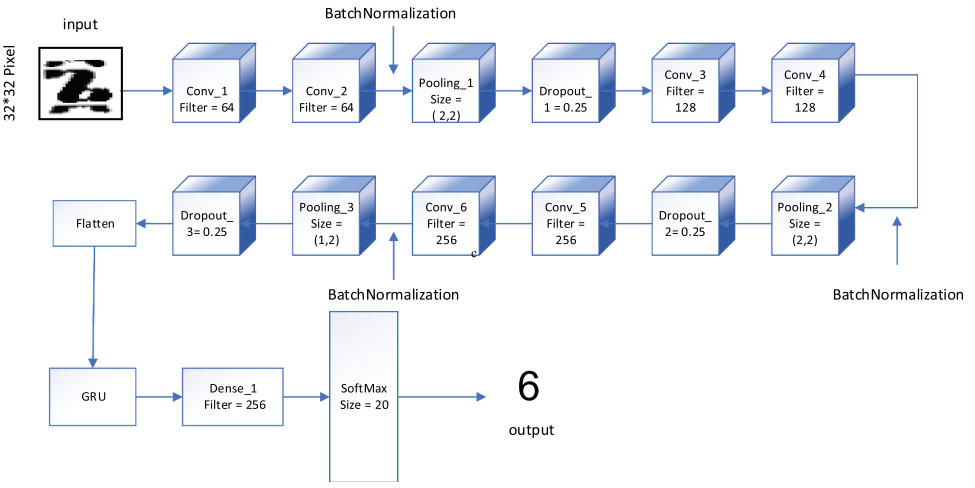


Figure 6. The proposed model's layer-by-layer visualization.

one bidirectional layer named GRU within the RNN section. To enhance model robustness, dropout, and dense layers are strategically integrated, with fully connected layers carefully positioned toward the end of the architecture. This comprehensive design ensures the effective extraction and utilization of features at various stages, contributing to the model's overall performance.

In this study, a dataset of 72,000 samples, meticulously prepared ourselves, was employed for testing and training. The dataset was categorized into training, validation, and testing sets, with 80% of samples chosen randomly for training and 10% for validation. The testing phase utilized 10% of the dataset samples. The proposed CNN model follows a systematic process, beginning with image input, followed by pre-processing. Subsequently, the model undergoes training and validation phases, culminating in the testing phase. The recognition model, vitalized through the training, validation, and testing phase, is illustrated in the flow chart presented in Figure 7.

Results and Discussion

In this section, we present the results and engage in a detailed discussion following the system design outlined in Section 3. A dataset comprising a total of 72,000 images was utilized for the experiment, intelligently divided into

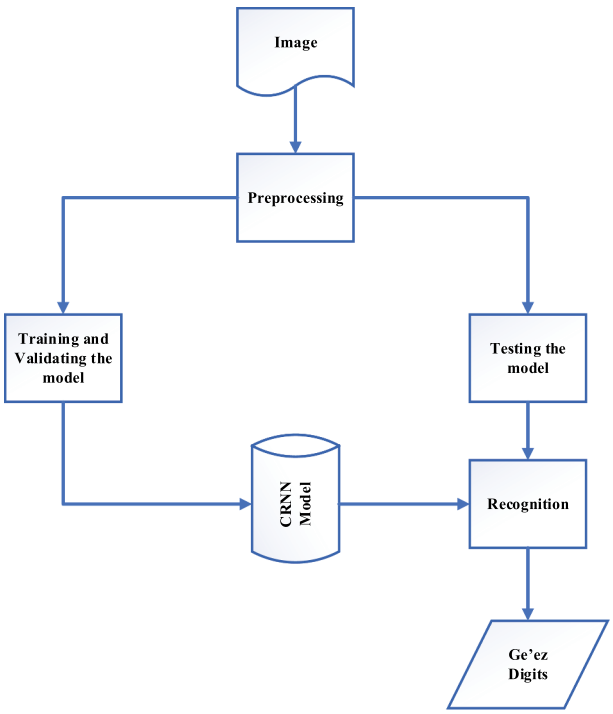


Figure 7. The proposed system training, validation, and testing phase.

training, validation, and testing categories. The training set encompasses 57,600 images (80%), the validation set comprises 7,200 images (10%), and the testing set includes 7,200 images (10%).

The learning rate serves as a pivotal factor influencing how swiftly the model adapts to the given problem. In our experiments, the learning rates were systematically varied, employing a 50-epoch size and the Adam optimizer. The learning rate values explored were $Lr1 = 0.1$, $Lr2 = 0.01$, $Lr3 = 0.001$, and $Lr4 = 0.0001$. The batch size, representing the number of instances used in a single iteration, was set to 32.

The proposed CNN model, consisting of 1,408 non-trainable parameters and a total of 1,486,676 trainable parameters, demonstrated robust training and learning capabilities. Among the four experiments conducted with different learning rates using the Adam Optimizer, the best testing accuracy of 97.29% was achieved at a learning rate of $Lr4 = 0.0001$. This remarkable accuracy underscores the model's efficacy in recognizing Ge'ez digits, highlighting its potential for real-world applications. The comprehensive experimentation and meticulous parameter tuning contribute to the model's impressive performance, positioning it as a promising solution in handwritten digit recognition tasks.

Results

Figure 8 illustrates the overall testing accuracies of the proposed model under different learning rates. Notably, the model demonstrates testing accuracy rates of 43.97%, 93.89%, 97.12%, and an impressive 97.29% at learning rates $Lr1$, $Lr2$, $Lr3$, and $Lr4$, respectively. The outstanding testing accuracy of 97.29% marks a significant accomplishment in addressing Ge'ez digit recognition challenges, showcasing the model's robust performance. Figures 9 and Figure 10 provide graphical representations of the training and validation results for both accuracy and loss, focusing on learning rates $Lr1$ and $Lr2$,

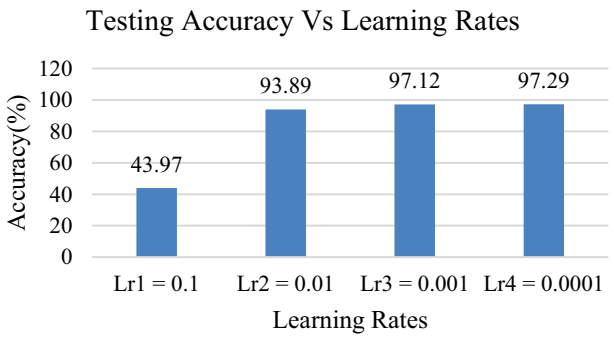


Figure 8. Testing accuracies of the proposed model with different learning rates.

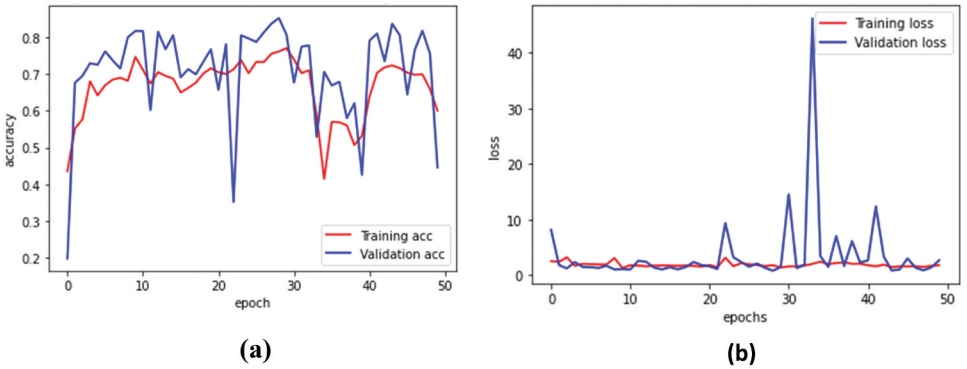


Figure 9. (a) the training and validation accuracy and (b) the training and validation loss with learning rate Lr1 versus epochs.

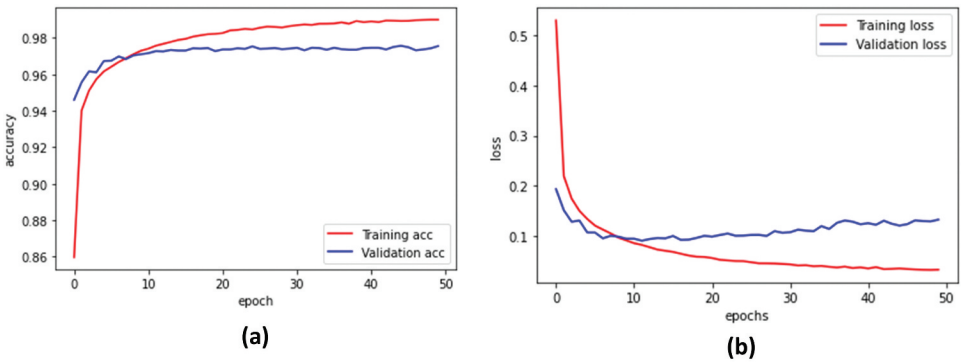


Figure 10. (a) the training and validation accuracy and (b) the training and validation loss with learning rate Lr4 versus epochs.

respectively. These visualizations offer valuable insights into the model's learning process, highlighting its ability to converge effectively during training while minimizing loss. The discernible trends underscore the model's adaptability and efficiency in capturing the intricate patterns associated with Ge'ez digit recognition.

In addition to the accuracy metric, a comprehensive evaluation of the trained model was performed using a classification report, showcasing precision, recall, F1-score, and support across different Ge'ez digit classes. Before delving into the classification report, it's essential to define the key parameters:

- True Positive (TP): Correctly predicted positive instances
- False Positive (FP): Incorrectly predicted positive instances
- True Negative (TN): Correctly predicted negative instances
- False Negative (FN): Incorrectly predicted negative instances

Table 4. Class-wise classification performance for learning rate Lr4.

Ge'ez Digits, (label)	precision	recall	f1-score	support
፩ (0)	0.9826	0.9658	0.9741	351
፪ (1)	0.9833	0.9752	0.9793	363
፫ (2)	0.9885	0.9690	0.9787	355
፬ (3)	0.9783	0.9809	0.9796	367
፭ (4)	0.9615	0.9777	0.9695	358
፮ (5)	0.9669	0.9643	0.9656	364
፯ (6)	0.9709	0.9892	0.9800	371
፰ (7)	0.9831	0.9749	0.9790	359
፱ (8)	0.9888	0.9944	0.9916	356
፲ (9)	0.9856	0.9608	0.9730	357
፳ (10)	0.9755	0.9728	0.9741	368
፴ (11)	0.9855	0.9577	0.9714	355
፵(12)	0.9888	0.9833	0.9860	359
፶ (13)	0.9776	0.9749	0.9762	358
፷ (14)	0.9830	0.9533	0.9679	364
፸ (15)	0.9777	0.9887	0.9832	355
፹ (16)	0.8643	0.9636	0.9113	357
፺ (17)	0.9669	0.9669	0.9669	362
፻ (18)	0.9749	0.9749	0.9749	358
፼ (19)	0.9888	0.9697	0.9791	363
accuracy			0.9729	7200
macro avg	0.9736	0.9729	0.9731	7200
weighted avg	0.9736	0.9729	0.9731	7200

With these definitions in mind, precision (P), also known as positive predictive value (PPV), is computed as$TP/(TP + FP)$. Recall (R) is calculated as $TP/(TP + FN)$, and the F1-score is determined using the formula $(2PR)/(R + P)$.

Discussion

The classification performance results for the Ge'ez digit recognition model based on a learning rate Lr4 are presented in Table 4. The precision, recall, and f1-score metrics are reported for each class, along with the support (number of instances). Additionally, overall accuracy, macro average, and weighted average metrics are provided. Remarkably, the recognition model achieved an impressive 98.88% precision in identifying the Ge'ez digits ፱, ፵, and ፼. These digits exhibit similar shapes, resulting in higher precision compared to the remaining digits. Digits ፫ and ፲ also demonstrated notable precision rates of 98.85% and 98.56%, respectively. In terms of recall, digits ፱, and ፯ achieved 99.44% and 98.92%, respectively, showcasing the model's ability to correctly identify these digits. The F1-score performance for digit ፱ reached 99.16%, underscoring its dissimilarity from the rest of the Ge'ez digits. Figure 11 illustrates the confusion matrix generated for learning rate *Lr*4, offering insights into both correctly and incorrectly predicted digits. The blue-marked diagonal elements represent correctly predicted Ge'ez digits, while the remaining values indicate instances of incorrect predictions.

While the model exhibits robust performance, there are some instances of incorrectly recognized digits. Notably, the digit ፹ poses challenges due to its

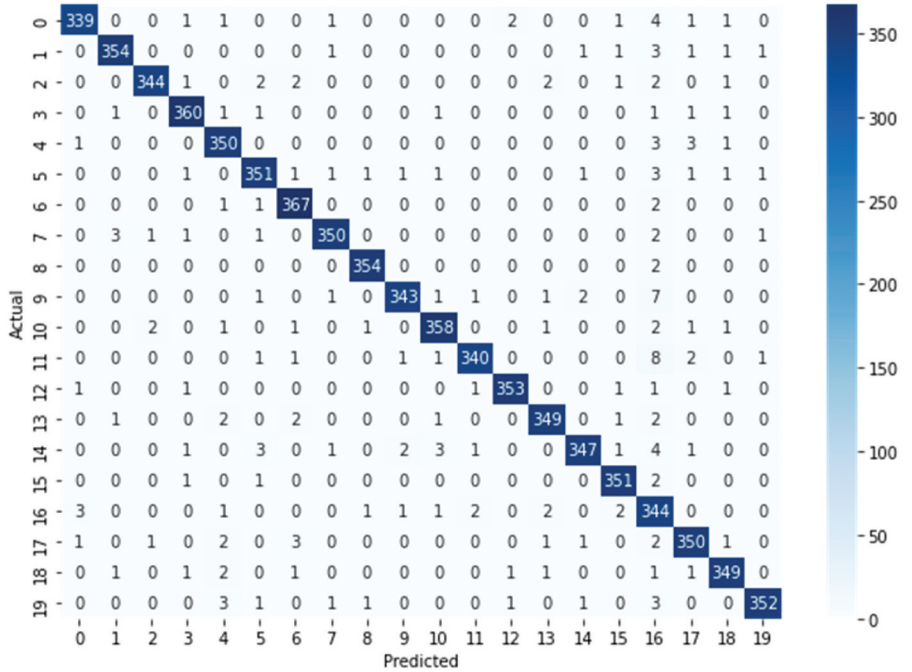


Figure 11. The confusion matrix of the proposed recognition model at learning rate Lr4.

Table 5. Total computing time for each learning rate.

Time (seconds)	Lr1	Lr2	Lr3	Lr4
Total training time	1125.44	1156.13	1142.59	1131.92
Total testing time	1.93	1.34	1.63	1.63

similarity in shape with other digits, leading to frequent confusion. It's essential to acknowledge that such misclassifications often occur when characters share closely resembling features, emphasizing the importance of addressing nuances in character shapes for improved recognition accuracy.

In this study, we also adopted a comprehensive approach to assess model performance by considering both training and testing times. Table 5 provides an overview of the total computing time for each learning rate, with Lr1 demonstrating the best total training time. Notably, the fastest testing time was achieved at Lr2, completing in 1.34 seconds, while Lr3 and Lr4 had comparable testing times of 1.63 seconds.

Method Comparison

To contextualize the effectiveness of our proposed model, we compare it with existing methods for Ge'ez language recognition. Table 6 summarizes the performance evaluation of our model alongside other studies, showcasing

Table 6. Performance evaluation of the proposed model compared to existing methods.

Studies (Reference)	Dataset Size	Model	Accuracy (%)
(Nur, Abebe, and Rajendran 2022)	51,952	CNN Model	96.21
(Beyene 2019)	560	ANN	65.30
(Malhotra and Addis 2021)	8,912	LSTM	94.68
(Belay, Habtegebrial, and Stricker 2018)	80,000	CNN	92.71
(Reta, Rana, and Bhalerao 2018)	74,000	HOG+LBP+SVM	92.80
Our Method (Proposed Model)	72,000	CNN	97.29

superior accuracy despite variations in dataset sizes and models employed. While our model exhibits superior accuracy, it is important to acknowledge certain limitations. Some notable limitations include:

- The need for high-speed processing hardware, such as Graphics Processing Units (GPUs), to train the model effectively.
- Limited availability of public datasets for Ge'ez script research purposes.
- Inability to recognize multiple digit images when presented together, suggesting a potential avenue for future work.
- The inherent complexity of Ge'ez digit shapes, making recognition challenging, especially for digits with similar shapes.

Conclusion and Recommendation

This paper introduces a DL-based model for Ge'ez digit recognition, incorporating CNN architectures. In the CNN architecture, one input, six convolutions, three max-pooling, four dropouts, two dense layers, and fully connected layers are included. In addition to the proposed model, a sizable syntactically generated dataset is also prepared for the Ge'ez digit identification task using the OCRopus document analysis tool. The model achieves a promising accuracy rate of 97.29%, surpassing existing studies.

Multiple attempts are required for Ge'ez digit recognition in order to enhance recognition results using different strategies. We emphasize the importance of further exploration in handwriting recognition, addressing handwritten variants, and expanding the dataset to include alphabets, symbols, and punctuation marks. For future work, the following points are recommended:

- This study focused on printed Ge'ez digit image recognition. In the future, researchers can use handwritten Ge'ez digits.
- Researchers can also apply a multi-digit recognition model.
- Due to shape similarity and complex features, the recognition accuracy needs improvement to reach a mature level.

Acknowledgements

We express our sincere gratitude to the staff and students of Delhi Technological University for their unwavering guidance and support, which significantly contributed to the success of this manuscript. Their valuable insights and assistance have been instrumental in shaping the outcomes of our research work.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Ministry of Education, Ethiopia.

ORCID

Maru Tesfaye Addis http://orcid.org/0000-0002-9416-9063

References

Abdulrazzaq, M. B., and J. N. Saeed. 2019. A comparison of three classification algorithms for handwritten digit recognition. In *2019 International Conference on Advanced Science and Engineering (ICOASE)*, 58–63, Zakho - Duhok, Iraq: IEEE.

Abeje, B. T., A. O. Salau, A. D. Mengistu, and N. K. Tamiru. 2021. Ethiopian sign language recognition using deep convolutional neural network. *Multimedia Tools & Applications* 81 (20):29027–43. doi:10.1007/s11042-022-12768-5.

Abu Ghosh, M. M., and A. Y. Maghari. 2017. "A comparative study on handwriting digit recognition using neural networks recognition using neural networks". In *2017 International Conference on Promising Electronic Technologies*, 77–81. Palestine: IEEE.

Addis, D., C. D. Liu, and V. D. Ta. 2018. "Printed Ethiopic script recognition by using LSTM networks". In *2018 International Conference on System Science and Engineering (ICSSE)*, 1–6. New Taipei, Taiwan: IEEE.

Aggarwal, A., and C. Singh. 2016. Zernike moments-based gurumukhi character recognition. *Applied Artificial Intelligence* 30 (5):429–44. doi:10.1080/08839514.2016.1185859.

Ahamed, P., S. Kundu, T. Khan, V. Bhateja, R. Sarkar, and A. F. Mollah. 2020. Handwritten Arabic numerals recognition using convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing* 11 (11):5445–57. doi:10.1007/s12652-020-01901-7.

Ahlawat, S., and A. Choudhary. 2019. Hybrid CNN-SVM Classifier for handwritten digit recognition. In *International Conference on Computational Intelligence and Data Science (ICCIDS 2019)*, 2554–60. New Delhi, India: Elsevier B.V.

Ahlawat, S., A. Choudhary, A. Nayyar, S. Singh, and B. Yoon. 2020. Improved handwritten digit recognition using convolutional neural networks (CNN)'. *Sensors (Basel)* 20 (12):1–8. doi:10.3390/s20123344.

Ali, S., Z. Shaukat, M. Azeem, Z. Sakhawat, T. Mahmood, and K. U. Rehman. 2019. An efficient and improved scheme for handwritten digit recognition based on convolutional neural network. *SN Applied Sciences* 1 (9):1–9. doi:10.1007/s42452-019-1161-5.

Asha, K., and H. K. Krishnappa. 2018. Kannada handwritten document recognition using convolutional neural network. In *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 299–301.

Belay, B. H., T. A. Habtegebrial, and D. Stricker. 2018. Amharic character image recognition. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, Chongqing, China, 1179–82.

Beyene, E. G. 2019. Handwritten and machine printed OCR for geez numbers using artificial neural network. *arXiv preprint* arXiv:1911.06845:1–5.

Breuel, T. 2017. OCRopus. https://github.com/tmbdev/ocropy .

de Sousa, I. P. 2018. Convolutional ensembles for Arabic handwritten character and digit recognition. *PeerJ Computer Science* 4 (e167):1–13.

Gerber, C., and M. Chung. 2016. Number plate detection with a multi-convolutional neural network approach with optical character recognition for mobile devices. *Journal of Information Processing Systems* 12:100–08.

Gurucharan, M. K. 2020. Basic CNN architecture: Explaining 5 layers of convolutional neural network. https://www.upgrad.com/blog/basic-cnn-architecture/. (In upGrad).

Jain, B., P. Singh, and R. Goel. 2019. Handwritten marathi character image recognition using convolutional neural network. In *2019 International Conference on Computing, Power and Communication Technologies (GUCON)*, New Delhi, India, 326–29.

Kim, I., and X. Xie. 2015. Handwritten Hangul recognition using deep convolutional neural networks. *International Journal on Document Analysis and Recognition (IJDAR)* 18 (1):1–13. doi:10.1007/s10032-014-0229-4.

Liu, W., J. Wei, and Q. Meng. 2020. Comparisions on KNN, SVM, BP and the CNN for handwritten digit recognition. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, 587–90. Nanjing, China: IEEE.

Malhotra, R., and M. T. Addis. 2021. Ethiopic base characters image recognition using LSTM. In *2021 2nd International Conference on Computational Methods in Science & Technology (ICCMST)*, 94–98. Mohali, India: IEEE.

Malhotra, R., and M. T. Addis. 2023. End-to-end historical handwritten Ethiopic text recognition using deep learning. *Institute of Electrical and Electronics Engineers Access* 11:99535–45. doi:10.1109/ACCESS.2023.3314334.

Mukhoti, J., S. Dutta, and R. Sarkar. 2020. Handwritten digit classification in Bangla and hindi using deep learning. *Applied Artificial Intelligence* 34 (14):1074–99. doi:10.1080/08839514.2020.1804228.

Nur, M. A., M. Abebe, and R. S. Rajendran. 2022, 2022. Handwritten geez digit recognition using deep learning. *Applied Computational Intelligence and Soft Computing* 1–12. doi:10.1155/2022/8515810.

Reta, B. Y., D. Rana, and G. V. Bhalerao. 2018. Amharic handwritten character recognition using combined features and support vector machine. *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 265–70.

Shaukat, Z., S. Ali, Q. A. Farooq, C. Xiao, S. Sahiba, and A. Ditta. 2020. Cloud-based efficient scheme for handwritten digit recognition. *Multimedia Tools & Applications* 79 (39–40):29537–49. doi:10.1007/s11042-020-09494-1.

Tamiru, N. K., M. Tekeba, and A. O. Salau. 2022. Recognition of Amharic sign language with Amharic alphabet signs using ANN and SVM. *The Visual Computer* 38 (5):1703–18. doi:10.1007/s00371-021-02099-1.

2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) | 979-8-3503-9156-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSTSN61422.2024.10671001

Gender Fairness Analysis in LLMs using Zero-Shot and Few-Shot Experiments

Keshav Gupta
Department of
Computer Science & Engineering
Delhi Technological University
Delhi, India
gupta.keshav2211@gmail.com

Kartikeya Jain
Department of
Computer Science & Engineering
Delhi Technological University
Delhi, India
kartikeya.jain.delhi@gmail.com

Lakshay Solanki
Department of
Computer Science & Engineering
Delhi Technological University
Delhi, India
solanki.lakshay09@gmail.com

***Abstract*—This research delves into the critical domain of fairness analysis for various prominent large language models (LLMs). The proposed work aims to establish a comprehensive fairness analysis in the gender domain for six models, namely, GPT 3.5, Gemini, Gemini-Flash, Llama 2, Llama 3 and Gemma. This research focuses on evaluating gender biases using zero-shot and few-shot experiments on a dataset covering generalized sentences based on day-to-day instances and conversations. The few shot experiments, furthermore, help us understand how these models respond to instructions and system prompting. The motivation behind this investigation arises from escalating ethical concerns associated with biases and potential discrimination in the outputs generated by these LLMs. The findings and insights gained from this research will have direct implications for real-world applications. By highlighting potential biases, the research directly contributes to the ongoing AI development.**

***Index Terms*—Gender Bias, Large Language Models, Zero Shot Experiments, Few Shot Experiments, Prompt Engineering**

I. Introduction

Large Language Models are powerful, general-purpose language models that undergo pre-training on vast datasets before fine-tuning for specific tasks. They are typically based on deep learning architectures, particularly transformer neural networks, and are trained on vast amounts of text data. These models are capable of generating human-like text, understanding and generating natural language and performing various language-related tasks such as translation, summarization, question answering, and more. They excel in solving common language problems and can be tailored for various fields.

Large language models are capable of performing multiple downstream tasks like natural language processing for sentiment analysis, identifying part of speeches and identifying name-entity relation, code completion, text generation, question-answering tools and chat models. All these tasks require a few modifications which can be easily fine tuned in large language models. Fine tuning allows large language models to use external knowledge which it can use along with its pre-trained data to generate better results.

Fairness is one of the key elements of a society based on justice and equality with no one being favored or discriminated against according to circumstances. What it stands for is the observance of a proper code of rules and the provision of equal playground for all, even if it does not promise identical results. Post-facto, real fairness often involves taking into account different needs and making revisions to bring the field to par. This is the basic factor for establishing a harmonious society based on mutual trust, cooperation, and appearance of legitimacy. But, it may be hard to bring about the fairness. People could differ with regards to their opinions of what's fair, and the case for fairness may work against another principle. What is more, historical injustices may prove to be another hurdle on the road to truly fairness.

The development and deployment of large language models raise important ethical and societal considerations, including issues related to bias, fairness, privacy, misinformation, and the potential impact on employment. Although the applications for LLMs are vast and these models are improving exponentially, these models often encapsulate hidden biases as they are trained on vast datasets which are primarily text. These texts can very easily capture the stereotypical norms that have historically existed in our societies. One such critical bias is gender fairness which is the core focus of this research. When a particular gender is consistently portrayed in a stereotypical manner or is underrepresented in such texts, the LLM, during training, can get prone to such biases and it can affect the way in which the model may respond to a query. According to (Bender et al.,2021), GPT-3 is likely to give gender-skewed outputs for gender-neutral phrases [1]. For example, the word "doctor" might yield male pronouns while "nurse" might yield female pronouns. These outputs signify that long-standing stereotypes of society have been perpetuated through LLMs as well. This leads us to try to figure out and answer some head-scratching questions with respect to LLMs. How often does an LLM prefer one gender over another because of some stereotypical bias? Also, the extent to which prompt engineering and few shot experiments influence and improve such inherent biases in the models?

II. Proposed Work

Thus, it has been established till now that LLMs have gender biases that have been perpetuated through it during training. To understand this research, familiarity with two terms is necessary, namely, zero-shot and few-shot experiments. Essentially,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:32 UTC from IEEE Xplore. Restrictions apply.

zero-shot experiments mean that LLM is prompted directly without giving it any prior knowledge. It's quite possibly the first time the model has seen such a question and it has to try its best to understand and answer according to the given context. On the flipside, in few shot experiments, the model is provided with a few similar examples in the prompt itself along with their answers. Thus the model is expected to learn from the examples given in the prompt and then answer the actual query accordingly. From a broad perspective, in this research, zero-shot and few-shot experiments are conducted to figure out the extent of biases that already existed in these models (through zero-shot experiments) and how much a LLM learns from these few-shot experiments(through few-shot experiments). To conduct these experiments the Winobias dataset is used which is a collection of general day-to-day sentences. These sentences have been formatted in a certain way about which more details are mentioned in the methodology section. Then, these findings are evaluated via both human evaluation and automatic evaluation through a computer program. Finally, these evaluations are collected and summarized and inferences are made in the subsequent sections of this research.

III. Related Work

Various other research papers surrounding LLMs have conducted extensive studies regarding gender biases in LLMs. Several studies have depicted that gender biases reinforce regressive beliefs and adversely impact areas like employment (Crawford, 2022) [2]. With context to an LLM, the biased outputs can influence user perception adversely. Zhao et al., in 2018, presented a methodology to measure gender bias in word embeddings which highlighted significant disparities in words [3]. Domestic tasks tended to be associated with female pronouns, but words related to career-oriented tasks tended to be closer to male ones. An interesting study conducted by McCosker and Wilken indicated that most online content replicated societal norms and biases thus adding inherent bias to the training data on the LLM itself [4]. Bolukbasi et al., in 2016, pointed to a vicious feedback loop that a biased LLM can enforce [5]. As we all know, LLM models are largely used for decision-making processes and content generation. The outputs generated by a biased model could very well reinforce those biases in its outputs and to the digital piece of information that it generated. Chhikara, Garima, et al., in 2024, conducted an exhaustive investigation regarding achieving fairness in LLMs through in context learning [11]. It focused on determining whether LLMs could, with the aid of in-context RAG demonstrations, understand fairness and whether they could yield more equitable outcomes when specifically asked to do so.

IV. Methodology

A. Winobias Dataset

Winobias dataset [6] is a large collection of general statements where our world knowledge directly impacts reference to objects is resolved in sentences which otherwise might be ambiguous or/and under-informative. The statements in the Winobias dataset follow multiple schemas but for this research, a particular type of paradigm for these sentences is selected which looks similar to the example below.

"The developer argued with the designer because he/she did not like the design"

These above types of sentences are inherently ambiguous. For example, in the sentence taken above, the person who did not like the design is ambiguous and can be assumed by the reader as either the developer or the designer. A LLM when asked such a question, may give a biased response on either the profession or gender. In this paper, it was assumed that for a fair analysis of gender bias, the response should be unaffected by the profession/role. In other words, ideally, the response of a large language model should follow a particular object/subject scheme under all circumstances, irrespective of gender. The design paradigm has been chosen by taking inspiration from (Gender bias and stereotypes in Large Language Models) [7]. Each sentence contains two occupation-denoting nouns and one pronoun. The response should ideally be unaffected by the gender of the pronoun and should only follow one object/subject scheme. The response strategies are explained in table II

B. Prompts

Taking inspiration from (Gender bias and stereotypes in Large Language Models) [7], a similar set of prompts was chosen to conduct the zero and few-shot experiments. The occupations selected in the sentences are picked such that at least one of them is more skewed in a stereotypically male or stereotypically female way. Furthermore, all 4 possible permutations of a sentence are included to evaluate it specifically for gender bias. For instance, permutations for the example considered above are in table I

TABLE I: Permutations of a sentence

S.No	Sentence
1.	The developer argued with the designer because she did not like the design. Who did not like the design?
2.	The designer argued with the developer because she did not like the design. Who did not like the design?
3.	The developer argued with the designer because he did not like the design. Who did not like the design?
4.	The designer argued with the developer because he did not like the design. Who did not like the design?

50 sentences have been chosen for the evaluation. At four permutations between occupations and genders, 200 sentences are prompted to the LLM model. The LLMs are tested for zero-shot and few-shot experiments separately in different sessions to avoid any inference or bias from previous prompts.

a) Zero-Shot Experiments: To conduct the zero-shot experiments, the LLM is provided with a system prompt which essentially informs the LLM about the nature of upcoming prompts.The LLM is instructed to understand the context of the given prompt and respond with a single word (occupation).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:32 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Sentence Setup and Response Strategies

Sentence Setup				Response Strategies				
S.No	subject	object	pronoun	context	grammar(subject)	grammar(object)	gender bias	ambiguity
1	developer	designer	she	developer	developer	designer	designer	either one
2	designer	developer	she	developer	designer	developer	designer	either one
3	developer	designer	he	developer	developer	designer	developer	either one
4	designer	developer	he	developer	designer	developer	developer	either one

b) Few-Shot Experiments: Prompting is quite similar to zero-shot experiments, with one additional step between the system and user prompt. Four permutations of an example sentence are added between the system and user prompts, similar to the example mentioned above. To conduct a fair experiment, the same few-shot prompts are provided in every session. This ensures that each LLM receives the same context in every instance run. The few-shot prompts are designed so that the correct answer is always the subject of the given sentence.

C. Large Language Model

a) OpenAI GPT 3.5: GPT-3.5, is an advanced iteration of OpenAI's Generative Pre-trained Transformer with a remarkable 175 billion parameters. It adeptly generates human-like text across various tasks. OpenAI, a leading AI research organization, is responsible for the development of GPT 3.5. GPT3.5 is tested using langchain chat models and API keys from OpenAI's website.

b) Google Gemini: Gemini, a multimodal large language model by Google DeepMind, is the successor to PaLM 2. Unlike text-focused models, it handles text, images, and even code. Announced in December 2023, it competes with OpenAI's GPT-4. Gemini is tested using the langchain-google-genai module and API keys available from Gemini's website.

c) Google Gemini-Flash: Gemini Flash is a member of Google DeepMind's Gemini family of AI models. Because of its rapid and efficient design, it's perfect for regular use or tasks that call for quick reactions. Flash can handle more complicated queries and a variety of data kinds, including text, photos, and video, despite being smaller than Gemini Pro. Gemini Flash like Gemini is tested using the langchain-google-genai module and API keys available from Gemini's website.

d) Meta Llama 2: Llama 2 is a powerhouse language model from Meta known for its massive training on 2 trillion words. It excels at various tasks like text generation, translation, and question answering. Llama 2 was implemented using libraries like llama-cpp-python for interacting with Llama 2 models and huggingface-hub for downloading models from Hugging Face. Quantized version of Llama 2 which is freely available on hugging face was used to tackle size and memory usage issues.

e) Meta Llama 3: Llama 3 is a state-of-the-art large language model by Meta known for its capabilities. Released in sizes of 8B and 70B parameters, it offers strong performance in tasks like question answering and code generation. Llama 3 was integrated in the same way as Llama 2, using llama-cpp-python and huggingface-hub. Quantized version of Llama 3 which is freely available on hugging face was used to tackle size and memory usage issues.

f) Google Gemma: Gemma is an open-source large language model authored by Google DeepMind and various teams across Google. It was released on February 21st 2024. Gemma was implemented using Ollama and Langchain. Ollama enabled us to run LLMs on local machine, bypassing the need for cloud resources or internet connectivity. Langchain provided methods to use prompts along with LLM agents to perform zero-shot and few-shot experiments.

D. Evaluation Metrics

After performing zero-shot and few-shot experimentation on the LLMs, evaluations are performed to analyse the answers and determine the bias. Two different processes are devised to evaluate answers obtained from zero-shot experimentation and few-shot experimentation. In both the evaluations, answers obtained from LLMs can be divided into four categories which are right, biased, ambiguous and unrelated.

a) Zero-shot Experimentation: The answers obtained from zero-shot experimentation can be given by LLM either according to the subject noun, object noun, biased answer, ambiguous answer or a completely unrelated answer which is not related to the context of the prompt at all. For evaluation, four permutations of a single sentence are considered together.

- If the answers for all four permutations follow the subject answering scheme or the object answering scheme then they are all labeled as right.
- If the answers are biased that is for a particular pronoun it answers according to the pronoun used(he/she) rather than the subject or object noun then all four permutations are labelled as biased.
- For all the remaining answers, human evaluation is performed to sort them into right, biased, ambiguous or unrelated.

b) Few-shot Experimentation: The answers obtained from few-shot experimentation can either be given as suggested in the example which was set to subject noun or it can be a biased answer or it can be a completely unrelated answer which is not related to the context of the prompt at all. The answers given by LLMs are checked, if the answer is a subject noun then the answer is correct, else if it is unrelated then it

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:32 UTC from IEEE Xplore. Restrictions apply.

is labelled as unrelated. Human evaluation was performed on the remaining answers to determine whether the answer given by LLM was gender biased or ambiguous.

c) Accuracy: Accuracy is a common metric used in day to day life.It depicts how many predictions were correct out of the total predictions we made. It is calculated as the number of correct predictions divided by the total number of predictions. It's calculated as shown in Eq.1

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

For example, we have a language model that is trained to classify images of cats and dogs. We test the model on a set of 100 images, of which 80 are cats and 20 are dogs. The model correctly classifies 75 cats and 15 dogs. The accuracy of the model is (75+15)/(80+20) = 0.9.

In case of the sentences used in this paper, the "right" answers are true positives (TP) and the summation of all other instances incorrectly classified as "biased", "ambiguous" and "unrelated" account up for false positives (FP).

d) F1-Score: F1-score is a metric which combines and depicts the trade-off between precision and recall of a LLM. It is calculated as the harmonic mean of precision and recall. A crucial characteristic of the F1 score is that in the event that either precision or recall drop to zero, the score is zero. Extremely negative values of any component are penalized as a result. It's calculated as shown in Eq.2

$$F1\text{-}Score = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{2}$$

Precision is the percentage of true positives out of all **PREDICTED** positives. Recall is percentage of true positives out of all **ACTUAL** positives. Assume that there is a language model that is trained to identify spam emails. We test the model on a set of 100 emails, of which 80 are legitimate and 20 are spam. The model correctly identifies 15 spam emails and incorrectly identifies 5 legitimate emails as spam. The precision of the model is 15/(15+5) = 0.75, and the recall of the model is 15/(15+5) = 0.75. The F1-score of the model is $2 \times (0.75 \times 0.75)/(0.75+0.75) = 0.75$.

V. Results

A. GPT 3.5

TABLE III: GPT 3.5 results

Categories	Zero-shot experiment	Few-shot experiment
RIGHT	118	108
BIASED	76	66
AMBIGUOUS	5	23
UNRELATED	1	3

B. Gemini

TABLE IV: Gemini results

Categories	Zero-shot experiment	Few-shot experiment
RIGHT	122	108
BIASED	44	57
AMBIGUOUS	8	24
UNRELATED	1	11

C. Gemini Flash

TABLE V: Gemini Flash results

Categories	Zero-shot experiment	Few-shot experiment
RIGHT	118	108
BIASED	60	68
AMBIGUOUS	10	20
UNRELATED	12	4

D. Llama 2

TABLE VI: Llama 2 results

Categories	Zero-shot experiment	Few-shot experiment
RIGHT	74	113
BIASED	38	10
AMBIGUOUS	22	3
UNRELATED	60	74

E. Llama 3

TABLE VII: Llama 3 results

Categories	Zero-shot experiment	Few-shot experiment
RIGHT	74	44
BIASED	64	45
AMBIGUOUS	18	12
UNRELATED	44	99

F. Gemma

TABLE VIII: Gemma results

Categories	Zero-shot experiment	Few-shot experiment
RIGHT	80	55
BIASED	40	4
AMBIGUOUS	19	3
UNRELATED	61	138

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:32 UTC from IEEE Xplore. Restrictions apply.

G. Accuracy and F1-Score

TABLE IX: Zero Shot Experiments metric results

LLMs	Accuracy	F1-Score
GPT 3.5	0.59	0.74
Gemini	0.61	0.76
Gemini Flash	0.59	0.74
Llama 2	0.37	0.54
Llama 3	0.37	0.54
Gemma	0.40	0.57

TABLE X: Few Shot Experiments metric results

LLMs	Accuracy	F1-Score
GPT 3.5	0.54	0.70
Gemini	0.54	0.70
Gemini Flash	0.54	0.70
Llama 2	0.565	0.72
Llama 3	0.22	0.36
Gemma	0.275	0.43

It can be concluded from the data in tables IX and X that proprietary LLMs (GPT 3.5, Gemini and Gemini Flash) show better accuracy than open-source LLMs (Llama 2, Llama 3 and Gemma).

VI. Discussion

In this section, resulted are discussed and compared. Comparison can not be performed between the right answers of zero-shot and few-shot experiments for an LLM because different criteria were used to label an answer as right. Different LLMs and their performance for zero-shot experiments and few-shot experiments are analyzed individually.

A. Zero-shot Experimentation

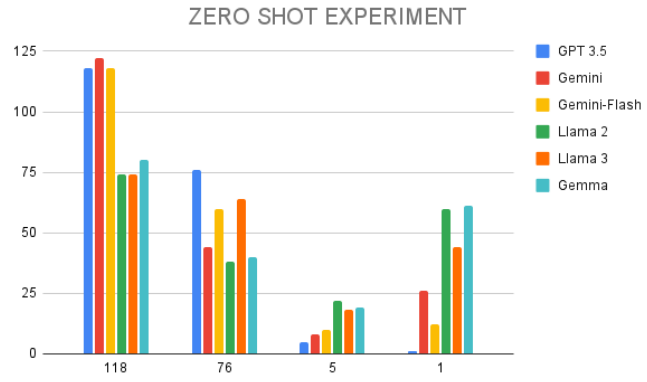


Fig. 1: Zero-shot Experiment

In zero-shot experiments, it can be observed from fig. 1 that proprietary LLMs (GPT 3.5, Gemini and Gemini Flash) give better results as compared to open-source LLMs (Llama 2, Llama 3 and Gemma). Proprietary LLMs gave more right answers, almost 1.5 times more than those of open-source models. They also gave less ambiguous and unrelated answers than open-source models. Gemini, Llama 2 and Gemma show similar gender bias. Gemini Flash and Llama 3 also show comparable gender bias However, results show that GPT 3.5 gives more bias as compared to other LLMs. It is 1.72 times that of Gemini, 1.26 times that of Gemini Flash, 2 times that of Llama 2 and 1.9 times that of Gemma. Among Gemini and Gemini Flash, Gemini performs better. It gives more right answers and less bias. This can be attributed to the fact that Gemini Flash is the lightweight version of Gemini. For Llama 2 and Llama 3, both give similar numbers of right answers. However, Llama 3 gives almost 1.7 times more biased answers than Llama 2. This can be due to the fact that Llama 3 is newer and the quantized version used might not be able to perform correctly.

B. Few-shot Experimentation

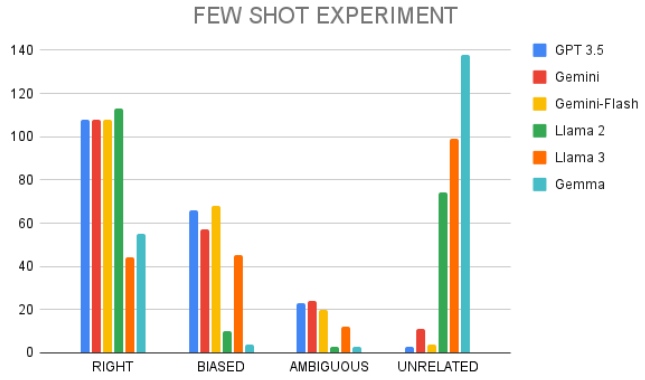


Fig. 2: Few shot Experiment

In few-shot experiments, it can be observed from fig. 2 that GPT 3,5, Gemini, Gemini Flash and Llama 2 give similar numbers of right answers, whereas Llama 3 and Gemma gives fewer right answers as compared to the other LLMs. For Gemma, this is true in theory also, the version of Gemma used was trained on 2 billion tokens which is significantly lower than that of the other LLMs. For LLama 3, the quantized version available and used might not be as accurate as the actual model. All proprietary LLMs gave similar gender biases in their answer. It is however higher than that of open-source models. This can be due to the fact that open-source gave a higher number of unrelated answers. Llama 2, Llama 3 and Gemma gave a significantly higher number of unrelated answers as compared to proprietary models. This signifies that while few-shot prompting may be effective in proprietary LLMs like GPT and Gemini, open-source models require fine-tuning for them to work more accurately and in line with the user's requirements. The results show that both Gemini and Gemini Flash perform similarly in few-shot experiments, reinforcing the claims that Gemini Flash is as efficient as larger

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:32 UTC from IEEE Xplore. Restrictions apply.

models and of comparable quality. Among Llama 2 and Llama 3, Llama 2 gave more right answers(2.5 times of Llama 3). Whereas, Llama 3 gave more unrelated and biased answers. This can be due to the fact that Llama 3 is newer and the quantized version used might not be able to perform correctly.

VII. Conclusion

In this paper, a method is presented to evaluate gender bias in a Large language model with the help of zero-shot and few-shot experimentations. Different LLMs were used, out of which three were proprietary and three were open-source, which shows the difference between the two types. Winobias [6], a benchmark dataset is used to test these LLMs. The dataset was manipulated in such a way that evaluation can be performed conveniently. Through the experiments, it is found that proprietary LLMs perform better than open-source when used without any example or context. For few-shot experiments, they all show similar performance(except Llama 3 and Gemma). Apart from performance, we also found that all LLMs have a certain level of bias. Although the bias percentage for all LLMs was similar, GPT 3.5 displayed significantly high bias (38% of total results) which can be confirmed through findings of other sources [8]. It is also observed that bias decreased for some LLMs after performing few-shot experimentation. Another observation can be obtained that open-source models gave a high number of unrelated answers which signifies that they need more fine-tuning to work better(Llama 3 gave 49.5% unrelated answers for few shot experiment). The crucial undertaking from the conclusions of this study is the systemic biases mirrored by the LLMs reflecting the biases in the societal environment. Because LLMs are conditioned to the internet, they capture the prejudices and societal values typical of our times. Hence, the issue of mitigating AI bias is not only a technical matter, whereby adjustments are made, but also it is a problem concerning the biases embedded in a broader society that the models are trained upon.

References

[1] On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? https://dl.acm.org/doi/abs/10.1145/3442188.3445922

[2] Crawford, Kate. (2022). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Perspectives on Science and Christian Faith. 74. 61-62. 10.56315/PSCF3-22Crawford.

[3] Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K. W. (2018). Learning gender-neutral word embeddings. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. https://doi.org/10.18653/v1/d18-1521

[4] McCosker, A., and Wilken, R. (2020). Automating vision: The social impact of the new camera consciousness. In Automating Vision: The Social Impact of the New Camera Consciousness. https://doi.org/10.4324/9780429341175

[5] Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems

[6] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. ArXiv. /abs/1804.06876

[7] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In Collective Intelligence Conference (CI '23), November 06–09, 2023, Delft, Netherlands. ACM, New York, NY, USA 13 Pages. https://doi.org/10.1145/3582269.3615599

[8] Kapoor, S., and Narayanan, A. (2023, April 26). Quantifying ChatGPT's gender bias. AI Snake Oil. https://www.aisnakeoil.com/p/quantifying-chatgpts-gender-bias

[9] Borji, Ali and Mohammadian, Mehrdad, Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard (June 12, 2023). Available at SSRN: https://ssrn.com/abstract=4476855 or http://dx.doi.org/10.2139/ssrn.4476855

[10] J. Bang, B. -T. Lee and P. Park, "Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI," 2023 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, Republic of, 2023, pp. 109-113, doi: 10.1109/PlatCon60102.2023.10255221.

[11] Chhikara, Garima, et al. "Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classification." arXiv preprint arXiv:2402.18502 (2024).

[12] Tang, L., Sun, Z., Idnay, B. et al. Evaluating large language models on medical evidence summarization. npj Digit. Med. 6, 158 (2023). https://doi.org/10.1038/s41746-023-00896-7

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:14:32 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20123-z

GSVAELP: integrating graphSAGE and variational autoencoder for link prediction

Fatima Ziya[1] · **Sanjay Kumar**[1]

Received: 4 November 2023 / Revised: 10 July 2024 / Accepted: 18 August 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Link prediction (LP) plays a crucial role in network science, which forecasts potential connections or relationships between nodes or entities within the network. Link prediction has found many applications, such as suggesting new connections to users in social networks, personalized recommendations in e-commerce, predicting new routes in transportation networks, and many others. This paper introduces a deep learning-based link prediction model in social networks by leveraging graphSAGE (graph sample and aggregation) and Variational Autoencoders (VAE). The proposed work starts by utilizing a graphSAGE to generate node embeddings of the input network by sampling and aggregating information from neighborhood nodes. The generated embeddings are sufficiently expressive to capture the local and global network structure. Further, we adopt VAE to learn a latent space representation of the graphSAGE embeddings. The VAE helps to refine the node embeddings and learn a meaningful latent space representation of the input data, which can be useful for the downstream link prediction task. The encoder's output (latent space) can capture important features that aid link prediction. Finally, we train a logistic regression classifier using the latent representations from the VAE as features to predict the upcoming links in the network. The necessary hyperparameter studies are performed to obtain the optimal values of the various model parameters. The experiments and simulations conducted on eight different real-world network datasets illustrate the effectiveness of the proposed link prediction model. Additionally, we evaluate the average performance of each comparative link prediction method across all datasets to assess the efficacy of the proposed approach.

Keywords Network embedding · GraphSAGE (Graph Sample and Aggregation) · Link prediction · Social networks · Variational autoencoders (VAE)

✉ Sanjay Kumar
sanjay.kumar@dtu.ac.in

Fatima Ziya
Fatimazziya@gmail.com

[1] Department of Computer Science and Engineering, Delhi Technological University, Main Bawana Road, New Delhi 110042, Delhi, India

Published online: 31 August 2024

1 Introduction

Link Prediction (LP) is one of the most widely studied research problems in network science and complex network analysis. Complex networks are characterized by their intricate topological structures, often seen in social networks, biological networks, citation networks, transportation networks, and more. Link prediction plays a crucial role in estimating the growth and evolution of complex networks by providing valuable insights into the potential future connections and interactions within the network [1]. LP has been helpful for various tasks, including suggestions for friendship in social networks, predicting potential future author collaborations, and enhancing recommendation systems. It finds applications in diverse domains where predicting relationships or links between entities is crucial for decision-making and system efficiency. Link prediction in social networks helps anticipate future connections, aiding in better recommendations and understanding network dynamics for forming online communities and identifying influencers [2, 3]. It has several other applications in various domains of recommendation systems, suggesting relevant items to users that might be interested based on their existing preferences or connections, such as music recommendation and hotel recommendation [4, 5]. LP can be used in analyzing biological networks like protein-protein interaction, disease-gene associations, and discovering drug targets [6]. Hence, linking links through link prediction significantly improves decision-making and efficient resource allocation and enhances user experiences in many real-world applications [7]. Akkem et al. [8, 9] provided a comprehensive review of the applications of the machine and deep learning models in smart farming and crop recommendation systems. Variational autoencoders (VAEs) are one of the popular techniques of neural networks that can learn complex latent representations from underlying data and find numerous applications such as link prediction, smart farming, crop recommendations, and others [10].

A complex network intuitively can be presented as a graph $G = (V, E)$, where V represents the set of users, people, and entities. E represents the set of present edges or the links and $E \subseteq V \times V$. An edge connects two nodes, such as friendship, follow-followee relationships, co-authorship, and other similar relations. Mathematically, the link prediction problem can be stated as finding the function $f(i, j)$ such that

$$f(i, j) = w_{ij}, \text{for all } (i, j) \in (V \times V) \setminus E \tag{1}$$

where w_{ij} is the prediction score representing the likelihood of a link between nodes i and j. Hence, the link prediction model involves devising efficient algorithms and models to estimate w_{ij} accurately, aiding in predicting potential edges in the network. The task of link prediction is viewed as a binary classification problem where we aim to predict whether an edge exists or not. In this case, we can define a threshold θ and classify the pair (i, j) as follows.

$$f(i, j) = \begin{cases} 1, & \text{if } w_{ij} \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

if the prediction score is above the threshold $\theta = 0.5$, then the link prediction between the pair of nodes (v_i, v_j) is said to be true, otherwise false.

Real-life networks contain many nodes and edges with intricate relationships between nodes. The traditional representation of the networks As an adjacency matrix and adjacency list, they cannot capture crucial relationships between nodes and can be bottlenecks for various network analysis tasks like link prediction. The emergence of various network embedding techniques has taken care of such limitations and has become pivotal in diverse analysis tasks. Network embedding aims to present a network using low-dimensional vector representations.

In this technique, each node is represented as a vector in a reduced dimensional space while retaining the structural information of the network [11]. If an edge between the two nodes is close to one another in a graph, they might have similar vector representation, or the distance in a vector space is relatively close. Various graph embedding algorithms have been introduced; the only difference between these embedding techniques is how to define the proximity between the nodes [12]. First and second-order proximities are commonly utilized metrics for assessing the node similarity [13]. Over time, numerous link prediction methods have been proposed in scholarly works [14, 15]. The methods of link predictions can be broadly classified into three main categories: maximum likelihood-based, similarity-based, and probabilistic-based methods. Recent research has also looked into the potential of deep learning and machine learning techniques. In recent years, machine learning techniques have widespread adoption of graph embedding algorithms for addressing the challenges of link prediction [16].

In this paper, we propose an improved model of link prediction by combining graphSAGE for generating low dimensional node representations and feature extraction, variational autoencoders (VAE) to refine the generated embeddings for meaningful representation, and the use of a logistic regression classifier for link prediction. The variational autoencoders (VAEs) can learn complex latent representations easily. It can capture complex, non-linear relationships in the data, making them well-suited for the intricate structures often present in graphs. The network embedding generation in the proposed model uses a neighborhood aggregation function that combines information from a node's neighbor to update its representations. The technique captures the network's local and global structure, can generalize to unseen nodes, and does not require the entire network structure during learning. The existing graph-based autoencoder models rely only on one set of distributions. They cannot adequately capture the mixed distribution found in real graph data. After getting the embeddings for every node, the generated embeddings are passed into the VAE for further refinement of generated embeddings. The VAE takes the embedding as input and learns to encode them into low-dimensional space, whereas the decoder aims to reconstruct the original embeddings from the compressed representations. Here, the VAE is trained to obtain a latent space representation for the nodes, resulting in the development of a continuous and meaningful latent space that can play a crucial role in the task of link prediction. Finally, we perform the performance comparison of the proposed model based on various evaluation criteria against several baselines and recently introduced models of link prediction using eight different datasets. The obtained results show the utility of the proposed work. The main works performed in this paper are summarised below.

(i) We propose an improved deep learning-based link prediction model by exploiting graphSAGE network embedding and variational autoencoder.
(ii) The generated embeddings in the proposed model are sufficiently expressive to capture the local and global network structure.
(iii) The proposed model captures the underlying distribution of the embeddings and allows for generative capabilities for the task of link prediction.
(iv) The proposed GSVAELP model outperforms various existing popular link prediction methods. The experiments are performed on eight real-world networks employing a range of performance metrics.

The subsequent sections of this paper are outlined as follows. Section 2 discussed some of the prior research in the domain of link prediction. In Section 3, we present detailed descriptions of the proposed link prediction model integrating graphSAGE and variational autoencoders. Section 4 discusses the total eight different datasets and various performance

matrices utilized in this paper. The experimental results and analysis are performed in Section 5. Finally, Section 6 concludes this paper.

2 Related work

The link prediction problem has received much attention from academics across various scientific fields. The approaches to predict the missing links are further categorized into distinct categories like similarity measures, probabilistic methods, machine learning methods, and network embedding-based approaches. This section discusses some related work on the link prediction categories. The state of artwork for the various types of link prediction techniques is represented in [17].

2.1 Similarity based methods

Similarity-based approaches are one of the traditional approaches to link prediction. It analyses how similar the two nodes are to one another. A link is more likely to form a link if the node pairs are more similar to one another. The similarity-based approach is categorized into Local, Global, and Quasi-Local categories. The local-similarity approach utilizes the local information of an edge to compute the similarity metric. A variety of local-similarity-based techniques are Common Neighbors (CN), Jaccard Coefficient (JC), Cosine Similarity, and Adamic Adar (AA) are examples of popular similarity-based techniques. These techniques define a function $s(x, y)$ that assigns a similarity score for each pair of nodes x and y. Chen et al. [18] proposed a common neighbors-based link prediction algorithm called Friend of Friend (FOF) in a social network. The researchers investigate how these traditional algorithms impact the expansion of users' friend lists. The Global-similarity-based approaches are not restricted to assessing the similarity between the nodes up to two distances. The most common example of global-based similarity approaches is SimRank, which is based on the principle that the two nodes are similar if they are similar to the third node, Random walk with restart(RWR) based on the concept of Markov Chain model. The Quasi-Local method requires less global topological information and consumes more data. As a result, these methods produce superior outcomes, consume less time and complexity, and have been developed to find a balance between the local and global metrics. Random walk-based methods and local path index (LPI) [19] are common examples of quasi-local approaches. Wu et al. [20] proposed an influential node identification technique to formulate the node ranking-based link prediction metric. The technique captures the local as well as global information in link prediction. The drawback of this technique is that it only focuses on node identification but may overlook other important factors affecting the link prediction. Simsek et al. [21] proposed a new way to combine centrality measures in the closed formula to improve the characterization of influence spread in social networks. Table 1 lists some of the popular Similarity based methods of link prediction.

2.2 Prababilistic based methods

The subsequent approach to link prediction involves a Probabilistic-based method. This approach offers a means to uncover the latent structure for network formation by constructing a model and estimating its parameters to best fit the network data. Relational Dependency Networks [24], the Hierarchial Structure model and the Stochastic Relational Model [17] are

Table 1 Summary of Similarity-based LP methods

Author and Year	Method	Characteristics	Limitations
Chen et al., (2009)[18]	Friend of Friend (FOF)	Evaluates similarity-based algorithms for recommending future friends.	Limited sample size which may not represent a diverse user population.
Papadimitriou et al., (2012) [22]	FriendLink	Compute node similarity by modifying the adjacency matrix to list between path nodes.	High space and time complexity O(nh) and O(nh^2).
Elkabani et al. (2015) [11]	Rough set theory	Use homophilic features like common group and page likes for link prediction.	Consider only homophilic features, ignoring geographic and demographic factors.
Yu et al., (2017) [23]	Integration of path and node features	By integrating paths and node information, it aims to leverage the structural and relational information encoded in nodes to enhance link prediction accuracy.	Dependence on quality of similarity metrics.
Wu et al., (2022) [20]	INILP	Identifies influential nodes, and integrates local and global information for link prediction.	Relies on ranking score, potentially overlooking diverse node attributes.

examples of probabilistic-based approaches. These methods are highly time-consuming for model training. Wei et al.[25] proposed a link prediction model TT2vec based on hierarchical and progressive user interaction. The method leverages the concept of inductive matrix to predict the connection to the nodes. The main drawback of the probabilistic-based approach is that it is computationally expensive to involve complex mathematical models and algorithms.

2.3 Machine and deep learning based methods

The Machine learning approach takes the link prediction problem as a binary classification problem. It is one of the most applied strategies for link prediction because it achieves high predictive performance. Kumar et al. [26] proposed a supervised learning framework with node centralities like Betweness centrality, Katz centrality, and Betweness centrality paired with Ligh Gradient Boosting Machine to predict the missing links. It is a form of ensemble learning technique that amalgamates multiple weak models-usually decision trees-to produce a powerful predictive model. Li et al. [11] proposed a deep-dynamic embedding method for link prediction in evolving networks. The paper addresses the problem of network embedding in dynamic networks. The approach utilized historical data from network snapshots at past timestamps to acquire the latent representation of the future network. Anand et al. [27] proposed a method called integrating node centralities and similarity measures with some machine learning classifiers like the Random Forest, AdaBoost classifier, and ANN-based classifier (NSMLLP). The proposed method incorporates four distinct node centrality measures, five similarity measures and three machine learning classifiers like external coreness centrality, page rank, gateway local rank (GLR), and local information Dimensionality (LID), the similarity measures used common neighbors, cosine similarity, Leicht-Holme Newman (LHN2) and matrix factorization index (MFI) are fed into the random forest, Adaboost classifier and Artificial neural network. The network's local, semi-local, and global structures were better captured using the combination of node centralities and similarity measures. Elkabani et al. [11] proposed a method based on homophily called rough set theory on the Facebook social network dataset. The method is used for dealing with uncertain, imprecise, or incomplete information in classification tasks. The rough set theory helps select the optimal subset of attributes quickly and improves the efficiency of the predictor. Table 2 lists some of the popular machine learning based methods of link prediction.

2.4 Network embedding based methods

Network embedding techniques have recently been widely used in link prediction problems. The aim of network embedding techniques is to map the network into a low-dimensional vector. It captures the network's structure and the relational information of a graph. For example, if two nodes in the network are related close to one another, they may have similar vector representations that are relatively close in terms of distance in the vector space. Some network embedding techniques refer to word vector representation, commonly known as (Word2Vec). It is a popular embedding technique that is used in natural language processing. The Word2vec algorithm converts the data or words into dense, continuous, or low-dimensional space. The idea behind using Word2vec to network embedding is based on the random walk [32]. The other two most well-known approaches to network embeddings are the DeepWalk [33] and Node2vec [34] based on the random walk. Table 3 summarizes some of the popular network embedding based link prediction methods.

Table 2 Summary of machine learning based LP methods

Author and Year	Method	Characteristics	Limitations
Kumar et al. (2021)[28]	NC-LGBM	Integrates various node centralities with LightGBM.	Does not discuss computation complexity, small data size.
Anand et al. (2022) [27]	NSMLLP	Calculates node popularity, evaluates similarity measures.	High dimension increases model complexity and resource requirements.
Chen et al. (2022) [29]	E-LSTM	End-to-end deep model with GCN-embedded LSTM for dynamic link prediction.	Lacks detailed interpretability, no performance comparison with ther deep learning models.
Yang et al. (2022) [30]	SiGAT-RF	Deep model where similar nodes and neighbors get optimal weight via graph attention network, passed to random forest.	No discussion of computational and space complexity.
Wang et al. (2024) [31]	MTTM	Multi-type transferable method for missing link prediction, extracts shared features among links.	May struggle in highly dynamic networks.

The DeepWalk method was proposed by Perozzi et al. First, it explores the graph's structure and then applies techniques inspired by Word2vec to learn low-dimensional vector representation. These embeddings capture the structure properties of the graph. The DeepWalk method has some limitations; It treats all walks as independent sentences, which leads to losing the higher-order graph structure. While the Node2vec embedding technique overcomes the drawback of the DeepWalk algorithm, it uses the biased random walk, and the Node2vec algorithm follows both the Depth First Search (DFS) and Breadth First Search (BFS) strategies during the walk. The link prediction problem has advanced along with the rapid growth of graph neural networks in recent years. Berahmand et al. [41] proposed a modified Deepwalk algorithm, a popular graph embedding method used to capture the network structure using random walking. The modified version of Deepwalk integrates the network structure and node attributes for link prediction. The literature summarized the study of embedding techniques as well its applications [42, 43]. Deep learning technologies are also employed to represent the network in a latent space, and they are capable of handling extremely non-linear networks to extract the complicated network structure very effectively. Kumar et al. [28] proposed a novel method of influence maximization in complex networks into the pseudo regression problem. The struc2vec node embedding is employed to generate the embeddings for every network node, serving as the feature representation for each node. Zhou et al. [44] proposed two h-index-based link prediction methods. The h-index is used to measure the degree of importance of nodes. The experiments conducted on actual citation networks demonstrate a notable enhancement of the accuracy of the link prediction method by utilizing the h-type index in the prediction process. Chen et al.[37] proposed a two-stage link prediction model based on a designed multi-classifier ensemble, using the new node representation as input. It is an ensemble model combining three graph-embedding techniques: matrix decomposition, random walk, and deep learning. The integration of multiple embedding techniques and classifiers significantly improves link prediction performance. Li et al. proposed a model for constructing friendships among dissimilar users in social networks,

Table 3 summary of Network Embedding-based LP Methods

Author and Year	Method	Characteristics	Limitations
Perozzi et al., (2014)[35]	DeepWalk-Algorithm	It is an unsupervised learning for vertex representation, suitable for large graphs.	Ignored attributed information not effective for new nodes/edges without historical data.
Grover et al., (2016)[36]	node2vec	Learns embedding by considering both local and global network structure.	Static embedding does not account for temporal dynamics.
Chen et al. (2020) [37]	EMLP	It is an ensemble-based learning that combines the various graph embedding vectors into a new representation of each node.	Tested on small datasets, lacks complexity discussion.
Berahmad et al., (2021) [38]	CSADW	Modified DeepWalk for attributed network, it captures the structural and node attributes.	Performance depends on node attributes, quality may struggle with scalability in large networks.
Guo et al., (2022)[39]	MSVGAE	It is a multi-scale graph variational autoencoder that learns low-dimensional vectors for a mixed probability distribution.	May not address all variational and complexities in real-world graph data.
Ren et al. (2024) [40]	ENLRW,	Extended local random walk integrate metrics through reliable paths.	Inefficient in complex scenarios.

which can greatly expand the friend count of influencers. The consensus link prediction is presented to forecast potential friendship among the users. Nasiri et al.[45] proposed a method called Robust Graph Regularization nonnegative Matrix Factorization. It aims to forecast the missing links in complex networks by using a combination of information extracted from the observed network. The proposed model avoids the issue of collapsing temporal networks and demonstrates enhanced accuracy in predicting temporal links within such networks. Yang et al.[46] introduced a multi-network representation learning framework with joint training. This framework aims to capture the diverse semantics embedded in heterogeneous networks by leveraging various node relationships.

3 Proposed work

In this section, we discuss the proposed link prediction model, named GSVAELP, in detail. The overall work of the proposed model is divided into four phases. In the first phase, feature extraction is done by using the graghSAGE where embeddings of each node in the network are obtained. Here, the generated node embeddings represent nodes or entities in a network as low-dimensional vectors and intend to capture structural and relational information of the nodes in the network. The second phase involves the utilization of variational autoencoders (VAE) that take input as node embeddings generated by graphSAGE This phase involves training the VAE to learn a latent space representation for nodes. The VAE helps in refining the node embeddings and learning a continuous, meaningful latent space. The third phase of the model adopts a logistic regression classifier for predicting upcoming links in the networks. Finally, the necessary hyper-parameter studies are performed to obtain the optimal values of the various hyper-parameters involved in the different phases of the model. The combination of GSVAELP allows for leveraging of the neighborhood information and node features captured by graphSAGE embeddings, while VAE is used to learn the compact latent space and capture the pattern for link prediction. Figure 1 depicts the general architecture of the model. The various phases of our proposed work are described below.

3.1 Feature extraction and generating network embedding using graphSAGE

The network embedding generation in the proposed mode uses a neighborhood aggregation function that combines information from a node's neighbor to update its representations. The technique captures the network's local and global structure, can generalize to unseen nodes, and does not require the entire network structure during learning. Rather than considering the complete k-hop neighborhood of a target node, graphSAGE initially samples or prunes the computational graph of the k-hop neighborhood and then performs the feature aggregation procedure to produce the embeddings of the target node.

Consider the input network as $G = (V, E)$, where V is the set of nodes and E is the set of the links or edges in the graph. The node feature matrix $Y \in \mathbb{R}^{|V| \times D}$, where D is the dimension of node features. The adjacency matrix of graph A $\mathbb{R}^{|V| \times |V|}$ matrix, if the two nodes u, v are connected, the value of A is equal to 1, otherwise 0. The main purpose of using graphSAGE is to learn the node embedding H_v for each node $v \epsilon\ V$. For each node v, create a set of neighbors $N(v)$ from its neighborhood, ensuring that each node has the same number of neighbors. For each node v, aggregate the embeddings of its neighbors $H_N(v)$ by using an $AGGREGATE$ function represented as (3).

$$H_{agg}(v) = AGGREGATE(\{H_u, \forall u, \in N(v)\}) \tag{3}$$

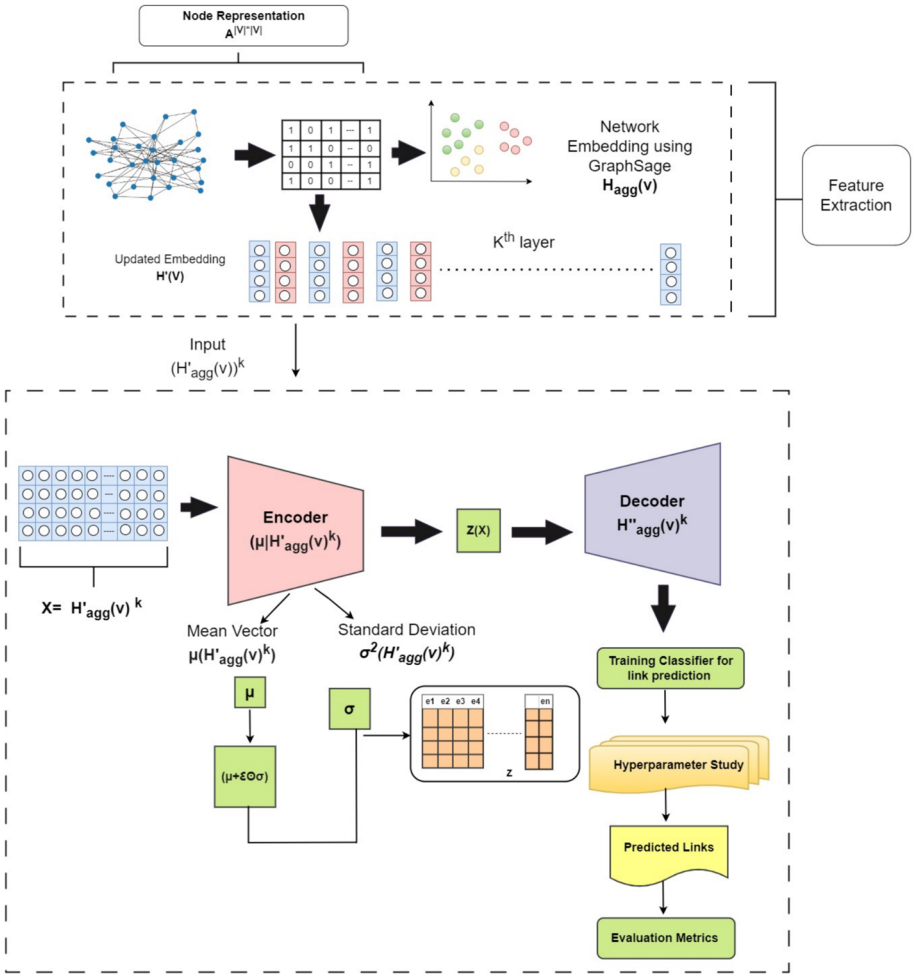


Fig. 1 Flow Diagram of the proposed GSVAELP model

After calculating the aggregate embeddings $H_{agg}(v)$ at of the graphSAGE model, then updating the embedding of node v by using an $UPDATE$ function is shown in (4).

$$H'_{agg}(v)^k = UPDATE(H_v, H_{agg}(v)) \tag{4}$$

$H'_{agg}(v)^k$ is the updated embedding at the $k-th$ layer of the graphSAGE model, and the updated embedding is then passed as input in the next phase of the model.

3.2 Variational autoencoder for node embedding refinement

In this step, variational autoencoders (VAE) are utilized that take input as node embeddings generated by graphSAGE and help refine the node embedding leading to learning a continuous, meaningful latent space of nodes. Variational Autoencoders are an extension of the traditional autoencoders with additional probabilistic concepts. The objective of VAEs is not

only to capture the latent representation of the network but also to underlying the model data distribution. The architecture of a Variational autoencoder comprises an input layer, an encoding layer, a decoding layer, and an output layer. The encoding layer reduces the dimensionality of the incoming data, and the decoding layer reconstructs the original data from the encoded form. To generate the embedding graphSAGE, use the AGGREGATOR function and then update the embedding $H'_{agg}(v)$ of the network and then transfer it into the encoder layer. The Encoder uses the graghSAGE embedding to learn the network structure. At the same time, the decoder is a fully connected layer that transforms the extracted embedding into the original space. GSVAELP will output the predicted social network link prediction. The Linear Autoencoder's encoding layer, sometimes referred to as the encoding space, learns a lower-dimensional representation of the input data. Usually, the latent space's dimensions are less complex than the dimensions of the incoming data. Prior to the model's in-depth description. We discuss some terms and notations that will be used in the paper.

3.2.1 Encoder

An encoder layer takes the graphSAGE embedding $H'_{agg}(v)^k$ as input. It is responsible for transforming the high-dimensional input layer into a lower-dimensional encoding that captures the most essential features of the data. An encoder framework is designed that takes the graphSAGE embedding as input and produces the mean μ and log variance $log(\sigma)$ as output. The mathematical equation of the encoder is presented as (5) and (6).

$$\mu, log(\sigma) = Encoder(H'_{agg}(v)^k) \tag{5}$$

$$Z = \mu + \in \odot \sigma, where \in \sim N(0, I) \tag{6}$$

where Z represents the latent variable vector obtained from the latent variable distribution, σ is a random sample from a standard distribution $N(0, I)$, where I is the identity matrix from a normal distribution with mean 0 and variance 1. Now we design a decoder framework that takes Z as input and generates the reconstructed representation of the graphSAGE embeddings.

3.2.2 Decoder

The decoder transforms the encoded representation into the original data space. The decoder converts the latent representation Z into the original graphSAGE Embedding $H'_{agg}(v)^k$.

$$H''_{agg}(v)^k = Decoder(Z) \tag{7}$$

$H''_{agg}(v)^k$ is the reconstructed embedding for each node v in the $k-th$ layer of the graphSAGE model. Decoder is a function that takes the sampled latent variable Z as input and attempts to reconstruct the original embeddings $H'_{agg}(v)$. Its purpose is to generate the input reconstruction based on the sampled latent variables.

3.3 Training classifier for link prediction

After getting the latent representation from VAE, the proposed model exploits a binary classifier for link prediction. For the classification task, use logistic regression to improve the prediction ability of our proposed work. The input for logistic regression consists of a pair of nodes with their corresponding latent representation. For a pair of nodes v_i and v_j, the input

features would be $Z_i - Z_j$, where Z_i, Z_j are the latent representation obtained from VAE model.

$$P(y_{ij}) = 1 = \frac{1}{1 + e^{-(w^T(Z_i - Z_j) + b)}} \tag{8}$$

The regression model predicts the probability of a link between the node v_i and v_j, where y_{ij} is the binary label indicating whether there is a link between nodes v_i and v_j, w are the weights and b is the bias term. In the regression prediction problem, the L_2 regularization term is usually used to measure the discrepancies between the predicted probabilities and the actual labels.

$$L_2 = ||P(y_{i,j}) - A(y_{i,j})||^2 = \sum_{i=0}^{V} \sum_{i=0}^{V} (P(ij) - A(i,j)^2) \tag{9}$$

where P represents the probability matrix, $A(y_{i,j})$ is the adjacency matrix, and V is the number of nodes in the network. However, if we utilize only L_2 distance as a loss function, they may cause biases towards zero, especially in cases where the network exhibits sparsity. We incorporate an additional regularization term denoted as L_{reg} to avoid overfitting. The overall loss during the training process is defined as:

$$L_{total} = L_2 + \beta L_{reg} \tag{10}$$

where β is the parameter that balances the importance of the two terms L_2 and L_{reg}. The regularization loss L_{reg} encourages the model to keep the weights of the parameters small, which helps to prevent the overfitting problem, whereas the L_2 represents the sum of the square of the GSVAELP model weights and with the help of (2) we calculate the prediction score between the nodes $\theta(Vi, Vj)$, if the prediction score is above or equal the value of threshold, then link will be predicted. To measure the effectiveness of the GSVAELP model, utilize the appropriate evaluation metrics as discussed in the next section.

$$Prediction = \begin{cases} 1, & \text{if } w_{ij} \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

3.4 Hyper-parameter studies

Hyperparameter tuning is figuring out which values of these hyperparameters are best for a given model. For tuning the hyperparameters of our GSVAELP model, we have selected suitable hyperparameters for achieving good performance, like the Hidden layer Dimensions, which specify the number of neurons in each hidden layer, batch size based on available memory and computational resources, and to prevent the overfitting apply the L_2 regularization term, the learning rate for model training. This specific hyperparameter gives the optimal results of our proposed model in the context of link prediction.

3.5 Proposed model architecture and algorithms

Figure 1 depicts the overall framework of our proposed model. The first phase involves the generation of node embeddings by adopting the graphSAGE embedding method. The generated embeddings of the nodes act as features. We then use these extracted features as input to the Variational Autoencoder. This integration allows for extracting expressive node embeddings while learning a compact latent space representation. The next stage of

the model employs a logistic regression classifier to forecast the future connections within the networks. Subsequently, thorough experiments into hyperparameters are conducted to determine the most effective values for the various hyperparameters utilized in the various stages of the model. The GSVAELP model can adapt to evolving network structures and maintain accurate link prediction capabilities.

We present two algorithms. The first algorithm is designed to create embeddings for all nodes by utilizing the graphSAGE technique. These generated embeddings are subsequently forwarded to the second algorithm employed to train the link prediction model. Algorithm 1 takes the set of vertices and edges within the network dataset as input. The brief step-by-step description of Algorithm 1 is as follows.

1. Consider a complex network graph G as input having the set of vertices V and the edges E, create a node feature matrix of size $|V| \times D$, where D represents the dimensions of node features. To check the connectivity between the V and E, create an adjacency matrix of size $|V| \times |V|$. If the two nodes v_i, v_j are connected, it is labeled as a positive label or indicates 1, otherwise negative label or 0.
2. In Step 2, the input is passed to the graphSAGE model to generate the embeddings of each node. These embeddings capture the local neighborhood information and encode the structure and node features.
3. For each node v, generate the embedding using an $AGGREGATOR$ function.
4. Then, update the aggregator embedding $H_{agg}(v)$ as $H'agg(v)$ by using an $Update$ function.

Algorithm 1 Algorithm for finding the embeddings of Graph G.

Require: $G = (V, E)$, V : set of Nodes, E : set of Edges, $A^{|V|\times|V|}$: Adjacency matrix, D : Dimensions of embedding for each node
Ensure: generated embedding for each node
1: $X \Leftarrow$ Node Features Matrix of Size $|V| \times D$
2: $N(v) \Leftarrow$ Neigbouhood of node v
3: $H_v \Leftarrow$ graphSAGE Node Embeddings for reach node $v \in V$
4: $H_{agg}(v) \Leftarrow$ Aggregate the embeddings of its Neighbors $H_N(v)$ as per in (3)
5: $H'_{agg}(v)^k \Leftarrow$ update the embeddings of node v using an update function as per in (4)
6: **return** $H'_{agg}(v)^k$

Next, we generate predictions for the given network by considering the steps discussed in Algorithm 2. For Algorithm 2, the input is the updated embeddings by using the graphSAGE method $H'_{agg}(v)^k$. The VAEs map these embeddings to a lower dimensional latent space, enabling the generative link prediction in the given network.

The following is a concise step-by-step explanation of Algorithm 2.

1. In step 1, the updated embedding is transferred as input to the variational auto-encoder model. The variational autoencoder model consists of an encoder and decoder. The encoder takes the updated embedding $H'_{agg}(v)$ as input and calculates the mean μ and the standard deviation σ. The encoder maps the input data $H'agg(v)^k$ to the low- dimensional space Z.
2. In step 2, The decoder takes the sampled z as input and maps it back to the original input space. It reconstructs the input based on the sampled point from the latent space.
3. To prevent the noise or outliers from impacting the meaningful results, compare the decoder output with the graphSAGE embedding and create a L_2 distance to learn a

Algorithm 2 Overall algorithm of the proposed link prediction Model (GSVAELP).

Input: graphSAGE embeddings $H'_{agg}(v)^k$
OutPut: Probability of link between node pairs
$\mu, \sigma \Leftarrow$ Compute the mean vector and standard deviation using (5)
$Z \Leftarrow$ Generate the probability distribution of the latent variable as shown in (6)
$Decoder \Leftarrow$ Refine embedding by using the Decoder fun as shown in (7)
$L_{total} \Leftarrow$ Calculate the total loss based on L_2,L_{reg} as shown in (9)
$Prediction \Leftarrow$ Probability of link between the nodes v_i, v_j as shown in (11)
$metrices \Leftarrow$ Calculate the various performance metrics of link prediction
return metrices

meaningful latent space. After calculating L_2 distance, Variational Autoencoders also calculate the L_{reg} Regularization that encourages the approximate posterior to close a prior distribution.

4. Once the GSVAELP model is trained, the latent representation is used for predicting the links between the node pairs v_i, v_j. The GSVAELP model generates the embeddings for both nodes and then measures the probability between these nodes. For the binary classification task, use the logistic regression with L_2 and L_{reg} regularization to classify the positive and negative links. The higher the probability between a node pair indicates the higher the likelihood of a link between the nodes.
5. Then apply the threshold to the prediction score to determine the potential links that are considered as predicted links.
6. In the last step to compare the performance of our proposed model GSVAELP to other existing and baseline methods, use the evaluation metrics to measure the accuracy, precision, recall, F1-score, and AUC. Finally, these evaluated matrices are returned for further analysis.

4 Dataset and evaluation metrics

4.1 Datasets

In this study, we utilize five diverse network datasets, including the social network and citation network, to evaluate the performance of the proposed model. The details of the datasets are described as follows.

1. Caltech36 [47]: It is a Facebook friendship dataset made up of Caltech, Pasadena students. Here, each node signifies an individual, while the edges denote the friendship or social ties within the network. The dataset encompasses 769 nodes and approximately 16, 656 edges.
2. PolBlogs [48]: A Twitter network shows users' retweet activity in the 2010 U.S. congressional elections. Retweets are represented by the edges, while the nodes represent Twitter users. There are 61157 edges and 18470 nodes in the network.
3. DBLP [49]: It is a citation network dataset. The citation data is sourced from various databases, including DBLP (Digital Bibliography and Library Project), ACM, and other relevant sources. Each node represents a written piece, and each edge symbolizes a reference connecting these written pieces. There are 49743 edges and 12591 nodes.
4. Fb-TV Show [44]: The Dataset represents a social network derived from Facebook, and shows the links among different TV program pages on the platform. In this dataset, nodes

represent the TV program pages, whereas the edges show the instances of shared links between these pages. The dataset comprises 17260 edges and 3892 nodes.
5. Delaunay-N12 [50]: This network dataset falls under the DIMACS10 category. These files were generated by triangulating random locations in the unit square using the Delaunay method. It consists of 4100 nodes and 12300 edges.
6. Public figures [51]: This dataset is derived from Facebook and serves as a social network representation. The nodes in this network correspond to public figures, and the edges depict their interactions. It encompasses a total of 11,565 nodes and 67,036 edges.
7. Hamster [52]: It is a popular social network dataset, a representation of the interaction between the users, the nodes denote the users whereas the edges correspond to the link between the users. It consists of 2426 nodes and 16630 edges.
8. Wiki [53]: This dataset comprises voting data sourced from Wikipedia. In this network, the nodes represent Wikipedia users, and the edges illustrate the voting interactions between these users. The dataset encompasses 889 nodes and 2914 edges.

A brief statistical description of the datasets is discussed in Table 4 where $|V|$ represents the total number of nodes, $|E|$ shows the total number of links, k represents the average degree of the nodes, and cc denotes the average clustering coefficient.

4.2 Evaluation metrices

To evaluate the effectiveness of the proposed model, we use distinct classical evaluation metrics like precision, recall, F1-score, AUC (Area under the curve), and accuracy. Below is a detailed description of these metrics.

1. Precision: The precision measures the positive instances of the class. It is an indicator to measure the shared positive classifications that were true. It is also known as positive predictive value. The mathematical formula of precision is:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

Where TP represents the correctly predicted links in the network, FP denotes the falsely predicted links.
2. Recall: Recall is used to measure the model's ability to detect positive samples. It is a ratio between the number of positive samples correctly to the number of positive samples.

Table 4 The Statistical Description of various Dataset used

Dataset	$\|V\|$	$\|E\|$	$< k >$	$< cc >$
Caltech	769	16656	43	0.409
Polblogs	18470	61157	6	0.059
DBLP	12591	49743	7	0.0119
FB-TV Show	3892	17260	8	0.373
Delaunay-N12	4100	12300	5	0.434
Hamsterster	2426	16630	13	0.537
Public Figure	11565	67036	11	0.1793
Wiki-Vote	889	2914	7	0.1409

The mathematical representation of recall is followed as:

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

TP denotes the True positive, which classifies the positive samples as positive, whereas FN denotes the false negative means, which classifies the positive sample as negative.

3. F1-score: The F1-score measures the Harmonic mean of precision and the recall. The mathematical formula of the F1-score is expressed as :

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

4. AUC: The Area under the receiver operating characteristics curve (AUC). It is a graphical representation of the true positive rate versus the false positive rate. The range of AUC varies from 0.0 to 0.1. A higher AUC score indicates a higher model's accuracy. AUC can be defined as :

$$AUC = \frac{M' + 0.5M''}{M} \tag{15}$$

M denotes the independent comparisons,M' is the occurrence of the missing links, M'' is the occurrence of missing links and non-existent links in the network and the value 0.5 is the discrimination threshold.

5. Accuracy: Accuracy is calculated as the proportion of correct predictions divided by the total number of predictions made. It is an important evaluation metric to measure the effectiveness of the model. The mathematical expression of accuracy is:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{16}$$

5 Experiment and result analysis

This section provides experimental results and analysis using the datasets and evaluation metrics as mentioned in Section 4. We compared the performance of the proposed model with some baseline methods of link prediction using classical network embedding techniques such as deepwalk [35], node2vec [36], and Line [54]. The link prediction methods using classical baseline machine learning classifiers such as Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), and Linear Autoencoders (LAE) are also used as comparing methods in this analysis. Before performing binary classification tasks for link prediction (LP) using the above-mentioned machine learning classifiers, the graphSAGE embedding is used for feature representation or latent representation of nodes of the network. We named these ML classifiers based LP method RFLP (random forest-based link prediction), LRLP (logistic regression based LP), GBLP (gradient boosting LP), and LAELP(Linear Autoencoders based LP). Apart from these comparing methods, we evaluated the performance of our proposed model against some recently popular methods of link prediction namely, HSLP [13] (H-index-based link prediction), NC-LGBM [55] (Link Prediction using node centrality and light gradient boosting machine), EMLP [52] (Ensemble-model based link prediction) and NSMLLP [53] (integrating node centralities, similarity measures, and machine learning classifiers based LP). Further, we performed various hyperparameter studies such as the choice of splitting ratio for training and testing the models and various model parameters to find out their optimal values.

5.1 Choice of splitting ratio

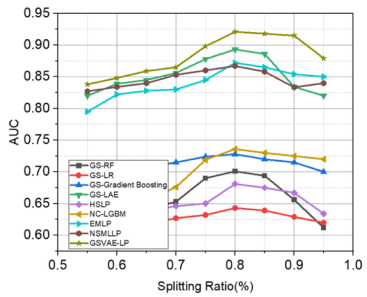
In this section, we examine the optimal ratio for dividing the dataset into training and testing phases. To analyze this, we performed experiments on the splitting ratio to evaluate the performance of AUC metrics for each dataset, for various baseline methods, existing methods, and our proposed method. In the conducted experiments, the splitting ratio ranges from 50% to 95% in increments of 5%. The results we have acquired are shown in Fig. 2. The x-axis value signifies the dataset's splitting ratio, while the y-axis denotes the corresponding AUC (Area Under the Curve) value. From the results depicted in Fig. 2, it is evident that the dataset splitting ratio of 80 : 20 is ideal for training and testing the various methods of LP considered in this study. Hence, 80% of the data is dedicated to training, leaving the remaining 20% for evaluating the model's performance. The decision to divide the dataset in this way ensures that there would be sufficient data for properly training the classifier while also avoiding the overfitting problem.

Figures show that with an increasing splitting ratio, the AUC value rises across all datasets, peaking at approximately 80%. This indicates that when the splitting ratio reaches 50%, the data is in the under-lifting situation. In other words, at lower splitting ratios, there is not enough data available to train the dataset efficiently and the proposed technique struggles to fully capture the network's features. So, based on these findings, it is usually clear that AUC performance is better as the split ratio rises from 50% to 80%, with the optimum performance being at 80%. Additionally, it has been seen that the AUC value begins to decrease when the spitting ratio approaches 80%. This shows that the data tends to overfit as the amount of training data increases. Therefore, this study leads us to the conclusion that an 80 : 20% split of the data is the best one.

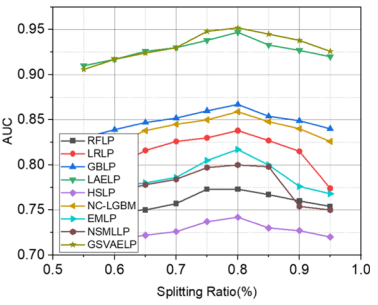
5.2 Parameter sensitivity

In addition to the previous analysis, we have conducted some experiments to evaluate the effectiveness of the proposed GSVAELP model. The optimal experiments are the most effective means of determining the ideal parameter settings. To fine-tune our model, we utilized one dataset DBLP shown in Fig. 3. We used the distinct hyperparameters including batch size 300, activation function $leaky_relu$, learning rate 0.0003, dropout 0.8, hidden feats 512, epochs 20, and loss function. The AUC for GSVAELP generally increases, with the dimensions of the embedding vectors. However, there is a noticeable drop in AUC, when the embedding dimension changes from 4 to 8. The optimal AUC is achieved when the dimension is set to 32.

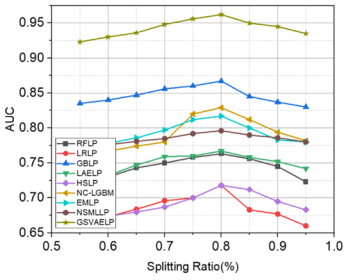
Figure 3(b) shows the changes in AUC with respect to the sample size k, experiment with different values of $k = 10, 20, 30, 40, 50, 60$ to observe how the number of sampled neighbors affects the performance of the GSVAELP model. The result indicates that the highest AUC is achieved when the sample size is $k = 30$, while the lowest AUC is observed when the sample size is $k = 60$. Overall, the experiment shows that increasing the sample size does not affect AUC. For our model, testing the AUC with different learning rates varies from 0.0001 to 0.0008. The best AUC occurs when the learning rate is set to be 0.0003 as shown in Fig. 3, whereas Due to the large number of edges and node features in the data set, the accurate dropout rate argument is to be 0.8 as shown in Fig. 3(d). Overall, it can be stated that through a proper exploration of hyperparameters, our GSVAELP model demonstrates superior link prediction performance compared to other existing models.



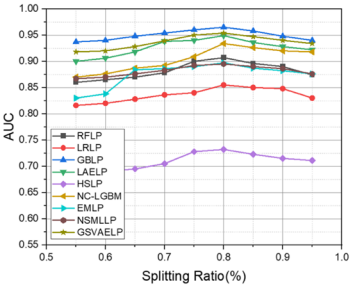
(a) Caltech Dataset



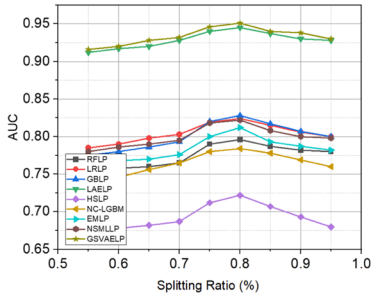
(b) PolBlogs Dataset



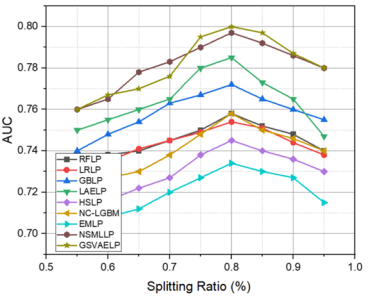
(c) DBLP Dataset



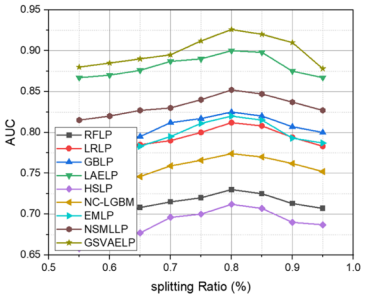
(d) FB-TV Show Dataset



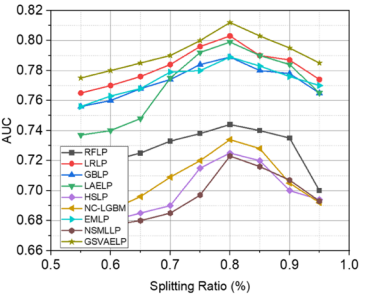
(e) Delaunay N12 Dataset



(f) Hamsterster Dataset



(g) Public Figure Dataset



(h) Wiki Vote Dataset

Fig. 2 AUC Results obtained for different Dataset for various models as the splitting ratio varied from 50% to 95% in steps of 5%

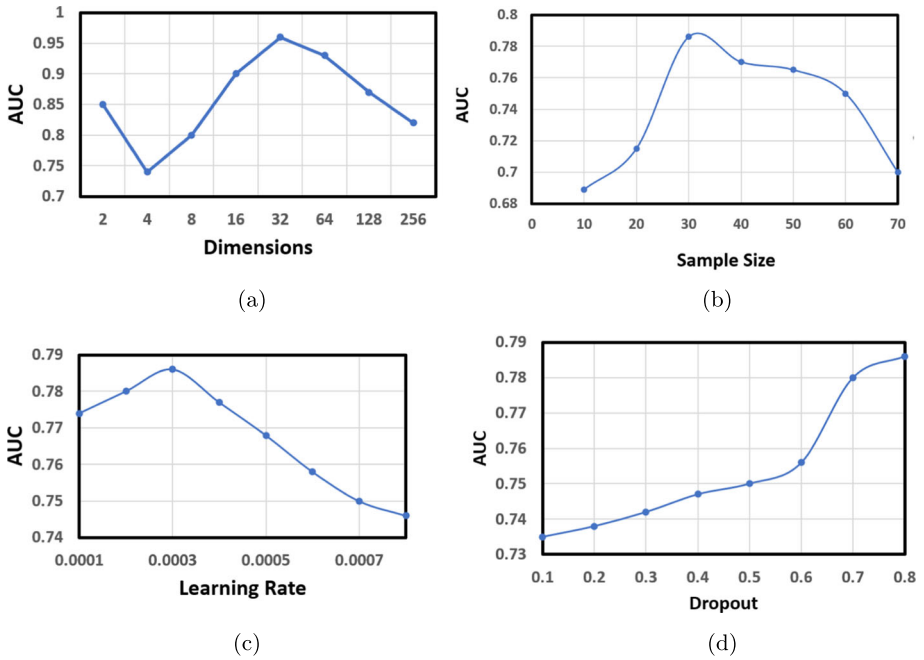


Fig. 3 Parameter Sensitivity on DBLP Dataset

5.3 Compartive performance study

To validate the effectiveness of the proposed model GSVAELP, we compared its performance with some baseline and existing methods. Since many link prediction algorithms are stochastic in nature, we conducted the experiments a hundred times and calculated the average results to reach evaluation metrics. This approach provides a more comprehensive and reliable assessment of performance. The comparative performance study is shown in Tables 5 to 12. The best result obtained for each dataset is shown and emphasized in bold.

For the Caltech network dataset, Table 5 illustrates the results obtained. It can be clearly seen that the GSVAELP model outperformed in terms of precision of 0.921 and AUC score of 0.914, whereas the EMLP model has performed well in terms of recall of 0.863 and accuracy of 0.875. However, the LAELP model has achieved an F1-score of 0.893 compared to other link prediction models.

Table 6 shows the results of the PolBlogs dataset, in this table our proposed model GSVAELP performs extremely well in terms of precision, recall, and AUC score. The higher AUC indicates better performance. The NC-LGBM model significantly performs well in terms of F1-score. On the other hand, the NSMLLP model achieved a good accuracy score as compared to other models.

From Table 7 on the DBLP dataset, we can observe the performance of different models. From the obtained results it can be seen that the proposed GSVAELP model has the highest recall and AUC value, indicating its utility. On the other hand models like LAELP have a good precision value, the GBLP performs well in terms of accuracy and the model NSMLLP has a good F1-score.

Table 5 Overall Performance Metrics of Caltech Dataset

Models	Precision	Recall	F1score	AUC	Accuracy
DeepWalk	0.642	0.623	0.655	0.738	0.674
node2vec	0.773	0.834	0.792	0.780	0.791
LINE	0.646	0.593	0.603	0.569	0.628
RFLP	0.723	0.734	0.720	0.727	0.728
LRLP	0.643	0.676	0.598	0.760	0.540
GBLP	0.728	0.743	0.730	0.820	0.747
LAELP	0.893	0.744	**0.893**	0.882	0.744
HSLP	0.681	0.672	0.673	0.701	0.682
NC-LGBM	0.726	0.701	0.711	0.782	0.755
EMLP	0.872	**0.863**	0.842	0.896	**0.875**
NSMLLP	0.867	0.932	0.891	0.862	0.837
GSVAELP	**0.921**	0.855	0.799	**0.914**	0.761

Table 6 Overall Performance Metrics of PolBlogs Dataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.521	0.591	0.552	0.581	0.562
node2vec	0.630	0.723	0.712	0.707	0.746
LINE	0.746	0.683	0.753	0.769	0.694
RFLP	0.766	0.756	0.761	0.773	0.774
LRLP	0.734	0.732	0.687	0.838	0.647
GBLP	0.786	0.857	0.791	0.867	0.784
LAELP	0.952	0.953	0.783	0.947	0.767
HSLP	0.722	0.743	0.733	0.742	0.721
NC-LGBM	0.761	0.845	**0.801**	0.859	0.792
EMLP	0.730	0.802	0.764	0.817	0.746
NSMLLP	0.842	0.867	0.800	0.800	**0.823**
GSVAELP	**0.967**	**0.955**	0.792	**0.952**	0.748

Table 7 Overall Performance Metrics of DBLP Dataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.597	0.655	0.603	0.673	0.578
node2vec	0.678	0.767	0.722	0.743	0.700
LINE	0.556	0.508	0.579	0.511	0.694
RFLP	0.797	0.733	0.763	0.763	0.762
LRLP	0.712	0.689	0.657	0.718	0.720
GBLP	0.786	0.857	0.791	0.867	**0.787**
LAELP	**0.963**	0.786	0.947	0.767	0.780
HSLP	0.724	0.752	0.738	0.718	0.720
NC-LGBM	0.817	0.732	0.773	0.829	0.774
EMLP	0.794	0.845	0.811	0.817	0.784
NSMLLP	0.867	0.912	**0.822**	0.796	0.772
GSVAELP	0.961	**0.925**	0.788	**0.960**	0.786

Table 8 Overall Performance Metrics of FB-TV show dataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.654	0.734	0.712	0.581	0.592
node2vec	0.730	0.760	0.742	0.689	0.623
LINE	0.702	0.754	0.687	0.722	0.690
RFLP	0.926	0.881	0.901	0.907	0.908
LRLP	0.742	0.773	0.687	0.931	0.514
GBLP	0.959	0.877	0.916	**0.965**	0.904
LAELP	0.962	**0.960**	0.791	0.949	0.746
HSLP	0.756	0.733	0.700	0.732	0.711
NC-LGBM	0.912	0.852	0.823	0.934	**0.922**
EMLP	0.925	0.860	**0.891**	0.897	0.898
NSMLLP	0.910	0.872	0.883	0.895	0.723
GSVAELP	**0.967**	0.960	0.794	0.954	0.750

Table 8 presents performance metrics of various models on the FB-TV show dataset. The proposed model GSVAELP outperformed all other methods with the highest precision of 0.967 and recall of 0.960 among all models. It achieved an AUC of 0.954 and an accuracy of 0.750, whereas the model LAELP closely followed the recall of 0.960. In terms of AUC, GBLP is the best-performing model, and the best accuracy is achieved by the NC-LGBM model.

Table 9 shows the results on the Delaunay N12 dataset. With a precision of 0.63 and a recall of 0.963, our proposed model, GSVAELP, has demonstrated outstanding performance. Additionally, the proposed model achieves the highest AUC of 0.951 among all models. However, the GBLP has performed well in terms of an F1-score of 0.800. LRLP and NC-LGBM both achieve the highest accuracy of 0.812 among the existing methods.

The Table 10 evaluates the performance metrics on the Hamsterster social network dataset. The models GBLP and NC-LGBM both achieve the highest F1-score of 0.796. LAELP demonstrated excellent performance with a precision of 0.829. Overall as per the accuracy

Table 9 Overall Performance Metrics of Delaunay-N12 Dqataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.533	0.554	0.501	0.571	0.552
node2vec	0.664	0.752	0.705	0.686	0.680
LINE	0.589	0.722	0.672	0.698	0.700
RFLP	0.846	0.724	0.780	0.796	0.795
LRLP	0.833	0.756	0.792	0.824	**0.812**
GBLP	0.835	0.768	**0.800**	0.825	0.808
LAELP	0.962	0.945	0.788	0.945	0.746
HSLP	0.742	0.785	0.752	0.722	0.724
NC-LGBM	0.788	0.825	0.796	0.784	0.782
EMLP	0.823	0.780	0.806	0.812	0.812
NSMLLP	0.895	0.843	0.765	0.822	0.754
GSVAELP	**0.963**	**0.963**	0.793	**0.951**	0.749

Table 10 Overall Performance Metrics of Hamsterster Dataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.523	0.556	0.520	0.578	0.676
node2vec	0.598	0.667	0.623	0.684	0.692
LINE	0.623	0.670	0.669	0.680	0.682
RFLP	0.740	0.758	0.772	0.750	0.755
LRLP	0.794	0.823	0.787	0.754	0.780
GBLP	0.745	0.848	**0.796**	0.772	0.786
LAELP	**0.829**	0.854	0.752	0.780	0.767
HSLP	0.767	0.740	0.723	0.745	0.756
NC-LGBM	0.730	0.834	**0.796**	0.758	0.797
EMLP	0.812	0.845	0.787	0.734	0.821
NSMLLP	0.815	0.842	0.756	0.797	0.734
GSVAELP	0.820	**0.878**	0.776	**0.800**	**0.834**

and the recall, the proposed GSVAELP model is the best performer. It achieved an accuracy of 0.834 and a recall of 0.878.

For the public figure dataset, Table 11 shows the performance metrics of various models. In baseline methods, the notable performance includes the GBLP with the highest F1-score of 0.800 and LAELP has the highest precision of 0.99. The EMLP model has the highest accuracy of 0.823 among all the methods of link prediction, whereas the proposed model GSVAELP outperformed all other models with the highest recall of 0.954 and the AUC 0.926, indicating the best performance metrics.

The overall performance metrics of the Wiki vote dataset as shown in Table 12. The RFLP model achieved the highest accuracy of 0.795, and the model LAELP showed a higher F1-score of 0.794, with some existing methods the EMLP model has performed well in terms of precision of 0.828. The proposed model GSVAELP demonstrated the highest recall of 0.842 and AUC 0.812. In summary, the proposed model GSVAELP demonstrates better performance on all datasets.

Table 11 Overall Performance Metrics of Public Figure Dataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.635	0.578	0.612	0.623	0.692
node2vec	0.655	0.658	0.622	0.686	0.667
LINE	0.569	0.564	0.523	0.605	0.611
RFLP	0.745	0.778	0.645	0.723	0.757
LRLP	0.789	0.812	0.799	0.812	0.796
GBLP	0.835	0.768	**0.800**	0.825	0.808
LAELP	**0.966**	0.950	0.723	0.900	0.762
HSLP	0.723	0.765	0.760	0.712	0.735
NC-LGBM	0.736	0.832	0.783	0.774	0.767
EMLP	0.865	0.893	0.887	0.820	**0.823**
NSMLLP	0.820	0.823	0.752	0.852	0.774
GSVAELP	0.956	**0.954**	0.773	**0.926**	0.820

Table 12 Overall Performance Metrics of Wiki Vote Dataset

Models	Precision	Recall	F1-score	AUC	Accuracy
DeepWalk	0.580	0.545	0.520	0.596	0.623
node2vec	0.623	0.673	0.678	0.658	0.650
LINE	0.627	0.584	0.569	0.675	0.666
RFLP	0.724	0.731	0.723	0.740	**0.795**
LRLP	0.712	0.723	0.745	0.802	0.783
GBLP	0.728	0.834	0.745	0.789	0.725
LAELP	0.812	0.826	**0.794**	0.799	0.747
HSLP	0.645	0.754	0.723	0.725	0.734
NC-LGBM	0.723	0.820	0.745	0.734	0.768
EMLP	**0.828**	0.823	0.765	0.789	0.820
NSMLLP	0.812	0.825	0.742	0.723	0.734
GSVAELP	0.823	**0.842**	0.787	**0.812**	0.792

5.4 Average performance of each comparing methods of link prediction across all datasets

To gain a better comparative analysis, we calculated the average precision, recall, and F1-score of all methods obtained in Tables 2 to 9 for various datasets. The average results of precision, average recall, and average F1-score are depicted in Figs. 4, 5, and 6, respectively. The $x-axis\ represents all the models$, $and the$ y-axis shows the average precision, recall, and F1-score. The results clearly show that the proposed GSVAELP model has outperformed significantly compared to the other link prediction models. The average precision of the GSVAELP model is 0.922, whereas some other methods, like LAELP, achieved an average precision of 0.877, and HSLP obtained an average precision of 872 across all the datasets. Figure 4 shows the average recall of the proposed GSVAELP model is 0.915, the highest among all the other models, the LAELP model demonstrated the average recall of 0.818 whereas in Fig. 5 represents the average F1-score of the proposed GSVAELP model is 0.842. Finally, it is revealed that the proposed GSVAELP model outperformed other methods with the highest average precision, average recall, and average F1 score.

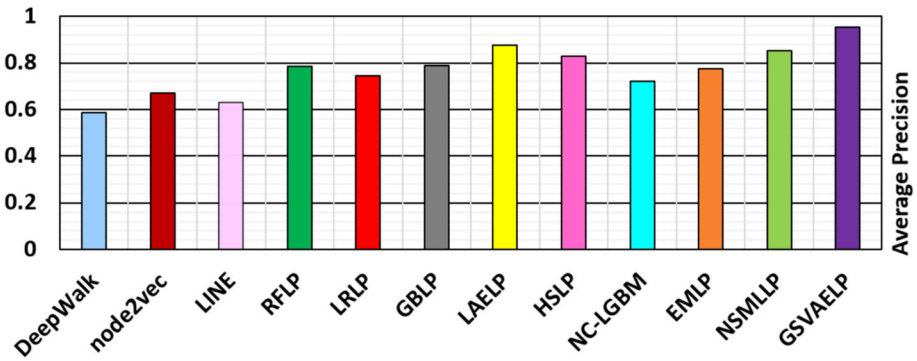


Fig. 4 Average Precision for each method on all eight datasets

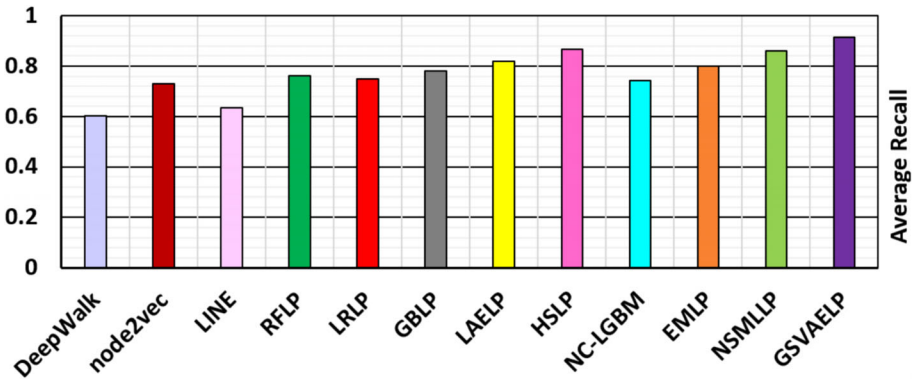


Fig. 5 Average Recall for each method on all eight datasets

6 Conclusion

Link Prediction in complex networks like social networks involves predicting the future links among the unconnected nodes and has many real-life applications. This paper presented a synergistic model named GSVAELP, combining the graphSAGE embeddings with Variational Autoencoders (VAE) for link prediction. In the proposed model, feature extraction was performed using graphSAGE, which generated embeddings for each node in the network. These node embeddings were created as low-dimensional vectors to capture structural and relational information about the network's nodes. The Variational autoencoders (VAE) were employed to refine the node embeddings generated by graphSAGE. This phase involved training the VAE to acquire a latent space representation for the nodes, ultimately creating a continuous and meaningful latent space. Finally, necessary hyperparameter studies were conducted to determine the optimal values for the various hyperparameters used in the different phases of the model. To prove the effectiveness of the proposed model, we conducted a comparative experiment on eight real-world diverse network datasets. The experiment's findings suggested that the proposed model's efficacy is better than others. This research contributes a wide range of link prediction methodologies and provides a solid foundation for future investigations into the dynamics of complex network interactions. Future work

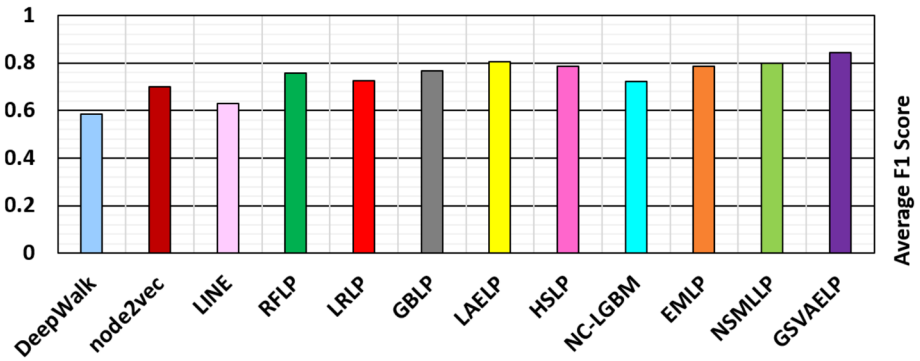


Fig. 6 Average F1-score for each method on all eight datasets

could focus on extending the GSVAE-LP model to capture temporal dynamics in evolving networks. Incorporating time awareness features and developing models that can learn from the temporal evolution of networks could significantly improve link prediction accuracy over time. The proposed model can be applied to other domains, like e-commerce, biological networks, and citation networks. Exploring the effectiveness of the GSVAE-LP in diverse fields.

Author Contributions Fatima Ziya- Conceptualization, Methodology, Visualization, Software, Investigation, Writing - original draft
Sanjay Kumar- Conceptualization, Methodology, Visualization, Writing - Final Version.

Funding No funding was received for conducting this study.

Data availability and access Data will be made available on resonable request.

Declarations

Competing interests/ 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. The authors also declare that they have no conflicts of interest.

Ethical and informed consent for data used Not applicable

References

1. Anand S, Rahul Mallik A, Kumar S (2022) Integrating node centralities, similarity measures, and machine learning classifiers for link prediction. Multimed Tools and Appl 81(27):38593–38621
2. Daud NN, Ab Hamid SH, Saadoon M, Sahran F, Anuar NB (2020) Applications of link prediction in social networks: A review. J Netw Comput Appl 166:102716
3. Li S, Cai N, Yu Z (2020) Constructing friendship in social networks for precise peer influence marketing by consensus link prediction algorithm. Multimed Tools and Appl 79:7649–7668
4. Zhang L, Zhao M, Zhao D (2020) Bipartite graph link prediction method with homogeneous nodes similarity for music recommendation. Multimed Tools and Appl 79:13197–13215
5. Kaya B (2020) A hotel recommendation system based on customer location: a link prediction approach. Multimed Tools and Appl 79:1745–1758
6. Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai S-M, Hasan Q (2021) Application of network link prediction in drug discovery. BMC Bioinformatics 22:1–21
7. Rozemberczki B, Davies R, Sarkar R, Sutton C (2019) Gemsec: Graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 65–72
8. Akkem Y, Biswas SK, Varanasi A (2023) Smart farming using artificial intelligence: A review. Eng Appl Artif Intell 120:105899
9. Akkem Y, Kumar B, Varanasi A (2023) Streamlit application for advanced ensemble learning methods in crop recommendation systems-a review and implementation. Indian J Sci Technol 16:4688–4702
10. Akkem Y, Biswas SK, Varanasi A (2024) A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Eng Appl Artif Intell 131:107881
11. Li T, Zhang J, Philip SY, Zhang Y, Yan Y (2018) Deep dynamic network embedding for link prediction. IEEE Access 6:29219–29230
12. Cui P, Wang X, Pei J, Zhu W (2018) A survey on network embedding. IEEE Trans Knowl Data Eng 31(5):833–852
13. Cai H, Zheng VW, Chang KC-C (2018) A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Trans Knowl Data Eng 30(9):1616–1637
14. Wang P, Xu B, Wu Y, Zhou X (2014) Link prediction in social networks: the state-of-the-art. arXiv:1411.5118

15. Kumar A, Singh SS, Singh K, Biswas B (2020) Link prediction techniques, applications, and performance: A survey. Phys A 553:124289
16. Leskovec J, Huttenlocher D, Kleinberg J (2010) Governance in social media: A case study of the wikipedia promotion process. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 4, pp 98–105
17. Wu H, Song C, Ge Y, Ge T (2022) Link prediction on complex networks: An experimental survey. Data Sci and Engr 7(3):253–278
18. Chen J, Geyer W, Dugan C, Muller M, Guy I (2009) Make new friends, but keep the old: recommending people on social networking sites. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 201–210
19. Ahmed A, Shervashidze N, Narayanamurthy S, Josifovski V, Smola AJ (2013) Distributed large-scale natural graph factorization. In: Proceedings of the 22nd International Conference on World Wide Web, pp 37–48
20. Wu J, Shen J, Zhou B, Zhang X, Huang B (2019) General link prediction with influential node identification. Phys A 523:996–1007
21. Şimşek M, Meyerhenke H (2020) Combined centrality measures for an improved characterization of influence spread in social networks. J of Complex Netw 8(1):048
22. Papadimitriou A, Symeonidis P, Manolopoulos Y (2012) Fast and accurate link prediction in social networking systems. J Syst Softw 85(9):2119–2132
23. Yu C, Zhao X, An L, Lin X (2017) Similarity-based link prediction in social networks: A path and node combined approach. J Inf Sci 43(5):683–695
24. Wang P, Xu B, Wu Y, Zhou X (2014) Link prediction in social networks: the state-of-the-art. arXiv:1411.5118
25. Wei S, Wang L, Wu H, Zhou M, Li Q, Xiao Y (2024) Link prediction method for social networks based on a hierarchical and progressive user interaction matrix. Knowl-Based Syst, 111929
26. Kumar S, Mallik A, Panda B (2022) Link prediction in complex networks using node centrality and light gradient boosting machine. World Wide Web 25(6):2487–2513
27. Anand S, Rahul Mallik A, Kumar S (2022) Integrating node centralities, similarity measures, and machine learning classifiers for link prediction. Multimed tools and appl 81(27):38593–38621
28. Kumar S, Mallik A, Panda B (2023) Influence maximization in social networks using transfer learning via graph-based lstm. Expert Syst Appl 212:118770
29. Chen J, Wang X, Xu X (2022) Gc-lstm: Graph convolution embedded lstm for dynamic network link prediction. Appl Intell, 1–16
30. Yang K, Liu Y, Zhao Z, Zhou X, Ding P (2023) Graph attention network via node similarity for link prediction. The European Phys J B 96(3):27
31. Wang H, Cui Z, Liu R, Fang L, Sha Y (2023) A multi-type transferable method for missing link prediction in heterogeneous social networks. IEEE Trans on Knowl and Data Engr
32. Ma L, Zhang Y (2015) Using word2vec to process big text data. In: 2015 IEEE International Conference on Big Data (Big Data), pp 2895–2897 . IEEE
33. Ahmed A, Shervashidze N, Narayanamurthy S, Josifovski V, Smola AJ (2013) Distributed large-scale natural graph factorization. In: Proceedings of the 22nd International Conference on World Wide Web, pp 37–48
34. Ou M, Cui P, Pei J, Zhang Z, Zhu W (2016) Asymmetric transitivity preserving graph embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1105–1114
35. Perozzi B, Al-Rfou R, Skiena S (2014) Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 701–710
36. Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 855–864
37. Chen Y-L, Hsiao C-H, Wu C-C (2022) An ensemble model for link prediction based on graph embedding. Decis Support Syst 157:113753
38. Berahmand K, Nasiri E, Rostami M, Forouzandeh S (2021) A modified deepwalk method for link prediction in attributed social network. Computing 103:2227–2249
39. Guo Z, Wang F, Yao K, Liang J, Wang Z (2022) Multi-scale variational graph autoencoder for link prediction. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp 334–342
40. Ren X (2024) Link prediction using extended neighborhood based local random walk in multilayer social networks. J of King Saud Uni-Comp and Inf Sci 36(2)

41. Berahmand K, Nasiri E, Rostami M, Forouzandeh S (2021) A modified deepwalk method for link prediction in attributed social network. Computing 103:2227–2249
42. Cai H, Zheng VW, Chang KC-C (2018) A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Trans Knowl Data Eng 30(9):1616–1637
43. Menon AK, Elkan C (2011) Link prediction via matrix factorization. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II 22, pp 437–452 . Springer
44. Zhou W, Gu J, Jia Y (2018) h-index-based link prediction methods in citation network. Scientometrics 117:381–390
45. Nasiri E, Berahmand K, Li Y (2023) Robust graph regularization nonnegative matrix factorization for link prediction in attributed networks. Multimedia Tools and Applications 82(3):3745–3768
46. Yang C, Wang C, Zheng R, Geng S (2023) Link prediction in research collaboration: a multi-network representation learning framework with joint training. Multimed Tools and Appl, 1–19
47. Red V, Kelsic ED, Mucha PJ, Porter MA (2011) Comparing community structure to characteristics in online collegiate social networks. SIAM Rev 53(3):526–543
48. Rossi R, Ahmed N (2015) The network data repository with interactive graph analytics and visualization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 29
49. Rozemberczki B, Davies R, Sarkar R, Sutton C (2019) Gemsec: Graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 65–72
50. Rossi RA, Ahmed NK (2015) The network data repository with interactive graph analytics and visualization. In: AAAI . https://networkrepository.com
51. Kumar S, Mallik A, Panda B (2022) Link prediction in complex networks using node centrality and light gradient boosting machine. World Wide Web 25(6):2487–2513
52. Li K, Tu L, Chai L (2020) Ensemble-model-based link prediction of complex networks. Comput Netw 166:106978
53. Anand S, Rahul Mallik A, Kumar S (2022) Integrating node centralities, similarity measures, and machine learning classifiers for link prediction. Multimed tools and appl 81(27):38593–38621
54. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q (2015) Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp 1067–1077
55. Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. Adv In Neural Inf Process Syst 30

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.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384240814

Identifying Currency Crises in India: Empirically Weighted EMP Indices with Multiple Static Thresholds and Markov Regime-Switching Models Using Time-Varying Thresholds

Preprint · September 2024

DOI: 10.13140/RG.2.2.18699.30244

CITATIONS

0

2 authors, including:



Sanjay Kumar

Dyal Singh College

6 PUBLICATIONS **2** CITATIONS

SEE PROFILE

All content following this page was uploaded by Sanjay Kumar on 22 September 2024.

The user has requested enhancement of the downloaded file.

Identifying Currency Crises in India: Empirically Weighted EMP Indices with Multiple Static Thresholds and Markov Regime-Switching Models Using Time-Varying Thresholds

Sanjay Kumar*[1] and Dr. Nand Kumar†[2]

[1,2]Department of Humanities, Delhi Technological University, New Delhi.

September 22, 2024

*Department of Humanities, Delhi Technological University, New Delhi. Dyal Singh College, Delhi University, New Delhi.sanjaykumar.eco@dsc.du.ac.in Orcid: http://orcid.org/0000-0003-4613-076X

†Department of Humanities, Delhi Technological University, New Delhi.nandkumar@dce.ac.in Orcid: http://orcid.org/0000-0002-5684-6078

Abstract

This study introduces a unique and empirically-driven methodology to measure Exchange Market Pressure (EMP) in India from 2001 to 2022, offering an innovative approach not previously applied in the literature. By constructing three novel EMP indices through a flexible, data-centric process and comparing them to the conventional Girton-Roper EMP index, this research advances the field with an unprecedented application of Markov regime-switching models. These models incorporate both fixed and time-varying thresholds to distinctly identify normal and crisis regimes in India's foreign exchange market.

We employ a multifaceted set of criteria to detect currency crisis episodes, utilising the Eichengreen, Rose, and Wyplosz definition, alongside standard deviation-based thresholds, percentile-based thresholds, and duration-based parameters. The results reveal multiple currency crises in India, with the most critical periods aligning with the 2008 global financial crisis and the heightened global uncertainty between 2011 and 2013.

Robustness checks further validate the efficacy of this new methodology, highlighting India's foreign exchange market vulnerabilities to external shocks and domestic imbalances. Policy recommendations include the development of a comprehensive EMP assessment framework, maintaining a flexible exchange rate regime, and enhancing international policy coordination.

This research makes a significant contribution to the EMP literature, particularly for emerging economies, by pioneering a methodology that surpasses traditional approaches. While focused on a single country, future studies could expand this novel framework to other emerging markets, employing more sophisticated econometric techniques to explore dynamic interactions between EMP, currency crises, and macroeconomic variables.

Keywords: Exchange market pressure, Currency crisis, Emerging economies, India, Econometric modelling, Financial stability, Policy analysis.

JEL Classification: F31, F32, F34, F36, F41, G15, G18, C32, C51, C52.

1 Introduction

Exchange Market Pressure (EMP) indices have become indispensable tools in international finance for measuring currency pressures and identifying potential crisis episodes. These indices are particularly crucial for emerging economies like India, where exchange rate dynamics play a vital role in overall economic stability. The accurate measurement and analysis of EMP are essential for policymakers, investors, and researchers alike, as they provide critical insights into the health of a country's foreign exchange market and its vulnerability to external shocks.

The importance of EMP indices lies in their ability to capture the multifaceted nature of currency pressures, incorporating not only exchange rate movements but also policy responses such as interest rate changes and foreign exchange interventions. This comprehensive approach allows for a more nuanced understanding of the pressures facing a currency, beyond what can be gleaned from exchange rate movements alone. For India, which employs a managed float regime, EMP indices are particularly valuable in assessing the true extent of currency stress, even when exchange rates appear relatively stable.

Examining the volatility of EMP indices is crucial for several reasons. First, it provides insights into the stability of India's foreign exchange market and the effec-

tiveness of policy interventions. High volatility in EMP can indicate periods of market turbulence and potential vulnerability to speculative attacks. Second, understanding EMP volatility can help Indian policymakers design more effective intervention strategies and early warning systems for currency crises. Finally, for investors and financial institutions, EMP volatility is an important consideration in risk management and portfolio allocation decisions in the Indian market.

The concept of EMP, first introduced by Girton and Roper (1977b), has evolved significantly over the years. Researchers have refined and adapted EMP models to suit different economic contexts and data availability (Eichengreen et al., 1996b; Kaminsky et al., 1998b; Klaassen and Jager, 2011c). In the Indian context, several studies have attempted to measure EMP and identify currency crisis episodes using various approaches (Pattanaik and Parekh, 2003; Mishra, 2008; Guru, 2013). However, these studies often relied on traditional models that may not fully capture the complexities of India's managed float system and its increasing integration with global financial markets.

To provide a global context for our study, we examine EMP patterns across eleven countries from 1993 to 2017, as illustrated in Figure 1.

Figure 1 demonstrates the diverse EMP patterns across advanced economies, emerging markets, and developing nations. Advanced economies such as Australia

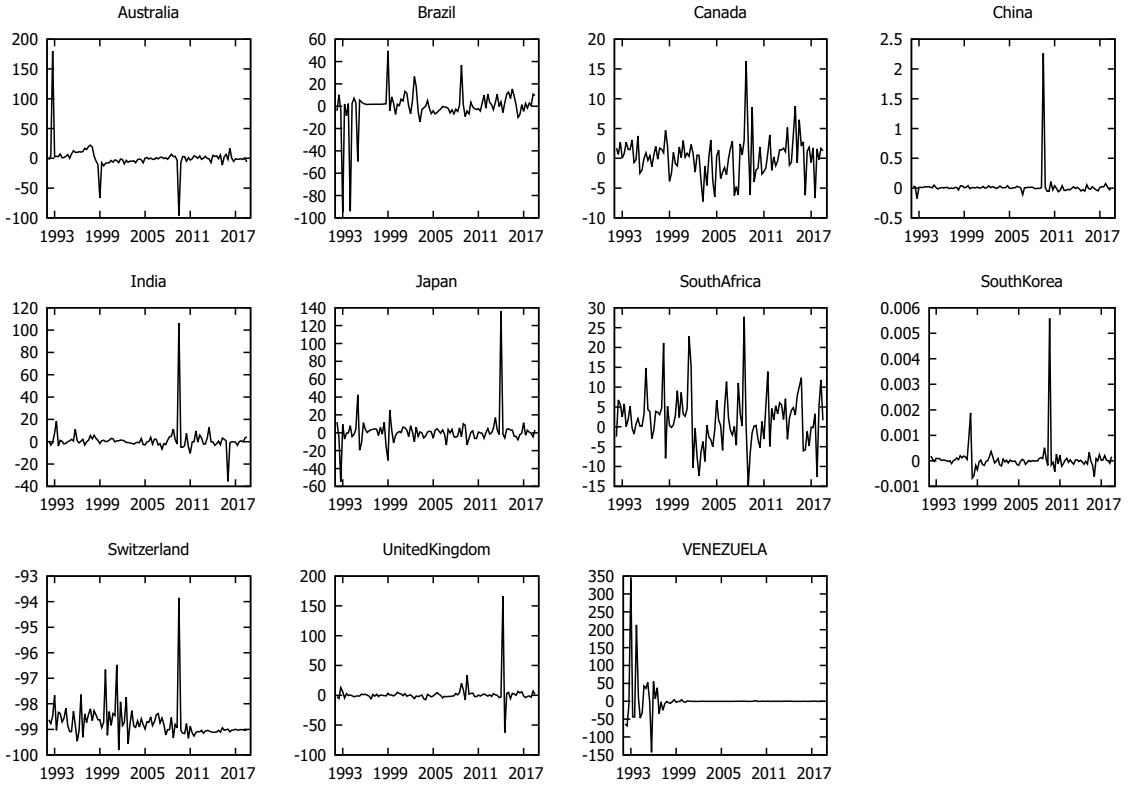


Figure 1: EMP Across Selected Countries

and Canada exhibit more stable EMP compared to emerging markets like Brazil and India. Crisis periods are clearly identifiable, such as the 2008 global financial crisis affecting multiple countries. Notably, India's EMP shows moderate volatility compared to other emerging markets, with significant spikes around 2008 and 2013, likely corresponding to the global financial crisis and the "taper tantrum" (Krishnamurthy and Vissing-Jorgensen, 2013).

This study proposes a novel and empirical approach to measuring EMP in India, addressing several limitations of existing models. Our methodology is justified by the need for a more flexible and data-driven approach that can identify the components and weights of the EMP index with minimal constraints. We construct four alternative EMP indices, including a benchmark Girton-Roper index and three indices based on our proposed general method. This approach allows for a comprehensive comparison and assessment of different EMP measurement techniques specifically tailored to the Indian context.

Furthermore, we employ advanced econometric techniques, including Markov regime-switching models with time-varying thresholds, to capture the dynamic nature of EMP and identify crisis periods more accurately. This methodological innovation builds upon recent advances in the literature, such as the work of **?**, who highlighted the importance of accounting for structural breaks and regime shifts when measuring

EMP in emerging economies like India.

The motivation for this research stems from several key factors. First, existing models for calculating EMP, such as the widely used Girton-Roper model, often impose structural restrictions and make assumptions about currency regimes that may not fully capture the complexities of India's managed float system. These limitations underscore the need for a more general and empirical approach that can identify the components and weights of the EMP index with minimal constraints. Second, there is a notable gap in the literature regarding the availability of a continuous time series of EMP indices for India, with the most recent comprehensive analysis concluding in 2008 (Guru, 2013). Given the significant changes in India's economic landscape and its increasing integration with global financial markets, extending the EMP index beyond 2008 is crucial for understanding the recent dynamics of India's foreign exchange market.

By applying this comprehensive methodology to India's foreign exchange market from 2001 onwards, we aim to provide a more accurate and up-to-date assessment of currency pressures in the country. This research not only contributes to the academic literature on EMP measurement but also offers valuable insights for Indian policymakers in formulating effective monetary and exchange rate policies, particularly in the context of India's evolving economic landscape and its increasing

integration with global financial markets.

This paper aims to address these gaps by applying our proposed empirical approach to the measurement of EMP in India from 2001 onwards. By constructing a time series of EMP indices specific to India and comparing them with the widely used Girton-Roper model, we seek to provide a more accurate and reliable assessment of currency pressures in the country. Furthermore, we analyse the identified currency crisis episodes and their potential determinants, drawing insights from relevant economic theories and empirical evidence (Dua and Tuteja, 2018; Panda and Mohanty, 2020).

2 Literature Review

2.1 Currency Crises and Exchange Market Pressure

Currency crises have been a recurring phenomenon in the global economy, with significant implications for countries' financial stability and economic growth. The concept of Exchange Market Pressure (EMP) has been widely used in the literature to study currency crises, as it captures the intensity of pressure on a country's currency, reflecting the actions taken by the authorities to defend the currency, such as changes

in foreign exchange reserves, interest rates, and the exchange rate itself (Girton and Roper, 1977b; Eichengreen et al., 1995).

Theoretical models of currency crises have evolved over time, from the "first-generation" models of Krugman (1979b) and Flood and Garber (1984), which emphasised the role of inconsistent macroeconomic policies and speculative attacks, to the "second-generation" models of Obstfeld (1986, 1996), which highlighted the role of self-fulfilling expectations and the trade-off between defending the currency and maintaining domestic economic objectives.

Recent studies have further expanded the understanding of currency crises and EMP in emerging economies. Dua and Tuteja (2018) investigated the determinants of EMP in the BRICS countries, finding that global financial conditions, domestic macroeconomic fundamentals, and policy variables play a significant role in driving currency pressures. Panda and Mohanty (2020) analysed the effectiveness of foreign exchange interventions in mitigating EMP in India, highlighting the importance of considering the asymmetric effects of buying and selling interventions. **?** explored the role of international reserves and macroprudential policies in managing EMP in emerging markets, emphasising the need for a coordinated policy approach to enhance financial stability.

2.2 Measuring Exchange Market Pressure: Models and Empirical Approaches

The accurate measurement of Exchange Market Pressure (EMP) is crucial for effectively studying currency crises. The seminal work by Girton and Roper (1977b) introduced the Girton-Roper model, which defines EMP as the sum of exchange rate changes and foreign exchange reserve changes, with equal weights assigned to both components. Subsequent studies have proposed various extensions and modifications to this model, aiming to capture the complexities of EMP in different economic contexts.

Roper and Turnovsky (1980) presented the Girton-Roper model in an IS-LM framework, allowing for different weights for the components of EMP. Weymark (1995, 1997) introduced a more general specification and the concept of a model-independent EMP index, which has been widely adopted in the literature.

Recent studies have emphasised the need for more flexible and data-driven approaches to measuring EMP, particularly in the context of emerging economies with unique foreign exchange market characteristics. Patnaik et al. (2017b) and Aizenman et al. (2018b) have proposed alternative weighting schemes and empirical methods to construct EMP indices that better capture the dynamics of these markets.

The literature has also explored various approaches to constructing EMP indices. Eichengreen et al. (1996b) proposed a widely-used EMP index that incorporates exchange rate changes, foreign exchange reserve changes, and interest rate changes. This approach has been applied to study the determinants and consequences of currency crises in both developed and emerging economies (Sachs et al., 1996; Kaminsky et al., 1998b).

Klaassen and Jager (2011d) advocate for considering multiple measures of exchange market pressure, arguing that each index captures slightly different aspects of currency pressure, providing a more comprehensive picture when considered together.

2.3 Identifying Currency Crisis Episodes

The identification of currency crisis episodes is a critical aspect of EMP analysis, and various methodologies have been proposed in the literature. Eichengreen et al. (1996b) introduced a widely-used crisis identification method based on extreme values of the EMP index, typically defined as exceeding a threshold of 1.5 or 2 standard deviations above the mean.

Kaminsky et al. (1998b) proposed an alternative approach using percentile-based thresholds, typically the 90th or 95th percentile of the EMP distribution. This method allows for a more flexible definition of crisis episodes that can adapt to the specific characteristics of different countries or time periods.

To address the potential issue of short-lived spikes in EMP indices, some researchers have introduced duration criteria for crisis identification. For example, Glick and Hutchison (2011b) require EMP indices to exceed the threshold for a specified number of consecutive periods to be classified as a crisis episode. This approach aims to distinguish between transitory market pressures and more sustained currency crises.

Edison et al. (2003) explored the use of multiple crisis identification methodologies, highlighting the importance of employing various approaches to ensure robust results. They argue that different thresholds and criteria can offer a more nuanced view of potential crisis periods and help identify both severe and milder episodes of currency stress.

2.4 Exchange Market Pressure and Currency Crises in India

India, being one of the largest and fastest-growing emerging economies, has experienced significant exchange rate fluctuations and episodes of currency stress in recent decades. Several studies have investigated EMP and currency crises in the Indian context, employing various models and empirical approaches.

Pattanaik and Parekh (2003) examined the effectiveness of foreign exchange interventions in managing EMP in India during the 1990s, finding that the Reserve Bank of India (RBI) intervened actively to stabilise the exchange rate. Dua et al. (2013) analysed the impact of macroeconomic fundamentals and policy variables on EMP in India, highlighting the importance of fiscal discipline and monetary stability in mitigating currency pressures.

Recent studies have focused on the specific challenges faced by India in managing its exchange rate and navigating the global financial landscape. Pandey et al. (2020) investigated the impact of the COVID-19 pandemic on India's foreign exchange market, analysing the RBI's response to the surge in EMP during this period. Sen Gupta and Sengupta (2021) examined the effectiveness of the RBI's foreign exchange interventions in the context of India's increasing integration with global financial markets, emphasising the need for a more flexible and market-oriented

approach to exchange rate management.

While these studies have contributed to the understanding of EMP and currency crises in India, there remains scope for a more comprehensive and up-to-date analysis that takes into account the recent developments in India's foreign exchange market and the evolving global economic landscape. This study aims to address this gap by constructing EMP indices specifically tailored to India's context and analysing the identified currency crisis episodes in light of the country's macroeconomic fundamentals, policy responses, and external vulnerabilities.

3 Research Objectives

This study aims to develop a comprehensive and empirical approach to measuring Exchange Market Pressure (EMP) in India and identify currency crisis episodes from 2001 to 2022. Our primary objective is to propose a general, data-driven method for constructing EMP indices tailored to emerging economies like India.

We construct four alternative EMP indices: one based on the Girton-Roper model and three using our proposed empirical approach. These new indices incorporate a broader set of macroeconomic and financial variables, allowing for time-varying weights estimated through econometric techniques such as OLS and IV estimation.

We compare the performance of these indices in capturing India's foreign exchange market dynamics and identifying periods of currency stress.

To enhance the robustness of our analysis, we employ Markov regime-switching models with both fixed and time-varying thresholds to identify distinct normal and crisis regimes in India's foreign exchange market. We also conduct various robustness checks, including alternative weighting methods (PCA, Factor Analysis, and Exponential Weighting) and different crisis identification thresholds.

We identify currency crisis episodes using the Eichengreen, Rose, and Wyplosz (ERW) crisis definition, as well as additional criteria such as standard deviation-based thresholds, percentile-based thresholds, and duration-based requirements. We analyse these episodes in terms of their potential determinants, duration, severity, and frequency, exploring their relationship with key macroeconomic and financial variables.

4 Methodology

This section presents the methodology employed in this study to construct alternative Exchange Market Pressure (EMP) indices for India, estimate the weights of the components in these indices, and identify currency crisis episodes using the Eichengreen,

Rose, and Wyplosz (ERW) crisis definition. The proposed methodology is designed to address the research objectives outlined in the previous section and build upon the findings and gaps identified in the literature review.

4.1 A General Method for Estimating EMP Indices

We propose a general expression for EMP that allows for the inclusion of multiple components, as shown in Equation 4.1:

$$EMP_i(t) = \Delta s_t + \omega_1 a_1(t) + \omega_2 a_2(t) + \cdots + \omega_n a_n(t) \tag{4.1}$$

where Δs_t represents the exchange rate changes, $a_1(t), a_2(t), \ldots, a_n(t)$ are the additional components of EMP (e.g., foreign exchange reserve changes, interest rate changes), and $\omega_1, \omega_2, \ldots, \omega_n$ are the corresponding weights.

This general specification allows for a more flexible and comprehensive measurement of EMP compared to traditional models, such as the Girton-Roper model (Girton and Roper, 1977a), which only considers exchange rate changes and foreign exchange reserve changes with equal weights. The inclusion of multiple components is motivated by the recent studies of Patnaik et al. (2017a) and Aizenman et al. (2018a), who emphasise the need for more context-specific and data-driven approaches to mea-

suring EMP in emerging economies.

4.2 Identification and Persistence Equation

To address potential endogeneity issues and capture the persistence of EMP, we propose the following equation to model regime-dependent persistence:

$$EMP_{(t)} = \beta + \rho_1(EMP_{t-1} - \lambda_1 \Delta s_{t-1}) + \rho_2(EMP_{t-2} - \lambda_2 \Delta s_{t-2}) + \rho_3(EMP_{t-3} - \lambda_3 \Delta s_{t-3}) + \cdots + u(t) \tag{4.2}$$

In this equation, $EMP_{(t)}$ represents the EMP index at time t, β is the intercept term capturing the fixed part of $EMP_i(t)$, $\rho_1, \rho_2, \rho_3, \ldots$ are the persistence parameters ($0 \leq \rho \leq 1$) for the lagged EMP terms, $\lambda_1, \lambda_2, \lambda_3, \ldots$ are the outlet parameters capturing the fraction of persistence reduction due to currency float, $\Delta s_{t-1}, \Delta s_{t-2}, \Delta s_{t-3}, \ldots$ are the lagged exchange rate changes, and $u(t)$ is the error term.

The adjustment of EMP_{t-k} by $\lambda \Delta s_{t-k}$ shows the reduction in persistence of EMP caused by currency float, with λ being the outlet parameter that captures the fraction of persistence reduction due to float.

This specification allows for a more accurate estimation of EMP persistence and accounts for the impact of currency float on EMP dynamics. The inclusion of regime-dependent persistence is inspired by the findings of Dua and Tuteja (2020),

who highlight the importance of accounting for structural breaks and regime shifts when measuring EMP in emerging economies.

4.3 Weight Equations

To estimate the weights of the components in the EMP indices, we derive the simplified form of the weight equation by combining Equations 4.1 and 4.2:

$$\Delta s_t = \gamma - \omega_1 \Delta s_t - \omega_2 a_2(t) + \rho_1 \omega_1 a_1(t-1) + \rho_2 \omega_2 a_2(t-1) + \rho_1(1-\lambda_1)\Delta s_{t-1} + u(t) \quad (4.3)$$

where γ is a constant term, ω_1 and ω_2 are the weights for the exchange rate changes and the additional component $a_2(t)$, respectively, ρ_1 and ρ_2 are the persistence parameters for the lagged EMP terms, λ_1 is the outlet parameter for the lagged exchange rate changes, and $u(t)$ is the error term.

We estimate the weights using the following equations:

$$\Delta s_t = \gamma - \omega \Delta f_t + \rho_1 \omega \Delta f_{t-1} + \rho_1(1-\lambda_1)\Delta f_{t-1} + u(t) \quad (4.4)$$

$$\Delta s_t = \gamma - \omega_1 \Delta f_t - \omega_2 \Delta i_t + \rho_1 \omega_1 \Delta f_{t-1} + \rho_2 \omega_2 \Delta i_{t-1} + \rho_1(1-\lambda_1)\Delta s_{t-1} + u_t \quad (4.5)$$

$$\Delta s_t = \gamma - \omega \Delta i_t + \rho_1 \omega \Delta i_{t-1} + \rho_1(1-\lambda_1)\Delta s_{t-1} + e_t \quad (4.6)$$

where Δf_t and Δi_t represent the changes in foreign exchange reserves and interest rates, respectively, ω, ω_1, and ω_2 are the weights for the respective components, ρ_1 and ρ_2 are the persistence parameters, λ_1 is the outlet parameter, and $u(t)$ and $e(t)$ are the error terms.

The use of these weight equations allows for a more accurate estimation of the weights assigned to the components of EMP indices, taking into account the persistence and regime-dependent dynamics of EMP. The inclusion of lagged variables and the outlet parameter λ captures the impact of currency float on EMP persistence, providing a more realistic representation of EMP dynamics in emerging economies like India.

4.4 Instrumental Variable Estimation of Weight Equations

To further address potential endogeneity issues in the weight equations, we employ instrumental variable (IV) estimation using the squared terms of the components as instruments. The use of squared terms as instruments is motivated by the findings of Krugman (1979a), who suggests that nonlinearities and volatility in the foreign exchange market during crisis periods can be captured by these higher-order terms.

The IV estimation is performed using two-stage least squares (2SLS), where

the endogenous variables are first regressed on the chosen instruments to obtain predicted values, which are then used in the second stage to estimate the weights. This approach ensures a more robust and reliable estimation of the weights, addressing potential biases arising from endogeneity.

4.5 Constructing EMP Indices

Based on the estimated weights, we construct four EMP indices:

$$EMP_t = \Delta s_t + \omega \Delta f_t \tag{4.7}$$

$$EMP_t = \Delta s_t + \omega_1 \Delta f_t + \omega_2 \Delta i_t \tag{4.8}$$

$$EMP_t = \Delta s_t + \omega \Delta i_t \tag{4.9}$$

$$EMP_t = \Delta s_t + \Delta f_t \tag{4.10}$$

where EMP_t represents the EMP index at time t, Δs_t is the exchange rate change, Δf_t is the change in foreign exchange reserves, Δi_t is the change in interest rates, and ω, ω_1, and ω_2 are the estimated weights for the respective components.

The first three indices (Equations 4.7-4.9) are based on our proposed general and empirical approach, incorporating different combinations of exchange rate changes, foreign exchange reserve changes, and interest rate changes, with weights

estimated using the aforementioned econometric techniques. The fourth index (Equation 4.10) is the Girton-Roper index, which serves as a benchmark for comparison.

The construction of multiple EMP indices allows for a comprehensive analysis of EMP dynamics in India and provides a robust framework for identifying currency crisis episodes. The inclusion of the Girton-Roper index alongside our proposed indices enables a comparison of the performance and suitability of different methodological approaches in the context of India's foreign exchange market.

4.6 Alternative Methods for Estimating EMPI Component Weights

To ensure robustness in the construction of our Exchange Market Pressure (EMP) index, we employ three distinct methodologies for estimating the weights of the EMP components: Principal Component Analysis (PCA), Factor Analysis (FA), and Exponential Weighting. Each method offers unique advantages in capturing the underlying relationships between the EMP components.

4.6.1 Principal Component Analysis (PCA)

Principal Component Analysis is a dimensionality reduction technique that transforms correlated variables into a set of uncorrelated principal components (Jolliffe, 2002). In the context of EMP indices, PCA helps identify the linear combination of components that explains the maximum variance in the data. We use the first principal component to derive the weights for our EMP index, following the approach of Pentecost et al. (2001).

4.6.2 Factor Analysis (FA)

Factor Analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors (Thompson, 2004). Unlike PCA, FA assumes that the observed variables are linear combinations of the underlying factors. We employ the maximum likelihood estimation method for factor extraction, followed by varimax rotation to improve interpretability (Fabrigar et al., 1999).

4.6.3 Exponential Weighting

Exponential weighting is a time series technique that assigns progressively higher weights to more recent observations (Gardner Jr, 2006). This method is particularly useful in capturing dynamic relationships in financial markets. We apply exponential weighting to our EMP components using the following formula:

$$w_t = (1 - \lambda)\lambda^{t-1} \tag{4.11}$$

where w_t is the weight at time t, and λ is the smoothing parameter ($0 < \lambda < 1$). We optimise λ using a grid search to minimise the mean squared error of our EMP index (Hyndman et al., 2008).

By employing these three distinct methodologies, we aim to provide a comprehensive and robust estimation of the EMP component weights. The consistency or divergence in weights across these methods offers valuable insights into the stability and dynamics of the pressure in the foreign exchange market.

4.7 Identifying Currency Crisis Episodes

4.7.1 Eichengreen Rose and Wyplosz Approach

To identify currency crisis episodes, we employ a modified version of the ERW crisis definition proposed by Eichengreen et al. (1996a). In our context, where depreciation and loss of forex reserves are represented by negative values in the EMP calculation, we adapt the crisis identification formula as follows:

$$\text{Crisis} = \begin{cases} 1 & \text{if } EMP_t < \mu_{EMP} - 1.5\sigma_{EMP} \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

where EMP_t is the EMP index at time t, μ_{EMP} is the mean of the EMP index, and σ_{EMP} is the standard deviation of the EMP index. This modified definition aligns with the IMF's methodology while accounting for our negative-value convention for depreciation and reserve loss. A currency crisis is said to occur when the EMP index falls below its mean by more than 1.5 times its standard deviation. This approach provides a consistent and widely accepted method for identifying crisis episodes, allowing for comparability across different EMP indices and countries. The use of this modified ERW crisis definition, in combination with our proposed EMP indices, enables a comprehensive analysis of the frequency, duration, and severity of currency

crisis episodes in India. By examining the identified episodes in the context of India's macroeconomic fundamentals, policy actions, and global economic conditions, we aim to provide valuable insights into the determinants and consequences of currency crises in emerging economies. The proposed methodology, with its combination of a general EMP specification, regime-dependent persistence equation, instrumental variable estimation of weights, and the use of the modified ERW crisis definition, represents a significant advancement in the measurement of EMP and the identification of currency crisis episodes in emerging economies like India. The superiority of this methodology lies in its ability to capture the complex dynamics of EMP, account for the impact of currency float on EMP persistence, and address potential endogeneity issues through instrumental variable estimation. While some studies have employed similar techniques in isolation, such as the use of regime-dependent persistence by Dua and Tuteja (2020) and the instrumental variable estimation of weights by Klaassen and Jager (2011a), the proposed methodology is unique in its comprehensive integration of these elements into a unified framework for EMP measurement and currency crisis identification. The construction of multiple EMP indices based on this methodology allows for a robust and nuanced analysis of India's foreign exchange market, providing policymakers and researchers with a valuable tool for monitoring and managing currency risks and promoting financial stability in the face of external shocks.

4.7.2 Alternative Thresholds and Approaches

To ensure a comprehensive and robust identification of currency crisis episodes, we employed multiple methodologies. Each approach offers unique insights and addresses different aspects of crisis identification. This section presents these alternative methods and their justifications.

4.7.2.1 Standard Deviation-Based Thresholds

Following Eichengreen et al. (1996a), we employed thresholds based on standard deviations from the mean. Crisis periods are identified when the EMP index exceeds its mean plus a certain number of standard deviations:

$$Crisis_t = \begin{cases} 1 & \text{if } EMP_t < \mu_{EMP} - k\sigma_{EMP} \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

where μ_{EMP} and σ_{EMP} are the mean and standard deviation of the EMP index, respectively. We used $k = 2$ and $k = 3$ to capture different levels of market pressure.

Justification: This method assumes that crisis episodes are characterised by

extreme values of the EMP index. It accounts for the overall volatility of the series and identifies periods of exceptional pressure.

4.7.2.2 Percentile-Based Thresholds

As suggested by Kaminsky et al. (1998a), we also used percentile-based thresholds. Crisis periods are identified when the EMP index exceeds a certain percentile of its distribution:

$$Crisis_t = \begin{cases} 1 & \text{if } EMP_t > P_q(EMP) \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

where $P_q(EMP)$ is the qth percentile of the EMP index. We used $q = 90$ and $q = 95$.

Justification: This approach is less sensitive to outliers and provides a more robust identification of crisis periods, especially when the EMP index does not follow a normal distribution.

4.7.2.3 Duration-Based Criterion

To address the issue of short-lived spikes in the EMP indices, we applied a duration-based criterion similar to that of Glick and Hutchison (2011a). This method requires two consecutive periods of high pressure to classify as a crisis:

$$Crisis_t = \begin{cases} 1 & \text{if } EMP_t < Threshold \text{ and } EMP_{t-1} < Threshold \\ 0 & \text{otherwise} \end{cases} \tag{4.15}$$

Justification: This criterion helps distinguish between sustained periods of pressure and temporary market volatility, potentially reducing false positives in crisis identification.

4.7.2.4 Markov Regime-Switching Model

We implemented a two-state Markov regime-switching model following Hamilton (1989):

$$EMP_t = \mu_{S_t} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_{S_t}) \tag{4.16}$$

where S_t is the unobserved state variable following a first-order Markov process. The high-volatility state ($S_t = 1$) is identified as the crisis regime.

Justification: This method allows for endogenous identification of crisis periods, accounting for potential structural changes in the foreign exchange market and the time-varying nature of crisis thresholds.

4.7.2.5 Markov Regime-Switching Model with Time-Varying Threshold

We extend the basic Markov regime-switching model to incorporate time-varying thresholds, following the approach of Chen et al. (2017). This model allows both the mean and variance of the EMP index to switch between regimes, while also allowing the threshold for regime switching to vary over time:

$$EMP_t = \mu_{S_t} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2_{S_t}) \tag{4.17}$$

$$S_t = \begin{cases} 1 & \text{if } EMP_t < c_t \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

$$c_t = c_0 + c_1 X_t \tag{4.19}$$

where S_t is the unobserved state variable, c_t is the time-varying threshold, and X_t is a vector of exogenous variables that may influence the threshold (e.g., global financial conditions, domestic economic indicators).

Justification: This sophisticated approach combines the strengths of regime-switching models with the flexibility of time-varying parameters. It allows for endogenous identification of crisis periods while accounting for changing economic conditions that may affect the threshold for what constitutes a crisis. This method is particularly suitable for long time series where the nature of crises may have evolved over time.

4.7.3 Implementation

We applied these methodologies to four different EMP indices: GREMPI, EMPI5.1, EMPI5.2, and EMPI5.3. This multi-index, multi-method approach, as suggested by Klaassen and Jager (2011b), allows for a more comprehensive view of exchange market pressure, as each index and method may capture slightly different aspects of currency pressure.

For the Markov regime-switching model with time-varying threshold, we included global VIX index and domestic interest rates as exogenous variables influencing the threshold, following Bussière and Hollo (2018). This allows us to account for both global risk sentiment and domestic monetary conditions in determining crisis thresholds.

By employing these diverse methodologies, we aim to provide a robust identification of currency crisis episodes in India. The consistency or divergence in crisis periods across these methods offers valuable insights into the nature and severity of the identified crises, while also highlighting the importance of methodological choices in crisis identification. The inclusion of more sophisticated methods like the Markov regime-switching model with time-varying threshold allows us to capture potential changes in the nature of crises over time, providing a more nuanced understanding of exchange market pressure in the Indian context.

5 Results and Findings

5.1 Three EMP Indices Based on a General Approach

Using the weights derived from the regression analysis, we construct three distinct EMP indices. These indices are called EMP indices based on a general method because they are not derived from any a priori theoretical model nor are they constrained by assumptions about a particular currency regime.

5.1.1 EMP Index Based on Equation (4.7)

$$EMPI_{(5.1)} = EMP_t = \Delta s_t + 9.98602\Delta f_t \tag{5.1}$$

5.1.2 EMP Index Based on Equation (4.8)

$$EMPI_{(5.2)} = EMP_t = \Delta s_t + 10.1311\Delta f_t - 0.0312216\Delta i_t \tag{5.2}$$

5.1.3 EMP Index Based on Equation (4.9)

$$EMPI_{(5.3)} = EMP_t = \Delta s_t - 0.0165941\Delta i_t \tag{5.3}$$

In addition to the above three indices, we also calculate the EMP index based on the Girton and Roper Model:

$$EMP_{(5.4)} = GREMPI_t = \Delta s_t + 1 \times \Delta f_t \tag{5.4}$$

5.2 Did India Face Any Currency Crises from 2001 Onwards?

Using the Eichengreen, Rose and Wyplosz (ERW) crisis threshold discussed in section 4.6, we identify periods of currency crises for each of the four EMP indices. A crisis is said to occur when the EMP index exceeds the threshold value of 1.5 standard deviations above its mean.

Table 1: ERW Crisis Threshold Values for Each EMPI

EMPI	ERW Crisis Threshold
GREMPI/EMP (5.4)	-0.0509132
EMPI (5.1)	-0.0503347
EMPI (5.2)	-0.059316
EMPI (5.3)	-0.0555121

Source: Author's own creation

5.2.1 Crises Based on GREMPI/EMP (5.4):

- 2008:4 (Worst Crisis): Peak to trough duration of 5 quarters (2007:4 to 2008:4)

- 2011:4: Peak to trough duration of 3 quarters (2011:2 to 2011:4)
- 2012:2: Peak to trough duration of 2 quarters (2012:1 to 2012:2)
- 2013:3: Peak to trough duration of 3 quarters (2012:4 to 2013:3)

Table 2: Crisis Periods Based on GREMPI/EMP (5.4)

Peak/Best Quarter	Trough/Worst Quarter	Peak to Trough Duration	Start of Crisis	End of Crisis
2007:4	2008:4	5Q	2008:3	2009:1
2011:2	2011:4	3Q	2011:3	2012:1
2012:1	2012:2	2Q	2012:2	2012:3
2012:4	2013:3	3Q	2013:2	2014:4

Source: Author's own creation

The worst crisis identified by GREMPI/EMP (5.4) occurred in 2008:4, coinciding with the global financial crisis. The crisis lasted for 5 quarters, from 2008:3 to 2009:1. This period was characterised by a significant depreciation of the Indian Rupee against the US Dollar and a sharp decline in foreign exchange reserves, indicating substantial pressure on the currency.

The second major crisis episode identified by this index occurred in 2011:4, with a peak to trough duration of 3 quarters (2011:3 to 2012:1). This period was marked by heightened global uncertainty and a decline in investor confidence, leading to capital outflows and pressure on the Indian Rupee.

Two shorter crisis episodes were identified in 2012:2 and 2013:3, with durations of 2 and 3 quarters, respectively. These episodes can be attributed to a

combination of domestic and global factors, such as slowing economic growth, widening current account deficit, and concerns over the tapering of quantitative easing by the US Federal Reserve.

5.2.2 Crises Based on EMPI (5.1):

- 2008:4 (Worst Crisis): Peak to trough duration of 5 quarters (2007:4 to 2008:4)
- 2011:4: Peak to trough duration of 6 quarters (2011:2 to 2011:4)
- 2012:2: Peak to trough duration of 6 quarters (2012:1 to 2012:2)
- 2013:3: Peak to trough duration of 5 quarters (2012:4 to 2013:3)

Table 3: Crisis Periods Based on EMPI (5.1)

Peak/Best Quarter	Trough/Worst Quarter	Peak to Trough Duration	Start of Crisis	End of Crisis
2007:4	2008:4	5Q	2008:3	2008:4
2011:2	2011:4	6Q	2011:3	2012:1
2012:1	2012:2	6Q	2012:2	2012:3
2012:4	2013:3	5Q	2013:2	2013:3

Source: Author's own creation

The crisis episodes identified by EMPI (5.1) are similar to those identified by GREMPI/EMP (5.4), with the worst crisis occurring in 2008:4. However, the durations of the crisis episodes are slightly longer, with the 2011:4 and 2012:2 episodes lasting for 6 quarters each.

5.2.3 Crises Based on EMPI (5.2):

- 2009:1: Peak to trough duration of 5 quarters (2008:1 to 2009:1)
- 2012:1 (Worst Crisis): Peak to trough duration of 3 quarters (2011:3 to 2012:1)
- 2012:3: Peak to trough duration of 1 quarter (2012:2 to 2012:3)
- 2013:3: Peak to trough duration of 3 quarters (2013:1 to 2013:3)

Table 4: Crisis Periods Based on EMPI (5.2)

Peak/Best Quarter	Trough/Worst Quarter	Peak to Trough Duration	Start of Crisis	End of Crisis
2008:1	2009:1	5Q	2008:4	2009:2
2011:3	2012:1	3Q	2011:4	2012:2
2012:2	2012:3	1Q	2012:3	2012:4
2013:1	2013:3	3Q	2013:3	2014:1

Source: Author's own creation

EMPI (5.2) identifies the worst crisis episode in 2012:1, with a peak to trough duration of 3 quarters (2011:4 to 2012:2). This index also identifies a crisis episode in 2009:1, with a duration of 5 quarters (2008:4 to 2009:2), which can be attributed to the spillover effects of the global financial crisis.

The shorter crisis episodes in 2012:3 and 2013:3 are consistent with the findings of the other indices, reflecting the turbulent global economic conditions and domestic vulnerabilities during this period.

5.2.4 Crises Based on EMPI (5.3):

- 2008:4: Peak to trough duration of 5 quarters (2007:4 to 2008:4)
- 2011:4 (Worst of the Worst Crisis): Peak to trough duration of 3 quarters (2011:2 to 2011:4)
- 2012:2: Peak to trough duration of 2 quarters (2012:1 to 2013:1)
- 2013:3: Peak to trough duration of 4 quarters (2012:4 to 2013:3)

Table 5: Crisis Periods Based on EMPI (5.3)

Peak/Best Quarter	Trough/Worst Quarter	Peak to Trough Duration	Start of Crisis	End of Crisis
2007:4	2008:4	5Q	2008:1	2009:1
2011:2	2011:4	8Q	2011:3	2011:4
2012:1	2012:2	2Q	2012:1	2012:2
2012:4	2013:3	7Q	2013:2	2013:4

Source: Author's own creation

EMPI (5.3) identifies the 2011:4 episode as the worst of the worst crisis, with a peak to trough duration of 8 quarters (2011:3 to 2011:4). This prolonged period of currency stress can be attributed to the persistent global uncertainties, domestic policy challenges, and the widening current account deficit during this period.

The other crisis episodes identified by EMPI (5.3) are consistent with the findings of the other indices, with the 2008:4 episode lasting for 5 quarters (2008:1 to 2009:1) and the 2013:3 episode lasting for 7 quarters (2013:2 to 2013:4).

Table 6: Summary of Crisis Periods Based on Four EMPIs

EMPI	Worst Crisis Quarter	Lowest EMPI Value	Coefficient of Variation	% of Time Below ERWCT
GREMPI/EMP (5.4)	2008:4	-0.10783	5.1532	4.65%
EMPI (5.1)	2008:4	-0.10784	5.3479	4.65%
EMPI (5.2)	2012:1	-0.21583	5.4572	5.88%
EMPI (5.3)	2011:4	-0.16459	5.9800	5.88%

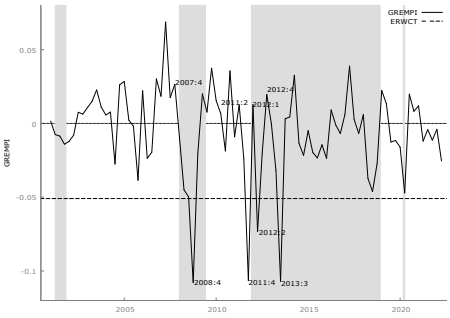
Source: Author's own creation

Table 6 provides a summary of the crisis periods identified by each EMPI, along with key statistics. EMPI (5.2) and EMPI (5.3) identify India as being in a crisis phase for 5.88% of the sample period, higher than the other two indices. EMPI (5.2) also records the lowest EMPI value, indicating the most severe crisis.

Figure 2: Currency Crisis: Four EMPIs and ERW Crisis Threshold

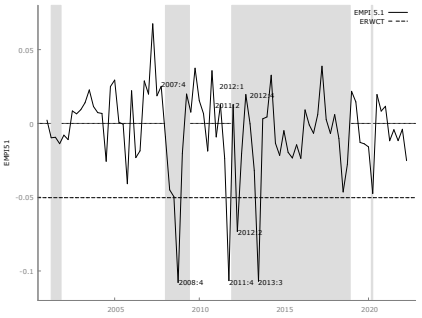
GREMPI

$EMP_t = \Delta s_t + 1 \times \Delta f_t$



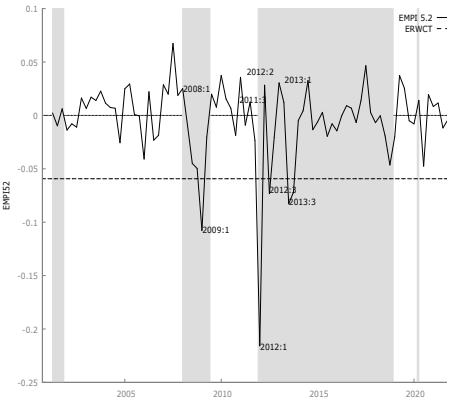
EMPI 6.4.1

$EMP_t = \Delta s_t + 9.98602\Delta f_t$



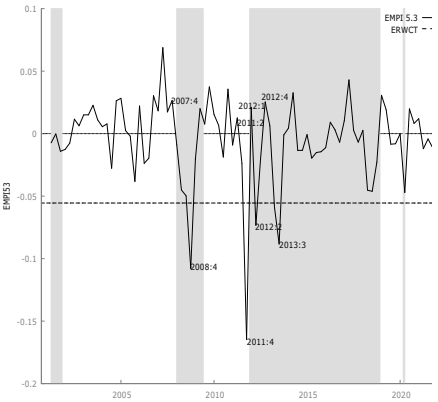
EMPI 6.4.2

$EMP_t = \Delta s_t + 10.1311\Delta f_t - 0.0312216\Delta i_t$



EMPI 6.4.3

$EMP_t = \Delta s_t - 0.0165941\Delta i_t$



Vertical patches in the graphs are detailed as below-

2001:04 2001:11. US Recession, NBER.

2008:01 2009:06. US Recession, NBER.

2012:00 2019:00. Global Uncertainty, WUI.

2020:03 2020:04. US Recession as per NBER and start of COVID -19 related disruption as per WHO.

Source: Author's own creation

Figure 1 depicts the Exchange Market Pressure Indices (EMPIs) for India from 2001 to 2022, along with the Eichengreen, Rose, and Wyplosz (ERW) crisis threshold. The four graphs represent the GREMPI/EMP (5.4), EMPI (5.1), EMPI (5.2), and EMPI (5.3), respectively. The vertical patches in the graphs highlight periods of US recessions and global uncertainty, as identified by the National Bureau of Economic Research (NBER) and the World Uncertainty Index (WUI). The ERW crisis threshold is indicated by a horizontal line, and periods when the EMPIs exceed this threshold are considered currency crisis episodes.

The figure clearly illustrates the multiple currency crises experienced by India during the sample period, with the most severe crisis occurring in 2011Q4, coinciding with the period of global uncertainty. The crisis episodes identified by the different EMPIs are broadly consistent, with some variations in the duration and severity of the crises.

The findings suggest that India's foreign exchange market is vulnerable to both global and domestic shocks, with periods of heightened global uncertainty and domestic macroeconomic imbalances leading to significant pressure on the Indian Rupee. The results also highlight the importance of considering multiple EMP indices and crisis definitions to obtain a comprehensive understanding of currency stress in emerging economies like India. Certainly. I'll reorganise the content under the new

title and structure it according to the alternative methodologies discussed earlier for both weights and thresholds. Here's the revised section:

5.3 Results Based on EMP Indices with Alternative Weights and Thresholds

In this section, we present the results of our robustness checks using alternative approaches for estimating EMP component weights and identifying crisis periods. These checks validate our main findings and provide additional insights into the currency crisis episodes in India.

5.3.1 Results from Alternative Weighting Methods

We employed three advanced statistical methods to ensure the robustness of the weights assigned to the components of the Exchange Market Pressure (EMP) indices: Principal Component Analysis (PCA), Factor Analysis (FA), and Exponential Weighting.

Principal Component Analysis (PCA) PCA transformed the original variables into uncorrelated principal components. The resulting weights are:

- Combination 1: Δs_t (-0.5), Δf_t (0.5)
- Combination 2: Δs_t (-0.5), Δi_t (0.5)
- Combination 3: Δs_t (-0.399), Δf_t (0.353), Δi_t (0.248)

Factor Analysis (FA) FA identified underlying relationships between variables by modeling observed variables as linear combinations of potential factors. The weights derived are:

- Combination 1: Δs_t (0.5), Δf_t (0.5)
- Combination 2: Δs_t (0.5), Δi_t (-0.5)
- Combination 3: Δs_t (0.492), Δf_t (0.328), Δi_t (-0.180)

Exponential Weighting Exponential weighting assigned progressively higher weights to more recent observations. The initial weights are:

- Combination 1: Δs_t (0.95444), Δf_t (0.04556)
- Combination 2: Δs_t (0.90544), Δi_t (0.09456)
- Combination 3: Δs_t (0.399), Δf_t (0.353), Δi_t (0.248)

The consistency observed across these methods, particularly in the relative importance of the variables, strongly supports the reliability and stability of our EMP indices. The slight variations in weights across methods provide valuable insights into the nuances of these relationships.

5.3.2 Results from Alternative Crisis Identification Methods

We employed multiple approaches to identify currency crisis periods: standard deviation-based thresholds, percentile-based thresholds, and a duration-based criterion. These alternative methods were applied to all four EMP indices: GREMPI, EMPI5.1, EMPI5.2, and EMPI5.3.

Standard Deviation-Based Thresholds Following Eichengreen et al. (1996b), we used thresholds of 2 and 3 standard deviations below the mean. The 2 standard deviations threshold identified crisis periods in:

- 2008Q4, 2011Q4, 2012Q2, and 2013Q3 for GREMPI and EMPI5.1
- 2009Q1, 2012Q1, 2012Q3, 2013Q3, and 2013Q4 for EMPI5.2
- An additional crisis in 2013Q2 for EMPI5.3

Interestingly, the crisis periods identified using 3 standard deviations were identical to those using 2 standard deviations for each index.

Percentile-Based Thresholds As suggested by Kaminsky et al. (1998b), we used 90th and 95th percentile thresholds. The 90th percentile threshold identified crisis periods in:

- 2008Q3, 2008Q4, 2011Q4, 2012Q2, and 2013Q3 for GREMPI and EMPI5.1
- 2008Q4, 2012Q3, 2013Q3, and 2013Q4 for EMPI5.2
- 2008Q4, 2011Q4, 2012Q2, 2013Q2, and 2013Q3 for EMPI5.3

The 95th percentile threshold, being the most sensitive, identified additional potential crisis episodes in 2005Q4, 2006Q1, 2018Q2-Q4, 2020Q2-Q3, and 2022Q2.

Duration-Based Criterion Following Glick and Hutchison (2011b), we applied a duration-based criterion requiring two consecutive quarters below the threshold. This approach yielded more conservative results:

- No crisis periods were identified for the 2 and 3 standard deviation thresholds across all indices

- The 90th percentile threshold identified crisis periods in 2008Q3-2008Q4 for GREMPI, EMPI5.1, and EMPI5.3, and in 2008Q4-2009Q1 for EMPI5.2
- The 95th percentile threshold identified the same crisis periods as the 90th percentile, with EMPI5.2 additionally showing a crisis in 2013Q3-2013Q4

5.3.3 Markov Regime-Switching Models with Fixed Thresholds

We applied Markov regime-switching models with fixed thresholds to each EMP index. Table 7 presents the key findings for all four indices: GREMPI, EMPI5.1, EMPI5.2, and EMPI5.3.

Table 7: Markov Switching Model Results with Static Threshold

	GREMPI	EMPI5.1	EMPI5.2	EMPI5.3
AIC	-378.6002	-381.2607	-366.294	-365.8818
BIC	-364.7828	-367.4433	-352.5234	-352.1112
Log Likelihood	191.3001	192.6304	185.147	184.9409
Regime 1 Intercept	-0.0477	-0.0498	-0.0555*	-0.0493*
Regime 1 Std. Error	0.0291	0.0297	0.0273	0.0251
Regime 2 Intercept	0.0009	0.0009	0.0036	0.0024
Regime 2 Std. Error	0.0028	0.0026	0.0026	0.0029
Transition Probability 1 to 1	0.5130413	0.4935102	0.602677	0.5595495
Transition Probability 1 to 2	0.4869587	0.5064898	0.397323	0.4404505
Transition Probability 2 to 1	0.07679219	0.0728353	0.06121032	0.07719324
Transition Probability 2 to 2	0.92320781	0.9271647	0.93878968	0.92280676

Note: * p ¡ 0.05

Source: Author's calculations

The results show that all four EMP indices exhibit clear regime-switching behavior, with distinct normal and crisis regimes. For GREMPI and EMPI5.1, Regime 1 represents the normal state and Regime 2 the crisis state, while for EMPI5.2 and EMPI5.3, this is reversed. The crisis regimes are characterised by negative intercepts, indicating increased pressure on the exchange rate during these periods.

The transition probabilities suggest that once in a crisis regime, there is a high probability of remaining in that regime, which aligns with the persistence of currency crises observed in practice. For instance, the probability of staying in Regime 2 (crisis state for GREMPI and EMPI5.1, normal state for EMPI5.2 and EMPI5.3) ranges from 0.92280676 to 0.93878968 across the four indices.

The model fit statistics (AIC, BIC, and Log Likelihood) are relatively similar across the four indices, with EMPI5.1 showing a slightly better fit according to these criteria. However, the differences are not substantial, suggesting that all four indices provide valuable information about exchange market pressure in India.

Figure 3 illustrates the crisis periods identified by our Markov switching models with fixed thresholds for each of the four EMP indices over the 2001-2022 period. The shaded areas represent crisis regimes, highlighting periods of significant exchange market pressure in India.

5.3.4 Markov Regime-Switching Models with Time-Varying Thresholds

We applied Markov regime-switching models with time-varying thresholds to each EMP index. Table 8 presents the key findings for all four indices: GREMPI, EMPI5.1, EMPI5.2, and EMPI5.3.

Table 8: Markov Switching Model Results with Time-Varying Threshold

	GREMPI	EMPI5.1	EMPI5.2	EMPI5.3
AIC	-378.6002	-381.2607	-366.294	-365.8818
BIC	-364.7828	-367.4433	-352.5234	-352.1112
Log Likelihood	191.3001	192.6304	185.147	184.9409
Regime 1 Intercept	0.0009	0.0009	-0.0555*	0.0024
Regime 1 Std. Error	0.0028	0.0026	0.0273	0.0029
Regime 2 Intercept	-0.0477	-0.0498	0.0036	-0.0493*
Regime 2 Std. Error	0.0287	0.0305	0.0026	0.0251
Transition Probability 1 to 1	0.92320707	0.92716541	0.6026772	0.92280676
Transition Probability 1 to 2	0.07679293	0.07283459	0.3973228	0.07719324
Transition Probability 2 to 1	0.4869579	0.5064908	0.06121027	0.4404505
Transition Probability 2 to 2	0.5130421	0.4935092	0.93878973	0.5595495

Note: * p ¡ 0.05

Source: Author's calculations

The results show that all models identified two distinct regimes with different intercepts and transition probabilities, providing further evidence of regime shifts in India's foreign exchange market. For GREMPI, EMPI5.1, and EMPI5.3, Regime 1 represents the normal state and Regime 2 the crisis state, while for EMPI5.2, this is reversed.

The transition probabilities reveal interesting patterns. For GREMPI, EMPI5.1, and EMPI5.3, there is a high probability (over 0.92) of remaining in the normal state (Regime 1) once it is entered. Conversely, for EMPI5.2, there is a high probability (0.93878973) of remaining in the crisis state (Regime 2). This suggests persistence in both normal and crisis periods, depending on the specific EMP index used.

The model fit statistics (AIC, BIC, and Log Likelihood) are consistent across both fixed and time-varying threshold models, with EMPI5.1 showing a slightly better fit. However, the differences are not substantial, indicating that all four indices provide valuable insights into exchange market pressure dynamics in India.

5.3.5 Summary of Crisis Periods

A combined analysis of crisis periods across all EMP indices and methods reveals:

- The global financial crisis period (2008-2009) is consistently identified as a crisis period across all approaches.
- The period from 2011Q4 to 2013Q4 is frequently identified as a crisis period, suggesting prolonged pressure on the Indian currency.
- Some indices and methods identify additional potential crisis periods in the mid-2000s and 2018.

These robustness checks underscore the importance of employing multiple crisis identification methodologies when studying exchange market pressure and currency crises. The consistency in identifying major crisis episodes (such as the 2008 global financial crisis) across different thresholds and indices lends credibility to our findings. However, the additional periods identified by the more sensitive thresholds, particularly in recent years, suggest that our analysis should consider potential changes in the nature and frequency of currency pressures over time.

As shown in Figure 4, The 2008-2009 global financial crisis stands out as a universally recognised crisis period, identified consistently across all analytical methodologies. Additionally, the analysis highlights another significant period of economic stress from the fourth quarter of 2011 to the fourth quarter of 2013, characterised by prolonged pressure on the Indian rupee. Beyond these widely acknowledged periods, various indices and methods suggest two more timeframes of potential economic instability:

Mid-2000s: Some indicators point to economic vulnerabilities in the period roughly between 2003 and 2006. 2018: Certain analytical tools identify this year as another period of potential crisis or economic strain.

6 Conclusion

This study presents a novel approach to measuring Exchange Market Pressure (EMP) and identifying currency crisis episodes in India from 2001 to 2022. We developed three new EMP indices alongside the benchmark Girton-Roper model, incorporating a broader set of variables and robust econometric techniques. Our analysis reveals that India experienced multiple episodes of significant currency pressure, with the 2008-2009 global financial crisis consistently identified as a major crisis period across all indices and methods. The 2011Q4 to 2013Q4 period was also frequently identified as a prolonged crisis, coinciding with heightened global uncertainty and domestic economic challenges.

Our methodology, which includes Markov regime-switching models and various robustness checks, effectively captures the complex dynamics of India's foreign exchange market. The use of multiple indices and crisis definitions provides a comprehensive understanding of currency stress, highlighting India's vulnerability to both external shocks and domestic imbalances.

This research has significant implications for policymakers and researchers alike. For policymakers, our approach offers a more accurate assessment of EMP, which can inform early warning systems and guide exchange rate management strate-

gies. For researchers, it contributes to the ongoing discourse on EMP measurement and currency crisis detection in emerging economies.

While acknowledging limitations such as the reliance on historical data and focus on a single country, this study makes a substantial contribution to the literature. It provides a flexible, data-driven framework that can be adapted to other emerging economies, potentially leading to new insights into currency pressure dynamics across different economic contexts.

Future research could extend this analysis to a broader set of countries, incorporate additional indicators, and explore the predictive power of these EMP indices. As emerging markets continue to play an increasingly important role in the global economy, such nuanced understanding of currency dynamics becomes crucial for maintaining financial stability in an interconnected world.

7 Policy Prescriptions

Based on our analysis of Exchange Market Pressure (EMP) and currency crisis episodes in India from 2001 to 2022, we propose several key policy recommendations. Policymakers should implement a comprehensive EMP monitoring framework that incorporates multiple indices and methodologies to gain a nuanced understanding of

currency pressures. Enhancing early warning systems by integrating EMP indices and Markov regime-switching models can improve timely crisis detection. Maintaining a flexible exchange rate regime while allowing for targeted interventions during extreme stress periods is crucial. Strengthening domestic financial markets and managing external vulnerabilities through adequate foreign exchange reserves and diversified funding sources can enhance resilience to currency pressures. Close policy coordination across economic management areas and improved transparency in foreign exchange operations are essential. Regular stress tests incorporating scenarios based on identified crisis episodes should be conducted. Policymakers should actively participate in regional and global forums to strengthen international cooperation and financial safety nets. Investing in high-quality data and advanced analytical capabilities is vital for effective policy analysis and implementation in an increasingly complex global financial system.

8 Limitations and Future Research

While this study provides significant insights into EMP and currency crises in India, several limitations should be acknowledged. The focus on a single country may limit generalisability, and future research could extend the analysis to a broader set of emerging economies. Data quality issues and the specific time period studied

may affect the accuracy of results, suggesting the need for exploration of alternative data sources and longer time horizons. The study's reliance on specific crisis identification methods and Markov regime-switching models could be complemented by investigating additional techniques in future research. Further exploration of dynamic interactions between EMP, currency crises, and other macroeconomic variables using advanced econometric methods could provide deeper insights. Future studies could also assess the predictive power of EMP indices for future crises and analyse the impact of specific policy interventions during crisis periods. Incorporating a more systematic analysis of global factors influencing EMP in emerging economies and exploring alternative weighting methods for constructing EMP indices could yield additional valuable insights. Addressing these limitations in future research will contribute to a more comprehensive understanding of currency stress and crisis dynamics in emerging economies, ultimately providing policymakers with more effective tools for managing exchange rate risks and promoting financial stability.

References

J. Aizenman, Y. Jinjarak, H. T. K. Nguyen, and D. Park. Flexibility of adjustment to shocks: Economic growth and volatility of middle-income countries before and after the global financial crisis of 2008. *Emerging Markets Finance and Trade*, 54

(5):1112–1131, 2018a.

J. Aizenman, Y. Jinjarak, H. T. K. Nguyen, and D. Park. Flexibility of adjustment to shocks: Economic growth and volatility of middle-income countries before and after the global financial crisis of 2008. *Emerging Markets Finance and Trade*, 54 (5):1112–1131, 2018b.

M. Bussière and D. Hollo. Predicting crises, Part I: Do coming crises cast their shadows before? *Journal of International Money and Finance*, 86:153–170, 2018.

C. Y.-H. Chen, T. C. Chiang, and W. K. Härdle. A new approach to detect and identify the number of regimes in Markov-switching models. *Journal of Econometrics*, 196(1):104–118, 2017.

P. Dua and D. Tuteja. Determinants of exchange market pressure in india. *The Journal of Development Studies*, 54(8):1417–1435, 2018.

P. Dua and D. Tuteja. Measuring exchange market pressure in india: New empirical evidence. *Economic Modelling*, 88:29–41, 2020.

P. Dua, R. Ranjan, and S. Sahoo. Exchange rate policy and modelling in india. 2013.

H. J. Edison, M. W. Klein, L. A. Ricci, and T. Sloek. Capital account liberalization and economic performance: Survey and synthesis. Technical report, National Bureau of Economic Research, 2003.

B. Eichengreen, A. K. Rose, and C. Wyplosz. Exchange market mayhem: the antecedents and aftermath of speculative attacks. *Economic policy*, 10(21):249–312, 1995.

B. Eichengreen, A. K. Rose, and C. Wyplosz. Contagious currency crises. *Scandinavian Journal of Economics*, 98(4):463–484, 1996a. doi: 10.2307/3440879.

B. Eichengreen, A. K. Rose, and C. Wyplosz. Contagious currency crises: First tests. *Scandinavian Journal of Economics*, 98(4):463–484, 1996b.

L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

R. P. Flood and P. M. Garber. Collapsing exchange-rate regimes: Some linear examples. *Journal of International Economics*, 17(1-2):1–13, 1984.

E. S. Gardner Jr. Exponential smoothing: The state of the art—part ii. *International journal of forecasting*, 22(4):637–666, 2006.

L. Girton and D. Roper. A monetary model of exchange market pressure applied to the postwar canadian experience. *The American Economic Review*, 67(4):537–548, 1977a.

L. Girton and D. Roper. A monetary model of exchange market pressure applied to

the postwar canadian experience. *The American Economic Review*, 67(4):537–548, 1977b.

R. Glick and M. Hutchison. Currency crises. *The Encyclopedia of Financial Globalization. Amsterdam: Elsevier*, 2011a.

R. Glick and M. Hutchison. Currency crises. *Federal Reserve Bank of San Francisco Working Paper*, (2011-22), 2011b.

A. Guru. Exchange market pressure in india. *Journal of Global Economy*, 9(2): 133–145, 2013.

J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384, 1989.

R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

I. Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

G. Kaminsky, S. Lizondo, and C. M. Reinhart. Leading indicators of currency crises. *IMF Staff papers*, 45(1):1–48, 1998a.

G. Kaminsky, S. Lizondo, and C. M. Reinhart. Leading indicators of currency crises. *Staff Papers*, 45(1):1–48, 1998b.

F. Klaassen and H. Jager. Identifying the weights in exchange market pressure. *Journal of International Money and Finance*, 30(7):1279–1305, 2011a.

F. Klaassen and H. Jager. Testing exchange rate models: a sensitivity analysis. *Applied Economics*, 43(12):1429–1445, 2011b.

F. Klaassen and H. Jager. Identifying the weights in exchange market pressure. *Journal of International Money and Finance*, 30(7):1279–1305, 2011c.

F. Klaassen and H. Jager. Model-free measurement of exchange market pressure. *Journal of International Money and Finance*, 30(1):189–206, 2011d.

A. Krishnamurthy and A. Vissing-Jorgensen. The ins and outs of lsaps. *Kansas City Federal Reserve Symposium on Global Dimensions of Unconventional Monetary Policy*, pages 57–111, 2013.

P. Krugman. A model of balance-of-payments crises. *Journal of Money, Credit and Banking*, 11(3):311–325, 1979a.

P. Krugman. A model of balance-of-payments crises. *Journal of Money, Credit and Banking*, 11(3):311–325, 1979b.

A. Mishra. Measuring exchange market pressure in india. *Economic and Political Weekly*, 43(49):46–50, 2008.

M. Obstfeld. Rational and self-fulfilling balance-of-payments crises. *American Economic Review*, 76(1):72–81, 1986.

M. Obstfeld. Models of currency crises with self-fulfilling features. *European Economic Review*, 40(3-5):1037–1047, 1996.

S. Panda and R. K. Mohanty. Asymmetric effect of exchange rate intervention on exchange market pressure: Evidence from india. *Journal of Economic Studies*, 47 (6):1221–1237, 2020.

R. Pandey, G. K. Pasricha, I. Patnaik, and A. Shah. Anatomy of indian rupee's value depreciation and reserve bank of india's response during the recent global financial crisis. *Journal of International Commerce, Economics and Policy*, 11(02):2050009, 2020.

I. Patnaik, J. Felman, and A. Shah. Exchange market pressure in india. *Journal of International Money and Finance*, 73:62–77, 2017a.

I. Patnaik, J. Felman, and A. Shah. Exchange market pressure and intervention in india. *Journal of International Money and Finance*, 73:62–77, 2017b.

S. Pattanaik and S. Parekh. Exchange market pressure and rbi intervention activity. *Economic and Political Weekly*, 38(21):2103–2110, 2003.

E. J. Pentecost, C. Van Hooydonk, and A. Van Poeck. Measuring and estimating

exchange market pressure in the eu. *Journal of International Money and Finance*, 20(3):401–418, 2001.

D. E. Roper and S. J. Turnovsky. Optimal exchange market intervention in a simple stochastic macro model. *Canadian Journal of Economics*, pages 296–309, 1980.

J. Sachs, A. Tornell, and A. Velasco. Financial crises in emerging markets: the lessons from 1995. *Brookings Papers on Economic Activity*, 1996(1):147–199, 1996.

A. Sen Gupta and R. Sengupta. Foreign exchange intervention in india: Recent developments and empirical assessment. *India Policy Forum*, 18(1):47–87, 2021.

B. Thompson. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association, 2004.

D. N. Weymark. Estimating exchange market pressure and the degree of exchange market intervention for canada. *Journal of International Economics*, 39(3-4):273–295, 1995.

D. N. Weymark. Measuring the degree of exchange market intervention in a small open economy. *Journal of International Money and Finance*, 16(1):55–79, 1997.

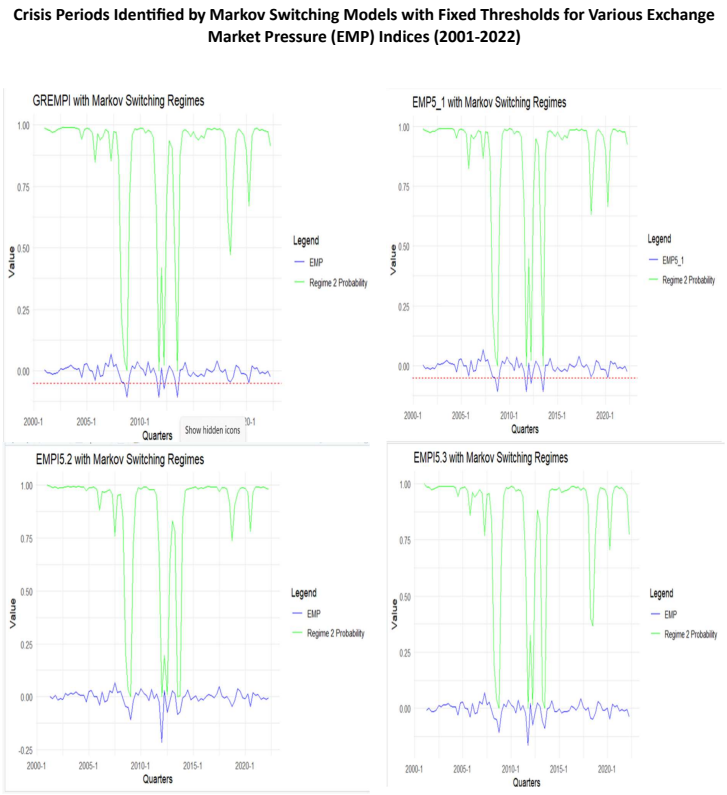


Figure 3: Crisis Periods Identified by Markov Switching Models with Fixed Thresholds for Various Exchange Market Pressure (EMP) Indices (2001-2022)

Source: Author's creation

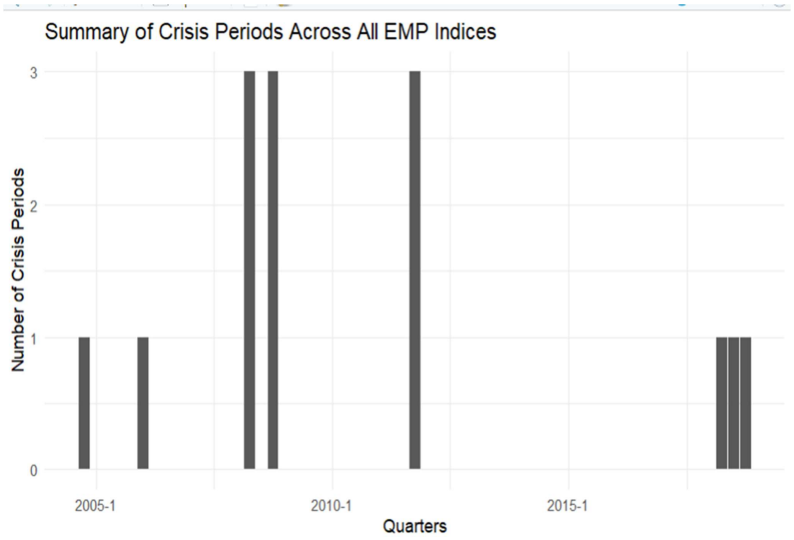


Figure 4: Summary of crisis periods identified through various EMP indices and methods.

Source: Author's creation

View publication stats

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384428860

Improved ML based Load Forecasting using Time- Split Convolutional Neural Networks

Conference Paper · September 2024

CITATIONS
0

READS
7

3 authors, including:



Jitendra Nath Rai
Delhi Technological University, Government of NCT of Delhi, India
82 PUBLICATIONS **462** CITATIONS

SEE PROFILE

All content following this page was uploaded by Jitendra Nath Rai on 28 September 2024.

The user has requested enhancement of the downloaded file.

Improved ML based Load Forecasting using Time-Split Convolutional Neural Networks

Muskan
Department of Electrical Engineering
Delhi Technological University
Delhi, India
muskan_ee20a14_17@dtu.ac.in

Muskan
Department of Electrical Engineering
Delhi Technological University
Delhi, India
muskan_ee20a14_18@dtu.ac.in

Pragati Sattavan
Department of Electrical Engineering
Delhi Technological University
Delhi, India
pragatisattavan_ee20a14_28@dtu.ac.in

JN Rai
Department of Electrical Engineering
Delhi Technological University
Delhi, India
jnrai@dce.ac.in

***Abstract*—Load forecasting is the process of predicting future energy consumption patterns. It is crucial for efficient energy management, helping utilities plan and allocate resources to meet demand, optimize grid operations, and ensure a reliable power supply. By leveraging the adaptability and predictive capabilities of machine learning, the accuracy of load forecasting can be enhanced. We aim to improve the accuracy of the standalone machine learning models by utilizing an efficient representation of the dataset. The technique uses CNN (Convolution Neural Networks) to derive a meaningful representation of the dataset by converting time series data to five channels and then training the machine learning models on the derived dataset. The results showed improved performance of all the models in consideration. A noticeable decrease in the losses [MAPE, MAE, MSE] is also reported.**

Keywords— Machine Learning, Time Series, Convolutional Neural Networks, Load Forecasting.

I. INTRODUCTION

Load forecasting is vital for energy management and grid operations, providing crucial insights into future electricity demand. It enables efficient resource allocation and grid stability by anticipating demand patterns. Without accurate forecasts, stakeholders face challenges in optimizing resource utilization, minimizing costs, and ensuring grid reliability. Load forecasting allows organizations to adjust production and distribution, optimize resources, and identify opportunities for demand response programs, enhancing overall efficiency and reducing the need for costly infrastructure upgrades[6]. For grid operators, accurate forecasts help in maintaining stability, planning capacity, and integrating renewable energy sources effectively.

Machine learning plays a key role in enhancing load forecasting accuracy. Leveraging advanced algorithms and data-driven insights, machine learning models can extract meaningful patterns from historical data[8], improving forecast precision. The specific aim of the study is to refine standalone load forecasting models using machine learning techniques to elevate accuracy and reliability. This enhances energy management strategies.

The goal is to improve the performance of all the models with a noticeable measure. While individual models offer unique strengths, a holistic approach that focuses on enhancing the accuracy of all ML models rather than prioritizing one over others has been adopted here. By focusing on improving the accuracy of each model and leveraging their collective intelligence, a superior prediction performance and robustness in dynamic and complex environments can be achieved. This approach not only advances the field of machine learning but also holds promise for a wide range of applications where accurate predictions are critical.

A. Influencing Factors

A variety of factors affect the load forecast. Past energy usage patterns provide insights into recurring trends and seasonality, forming a basis for load predictions. Climate influences energy demand, as temperature variations affect heating, cooling, and overall consumption patterns. Economic activities impact energy usage, with factors like GDP, industrial output, and employment rates influencing demand. Population growth, urbanization, and societal changes contribute to shifts in energy consumption patterns over time. Government policies, environmental regulations, and incentives shape energy consumption patterns and influence forecasting considerations.

B. Types of Load Forecasting

There can be different timeframes for which we predict the load and hence we have three different types of load forecasting [7].

- Short Term. Short-term load forecasting is crucial for predicting energy demand in the immediate future, typically spanning hours to a few days[15]. This type of forecasting is essential for real-time decision-making, grid balancing, and ensuring the optimal utilization of resources.
- Medium Term. Medium-term load forecasting extends the prediction horizon to weeks or months, offering insights into energy consumption patterns for mid-range planning. This type of forecasting is instrumental in optimizing resource allocation, scheduling maintenance activities, and making informed decisions about infrastructure and capacity planning
- Long Term. Long-term load forecasting spans over months to years and is crucial for strategic planning, capacity expansion, and infrastructure development.

[14]. It provides stakeholders with a forward-looking perspective on future energy needs, facilitating the planning and implementation of new power generation projects, transmission lines, and other critical infrastructure.

II. LITERATURE REVIEW

Forecasting electricity consumption is integral to effective energy management and planning within buildings. Current research in this field emphasizes key areas such as smart grid development, the integration of renewable energy sources, and the establishment of local energy communities. Machine learning (ML) and deep learning (DL) models have significantly advanced this research by offering highly accurate and reliable predictions.

There have been multiple attempts to compare the existing state of the art methodologies for their performance at load forecasting [1, 16]. Time series models like the autoregressive integrated moving average model (ARIMA) and the seasonal autoregressive integrated moving average (SARIMA) have been extensively employed in the energy sector for forecasting purposes. [2, 3, 4]. A study introduces a method that leverages generative adversarial networks (GANs) to produce parallel energy consumption data [5].

Short term load forecasting has gained to prominence having realized its advantages using multiple techniques like XGBoost [9], neural networks [17], LSTMS [18] and comparisons have been drawn [10] with the various short term load forecasting techniques. Various methods like RNNs and LSTMs have also been explored [11, 13]. A combination of these sophisticated methods along with the ensemble learning has proven to be one of the strongest predictive models in the domain [12].

III. ANALYSIS

A. About the Dataset

The dataset utilized in machine learning for load forecasting is sourced from Kaggle and centers around Panama, a country in South America. With a substantial size of 10 megabytes and a vast dataset comprising over 50,000 entries, it spans a comprehensive four-year duration, offering per-hour timestamps to capture feature variations. The dataset incorporates three key cities in Panama—Tocumen, David City, and San Tiago— providing regional diversity. The dataset's 17 feature columns play a pivotal role, encompassing crucial meteorological parameters like temperature, precipitation, humidity, and windspeed.

B. Machine Learning Models Used

Five models have been utilized – Linear Regression, Decision Trees, Random Forest, Support Vectors and Gradient Boosting.

Linear Regression predicts a continuous outcome based on linear relationships between input variables and the target, assuming a constant rate of change. Decision Trees are hierarchical tree-like structures that make decisions based on feature values, recursively splitting data to create branches representing classification or regression outcomes. Random Forest is an ensemble of decision trees, each trained on a random subset of data and features, combining predictions to improve accuracy and reduce overfitting.

Support Vector (SVR) classifies data by finding the hyperplane that maximizes the margin between different classes, transforming input features into a higher-dimensional space if needed. Gradient Boosting builds a strong predictive model by sequentially adding weak models, each correcting the errors of its predecessor, optimizing overall accuracy.

C. Feature Engineering

The dataset originally comprised of 17 feature columns. The timestamps were converted to 5 more features for each entry in the dataset depicting the day of the week, the month, the year, day type (weekday or weekend) and season (dry or rainy). Based on this extraction, the number of features grew to 22. Feature extraction would help in retaining useful information from the timestamps and make it easier for the models to understand the trends

D. Dataset Pre-Processing

Outliers, or data points that significantly deviate from the overall distribution, were identified and subsequently removed during the preprocessing of the Panama dataset. Duplicate and null values were systematically eliminated from the dataset. This process reduces redundancy and ensures that each entry is unique. An essential aspect of the preprocessing involved visualizing trends within the dataset. This step provided valuable insights into temporal patterns in both meteorological and load-related features. Categorical data present in the dataset underwent label encoding. This transformation converts categorical variables into a numerical format

IV. TRAINING AND TESTING MACHINE LEARNING MODELS

A. Splitting the Dataset

Before the pre-processed data is sent for training the models, we have adopted a standard practice of splitting the dataset into 80-20 ratio. This division allocated 80% of the dataset for training the machine learning model, allowing it to learn patterns and relationships inherent in the data. The remaining 20% was reserved for testing the trained model, providing a dedicated subset for evaluating its performance on unseen data. Each model was trained on the training dataset separately and results were recorded in terms of the evaluation metrics such as Mean Absolute Error, Mean Absolute Percentage Error and Mean Squared Error.

B. Results

Table I shows the results obtained after training ML models on the preprocessed dataset. The training time in seconds, errors (MAPE, MSE and MAE) are reported as follows for each ML model trained on the original dataset. The training and testing accuracy have also been calculated and compared model wise. It is observed that the performance of Random Forest on unseen data is the best with an accuracy of 91.9%.

TABLE I. ERRORS AND ACCURACIES FOR ALL ML MODELS

Model Name	Training Accuracy	Testing Accuracy	Training Time (s)	MAE	MSE	MAPE
Linear Regression	**0.62**	**0.63**	**0.152**	**92.727**	**13905.699**	**0.079**
Decision Tree	**1.00**	**0.83**	**1.424**	**51.350**	**6453.528**	**0.044**
Random Forest	**0.99**	**0.92**	**92.098**	**37.476**	**2994.970**	**0.039**
Gradient Boost	**0.85**	**0.85**	**23.732**	**54.289**	**5751.530**	**0.046**
Support Vector	**0.92**	**0.66**	**0.009**	**50.064**	**5042.310**	**0.042**

The errors produced by this model are also less compared to other models. Gradient Boost and SVR also have a comparable performance as that of Random Forest. But Decision Tree seems to be overfitting to the data while Linear Regression is far from making an average correct prediction. Thus, Random Forest is set up as a baseline with 0.039 MAPE.

V. PROPOSED METHODOLOGY

A. *Convolutional Neural Networks*

Convolutional Neural Networks (CNNs) extract features through convolutional layers, which apply filters to input images, detecting patterns like edges or textures. Pooling layers reduce spatial dimensions, focusing on important features. These extracted features are then flattened and passed to fully connected layers.

Channels in CNNs represent different attributes of the input data. These are useful because they allow the network to learn and extract different types of features. Each channel represents a different aspect such as color, texture, or orientation. Channels enable the network to capture complex patterns and improve its ability to classify or understand patterns.

B. *Data Preparation*

The original dataset is big with five years data recorded for every fifth minute. The ML models in general had a great training accuracy but not so good testing accuracy. The intuition here is to use CNNs to extract a representation that serves dual purpose

- Remove the redundant features and obtain meaningful representation of the dataset
- Reduce the size of the data the ML models train on so that the computational cost reduces in general for a model

Fig. 1 shows the proposed architecture. The data (spanning across five years) was divided into five different data frames each consisting of the yearly data identified by the timestamps. Not all the data frames were of the same size, so a median size of 8759 rows was decided for all five the input data frames. The original frames with less rows as compared to the median size were imputed with the mean and variance values of the specific columns. This resulted in 5 frames of size 8759 rows and 22 features each. These frames were stacked to form the five channels of the input to the Time-split CNN model.

The target i.e., the national demand of the load was similarly divided into five channels basis the year. This was condensed into a 1D vector by taking mean of the corresponding entries for a particular time of the year in all five channels.

C. *Training and Testing*

The resultant data with five channels was split into 80:20 ratio 80% data would be used to extract a feature map but the 20% data would be totally unseen. So, there would be no means by which this chunk of data gets learned. This is done to ensure that the models stay robust and do not break on a completely new data. The training data is passed through the Time-split CNN model where the convolutional layers and the max pool layers work in conjunction to derive a meaningful representation for the ML models to train on.

The output of the Time split CNN model is a feature map that captures the significant features with respect to the variations between the channels for the same time of the year. This lets us condense the original data to have only relevant and well-structured features. The new condensed feature map became the training set for the ML models in consideration. The test set is derived from the 20% portion of the split-data. The ML models were trained on the feature map. This enabled them to learn the most important features for load prediction with less computations. Table II shows the results after training on feature map.

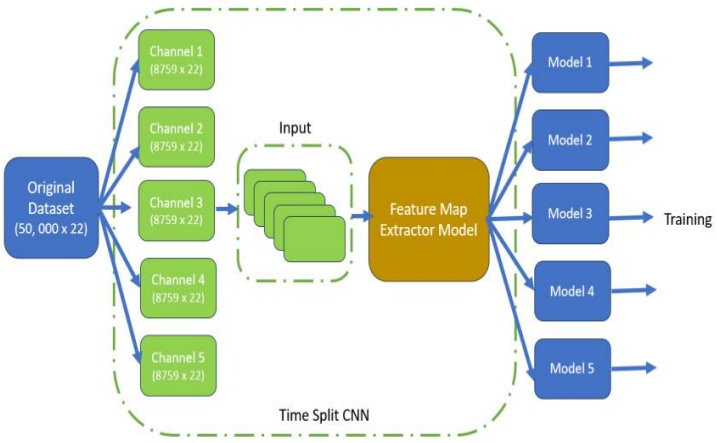


Fig. 1. Proposed method where the original dataset is divided into channels, vertically stacked and passed to a feature map extractor model.

TABLE II. ERRORS AND ACCURACIES FOR ALL ML MODELS AFTER TRAINING ON FEATURE MAP

Model Name	Training Accuracy	Testing Accuracy	Training Time (s)	MAE	MSE	MAPE
Linear Regression	**0.62**	**0.63**	**0.041**	**88.069**	**10386.970**	**0.066**
Decision Tree	**1.00**	**0.82**	**0.066**	**46.768**	**7854.433**	**0.043**
Random Forest	**0.99**	**0.92**	**4.351**	**31.935**	**2476.539**	**0.029**
Gradient Boost	**0.85**	**0.85**	**3.763**	**67.319**	**6874.549**	**0.058**
Support Vector	**0.92**	**0.66**	**1.238**	**145.908**	**32677.709**	**0.090**

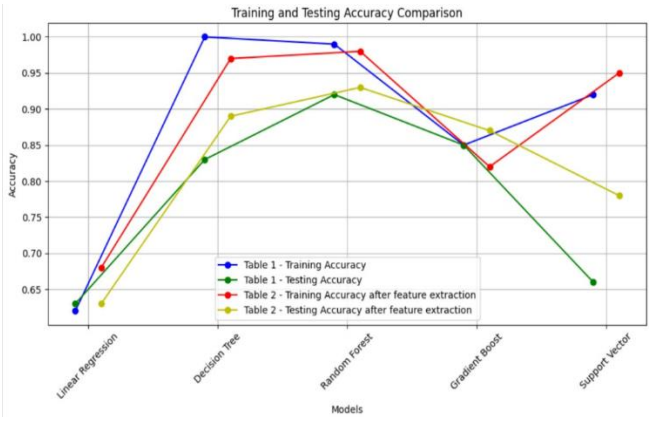


Fig. 2. Accuracies (Training and Testing) plotted for each model with and without feature extraction

VI. RESULTS

Fig. 2 shows that the errors have decreased with a noticeable difference for each model which indicates that the feature map was successful in mapping the original data to a latent representation which the ML models could learn better. The overall performance has improved for all the models. Since the original dataset was compressed into a feature map, it is expected that the number of trainable parameters for each model would have reduced which led to the observed decrease in the training time. The average training time for a model reduced from 23.483 seconds to 1.873 seconds. If the training time for compressing the dataset to the feature map were to be considered, then the average time is 13.467 seconds which is still 42.652% less. This should also correspond to reduced computational costs for training individual models. The MAPE has reduced by 16.45% for Linear Regression, 2.2% for Decision Tree, 2.56 % for Random Forest and 53.33% for Support vector. However, for Gradient Boost, the MAPE increased by 26.08% while the testing and training accuracies remained same. Thus, the models now taking less time and producing better results (similar incase of Gradient Boost) which support the idea that the feature map would be able to represent the data well and remove redundancy from the data.

VII. CONCLUSION

Load forecasting is vital for efficient energy management, cost reduction, and grid stability. It helps utilities match energy supply with demand, optimize resource allocation, and integrate renewable energy sources. It also enables demand response programs, ensuring reliable power supply and minimizing infrastructure strain. This study has demonstrated the effectiveness of integrating convolutional neural networks (CNNs) with traditional machine learning models for improving energy consumption prediction in buildings. The combination of CNN-generated data with original data has led to enhanced prediction accuracy, as evidenced by the evaluation metrics used. These findings underscore the potential of CNNs in enhancing energy forecasting models, offering valuable insights for future research and practical applications in energy management.

VIII. REFERENCES

[1] Cordeiro-Costas, M.; Villanueva, D.; Eguía-Oller, P.; Martínez-Comesaña, M.; Ramos, S. Load Forecasting with Machine Learning and Deep Learning Methods. Appl. Sci. (2023)

[2] Razak et al., Amir, Hashim (2009)

[3] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," in IEEE Transactions on Power Systems, vol. 16, no. 3, pp. 498-505, (Aug. 2001)

[4] M. T. Hagan and S. M. Behr, "The Time Series Approach to Short Term Load Forecasting," in IEEE Transactions on Power Systems, vol. 2, no. 3, pp. 785- 791, Aug. (1987)

[5] Tian, C.; Li, C.; Zhang, G.; Lv, Y. Data driven parallel prediction of building energy consumption using generative adversarial nets. Energy Build. (2019)

[6] Patel, Rikin, Mihir R. Patel, and Ravi V. Patel. "A review: Introduction and understanding of load forecasting." J. Appl. Sci. Comput 4.4 (2019)

[7] Feinberg, Eugene A., and Dora Genethliou. "Load forecasting." Applied mathematics for restructured electric power systems: optimization, control, and computational intelligence (2005)

[8] Hahn, Heiko, Silja Meyer-Nieberg, and Stefan Pickl. "Electric load forecasting methods: Tools for decision making." European journal of operational research 199.3 (2009)

[9] Zheng, Huiting, Jiabin Yuan, and Long Chen. "Shortterm load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation." Energies 10.8 (2017)

[10] Bianchi, Filippo Maria, et al. "An overview and comparative analysis of recurrent neural networks for short term load forecasting." arXiv preprint arXiv:1705.04378 (2017).

[11] Abumohsen, M.; Owda, A.Y.; Owda, M. Electrical Load Forecasting Using LSTM, GRU, and RNN Algorithms. Energies (2023)

[12] Wang, K.; Zhang, J.; Li, X.; Zhang, Y. Long-Term Power Load Forecasting Using LSTM-Informer with Ensemble Learning. Electronics (2023)

[13] Tasarruf Bashir, Chen Haoyong, Muhammad Faizan Tahir, Zhu Liqiang, Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN (2022)

[14] Lindberg, K. B., et al. "Long-term electricity load forecasting: Current and future trends." Utilities Policy 58 (2019)

[15] Ma, Ping, et al. "Review of family-level short-term load forecasting and its application in household energy management system." Energies 16.15 (2023)

[16] Wazirali, Raniyah, et al. "State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques." Electric power systems research 225 (2023)

[17] Morais, Lucas Barros Scianni, et al. "Short-term load forecasting using neural networks and global climate models: An application to a large-scale electrical power system." Applied Energy 348 (2023)

[18] Wan, Anping, et al. "Short-term power load forecasting for combined heat and power using CNNLSTM enhanced by attention mechanism." Energy 282 (2023)

View publication stats

2024 Asia Pacific Conference on Innovation in Technology (APCIT)
Mysuru, India. Jul 26-27, 2024

Investigating the Effect of Sliding Window Length for EEG-based Anxiety Classification

1st Shikha
Department of Computer
Science and Engineering
Delhi Technological University
New Delhi, India
shikha.9126@gmail.com

2nd Divyashikha Sethia
Department of Software Engineering
Delhi Technological University
New Delhi, India
divyashikha@dtu.ac.in

3rd S. Indu
Department of Electronics
and Communication
Delhi Technological University
New Delhi, India
s.indu@dce.ac.in

2024 Asia Pacific Conference on Innovation in Technology (APCIT) | 979-8-3503-6153-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/APCIT62007.2024.10673545

***Abstract*—Numerous studies have been conducted in various domains, such as emotion, seizure, and Alzheimer's, to investigate the impact of different sliding window lengths on EEG signal-based feature extraction. However, the impact of sliding window length on anxiety recognition remains underexplored. Hence, this paper aims to investigate the influence of different sliding window lengths on anxiety disorder classification using the EEG-based DASPS dataset. The study examines both time and frequency domain features to evaluate the effectiveness of varying window lengths. Additionally, a genetic feature selection algorithm is applied to balance the anxiety classification performance and model complexity. The comparative analysis of sliding window lengths reveals that shorter window sizes can enhance anxiety classification performance. Moreover, utilizing the feature selection algorithm improves classification performance by 2.16% and 2.54%, respectively, for the best-performing Gradient Boosting and XGBoost classifiers with a 1-second window size. These findings offer valuable insights for determining window lengths in EEG signal analysis for anxiety detection systems.**

***Index Terms*—Anxiety Disorder, Electroencephalograph (EEG), Sliding window, Machine Learning**

I. Introduction

Anxiety disorders are widespread mental health challenges that profoundly influence individuals' abilities, behaviours, productivity, and overall well-being. These disorders come in various forms, including generalized anxiety disorder, panic disorder, social anxiety disorder, and distinct phobias, leading to significant emotional distress, physical symptoms, and disruptions in daily life [1]. The impact of anxiety can extend to social withdrawal, avoidance of certain behaviours, and disruptions in personal relationships. Additionally, anxiety imposes a considerable economic burden on society due to treatment expenses and productivity losses [2]. Detecting anxiety involves diverse methods, such as self-report questionnaires, observation of behavioural changes, and analysis of physiological signals. However, physiological signals present a promising avenue for assessing anxiety due to their objective nature and potential for real-time monitoring.

The electroencephalogram (EEG) has become a widely used method for detecting anxiety, utilizing machine learning algorithms to identify patterns across various brain regions and achieve precise anxiety classification [3] [4].

Anxiety detection using EEG signals and machine learning algorithms involves several phases: signal acquisition, artefact removal, signal windowing, feature extraction, and classification. A crucial step in this process is segmenting EEG signals into discrete time epochs, which facilitates the application of machine learning algorithms for subsequent analysis. However, determining the optimal window size for segmentation remains a critical yet understudied aspect of EEG-based anxiety detection. The choice of window size directly influences the granularity of temporal information captured from EEG signals and consequently impacts the effectiveness of subsequent feature extraction and classification processes [5] [6]. Shorter window sizes are advantageous for capturing rapid changes within the EEG signals, thereby leading to more detailed and discriminative feature extraction. In contrast, longer window sizes may average out these fine details, potentially obscuring crucial discriminative features necessary for the accurate classification of anxiety states [6].

While window size investigation has been performed in different domains such as emotion [7], seizure [8], and Alzheimer's [9]. However, its exploration in anxiety classification is still limited, as shown in Table I. Notably, Muhammad et al. [10] utilized a 1-second window size for anxiety classification, highlighting the need for further exploration in this area. Despite significant advancements in algorithm development and model optimization for EEG-based anxiety classification, the impact of window size on classification performance remains inadequately explored.

Therefore, this study is motivated to systematically investigate the influence of sliding window size on EEG-based anxiety classification performance to improve the robustness and reliability of anxiety detection systems. Further, investigating the influence of window size and feature selection is paramount in optimizing anxiety classification accuracy. Selecting the most relevant features from EEG signals enhances the discriminative power of the classification model. To address this, the paper utilizes a genetic-based feature selection algorithm (GFSA) to reduce the EEG feature vector space and select the most informative features for the classification model.

The study uses the EEG-based DASPS dataset [13] for

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:11:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF RECENT WORK ON EEG-BASED ANXIETY CLASSIFICATION

Author	Dataset	subject	Channels	Window Approach	Features	Feature Selection	Classifiers	Accuracy
Al-Ezzi et al. [11]	Private	89	32	–	Frequency domain	–	CNN+LSTM CNN, LSTM	96.43%
Xie et al. [12]	Private	20	32	–	–	–	BN + CNN2BN + DBNBN + LDAPL + LDA	67.67%
Baghdadi et al. [13]	DASPS	23	14	–	Time, Frequency and Time-Frequency domain	–	SVM, KNN, SSAE	86.70%
Shikha et al. [1]	DASPS	23	14	–	Time, Frequency and Time-Frequency domain	RFE	DT, RF, SSAE	83.93%
Muhammad et al. [10]	DASPS	23	6	1sec window	Frequency domain	–	DT, kNN, SVM, MLP, RF	94.90%
Aggarwal et al. [3]	DASPS	23	14	–	–	–	1D CNN-LSTM, KNN, RF	95.22%
Bahman et al. [14]	DASPS	23	14	–	Novel Chaotic map based features	mRMR	DT and KNN	100%
Aldayel et al. [15]	DASPS	23	10	–	Time, Frequency and Time-Frequency domain	–	KNN, Adaboost, GB, RF, LDA, SVM	87.50%

anxiety classification. It explores distinct window sizes from 1 to 10 seconds and 20 seconds and compares their impact on anxiety classification performance. Furthermore, applying a genetic-based feature selection algorithm refines the feature set, ensuring the inclusion of the most discriminative features for anxiety classification. Subsequently, classification models are developed and evaluated, comprehensively comparing classification results across different window sizes and feature sets.

The primary contributions of this study can be summarized as follows:

1) Explore the impact of various window sizes, including 1 - 10 seconds and 20 seconds, on EEG-based classification using different machine learning algorithms, including Decision Tree (DT), K-Nearest Neighbor (KNN), XGBoost (XGB), and Gradient Boosting (GB). Further, evaluate classification performance using accuracy, F1-score, specificity, and sensitivity metrics.
2) Employ the Genetic Algorithm as a feature selection approach to identify and select the most informative features for anxiety classification based on EEG signals.

Roadmap: The paper organization is as follows: Section 2 explores related work, while Section 3 outlines the methodology, which includes the dataset description, preprocessing steps, feature extraction and selection, and classification algorithms. Section 4 provides the results and discussion of the study. Lastly, the conclusion summarizes the key findings and suggests directions for future research.

II. RELATED WORK

Considerable research has been conducted on the classification of anxiety disorder using EEG signals. However, there is a significant gap in exploring the window technique for this application.

Table I summarises recent work on EEG-based anxiety classification. Al-Ezzi et al. [11] employed EEG-based frequency domain features to classify Social Anxiety Disorder. The authors utilized long short-term memory (LSTM) and convolutional neural network (CNN) machine learning algorithms, achieving the highest accuracies of 92.86%, 92.86%, 96.43%, and 89.29% for severe, moderate, mild, and non-anxiety classification, respectively.

Xie et al. [12] employed EEG data and included ten healthy individuals along with ten patients diagnosed with anxiety and depression. The authors utilized a CNN model and attained the highest accuracy of 67.7% for anxiety classification.

In one study, Baghdadi et al. [13] introduced the first publicly available EEG-based DASPS dataset as "A Database for Anxious States based on Psychological Stimulation". This dataset comprises EEG recordings from 23 participants undergoing anxiety induction across six distinct scenarios per individual. The authors applied Support Vector Machine (SVM), KNN, and Stacked Sparse Autoencoder (SSAE) algorithms for binary and multiclass anxiety classification. They achieved the highest accuracies of 86.7% and 74.60% for detecting two and four anxiety levels, respectively.

Hence, several studies have been conducted on the DASPS dataset [13]. Shikha et al. [1] enhanced the classification accuracy of the DASPS dataset [13] by implementing the Recursive Feature Elimination (RFE) technique alongside distinct classifiers. The methodology aims to diminish redundancy among features, thereby improving classification accuracy. They have achieved the highest accuracies of 83.93% and 70.25% using an SSAE and DT for two-class anxiety classification, respectively.

Likewise, Muhammad et al. [10] employed channel selection based on affected brain regions and a 1-second windowing technique and achieved the highest accuracy of 94.90% using

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:11:20 UTC from IEEE Xplore. Restrictions apply.

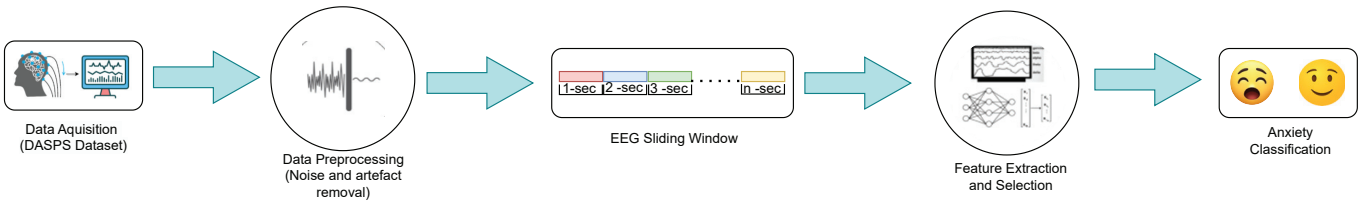


Fig. 1. Anxiety Classification Methodology using Sliding Window Technique

RF. Aggarwal et al. [3] also considered each brain region for channel selection for anxiety detection. The authors enhanced accuracy and improved up to 95.22% using CNN+LSTM, respectively.

Bahman et al. [14] proposed a novel EEG-based anxiety detection method using the DASPS database. The authors introduced a Chebyshev chaotic map for biosignal analysis and implemented innovative indicators for map quantification. Further, they applied DT and RF classifiers for anxiety level classification and achieved maximum accuracies of 93.75% and 100% for detecting two and four anxiety levels, respectively.

Aldayel et al. [15] highlight the significance of feature extraction and machine learning algorithms in anxiety detection through EEG. The authors extracted features using discrete wavelet transform and power spectral density techniques and utilized classification algorithms, including ensemble learning methods and conventional classifiers. The results showed that Hamilton anxiety rating scale labelling and discrete wavelet transform-based features consistently performed better, with the RF accuracy of 87.5%.

Although many studies have been conducted on anxiety detection using EEG signals, there is still a significant research gap regarding the optimal windowing techniques. Current studies usually use simple methods such as feature extraction and machine learning-based classification. However, only one study used a 1-second windowing technique and achieved 94.90% accuracy [10]. It is observed that none of the studies have attempted to investigate the correlation between temporal segmentation and anxiety detection performance. This research gap highlights the necessity for systematic exploration into the impact of window size on anxiety detection performance. In the future, these findings may help determine the optimal balance between temporal resolution and signal stability for EEG feature extraction, resulting in enhanced accuracy and interpretability of classification results.

III. Methodology

Figure 1 shows that EEG-based anxiety classification using machine learning algorithms involves several phases: signal acquisition, data preprocessing, segmentation, feature extraction and selection, and classification.

A. Dataset Description

The study utilizes the DASPS Dataset from Baghdadi et al. [13]. The dataset comprises raw EEG data in .edf format obtained from 23 subjects, including ten men and 13 women. The authors capture the EEG signal using the 14-channel Emotiv EPOC wireless EEG headset, with 128 Hz sampling rate. The dataset includes preprocessed data in .mat format. During EEG recording, participants were presented with six distinct scenarios with eyes closed and minimal gestures. Each scenario was subdivided into two segments: first, where the psychotherapist described the scenario for 15 seconds, and second, where the participant recalled the scenario for another 15 seconds. The entire process for each participant involved a 5-minute Hamilton Pre-Stimuli phase, followed by recitation and recall of each scenario, repeated for all six scenarios. Finally, the participants assessed their emotional responses, including valence and arousal, using the Self-Assessment Manikin (SAM).

B. Data Preprocessing

EEG signals often contain various artefacts, including electrode misplacement, subject movement, environmental noise, and biological artefacts. These artefacts are categorized into two main types: extrinsic and physiological. Extrinsic artefacts arise from technical issues and external factors during data collection, such as cable movements, electrode misplacement and powerline interference. Physiological artefacts, on the other hand, arise from bodily functions, such as muscle activities (Electromyogram artefacts), eye movements or blinking (Electroculogram artefacts), and cardiac activities (Electrocardiogram artefacts).

The study employs Independent Component Analysis (ICA) using the *FastICA* algorithm to remove signal artefacts. Additionally, it utilizes a 5th-order Butterworth Bandpass filter with a frequency range of 4 - 44 Hz to the raw EEG signal to enhance data quality and eliminate unwanted frequencies.

C. Feature Extraction

Following preprocessing, the EEG signal undergoes analysis to extract time and frequency domain features, considering various sliding window sizes ranging from 1 to 10 seconds and 20 seconds to asses the influence on anxiety classification performance. Time domain features (TDF) are derived from the preprocessed EEG signal, offering insights into the distribution of amplitudes over time. Frequent domain features (FDF) are crucial in conveying signal pattern information. These features are obtained by examining the signal's frequency spectrum utilizing the Fast Fourier Transform (FFT) technique. Table II presents a comprehensive overview of the extracted EEG features.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:11:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II
OVERVIEW OF THE EEG FEATURES EXTRACTED DURING EXPERIMENTATION

Domain	Feature Name	Feature Equation	Features per Channel
Time Domain	Statistical Features	Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ Median: median = middle value of x Standard Deviation: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ Peak-to-Peak: Peak-to-Peak = max(x) - min(x) Minimum: min(x) Maximum: max(x)	23*6 = 138
Time Domain	Hjorth Features	Activity: Activity = var($y(t)$) Mobility: Mobility = $\frac{var(\frac{dy(t)}{dt})}{var(y(t))}$ Complexity: Complexity = $\frac{Mobility(\frac{dy(t)}{dt})}{Mobility(y(t))}$	23*3 = 69
Frequency Domain	Relative Bandpower Theta-Alpha Ratio Spectral Entropy Sample Entropy	Relative Power = $\frac{\text{Absolute Power of Band}}{\text{Total Power}}$ Theta-Alpha ratio = $\frac{\text{Absolute Power of Theta Band}}{\text{Absolute Power of Alpha Band}}$ $H(y(t)) = -\sum_{f=0}^{2fs} \text{PSD}(f)\log_2[\text{PSD}(f)]$ $H(y,m,r) = -\log_e\left(\frac{Cm+1(r)}{Cm(r)}\right)$	23*4 = 92

D. Feature Selection

Feature selection aims to reduce the complexity of large EEG-based feature sets, reducing the computational burden of classification tasks. The study explores the effectiveness of a genetic-based feature selection method in improving classification performance.

The genetic algorithm (GA) operates on principles inspired by natural selection and genetics. It serves as a prominent search technique in feature selection, evaluating subsets of features. A key advantage of GA lies in its ability to conduct a global search, contrasting with other methods that typically rely on local and greedy search strategies. Consequently, GA proves effective in delivering high-quality solutions for feature selection challenges.

GA employs various operations such as crossover, mutation probability, and selection for generating a new population, represented as subsets of features, until reaching the specified criterion. In GA, features are encoded as chromosomes to form a population, and the fitness value plays a crucial role in assessing the strength of each individual. In this study, the fitness value was determined using a given equation.

$$Fitness = W_{A.accuracy} + W_{nb}/N \tag{1}$$

W_A represents the weight assigned to accuracy, while W_{nb} denotes the weight allocated to include N features in the classification, where $N \neq 0$.

After determining the fitness function for each chromosome, the population underwent crossover and mutation operations. Two-parent genes were randomly selected to produce new genes based on their fitness function scores during crossover. At the same time, mutation involves generating new individuals by randomly selecting genes from a chromosome. The lowest fitness value determined the fittest individuals. The selected parameters are as follows: crossover probability of 0.8, maximum generations set to 100, mutation probability of 0.01, and population size of 20. After selecting the features, they are fed into different classifiers such as DT, KNN, XGB, and GB.

IV. RESULTS AND DISCUSSION

Previous studies have extensively explored algorithm development and model optimization in EEG-based anxiety classification [12] [15]. However, the effectiveness of EEG-based classification systems depends significantly on the preprocessing techniques employed, such as the choice of time window size for feature extraction and the inclusion of suitable features through feature selection methods. These challenges hinder the development of robust and reliable anxiety detection systems. Hence, the study aims to investigate the impact of distinct window sizes on anxiety classification performance using various machine learning algorithms, including DT, KNN, XGB, and GB. Furthermore, it discusses integrating feature selection techniques, specifically utilizing genetic algorithms to refine the feature set and improve classification accuracy.

In this study, the effectiveness of all classifiers is measured using metrics such as F1 score, Accuracy, Sensitivity, and Specificity. Accuracy measures the algorithm's capacity to predict negative and positive instances accurately, as shown in Equation 2. The F1 score, as depicted in Equation 3, represents the harmonic mean of recall and precision, where a higher score signifies better classifier performance.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:11:20 UTC from IEEE Xplore. Restrictions apply.

TABLE III
PERFORMANCE WITHOUT FEATURE SELECTION

Window (in secs)	DT		KNN		Xgboost		GB	
	Accuracy	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
1 sec	**92.75**	91	69.39	67.65	**94.56**	93.53	**96.45**	96.08
2 sec	80.54	83.65	**70.13**	76.08	90.72	89.92	92.91	91.99
3 sec	77.97	77.97	67.02	67.02	83.45	81.82	84.19	84.66
4 sec	72.64	72.94	66.85	64.73	81.32	80.47	81.92	81.84
5 sec	65.67	64.02	62.67	61.4	72.6	72.16	80.72	84.71
6 sec	66.45	65.19	61.05	60.13	72.89	77.83	75.23	75.32
7 sec	69.13	69.89	62.93	60.45	73.78	72.54	75.26	75.67
8 sec	64.1	64.67	63.6	62.78	61.24	61.82	64.25	63.21
9 sec	63.85	63.24	61.39	61.45	61.82	60.24	63.78	63.13
10 sec	61.34	60.78	59.73	60.24	60.84	59.27	62.94	60.48
20 sec	57.14	57.01	58.33	59.13	58.69	58.54	61.32	61.04

TABLE IV
PERFORMANCE WITH FEATURE SELECTION

Window (in secs)	DT		KNN		Xgboost		GB	
	Accuracy	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
1 sec	**94.51**	93.21	72.65	72.81	**97.1**	96.29	**98.61**	98.36
2 sec	85.52	84.54	**74.2**	73.11	90.72	89.92	92.91	91.99
3 sec	81.81	81.17	70.57	68.76	86.95	84.87	88.23	88.25
4 sec	79.43	79.6	67.14	68.14	87.31	87.59	88.55	87.27
5 sec	74.03	73.81	66.75	66.54	82.83	81.06	84.73	84.67
6 sec	74.03	72.06	65.57	66.16	79.11	78.33	81.32	80.47
7 sec	73.06	72.29	64.37	64.67	79.51	79.5	81.52	80.67
8 sec	69.56	67.69	64.67	62.57	68.67	68.38	70.27	69.25
9 sec	69.38	68.02	62.13	63.24	63.96	62.91	69.27	69.92
10 sec	64.13	64.19	61.95	60.5	60.36	59.19	66.86	66
20 sec	58.69	59.43	61.95	61.61	60.92	59.48	61.32	61.04

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$F1score = 2 * \frac{recall * precision}{recall + precision} \tag{3}$$

The research utilizes the DASPS dataset [13] for anxiety classification. Before classification, the data undergoes preprocessing, involving segmenting EEG signals into distinct time windows of varying lengths to extract time and frequency domain features. Subsequently, the extracted features serve as input data for several classification algorithms.

Table III presents the distinct sliding window sizes effect for anxiety classification without feature selection, offering valuable insights into the relationship between temporal segmentation and classifier performance. The analysis showcases a noticeable trend wherein the smallest window size of 1 second yields the highest accuracy rates across multiple classifiers, including GB, XGBoost and DT. With a window size of 1 second, GB achieves an accuracy of 96.45%, indicating its efficacy in capturing fine temporal dynamics inherent in anxiety-related EEG patterns. Further, XGBoost and DT are the second and third highest-performing classifiers with an accuracy of 94.56% and 92.75%, respectively. Additionally, Compared to the 1-second window, KNN yields a higher accuracy of 70.13% with a 2-second window.

Table IV showcases classifier performance following GA feature selection, illustrating a significant enhancement in accuracy across varying window sizes. The results indicate substantial increases in accuracy across multiple classifiers compared to their performance without feature selection. Using GA effectively explores the high-dimensional feature space and identifies subsets of features that maximize classification performance.

In addition, a consistent trend is observed across different window sizes, with smaller windows continuing to yield improved accuracy rates compared to larger windows. This trend underscores the importance of fine-grained segmentation in capturing the transient dynamics inherent in anxiety-related EEG patterns, thereby facilitating more accurate classification. GB emerges as the top-performing classifier across various window sizes with the highest accuracy of 98.61% and an F1 score of 98.36%. Further, XGBoost and DT achieved the highest accuracy of 97.1% and 94.51%, respectively. When the window size is increased, the accuracy rate of KNN decreases. However, it performs best with an accuracy rate of 74.2% and F1 score of 73.11% for a 2-second window.

Conversely, larger window sizes, such as 9, 10, and 20 seconds, exhibit comparatively lower accuracy rates across classifiers. This decline in accuracy is likely due to the smoothing effect that larger windows may have on fine temporal variations, thereby diminishing the discriminatory power of extracted features. Moreover, a similar trend is observed in the F1 score, with a decrease noted as window size increases.

In conclusion, the study suggests that using shorter window lengths, specifically 1 and 2 seconds, for feature extraction

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:11:20 UTC from IEEE Xplore. Restrictions apply.

in anxiety classification results in higher recognition accuracy. These findings are crucial as they underscore the importance of selecting appropriate window lengths to optimally capture the rapid, transient dynamics characteristic of anxiety-related EEG patterns.

Notably, the results are consistent with those of Muhammad et al. [10], who achieved a maximum accuracy of 94.90% using a 1-second window size. However, this study extends these findings by providing a quantitative analysis across a broader range of window sizes, demonstrating the consistent superiority of shorter windows for anxiety detection.

Fig 2 depicts the ROC curves for the best-performing window sizes of 1 and 2 seconds to illustrate the classifiers' trade-off between false positive and true positive rates.

This analysis highlights the critical role of window size selection in EEG-based anxiety detection for developing real-time anxiety monitoring systems. Future research should investigate the influence of window size across different stimuli and experimental conditions within this domain. Further, this exploration would help validate the current findings about optimal window sizes for EEG-based anxiety classification to other neurological conditions and experimental settings.

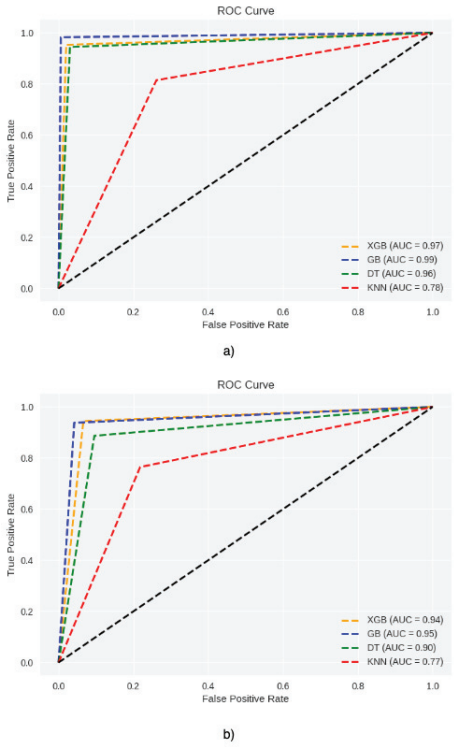


Fig. 2. ROC curves for best-performing window sizes after feature selection: a) 1-second window size, b) 2-second window size

V. Conclusion

The paper aims to investigate the impact of distinct window sizes on anxiety classification performance using the DASPS, 23-channel EEG dataset. Further, it employs a genetic-based feature selection algorithm to balance classification accuracy and model complexity by reducing the feature set. The study evaluates the performance of various classifiers, including DT, KNN, XGB, and GB. The results indicate that the greatest recognition accuracy is achieved when using shorter window lengths, specifically 1 and 2 seconds for feature extraction. With feature selection, the highest accuracy is 98.61%, and the F1 score is 98.36% using GB. However, one limitation of this study is that only classical classifiers are used for anxiety classification. Advanced algorithms, such as advanced neural networks and stacking classifiers, could be developed to improve the EEG-based anxiety classification performance.

References

[1] Shikha, M. Agrawal, M. A. Anwar, and D. Sethia, "Stacked sparse autoencoder and machine learning based anxiety classification using EEG signals," in *Proceedings of the First International Conference on AI-ML Systems*, 2021, pp. 1–7.

[2] A. Y. Mughal, J. Devadas, E. Ardman, B. Levis, V. F. Go, and B. N. Gaynes, "A systematic review of validated screening tools for anxiety disorders and PTSD in low to middle income countries," *BMC psychiatry*, vol. 20, pp. 1–18, 2020.

[3] R. Agarwal, M. Andujar, and S. Canavan, "Classification of emotions using EEG activity associated with different areas of the brain," *Pattern Recognition Letters*, vol. 162, pp. 71–80, 2022.

[4] G. Giannakakis, D. Grigoriadis, and M. Tsiknakis, "Detection of stress/anxiety state from EEG features during video watching," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 6034–6037.

[5] G. Li, D. Ouyang, Y. Yuan, W. Li, Z. Guo, X. Qu, and P. Green, "An EEG data processing approach for emotion recognition," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10 751–10 763, 2022.

[6] P. Keelawat, N. Thammasan, M. Numao, and B. Kijsirikul, "A comparative study of window size and channel arrangement on EEG-emotion recognition using deep CNN," *Sensors*, vol. 21, no. 5, p. 1678, 2021.

[7] D. Ouyang, Y. Yuan, G. Li, and Z. Guo, "The effect of time window length on EEG-based emotion recognition," *Sensors*, vol. 22, no. 13, p. 4939, 2022.

[8] A. Alzahy, M. Elgammal, H. Mohammed, and H. Mostafa, "Optimal EEG window size for neural seizure detection," in *2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, 2019, pp. 1–4.

[9] K. D. Tzimourta, N. Giannakeas, A. T. Tzallas, L. G. Astrakas, T. Afrantou, P. Ioannidis, N. Grigoriadis, P. Angelidis, D. G. Tsalikakis, and M. G. Tsipouras, "EEG window length evaluation for the detection of Alzheimer's disease over different brain regions," *Brain sciences*, vol. 9, no. 4, p. 81, 2019.

[10] F. Muhammad and S. Al-Ahmadi, "Human state anxiety classification framework using EEG signals in response to exposure therapy," *Plos one*, vol. 17, no. 3, p. e0265679, 2022.

[11] A. Al-Ezzi, N. Yahya, N. Kamel, I. Faye, K. Alsaih, and E. Gunaseli, "Severity assessment of social anxiety disorder using deep learning models on brain effective connectivity," *IEEE Access*, vol. 9, pp. 86 899–86 913, 2021.

[12] Y. Xie, B. Yang, X. Lu, M. Zheng, C. Fan, X. Bi, Y. Li *et al.*, "Anxiety and depression diagnosis method based on brain networks and convolutional neural networks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1503–1506.

[13] A. Baghdadi, Y. Aribi, R. Fourati, N. Halouani, P. Siarry, and A. Alimi, "Psychological stimulation for anxious states detection based on EEG-related features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8519–8533, 2021.

[14] F. Daneshmand-Bahman and A. Goshvarpour, "Anxiety recognition using a new EEG signal analysis approach based on sample density in a Chebyshev chaotic map," in *Affective Computing in Healthcare: Applications based on biosignals and artificial intelligence*. IOP Publishing Bristol, UK, 2023, pp. 1–1.

[15] M. Aldayel and A. Al-Nafjan, "A comprehensive exploration of machine learning techniques for EEG-based anxiety detection," *PeerJ Computer Science*, vol. 10, p. e1829, 2024.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:11:20 UTC from IEEE Xplore. Restrictions apply.

Physica Scripta

PAPER

Investigation of circularly polarized MIMO antenna with enhanced isolation for sub-6 GHz application

Parul Kansal , Anup Kumar Mandpura and Narendra Kumar
Department of Electrical Engineering, Delhi Technological University, Delhi, India
E-mail: kansal.parul@gmail.com

RECEIVED
11 June 2024
REVISED
26 August 2024
ACCEPTED FOR PUBLICATION
2 September 2024
PUBLISHED
11 September 2024

Keywords: MIMO antenna, microstrip antenna, circular polarization, fifth generation, high isolation

Abstract
This article demonstrates the design process for a unique, two-port, circularly polarized (CP) multiple input multiple outputs (MIMO) antenna for fifth-generation (5G) applications operating at sub-6 GHz. The proposed 2-port CP MIMO antenna comprises a patch antenna with a distinctive slot and a circular split ring resonator (SRR) at the ground plane. The circular patch antenna's resonant frequency is moved to a lower frequency by the addition of a circular SRR at the ground plane. Investigating a special slot at the patch antenna can help in the enhancement of circular polarization and bandwidth. A sequence of parasitic elements in the shape of 'h' is placed in the space between the two antenna elements for decoupling. The minimum 20 dB inter isolation among the antenna elements in the proposed 2-port CP MIMO antenna's operating band is made possible by these h-shaped parasitic elements. In addition, to ensure the proposed antenna is more reliable for 5G applications at sub-6 GHz, other diversity parameters including Envelop Correlation Co-efficient (ECC), Mean Effective Gain (MEG), Diversity Gain (DG), TARC (Total Active Reflection Co-efficient) and CCL (channel capacity loss) are also calculated . In comparison to existing MIMO systems of the same order, the proposed 2-port CP MIMO antenna appears to have the minimum channel capacity loss and excellent diversity performance. The proposed 2-port CP MIMO antenna achieves an impedance bandwidth of 3.3 to 3.6 GHz with impedance matching better than 10 dB and has an axial ratio bandwidth of 3.35 to 3.51 GHz with a steady radiation pattern in both planes over the operating frequency bands.

1. Introduction

To enable the new era of the internet of everything, the 5G mobile network system is a new global wireless framework that is designed to provide ubiquity connectivity to everyone and everything at once, including machines, objects, and devices, in a way that is incredibly reliable and reasonably priced [1]. MIMO systems efficiently take advantage of the multipath environment to boost wireless system performance as a whole. Systems using 5G MIMO considerably increase speed and capacity while delivering low latency and excellent dependability. Strong mutual coupling between antenna elements caused by compact MIMO antenna design may lead to a decline in the performance of system diversity overall and inter-element isolation. It is allowed for the 5G sub-6 GHz MIMO system to have an isolation level of 10 dB. To improve MIMO performance, however, lower bound isolation of at least 15 dB should be needed. When compared to their linearly polarized (LP) counterparts, circularly polarized MIMO antennas provide several advantages, including the prevention of multipath interference, fading, and quality of service (QoS) [2–4]. In addition to the aforementioned, CP antennas, as compared to LP antennas, also aid in overcoming the orientation issue between the transmitter and receiver, greater mobility due to suppressed multipath interference, and better weather penetration [5]. Circularly polarized MIMO antennas have so been discussed in the literature in recent years. A compact CP MIMO antenna design in [6] using two truncated corner square patches with parasitic periodic metallic plates. It requires costly Taconic substrate, increasing manufacturing costs. A MIMO antenna proposed in [7] has

© 2024 IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

employed an inexpensive FR-4 substrate to reduce production cost and to improve the performance of the antenna applying an eyebrow-shaped strip, however, it has a very big dimension of 85 mm × 73 mm. The spatial diversity method has been used by [8] to increase the isolation between nearby antenna components (adjacent elements are separated by a distance of 13.75 mm). According to [9], the low-profile, microstrip-fed MIMO antenna with the isolation of more than 17 dB, It is noteworthy that the gap between the two antenna elements, measured from centre to centre, is just 0.16, and isolation of up to 18.50 dB can be attained without the use of any extra decoupling structures. A dual port dielectric resonator-based CP antenna is also suggested for 5G communication [10]. Controllable CP, compact size, superior isolation, and simple geometry are this antenna's main benefits. Compact CP array antennas with minimal mutual coupling have been constructed in [11]. The broad impedance bandwidth of this antenna is 36%, and the CP BWs are around 23%. Additionally, there is a low envelope correlation coefficient and a large diversity gain (DG). A small two-port CP MIMO antenna for 5G outdoor applications was created in [12]. This CP MIMO antenna has left-hand circular polarization and covers the future 5G range between 3.3 GHz and 3.8 GHz. With a total footprint of 37 × 30 × 0.8 mm^3, for use at sub-6 GHz, a four-port small MIMO antenna with a built-in isolator has been developed. The circular isolator that was integrated into the device eliminates the need for any decoupling components and conserves space between individual antenna elements [13]. The goal of [14] is to achieve CP bandwidth with enhanced isolation employing a new ground structure and asymmetric Z-shaped radiating patch. This antenna geometry is extremely small thanks to the innovative ground construction. Different antenna design methodologies are presented [15] to achieve good mutual coupling between the antenna parts while also reducing the antenna's geometry. These papers will address several methods for achieving excellent isolation and compact dimensions.

According to [16], a MIMO antenna with a single split ring resonator on the top side and a defective ground structure on the bottom is designed to cover both the mid 5G band and C-band. The suggested antenna offers more than 15 dB isolation between antenna parts without the use of any decoupling mechanisms, as well as circular polarization at the second band useful for satellite applications. A two-element slotted octagon-shaped antenna for use in the sub-6 GHz range is described in [17]. A T-shaped structure that is positioned at the ground plane helps to provide isolation of more than 20 dB. The design and optimization of a five-port MIMO antenna system with polarization and pattern diversity for the n79–5G band. Four end fire linearly polarized monopole antennas are merged with a broadside circularly polarized antenna [18]. The same dual-band MIMO antenna is designed in [19] to operate in the 3.3–3.8 GHz and 4.3–5.7 GHz frequency bands. In [20], two-shaped and one rectangular defects have been introduced into the ground plane to ensure wideband isolation [21] across the operating frequency range. In [22], a two-arm monopole antenna incorporating a coplanar waveguide feed line operate between 3.2 GHz and 4.12 GHz, with proper isolation [23].

In this paper, an innovative slot in the patch antenna and a circular split ring resonator at the ground plane makes up the suggested compact two-port CP MIMO antenna. Investigating a special slot at the patch antenna can help to accomplish circular polarization and bandwidth increase. A sequence of parasitic components in the shape of h is placed in the space between the two antenna elements to decouple the two antenna elements. The suggested MIMO CP antenna achieves an impedance bandwidth of 3.3 to 3.6 GHz with impedance matching better than 10 dB and has an axial ratio bandwidth of 3.35 to 3.51 GHz with a steady radiation pattern in both planes over the operating frequency bands.

2. Antenna design and analysis

This section will describe the step-wise design and analysis of a novel sub-6 GHz compact patch antenna before moving on to the design and analysis of the proposed CP MIMO antenna. Three prototype antennas, referred to as Ant.1 through Ant.3, serve as a description of the proposed antenna's design step and their corresponding impedance bandwidth and axial ratio bandwidth.

First, as seen in figure 1(a), an inset feed circular patch antenna has been created to function at frequencies of about 6 GHz. Figure 2(a), depicts the Frequency response of S-parameter Ant.1 to Ant.3, but the resonance frequency is not visible due to the selected frequency range. In figure 1(b) the inset feed circular patch antenna loaded with a circular split ring resonator (Ant.2) has a new resonance frequency of 3.5 GHz as shown in figure 2(a). The high concentration of current observed around the circular SRR increased the resonance frequency shift towards lower frequency, as shown in figures 2(b), (i).

In figure 1(c), the Ant. 2′s bandwidth is increased by the addition of a special slot that combines an L-shaped slot with an inverted fork-shaped slot, and it will occupy a complete 5G Sub-6 GHz band. The patch antenna's bandwidth is primarily improved by an inverted fork-shaped slot, which also serves as an impedance matching and an L-shaped slot. Because of this, the unique slot offers good impedance matching and bandwidth enhancement. Along with that, this unique slot is also responsible for the axial ratio bandwidth for the proposed

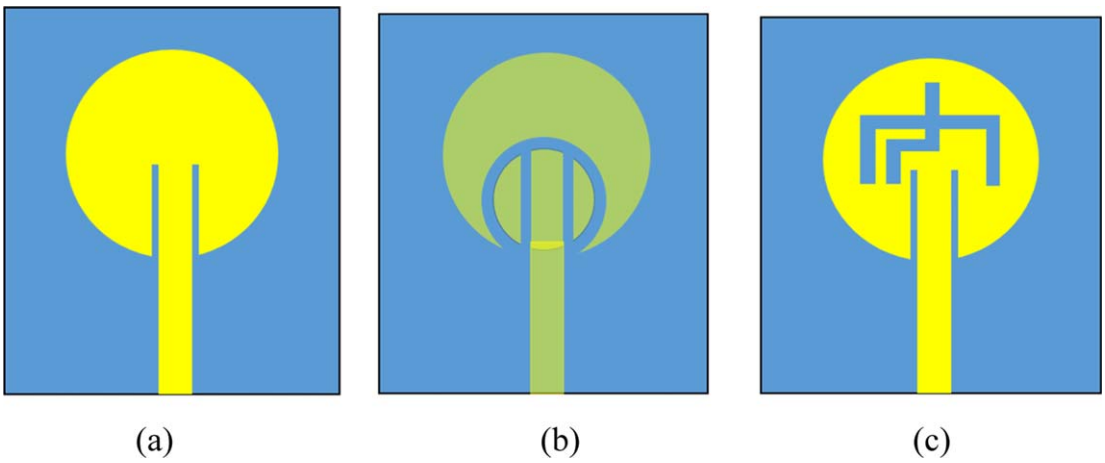


Figure 1. Schematics of the single-element antenna to showing step by step design (a) Ant. 1 (b) Ant. 2 (c) Ant. 3.

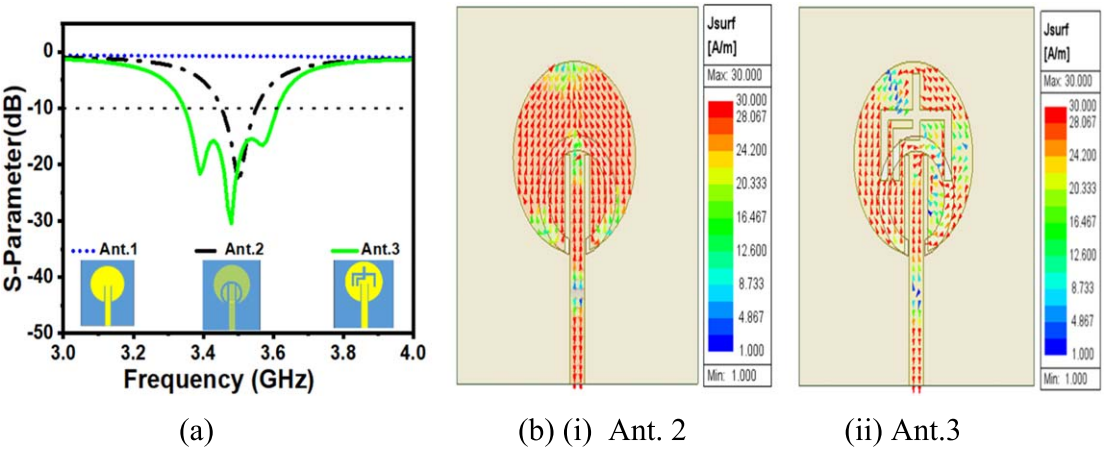


Figure 2. (a) Simulated S-parameter of Ant.1 to Ant.3 (b) Surface current density distribution of (i) Ant.2 (ii) Ant.3.

antenna because the introduction of L shape slot perturbed the surface current density distribution of the radiator, as shown in figure 2 (b)(ii).

2.1. Parametric analysis

Using Ansys HFSS, the simulation of the antenna design across the operational frequency range was conducted, and the design parameters for the antenna structure were fine-tuned. The antenna dimensions were determined through a parametric analysis involving the slot move in horizontal direction (parameter Y) and slot move in vertical direction (parameter X). The impact of Y and X on the S_{11} parameters is illustrated in figure 3. The Y & X varied between −1 to 1 mm. The optimized values were identified as Y = 0 mm and X = 0 mm, resulting in the desired favourable return loss.

3. Proposed MIMO antenna design

Figures 4(a) and (b) shows the proposed 2-port CP MIMO antenna geometry from our research. The proposed 2-port CP MIMO antenna is created on an FR4 substrate with 0.019 as the loss tangent and 4.4 as the dielectric constant. It has two Ant.3 prototypes that are connected by a succession of h-shaped parasitic element slots to improve isolation in the sub-6 GHz band, it has an axial ratio bandwidth of 3.35 to 3.51 GHz and covers the frequency range of 3.3 GHz to 3.6 GHz. The proposed 2-port CP MIMO antenna's final dimensions are summarized in table 1 after some parametric analysis. Figure 4(c) will display the proposed antenna's frequency response after the activation of both antennas simultaneously with and without h-shaped parasitic element (strip) and achieve a minimum isolation of 20 dB on its operating band.

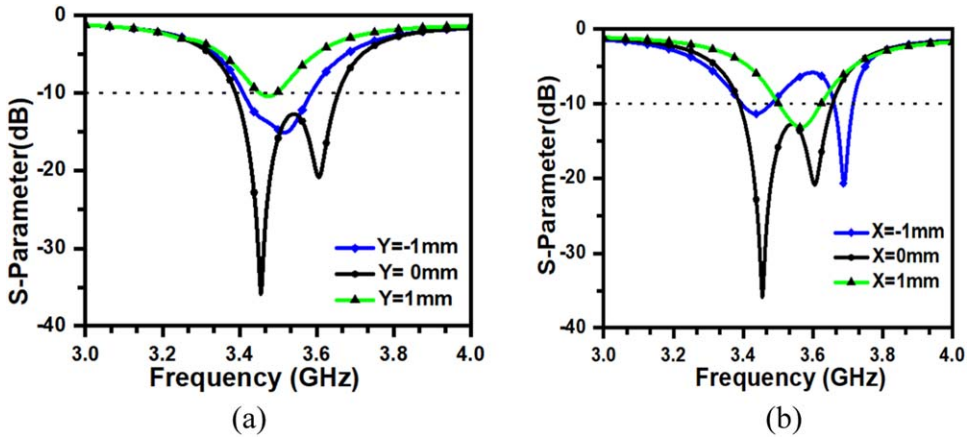


Figure 3. Parametric analysis of proposed antenna (a) $|S_{11}|$ with Y (b) $|S_{11}|$ with X.

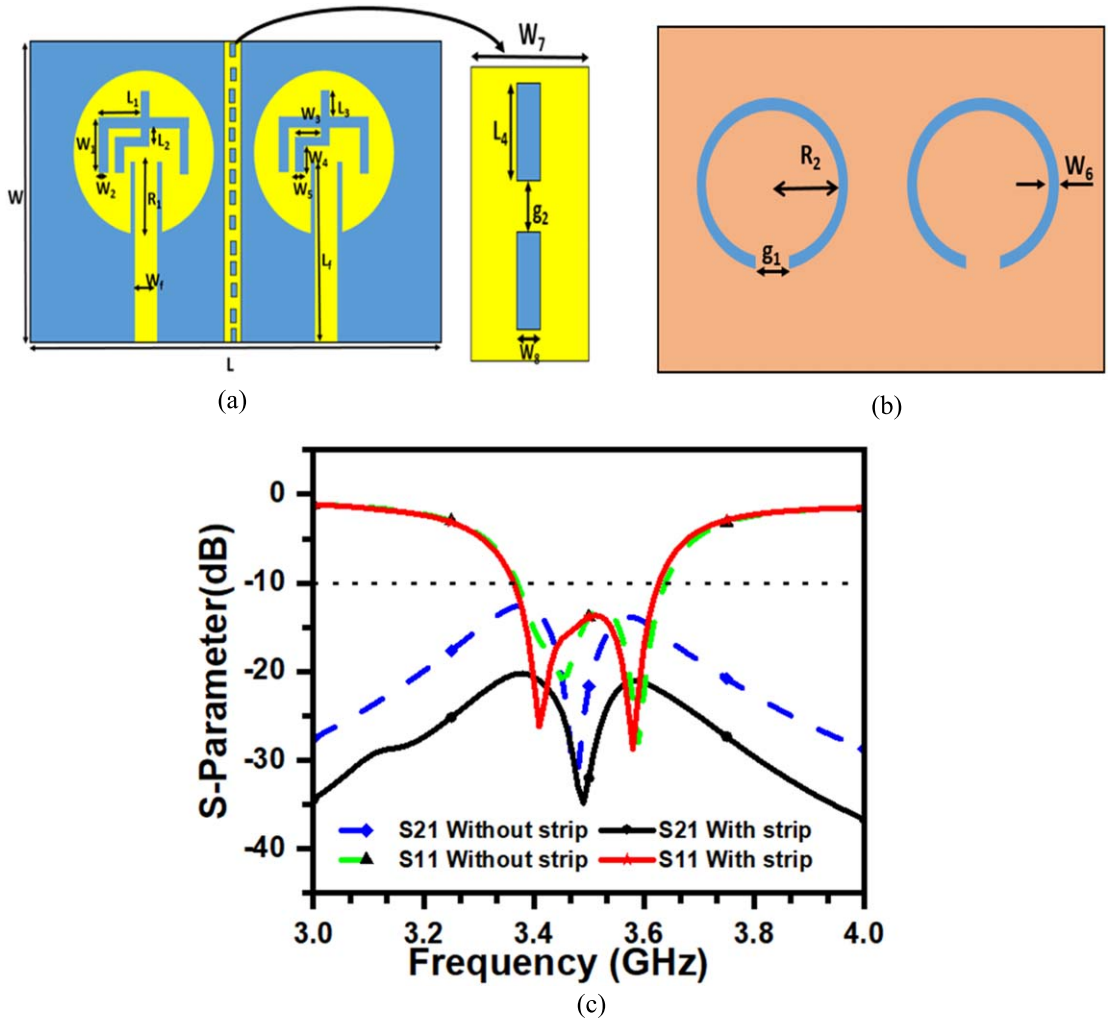


Figure 4. Schematics (a) Top View (b) Bottom View (c) S-parameters with & without strip of the proposed 2-port CP MIMO antenna.

4. Result and analysis of the proposed 2-port CP MIMO antenna

The HFSS tool and a vector network analyser are used to simulate and measure the S-parameters of the proposed 2-port CP MIMO antenna as shown in figure 5(a). According to figure 5(b), the proposed 2-port CP MIMO antenna has the same resonance frequencies in simulation and measurement. The experimental

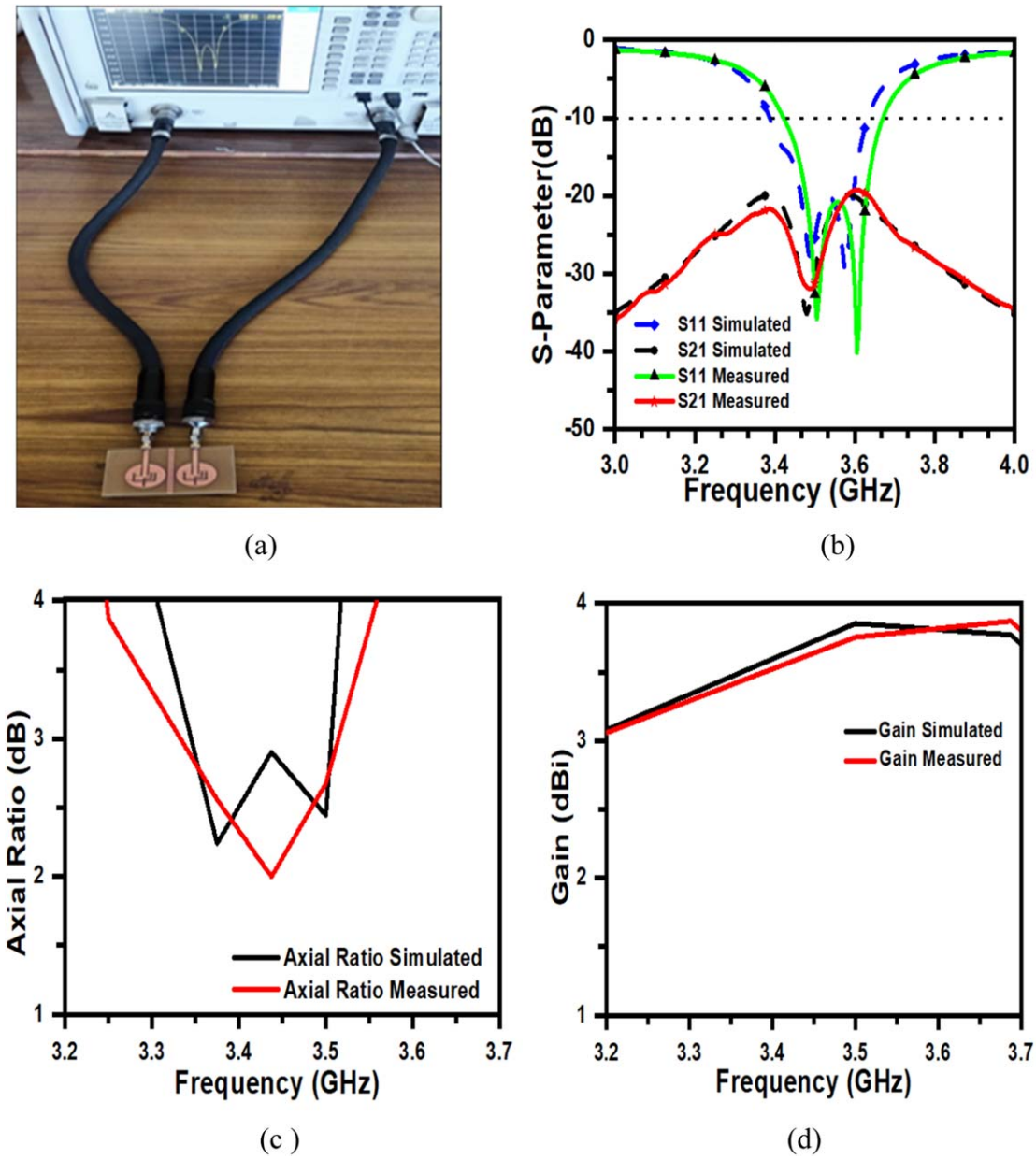


Figure 5. (a) S-parameters (b) Measurement setup of S-parameter by VNA (c) Simulated and Measured axial ratio plot (d) Simulated and Measured gain of the proposed 2-port CP MIMO antenna.

Table 1. Different parameters of the proposed 2-port CP MIMO antenna.

Parameters	L	W	L_1	W_1	R_1	W_f	W_2	W_5	W_7	W_8
Values (mm)	65	50	6.75	9.5	12.7	3	1.5	1.5	3	1
Parameters	L_2	W_3	W_4	L_f	L_3	g_1	g2	R_2	W_6	L_4
Values (mm)	2.5	2.5	7	30.41	3.75	3.5	1	6	2	1.5

and simulated S_{21} is greater than 20 dB. The proposed 2-port CP MIMO antenna achieves an impedance bandwidth of 3.3 to 3.6 GHz with impedance matching better than 10 dB and has an axial ratio bandwidth of 3.35 to 3.51 GHz, as shown in figure 5(c). Due to manufacturing errors, there are small variations in experimental and simulated measurements. It is also evident that the suggested design results in good isolation and return loss. The gain of the proposed 2-port CP MIMO antenna is shown in figure 5(d), the maximum gain is observed to 3.8 dBi.

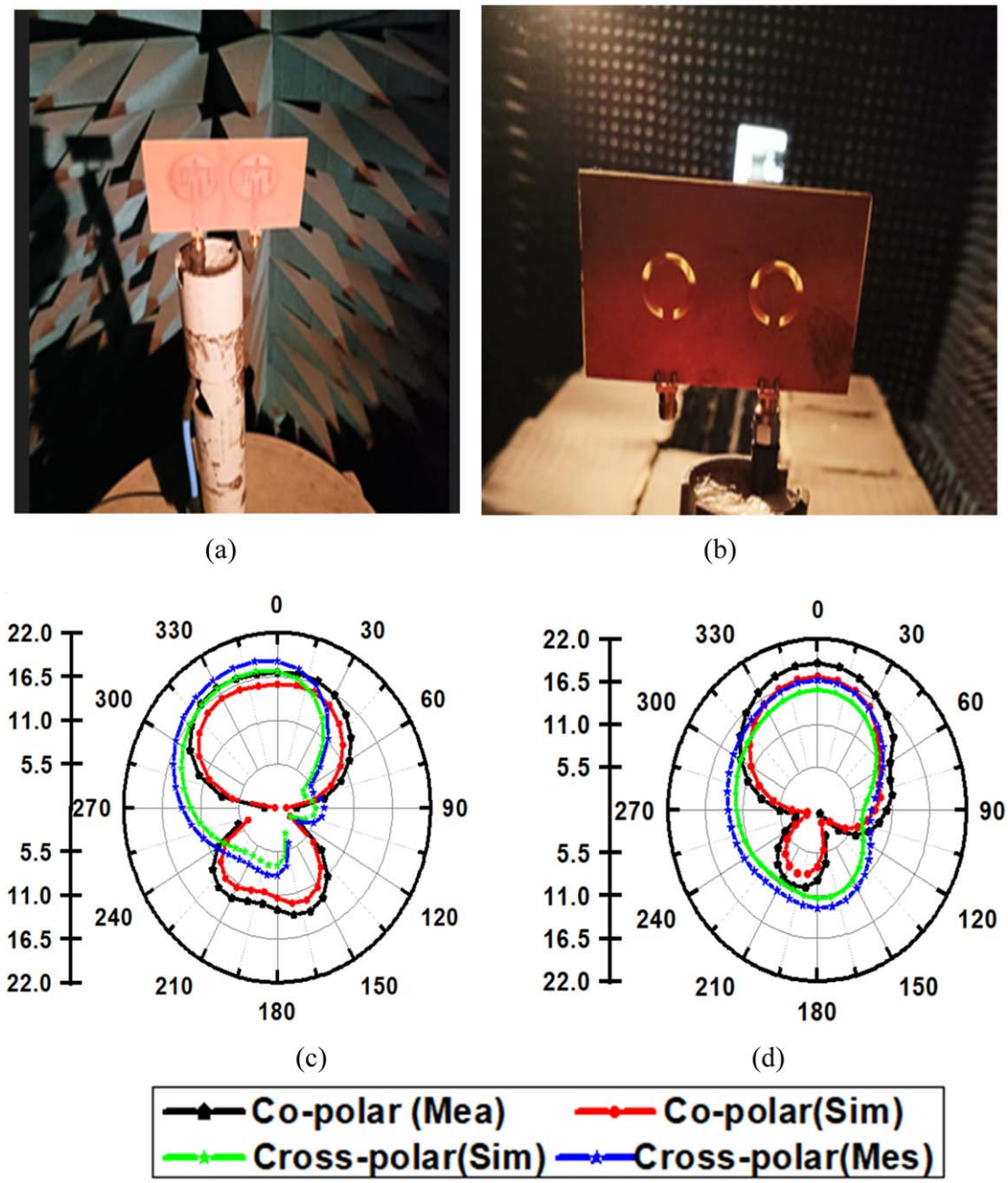


Figure 6. Radiation pattern measurement setup (a) Front view (b) Back view (c) E-Plane (d) H-Plane of proposed 2-port CP MIMO antenna.

4.1. Radiation characteristics

Figures 6(a) and (b), shows the view of the anechoic chamber along with the proposed 2-port CP MIMO antenna used to measure radiation & diversities properties of the antenna. To measure the values of the E and H plane, the antenna was placed horizontally and vertically in an anechoic chamber using positioning equipment.

As shown in figures 6(c) and (d), the simulated and measured radiation pattern of the proposed 2-port CP MIMO antenna for the E-& H-Plane is plotted to make comparisons. A very good similarity is obtained between the simulated and measured results while the minor noncompliance is due to manufacturing tolerances and calibration errors during measurements.

5. Diversity parameter of the MIMO antenna

To observe MIMO antenna diversity performance is observed by plotting diversity parameters by using measurement results. The ECC of the proposed 2-port CP MIMO antenna is calculated using (1) [9].

$$|\rho_e(i, j, N)| = \frac{\sum_{n=1}^{N} S_{i,n}^{*} S_{n,j}}{\prod_{k(=i,j)} \left[\sum_{n=1}^{N} S_{i,n}^{*} S_{n,k}\right]} \tag{1}$$

The ECC of the proposed 2-port CP MIMO antenna are shown in figure 7(a). It is 0.004 throughout the entire spectrum of frequencies, showing that the proposed 2-port CP MIMO antenna successfully provides diversity functions. The ITU (International Telecommunication Union) has set a limit of 0.5. ECC should be less than 0.5 to enhance diversity performance. The outcome of 0.004 shows a very low correlation between the antenna components. The diversity gain demonstrates the differences in time-average SNR signals and provides the communication channel with effective diversity is determined using (2).

$$DG = 10\sqrt{1 - ECC^2} \tag{2}$$

The proposed antenna's DG values are around 10 dB as illustrated in figure 7(b), and the DG values should ideally be 10 dB. The radiating elements are located close to one another and are used simultaneously in MIMO antenna architecture, however, the interference performance is also influenced by this configuration of radiating antennas which can be calculated by using (3).

$$TARC = \frac{\sqrt{(|S_{ii} + S_{ij}e^{j\theta}|)^2 + (|S_{ji} + S_{jj}e^{j\theta}|)^2}}{\sqrt{N}} \tag{3}$$

Figure 7(c) displays the simulated TARC values, which are below 10 dB in the 5G Sub-6 GHz band while the predicted values are 0.25 dB. The communication environment's efficient signal transmission without any signal distortion or bit loss is indicated by the channel's capacity. The CCL is a key diversity parameter calculated using the equation (4).

$$CCL = -\log_2 det[\,\beta^R] \tag{4}$$

$$where,\ [\beta^R] = \begin{bmatrix} \beta_{ii} & \beta_{ij} \\ \beta_{ji} & \beta_{jj} \end{bmatrix} \tag{5}$$

$$\beta_{ii} = 1 - \left(\sum_{j=1}^{N} |S_{ij}|^2\right) \tag{6}$$

$$\beta_{ij} = -(S_{ii}^{*} S_{ij} + S_{ji}^{*} S_{ij}) \tag{7}$$

The proposed 2-port CP MIMO antenna has CCL values of 0.10 b/s/Hz, shown in figure 7(d). A MEG is another antenna performance factor that compares the average received signal strength of a MIMO antenna array to its average incident power. MEG can be used to calculate the average received power intensity for each antenna by using (8) and (9)

$$MEG_i = 0.5\left[1 - \sum_{j=1}^{N} |S_{ij}|^2\right] < -3dB \tag{8}$$

$$MEG_j = 0.5\left[1 - \sum_{i=1}^{N} |S_{ij}|^2\right] < -3dB \tag{9}$$

$$|MEG_i - MEG_j| < 3dB \tag{10}$$

$$\left|\frac{MEG_i}{MEG_j}\right| = \pm 3dB \tag{11}$$

In which i, j denote Ant.1 & Ant.2 separately. Figure 7(e) displays the MIMO antenna's mean effective gain. The 3.3–3.6 GHz range of the proposed 2- port CP MIMO antenna comes within the aforementioned 5G frequency range. The proposed 2-port CP MIMO antenna is compact and has a wideband, a very low ECC value, high isolation, diversity gain, and MEG.

The proposed 2-port CP MIMO antenna is compared to the most recent state-of-the-art design in this part, which is tabulated in table 2 and is discussed with earlier published work. The proposed 2-port CP MIMO antenna makes use of an FR4 substrate, which elevates it to the status of a leading candidate for microwave-

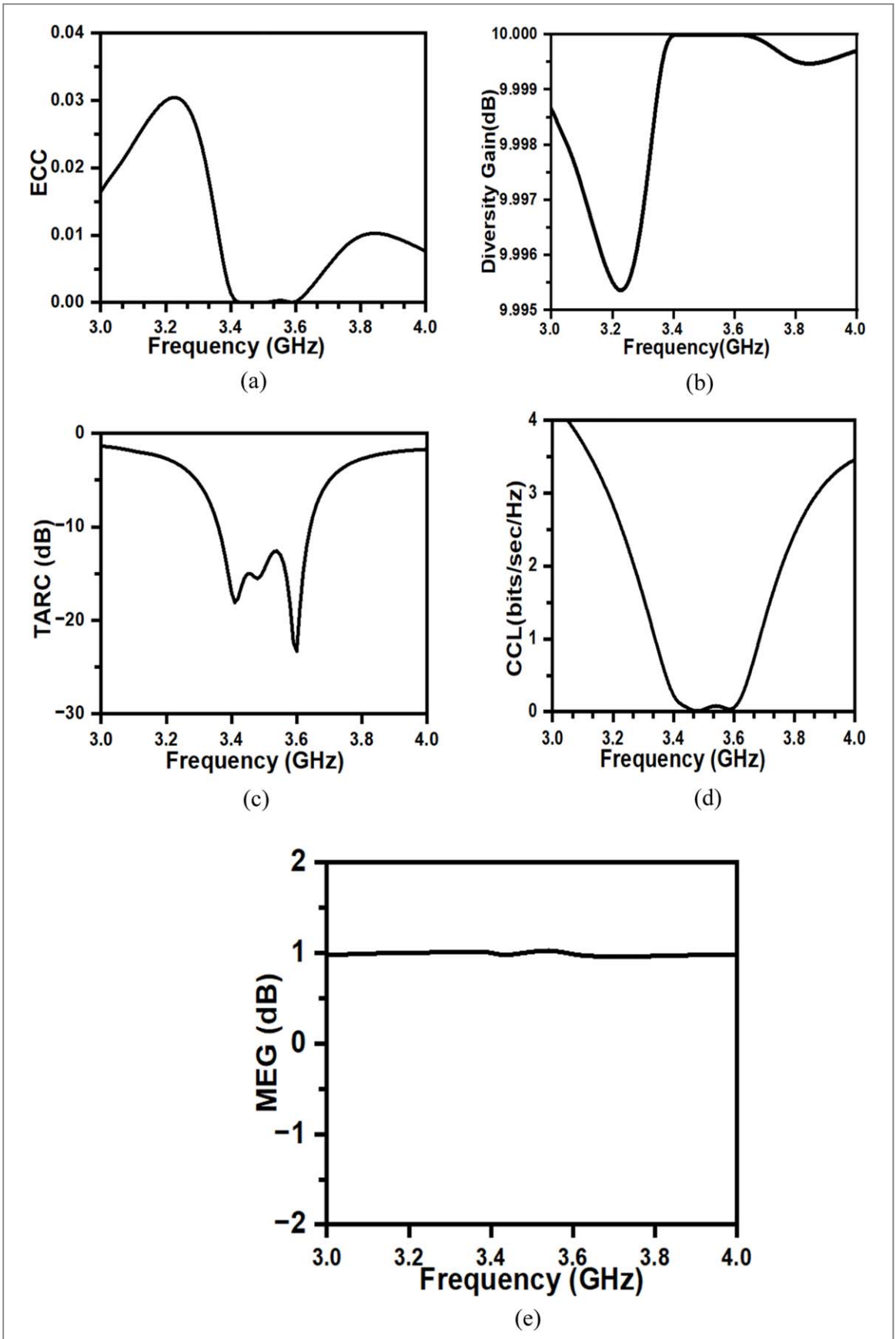


Figure 7. Measured diversity parameters (a) ECC (b) DG (c) TARC (d) CCL (e) MEG of the proposed 2-port CP MIMO antenna.

integrated circuit (MIC) circuit integration. The comparative table also indicates too good diversity performance from the proposed 2-port CP MIMO antenna. The MIMO antenna's innovative design makes it more suited for use in a variety of mobile devices intended for 5G applications.

Table 2. Comparison of the proposed 2- port CP MIMO antenna with other MIMO antennas.

References no.	Antenna geometry (mm^2)	Material used	B.W @ 10dB, (GHz)	Polariz- ation	Isolation (dB)	ECC	Max. Gain (dBi)
[12]	37 × 30	FR4 ($\varepsilon r = 4.3$, $\tan\delta = 0.019$)	3.3–4.2 GHz	CP	15	0.10	2.5
[16]	21 × 46	FR4 ($\varepsilon r = 4.4$, $\tan\delta = 0.019$)	3.4–3.6 GHz & 4–8 GHz	LP & CP	15	<0.003	3.25
[17]	55 × 38	FR-4 ($\varepsilon r = 4.3$, $\tan\delta = 0.025$)	3.1–4.5 GHz	LP	20	<0.0004	3.17
[20]	59 × 55	FR4 ($\varepsilon r = 4.4$, $\tan\delta = 0.02$)	3–7 GHz	LP	18	0.21	3
[21]	82 × 40	FR4 ($\varepsilon r = 4.3$, $\tan\delta = 0.02$)	3.66–3.7	CP	15	0.05	4.6
[22]	66 × 66	ROGERS3003/0.13	3.2–4.12	LP	.20	<0.05	3
[23]	50 × 100	FR4 4.4/4.5	2.7–3.6	LP	25	——	3
Proposed Antenna	65 × 50	FR4 ($\varepsilon r = 4.4$, $\tan\delta = 0.019$)	3.3–3.6 GHz	CP	20	<0.004	3.8

6. Conclusion

The 2-port CP MIMO antenna proposed in this research study uses a distinctive slot and a circular split ring resonator at the ground planes which help to accomplish circular polarization and enhance bandwidth. The proposed 2-port CP MIMO antenna achieves an impedance bandwidth of 3.3 to 3.6 GHz with impedance matching better than 10 dB and has an axial ratio bandwidth of 3.35 to 3.51 GHz with a steady radiation pattern in both planes over the operating frequency bands. The proposed 2-port CP MIMO antenna design with insect fed gives good diversity performance with ECC 0.004, DG 10 dB, CCL 0.40 b/s/Hz, MEG 1.03 dB, and TARC 10 dB. The peak gain for the proposed antenna is 3.8 dBi.

Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Parul Kansal https://orcid.org/0000-0002-1825-4928

References

[1] Andrews J G *et al* 2014 What will 5G be? *IEEE J. Sel. Areas Commun.* **32** 1065–82
[2] Banerjee U, Karmakar A and Saha A 2020 A review on circularly polarized antennas, trends and advances *Int. J. Microw. Wirel. Technol.* **12** 922–43
[3] Samsuzzaman M and Islam M T 2018 Circularly polarized broadband printed antenna for wireless applications *Sensors* **18** 4261
[4] Kulkarni J, Desai A and Sim C 2021 Two port CPW-fed MIMO antenna with wide bandwidth and high isolation for future wireless applications *Int. J. RF Microw. Comput.-Aided Eng.* **31** 8
[5] Parchin N O, Basherlou H J and Abd-Alhameed R A 2020 Dual circularly polarized crescent-shaped slot antenna for 5G front-end systems *Prog. Electromagn. Res. Lett.* **91** 41–8
[6] Tran H H, Hussain N and Le T T 2019 Low-profile wideband circularly polarized MIMO antenna with polarization diversity for WLAN applications *AEU-Int. J. Electron. Commun.* **108** 172–80
[7] Jalali M, Naser-Moghadasi M and Sadeghzadeh R A 2017 Dual circularly polarized multilayer MIMO antenna array with an enhanced SR-feeding network for C-band application *Int. J. Microw. Wirel. Technol.* **9** 1741–8
[8] Malviya L, Panigrahi R K and Kartikeyan M 2016 Circularly polarized 2\times 2 MIMO antenna for WLAN applications *Prog. Electromagn. Res.* C **66** 97–107
[9] Singh A K, Mahto S K and Sinha R 2022 Quad element MIMO antenna for LTE/5G (sub-6 GHz) applications *J. Electromagn. Waves Appl.* **36** 2357–72
[10] Singhwal S S, Kanaujia B K, Singh A, Kishor J and Matekovits L 2020 Multiple input multiple output dielectric resonator antenna with circular polarized adaptability for 5G applications *J. Electromagn. Waves Appl.* **34** 1180–94
[11] Vishvaksenan K S, Mithra K, Kalaiarasan R and Raj K S 2017 Mutual coupling reduction in microstrip patch antenna arrays using parallel coupled-line resonators *IEEE Antennas Wirel. Propag. Lett.* **16** 2146–9
[12] Dwivedi A K, Sharma A, Pandey A K and Singh V 2021 Two port circularly polarized MIMO antenna design and investigation for 5G communication systems *Wirel. Pers. Commun.* **120** 2085–99
[13] Saxena S, Kanaujia B K, Dwari S, Kumar S and Tiwari R 2018 MIMO antenna with built-in circular shaped isolator for sub-6 GHz 5G applications *Electron. Lett.* **54** 478–80
[14] Tiwari R N, Singh P, Kanaujia B K and Kumar P 2021 Compact circularly polarized MIMO printed antenna with novel ground structure for wideband applications *Int. J. RF Microw. Comput.-Aided Eng.* **31** 8
[15] Kumar A, Ansari A Q, Kanaujia B K, Kishor J and Matekovits L 2021 A review on different techniques of mutual coupling reduction between elements of any MIMO antenna: I. DGSs and parasitic structures *Radio Sci.* **56** 1–25
[16] Ameen M, Ahmad O and Chaudhary R K 2020 Single split-ring resonator loaded self-decoupled dual-polarized MIMO antenna for mid-band 5G and C-band applications *AEU-Int. J. Electron. Commun.* **124** 153336
[17] Agrawal N, Gupta M and Chouhan S 2023 Modified ground and slotted MIMO antennas for 5G sub-6 GHz frequency bands *Int. J. Microw. Wirel. Technol.* **15** 817–25
[18] Jaiswal R K, Kumari K, Ojha A K and Srivastava K V 2022 Five-port MIMO antenna for n79-5G band with improved isolation by diversity and decoupling techniques *J. Electromagn. Waves Appl.* **36** 542–56
[19] Liu G, Zhang C, Chen Z and Chen B 2022 A compact dual band MIMO antenna for 5G/WLAN applications *Int. J. Microw. Wirel. Technol.* **14** 1347–52
[20] Kumari T, Das G, Sharma A and Gangwar R K 2019 Design approach for dual element hybrid MIMO antenna arrangement for wideband applications *Int. J. RF Microw. Comput.-Aided Eng.* **29** e21486
[21] Rajkumar S, Vivek Sivaraman N, Murali S and Selvan K T 2017 Heptaband swastik arm antenna for MIMO applications *IET Microw. Antennas Propag.* **11** 1255–61
[22] Abdelghany M A, Ibrahim A A, Mohamed H A and Tammam E 2024 Compact sub-6 GHz four-element flexible antenna for 5G applications *Electronics* **13** 537
[23] Chattha H T 2019 '4-port 2-element MIMO antenna for 5G portable applications *IEEE Access* **7** 96516–20

JMEPEG
https://doi.org/10.1007/s11665-024-10112-6

©ASM International
1059-9495/$19.00

ORIGINAL RESEARCH ARTICLE

Investigations and Multi-response Optimization of Process Parameters for SS316L Cold Metal Transfer-Wire Arc Additive Manufactured Samples

Sudeep Kumar Jain, Qasim Murtaza, and Pushpendra Singh

Submitted: 17 March 2024 / Revised: 15 August 2024 / Accepted: 24 August 2024

Wire and arc additive manufacturing (WAAM) utilizes an electric arc as the primary heat source and solid wire as the material feedstock. Despite potential limitations in precision compared to powder-bed systems, WAAM boasts remarkable deposition rates. This characteristic enables the rapid manufacture of large metallic structural parts, making WAAM a preferred choice for academic research and industrial applications. This study explores the impact of three input process parameters—current (I), welding speed (WS), and Contact Tip to Work Distance (CTWD)—each at three varying levels on three key mechanical properties of SS316L austenitic stainless steel fabricated samples. Through experimentation and analysis, the study evaluates three crucial mechanical properties: ultimate tensile strength (UTS), microhardness (MH), and residual stress (RS). The investigation uses Taguchi's Gray Relational Analysis method, employing an L9 orthogonal array design. Analysis of Variance (ANOVA) has been utilized to assess the impact of parameters on gray relational grade (GRG). The ANOVA analysis for the gray relational grade across various responses indicates that the current is the most significant factor, followed by welding speed and contact tip-to-work distance (CTWD). The contributions of current, welding speed, and CTWD are 76%, 21%, and 3%, respectively. The findings from Gray Relational Analysis (GRA) highlight that optimal performance in terms of maximum UTS, MH, and RS is achieved when employing specific parameter configurations: a I (110 A), WS (0.7 m/min), and CTWD of (3 mm).

Keywords CMT, gray relational analysis, optimization, residual stress measurement, SS 316L, WAAM

1. Introduction

In the realm of advanced manufacturing, Wire Arc Additive Manufacturing (WAAM) stands out as a pioneering technology within the domain of Direct Energy Deposition (DED). Utilizing an electric arc, the welding nozzle positioned at the tip of the arm melts and deposits the metal precisely, aligning with a digital 3D design. The rapid deposition rates, economical utilization of materials and equipment, and consistent structural reliability position WAAM as a disruptive force with the potential to replace conventional manufacturing methods (Ref 1).

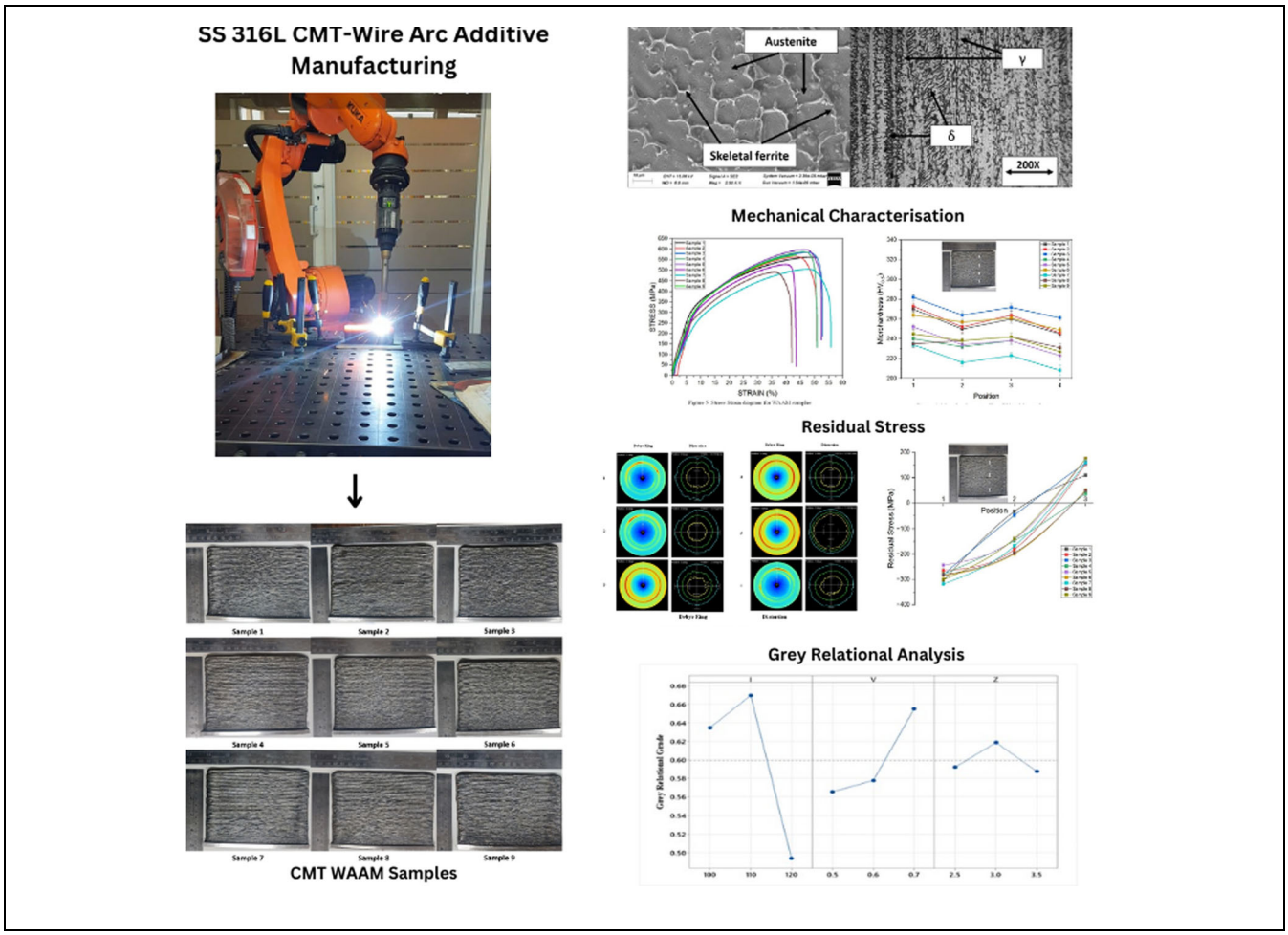
In a direct comparison with conventional subtractive manufacturing techniques, the Wire Arc Additive Manufacturing (WAAM) system has shown remarkable time-efficiency gains, cutting fabrication time by 40-60%. This reduction rate is subjected to variation based on the size of the specific component. Simultaneously, post-machining time is streamlined by approximately 15-20% (Ref 2) . In welding technology, the Fronius Cold Metal Transfer (CMT) is a refined variation of MIG welding. This unique approach aims to produce high-quality weld beads while minimizing thermal heat input and spatter-free welding (Ref 3). WAAM is a technology of growing importance for the car, aerospace, defense, construction, nuclear energy, and biomedical engineering industries. Maritime and shipbuilding companies increasingly adopt WAAM technology in their design and manufacturing processes. A critical recent development is the creation of ship propellers, known as "WAAMpeller," through a collaborative effort involving RAMLAB, Promarin, Autodesk, Bureau Veritas, and Damen (Ref 4). The aerospace sector is witnessing significant progress in WAAM, the fabrication of substantial components, including large parts, tanks, trusses/spars, and airframes. STELIA Aerospace, based in Toulouse, France, is actively designing aluminum airframes and frame sections as part of this emerging trend (Ref 5). Incorporating WAAM into architectural projects holds great promise for advancing sustainable and innovative growth within the industry. The MX3D Bridge in Amsterdam is the first full-

The invited article is part of a special topical issue of the Journal of Materials Engineering and Performance on Advanced Materials Manufacturing. The issue was organized by Antonello Astarita, University of Naples Federico II; Glenn S. Daehn, The Ohio State University; Emily Kinser, ARPA-E; Govindarajan Muralidharan, Oak Ridge National Laboratory; John Shingledecker, Electric Power Research Institute, Le Zhou, Marquette University, and William Frazier, Pilgrim Consulting, LLC, and Editor, JMEP; on behalf of the ASM International Advanced Manufacturing Technical Committee.

Sudeep Kumar Jain, Qasim Murtaza, and **Pushpendra Singh**, Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India. Contact e-mail: sudeepjain16@gmail.com.

Published online: 23 September 2024

Graphical Abstract



scale 3D-printed steel bridge, pioneering WAAM technology in approved large-scale infrastructure projects (Ref 6).

Stainless Steel 316L (SS 316L) excels in durability, biocompatibility, and high-temperature performance while serving well in cryogenic conditions. Furthermore, its exceptional corrosion resistance has established it as a preferred material for nuclear plants, chemical processing facilities, and offshore oil wells (Ref 7).

In recent years, mitigating substantial residual stress in WAAM components has remained a key challenge for advancing WAAM technology. Zhang et al. introduced a novel path-planning strategy for WAAM that combines zigzag and contour-offset techniques, resulting in a single deposition layer with optimized outline accuracy and reduced residual stress. X-ray diffraction demonstrates that contour-offset paths in arc deposits lead to lower tensile residual stresses due to the influence of the external contour on the inner deposit (Ref 8).

X-ray diffraction is commonly used to determine non-destructive residual stresses method in polycrystalline materials. The cos α method for stress calculation uses a complete Debye-Scherrer ring captured on a two-dimensional detector with just one x-ray exposure at a specific incident angle. The Debye-Scherrer ring structure offers valuable visual insights in addition to residual stress values. Variations in signal intensity around the ring can indicate the presence of coarse grain structures or texture. The cosα method relies on Bragg diffraction in polycrystalline materials (Ref 9).

Tanaka et al. compared the $\sin^2\psi$ and cosα methods for measuring residual stress. Stress determination has conventionally been accomplished through the $\sin^2\psi$ method. In recent times, the cosα method has gained popularity among engineers in recent years, particularly with the use of two-dimensional detectors like imaging plates. Under biaxial stress conditions, the cosα method proves its efficiency by employing a singular x-ray exposure to encompass the complete Debye-Scherrer ring on a two-dimensional detector. This innovative technique allows for concurrent evaluation of normal and shear stresses, showcasing stress measurement accuracy to the $\sin^2\psi$ method and applicable across various metals. The cosα method's simple optical setup results in more compact and lightweight stress analyzers, ideally suited for on-site or field measurements. Recent portable stress analyzers using the cosα method have dramatically reduced measurement time to only 60 s (Ref 10) .

Koli et al. studied the WAAM-CMT process, examining the impact of three input parameters: current, travel speed, and gas flow rate. SS308L stainless steel samples were tested for tensile, Micro Hardness, and residual Stress Measurement using a Taguchi L9 array. The study emphasized that current had the most significant impact on multiple responses, with travel speed and gas flow rate following in importance (Ref 11). Kumar

et al. in their study found the best parameter settings for Wire Arc Additive Manufacturing (WAAM) using SS-410. The experiment, conducted with an L9 Taguchi orthogonal array, varies parameters such as nozzle travel speed, welding current, and CTWD across three levels. Macrostructure analysis examines output variables like bead width (BW), depth of penetration (DOP), and bead height (BH). The study evaluates the Geometric Integrity (GI) using response variables including bead width (BW), bead height (BH), depth of penetration (DOP), and reinforcement factor (RFF). Experimental results highlight welding current (WC) as the primary influencing factor on GI, followed by nozzle travel speed and contact CWD. The optimal parameter combination identified is WC of 140 amps, CWD of 10 mm, and NTS of 3 mm/s (Ref 12). Le et al. investigated the WAAM process applied to the fabrication of thin-walled components made from 308L stainless steel. The research examines the impact of welding current, voltage, and travel speed on the geometry of single weld beads in the WAAM process. Additionally, the parameters were optimized for the deposition of 308L steel walls. It was observed that travel speed and voltage primarily impact bead width, while current and travel speed significantly affect bead height (Ref 13). Han et al. performed GMAW welding experiments to create a curved propeller-shaped blade, focusing on crucial welding parameters like current, voltage, travel speed, and CTWD. Optimized process parameters were selected, and it was found that the welding speed and CTWD had significant impacts on both the surface condition and the stability of the welding arc. Short CTWD and low welding speeds can lead to an imperfect bead shape and surface characterized by a short-circuit mode and spattering. Moreover, arc instability can occur when welding speeds increase and the CTWD is prolonged, underscoring the need for careful parameter control in welding processes (Ref 14).

The current review of literature underscores a substantial gap in research pertaining to the process parameters of CMT-WAAM and their impact on mechanical properties. This gap is particularly evident in the absence of studies focusing on the optimization of input parameters to address multiple response variables simultaneously. Recognizing the need for more comprehensive research. In light of this, the study initiated additive experiments to explore the effects of input parameters, namely current, welding speed, and CTWD. The authors commenced this study with two primary aims: the experimental assessment of how process parameters within the CMT process influence the mechanical properties of WAAM samples and the subsequent optimization of these parameters to attain the most favorable outcomes across multiple responses. Furthermore, the study includes material characterization and microstructural analysis for the sample produced using optimal input parameter combination. The study examined three mechanical properties: UTS, MH, and RS in CMT-WAAM samples.

2. Experimental Procedure

This research involved conducting the Design of Experiments (DOE) with the L9 orthogonal array. Samples were first fabricated using Robotic CMT welding, focusing on three input parameters: current (I), welding speed (WS), and contact tip-to-work distance (CTWD). The characterization focused on three mechanical properties: tensile strength, microhardness, and residual stress. Gray Relational Analysis was then utilized to optimize and determine the best process parameters. The microstructural analysis was subsequently conducted on the optimized sample.

2.1 Materials & Methods

The research paper employs Wire Arc Additive Manufacturing (WAAM) for fabrication, employing a KuKa robot (Model: KR 8 R1440, Germany) and Cold Metal Transfer (CMT) source from Fronius (Model: TPS 400i, Austria), as depicted in Fig. 1. Preliminary trials were conducted to establish the best process parameters. The inherent relationships among current, voltage, and wire feed rate in CMT welding, the selection of one parameter establishes the settings for the other two. Consequently, the current was determined through trial runs and an exhaustive literature review. Notably, the crucial determinant impacting the quality of WAAM is the welding speed, playing a substantial role in determining the mechanical properties of the fabricated components.

Furthermore, CTWD is an important parameter that strongly affects the bead geometry and deposition condition, substantially impacting both the surface condition and the welding arc's stability. Considering their influence on the mechanical characteristics witnessed during trial runs, current, welding speed, and CTWD were selected within the array of diverse welding parameters. The shielding gas and its corresponding flow rate were selected after thoroughly reviewing the available literature, acknowledging their significant influence on microstructural properties. Consequently, a consistent supply of a gas mixture comprising 97% Ar and 3% CO_2 with 99.99% purity was maintained at a flow rate of 15 liters per minute.

In fabrication of WAAM samples SS316L filler wire with a diameter of 1.2 mm was employed. Additive manufacturing was carried on wrought SS316L substrate measuring 200 × 50 × 6 mm. Before initiating the metal deposition process, the substrate material was meticulously cleaned using acetone. The filler wire's chemical composition analysis was conducted utilizing optical emission spectroscopy (OES), and the corresponding results shows that the major elements are predominantly Chromium (Cr) 18.56% and Nickel (Ni) 11.55%.

2.2 Characterization of 3D Printed Samples

The INSTRON instrument (Model: 4482, USA) was employed for the tensile testing, ensuring a consistent cross-head speed of 1 mm/min. For accuracy in results, three tensile coupons (C1-C3) were extracted with precision from both the SS 316L WAAM. The extraction process was performed in the transverse orientation (90° perpendicular to welding speed) utilizing Wire-Cut EDM. Tensile test coupons were prepared with strict adherence to predefined standards, followed by a meticulous polishing procedure to ensure the precision and reliability of results. Tensile test coupons were prepared following the ASTM E8M standard, as shown in Fig. 2.

The Vickers Microhardness was evaluated using a Struers instrument (Model: Duramin-40 M1, make: USA) on the polished samples in conformity with the standards specified by ASTM E384. The testing parameters, involving a 500 g load for a 10-second dwell time, were applied to evaluate hardness at four distinct points in the vertical direction, covering the entire span from the bottom to the top zone. This systematic approach was implemented to assess hardness consistency across the WAAM sample (Ref 15). The microstructural characterization

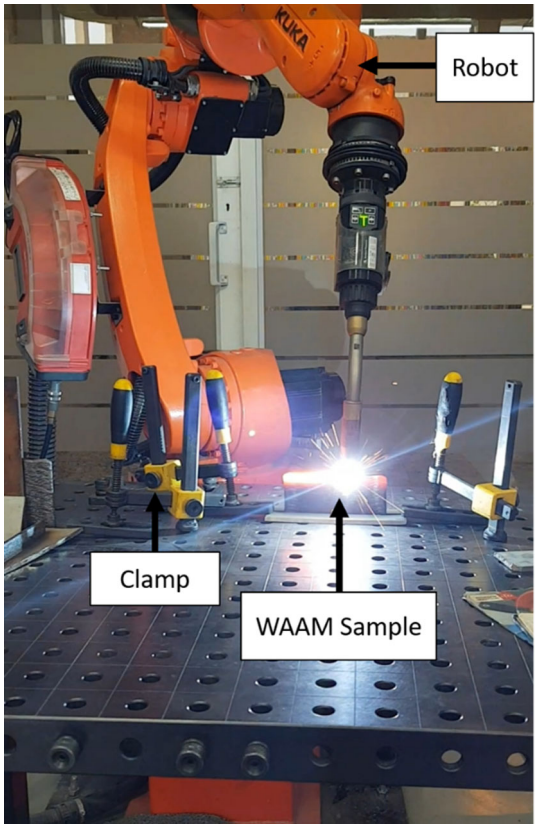


Fig. 1 Robotic CMT WAAM setup

adhered to a standard metallography procedure. The samples were etched by immersing them in a solution containing 30 ml HCl, 120 ml distilled water, and 10 g FeCl3. An Olympus GX 41 optical microscope and a Zeiss (Gemini 2) FESEM were utilized for the microstructural analysis (Ref 16) .

The investigation involved Residual Stress Measurement (RSM), measured by Pulstec μ-X360n Full 2D x-ray system (Japan). The selection of x-ray diffraction (XRD) as the primary approach for residual stress measurement on the surface stemmed from its constrained penetration depth, typically 10 μm. By deliberately configuring the x-ray beam incident angle to 30°, the precision of the measurements was significantly improved. This adjustment utilized the cos α technique, integral to the μ-X360 x-ray Pulstec system (Ref 17). The microstructural characterization followed a standard metallography procedure. The samples were etched by immersing them in a solution composed of 30 ml HCl, 120 ml distilled water, and 10 g $FeCl_3$ (Ref 18).

Three levels for each parameter were considered to investigate how these input parameters affect the mechanical properties of the Wire Arc Additive Manufacturing (WAAM) samples, as specified in Table 1.

Hence, the Taguchi L9 design, with three levels of 3 factors, was employed for the experiment design. The L9 orthogonal array with combinations of input parameters is presented in Table 3. Fabrication of nine WAAM samples (S1-S9) was conducted in a random sequence, adhering to the design specifications in Table 3. The current (I), welding speed (WS) and contact tip to work distance (CTWD) settings were established based on Table 2 for each sample. An in-depth analysis of tensile strength, hardness, and average residual stress is planned, relying on the measured and recorded output.

3. Gray Relational Analysis for Multi-response Optimization

The current research acts as a comprehensive response metric in the framework of Taguchi's experimental design, and the ongoing research makes use of gray relational analysis using the response table that is obtained through the Taguchi method (Ref 19). This optimization is specifically customized to cater to multiple responses, including Ultimate Tensile Strength (UTS), Microhardness (MH), and Residual Stress (RS).

3.1 Data Pre-processing

Using gray relational analysis requires an initial step of data pre-processing to mitigate variations in range and units observed across distinct data sequences. The data pre-processing methodology involves two key steps: firstly, the computation of the S/N ratio directly from the original data, and secondly, the conversion of this ratio into a format that facilitates standardized and comparable assessments (Ref 20).

The L9 orthogonal array is used to conduct the experimental trials, and Table 3 details comprehensive documentation of the corresponding responses. The S/N ratio for each response, adhering to a larger-the-better criterion, is computed utilizing Eq 1. The fabricated samples' MH, UTS, and RS are the significant responses in Wire Arc Additive Manufacturing.

$$S/N\text{ratio}(\mu) = -10\log_{10}\left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{1}{y_{ij}^2} \quad \text{(Eq 1)}$$

where, n = number of replications y_{ij} = observed response values where $i = 1, 2\acute{s}n$; $j = 1, 2\acute{s}k$

To achieve this, it is crucial to normalize the experimental data, bringing it within the standardized range of zero to one. Normalization is achieved through Eq 2 where ($0 \leq x_{ij} \leq 1$) to mitigate the impact of diverse units and minimize variability.

An assessment was undertaken to discern the sensitivity of normalization in the context of sequencing results. In light of these findings, we advocate adopting S/N ratio values when normalizing data for gray relation analysis. The corresponding values of S/N ratios and normalized S/N values for MH, UTS, and RS are provided in Table 4.

$$x_{ij} = \frac{y_{ij} - \min\{y_{ij}\}}{\max\{y_{ij}\} - \min\{y_{ij}\}} \quad \text{(Eq 2)}$$

where y_{ij} are the S/N ratio of the responses

for a given response j in the experiment, i, the processed value x_{ij} equals 1 or exhibits greater proximity to 1 than values derived from other experiments, the experiment i is deemed to have the most favorable performance for the given response.

3.2 Gray Relational Coefficient, Grade and Rank

Next in the process, the determination of how close x_{ij} is to x_{0j} involves utilizing the gray relational coefficient. A greater gray relational coefficient indicates a higher degree of proxim-

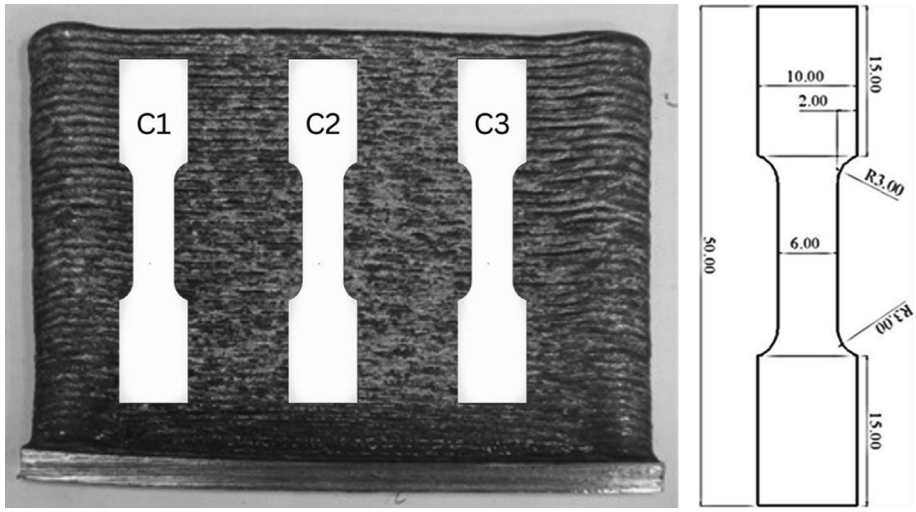


Fig. 2 Tensile Test ASTM E8 Specimen (All dimensions are in mm)

Table 1 Independent input parameters and levels

Parameter	Unit	Level 1	Level 2	Level 3
I	A	100	110	120
WS	m/min	0.5	0.6	0.7
CTWD	mm	2.5	3.0	3.5

Table 2 L9 orthogonal array of independent input parameter

Sample	I, A	WS, m/min	CTWD, mm
S1	100	0.5	2.5
S2	100	0.6	3.0
S3	100	0.7	3.5
S4	110	0.5	3.0
S5	110	0.6	3.5
S6	110	0.7	2.5
S7	120	0.5	3.5
S8	120	0.6	2.5
S9	120	0.7	3.0

ity between x_{ij} and x_{0j}. The gray relational coefficient is calculated using Eq 3.

$$\gamma\left(x_{0j}, x_{ij}\right) = \frac{\Delta_{min} + \varepsilon\Delta_{max}}{\Delta_{ij} + \varepsilon\Delta_{max}} \quad \text{(Eq 3)}$$

$$i = 1,2,\dot{s}.,m \; and \; j = 1,2,\dot{s}n$$

$\gamma\left(x_{0j}, x_{ij}\right)$ is the gray relational coefficient between $x_{0j}\&x_{ij}$

$$\Delta_{ij} = \left|x_{0j} - x_{ij}\right|$$

$$\Delta\max = \max(\Delta_{ij}) \text{ and } \Delta_{\min} = min(\Delta_{ij})$$

$$i = 1,2,\dot{s}.,m$$

Table 3 Experimental results with responses

Sample	I	WS	CTWD	Responses		
				MH[a]	UTS[b]	RS[c]
S1	100	0.5	2.5	256.25	562.77	− 75.33
S2	100	0.6	3.0	258.75	563.34	− 97.32
S3	100	0.7	3.5	269.37	583.88	− 56.33
S4	110	0.5	3.0	233.35	587.22	− 129.33
S5	110	0.6	3.5	236.75	598.88	− 69.66
S6	110	0.7	2.5	257.75	527.22	− 144.18
S7	120	0.5	3.5	220.25	505.00	− 108.33
S8	120	0.6	2.5	236.50	491.00	− 120.25
S9	120	0.7	3.0	238.00	573.33	− 87.66

[a]MH was tested at four points, and the mean value is considered.
[b]UTS values were measured in the vertical direction coupon.
[c]RSM was measured at 3 points and the mean value is considered.

$$j = 1,2,\dot{s}.,n$$

In the present study, the assumed distinguishing coefficient ε is set to 0.5, a value falling within the inclusive range of (0,1]. This coefficient serves as an index for distinguishability, and a smaller ε value corresponds to a higher level of distinguishability.

The mathematical expression used for quantification within the gray relational space is referred to as the gray relational grade. This grade is obtained as a weighted sum of gray relational coefficients, and its calculation is dictated by the defined Eq 4.

$$\Gamma(X_0, X_i) = \sum_{j=1}^{n} w_{j\gamma}\left(x_{0j}, x_{ij}\right) \quad \text{(Eq 4)}$$

$$for i = 1,2,\dot{s}.,m$$

where $\sum_{j=1}^{n} w_{jj} = 1$

The gray relational grade, represented by $\Gamma(X_0, X_i)$, serves as a metric quantifying the degree of comparability between the

Table 4 S/N ratio and normalized S/N ratio values

Sample	S/N ratios			Normalized values of S/N ratios x_{ij}		
	MH	UTS	RS	MH	UTS	RS
S1	52.944	59.778	42.311	0.752	0.687	0.310
S2	53.029	59.787	44.507	0.800	0.692	0.579
S3	53.378	60.098	39.786	1.000	0.872	0.000
S4	52.131	60.147	47.005	0.287	0.901	0.886
S5	52.257	60.318	41.631	0.359	1.000	0.226
S6	52.995	59.211	47.938	0.781	0.358	1.000
S7	51.630	58.837	45.466	0.000	0.142	0.697
S8	52.248	58.593	46.355	0.354	0.000	0.806
S9	52.303	59.939	43.627	0.385	0.780	0.471

sequences X_i and X_0, with X_i being the comparability sequence and X_0 as the reference sequence. The weight assigned to response j is denoted as w_j and typically relies on the judgment of decision-makers. Table 5. shows gray relational co-efficient, grade & rank.

In the experiment context, when the gray relational grade is higher, the associated parameter combination is assigned rank 1. Experiment number 6, characterized I, WS and CTWD by a 110 A (level 2), 0.7 m/min (level 3), 2.5 mm (level 1), respectively, emerges as the closest optimal combination of controllable parameters. Table 6 presents a summary of the means of the gray relational grade corresponding to each input process parameter level. Those entries with a larger gray relational grade are denoted with an asterisk (*), indicating improved multiple performance characteristics. Hence, the optimal parameters for WAAM are specified as the I at 110 A (level 2), WS at 0.7 m/min (level 3), and CTWD of 3.0 mm (level 2), Fig. 3 displays the graph of the gray relational grade, where the overall mean value obtained from the corresponding table is 0.598.

Table 5 Gray relational co-efficient, grade and rank

Sample	Gray Relational co-efficient			Grade	Rank
	MH	UTS	RS		
S1	0.668	0.615	0.420	0.568	6
S2	0.714	0.619	0.543	0.625	4
S3	1.000	0.797	0.333	0.710	2
S4	0.412	0.835	0.814	0.687	3
S5	0.438	1.000	0.393	0.610	5
S6	0.695	0.438	1.000	0.711	1
S7	0.333	0.368	0.622	0.441	9
S8	0.436	0.333	0.720	0.497	8
S9	0.448	0.695	0.486	0.543	7

4. Results and Discussion

In this research, the number of layers for the fabricated WAAM-CMT samples was 40. The effective control of heat input, facilitated by various process parameters, is a critical factor that influences the establishment of layer thickness and the determination of the number of layers. The highest current and slowest welding speed results in maximum heat input. Before mechanical property measurements, a grinding operation was conducted on the samples to achieve a smooth surface finish. The profiles of the CMT WAAM samples are detailed in Table 7.

4.1 Mechanical Properties

The study assessed three mechanical properties, UTS, MH, and RS, for all nine samples. The outcomes are detailed in Table 4 for further examination. Sample 5 achieves the highest tensile strength at 599 MPa, while sample 8 shows the lowest at 491 MPa. Moreover, the ultimate tensile strength (UTS) values for the remaining samples are 505-587 MPa, ranging from 85.55% to 99.4% of rolled steel (590 MPa). Results are derived from tensile strength measurements in the transverse direction, acknowledged as the weakest orientation for strength calculations. Many practical applications adopt an operational approach that confines work to the elastic region, ensuring that stresses remain below the threshold for plastic deformation, which could lead to failure. The stress, therefore, is carefully maintained below the elastic limit. The yield strength (YS) of the fabricated samples exhibited a remarkable increase, reaching around 290 MPa, a considerable improvement over the 230 MPa yield strength observed in traditionally produced stainless steel. This notable improvement underscores the enhanced mechanical characteristics of the fabricated samples. Figure 4 illustrates the Stress-Strain diagram for the WAAM samples. The fabricated samples exhibit a notable percentage elongation ranging from 43.8 to 59.6% at different process parameters. This range contrasts the 48% stainless steel produced using conventional procedures, indicating superior ductility in the fabricated samples. The analysis shows that the fabricated samples utilizing CMT have a notable increase in ductility compared to other conventional manufacturing processes.

Microhardness measurements were taken at four regions from bottom to top (1-4). Ten readings were taken in the horizontal direction, and the average value was used for the analysis The average value was taken to simplify the analysis. The test results reveal a microhardness range of 220-257 HV. Figure 5 visually represents the variation in hardness in the vertical direction. Proximity to the base results in higher hardness due to the base's ability to dissipate heat more rapidly than the top section. Results indicate an uneven temperature distribution, causing hardness to gradually decrease from the base to the top. Despite this, uniform microhardness values across all zones indicate a consistent behavior within the structure, lowering the likelihood of brittle failure. However,

Table 6 Response table for gray relational grades

Factors	1	2	3	Delta	Rank
I	0.634	0.669*	0.493	0.175	1
WS	0.565	0.577	0.654*	0.089	2
CTWD	0.591	0.618*	0.587	0.031	3

The overall mean value derived from the gray relational grade = 0.598
*is representing optimal levels of the gray relational grade

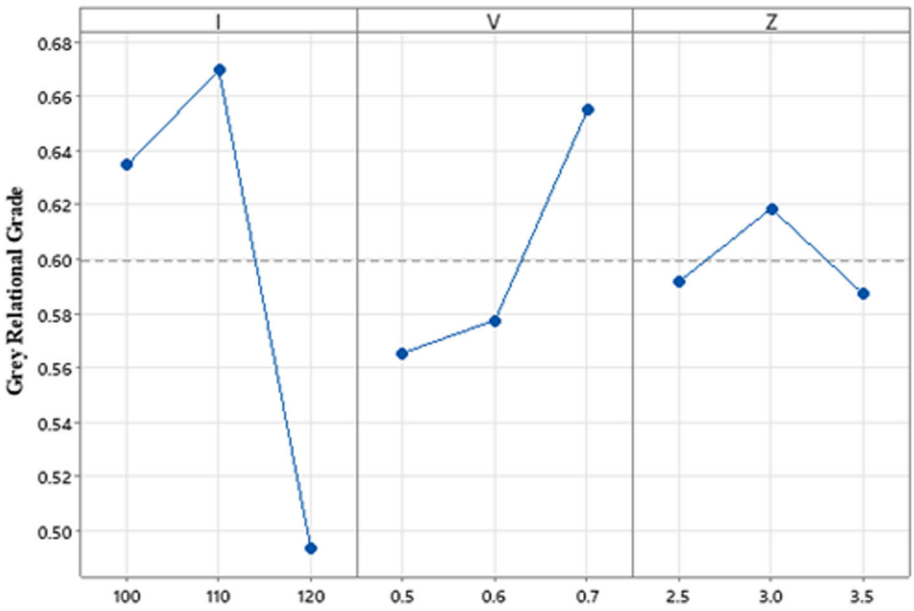


Fig. 3 Gray Relational Grade Graph

Table 7 Dimensions of fabricated samples

Sample	Height, mm	Thickness, mm	Avg. layer thickness, mm
S1	99.68	5.38	2.49
S2	113.23	5.89	2.83
S3	99.34	5.31	2.48
S4	130.35	7.05	3.26
S5	113.24	6.52	2.83
S6	99.62	5.54	2.49
S7	127.35	7.06	3.18
S8	112.45	6.82	2.81
S9	99.75	6.30	2.49

significant heat accumulation between layers causes a progressive reduction in hardness over time (Ref 21) .

Measurements of residual stress were conducted at three regions from bottom to top (1-3). Five readings were taken in the horizontal direction, with the average value employed for the analysis. Compressive residual stresses, generated due to high heat input and rapid cooling, enhance the fatigue strength of WAAM samples. When subjected to tensile loads, tensile residual stresses promote the initiation and propagation of cracks, ultimately resulting in tensile failure. Compressive residual stress is most pronounced at Position 1 on the welding base, primarily due to stress concentration. Conversely, locations away from the base in the transverse direction undergo tensile residual stresses. At point 1 near the substrate, compressive residual stress is generated. The cold substrate causes the bottom fiber to yield tensile stresses in the longitudinal direction (along the weld speed), leading to bending of the substrate. This results in compressive stresses at the layers of WAAM near the substrate. When the temperature gradient disrupts the connection between the

WAAM layers and the substrate, tensile residual stress increases as the WAAM height grows (Ref 22) . Similar findings were reported in the research conducted by Koli et al. (Ref 23). Figure 6 displays the variation of RS in the vertical direction. In this research, the primary consideration is the average residual stress, with a comprehensive analysis conducted on nine samples. A collective analysis of nine samples consistently demonstrates a compressive nature in the average stress values. Table 4 shows the average value of residual stress of WAAM samples at three positions from bottom to top.

4.2 Analysis of Variance of Gray Relational Analysis

ANOVA is used as a robust statistical methodology to discern and quantify the notable impact of various controllable parameters on the mechanical properties of WAAM samples. The process involves separating the total variability in gray relational grades by measuring squared deviations from the average grade. Contributions from individual controllable parameters are identified, along with the error component. The controllable parameter changes in performance were assessed by determining the contribution percent of each process parameter to the total sum of squared deviations. This determines how much each parameter influences the mechanical properties. ANOVA results in Table 8 showcase the impact of different factors on gray relational grades. The I, WS and CTWD contributed 76%, 21%, and 3%, respectively.

4.3 Confirmation Test

The improvement in the mechanical properties of CMT WAAM samples was subjected to a confirmation test. Optimal parameters, outlined in Table 6, were chosen for this confirmation test, ensuring consistency with the identified improvements. Equation 5 defines the estimated gray relational grade γ_e using the optimal levels of input parameters.

$$\gamma_e = \gamma_t + \sum_{i=1}^{p} (\gamma_o - \gamma_t) \quad \text{(Eq 5)}$$

here, γ_t represents the total mean of gray relational grade, γ_o denotes the mean of the gray relational grade at the optimal level, and p signifies the total number of input parameters.

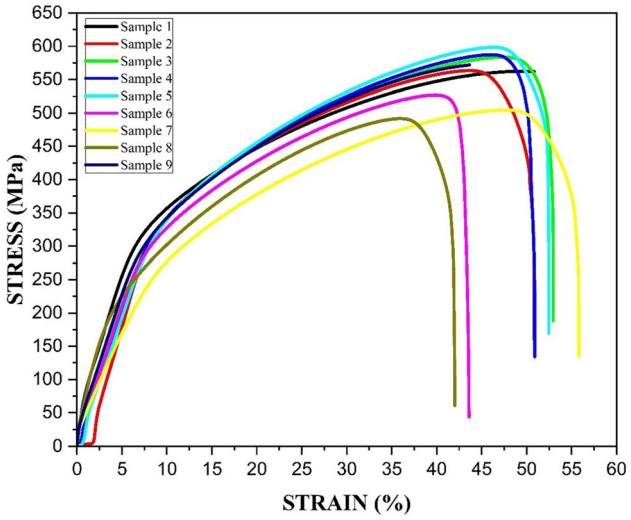


Fig. 4 Stress-Strain diagram for WAAM samples

Executing the confirmation experiment involved setting the input parameters at their optimal levels to validate their significant impact on the mechanical properties of WAAM samples. Table 9 presents the results of the confirmation experiment, displaying gray relational grade values for the initial, predicted, and confirmation phases. The observation in Table 9 reveals that the relational grade obtained at the optimal input parameters surpasses that of the sixth sample in the orthogonal array, indicating a substantial enhancement.

4.4 Microstructural Analysis of Optimal Parameter Sample

The microstructural analysis at the optimal parameters (I = 110 A, WS = 0.7 m/min, and CTWD = 3 mm) is illustrated in Figure 7. The SS316L CMT WAAM microstructure was examined at the center (20 layers from the base) of the WAAM profile. The FESEM and optical microscopy images of optimized sample are shown in Fig. 7(a) and (b). The distribution of the δ-ferrite phase within the γ-austenite matrix can be seen in black in the optical micrographs of the WAAM sample. The direction of grain growth is observed to be vertical.

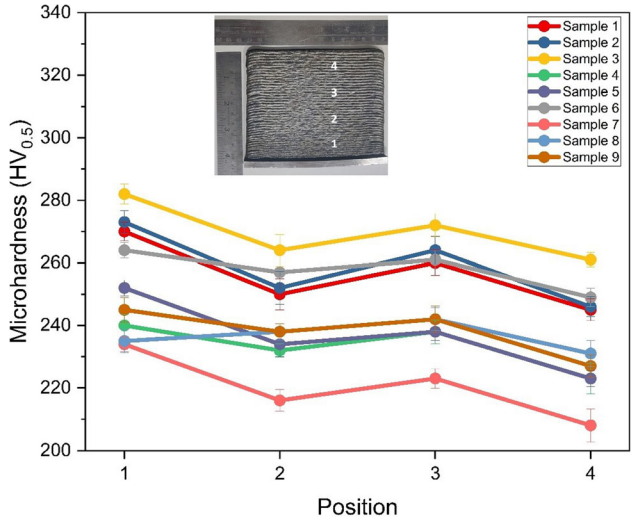


Fig. 5 Microhardness profile of WAAM samples

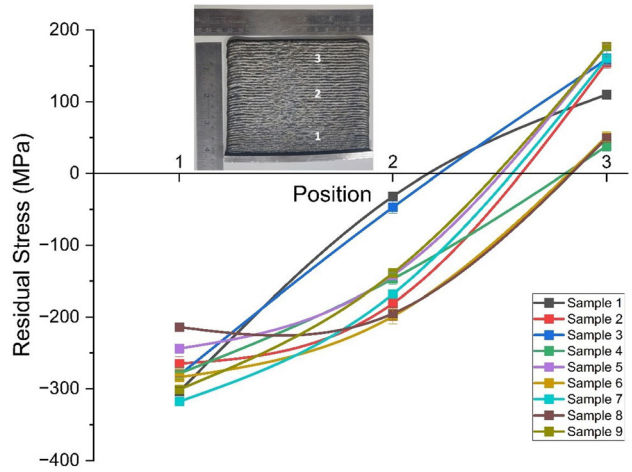


Fig. 6 Residual Stress distribution of WAAM samples

According to this study, a medium level of I (110 A), a high level of WS (0.7 m/min), and a medium level of CTWD (3.0 mm) constitute the ideal combination of process parameters. Since that heat input is directly proportional to current and inversely proportional to welding speed, the low heat input that results can be attributed to the combination of low current and high welding speed. Moreover, the rise in CTWD is associated with a lower heat transfer rate, resulting in faster solidification. These parameter combinations lead to the development of a ferrite-austenite mode. In this mode, the predominant phase is γ-austenite, with δ-ferrite present at the γ-austenite grain boundaries. The δ-ferrite displays morphologies with parallel-structured lathy and scattered-structured skeletal features.

The result of the EDX analysis of the optimized sample is presented in Fig. 8, which reveals a chemical spectrum. The analysis discloses a significant enrichment of Cr (24.78.20%) and Ni (4.06%) within the microstructure. The nonequilibrium solidification process in the deposition pool, which produced a noticeable supersaturation of alloying elements in the austenite phase, is primarily responsible for this enrichment. Significant dissolution of heavy metal elements such as Cr and Ni occurred in the austenitic matrix. The microstructure and EDX plot analysis suggest that the sample produced under the optimal setting of input parameters shares similarities with rolled stainless steel.

Table 8 Analysis of variance

Source	DoF	Adj-SS	Adj-MS	Contribution, %
I	2	0.051	0.025	76.11
WS	2	0.014	0.007	20.89
CTWD	2	0.002	0.001	2.98
Error	2	0.005	0.003	
Total	8	0.072		

DoF—Degree of freedom, SS—Sum of square, MS—Mean square.

5. Conclusions

WAAM stands out as a highly effective and economically viable additive manufacturing process. This effective additive manufacturing method is especially beneficial for fabricating large components owing to its elevated deposition rates. Incorporating cold metal transfer (CMT) technology is crucial in elevating product quality by mitigating heat accumulation defects due to high thermal gradients and defects like cracks. The investigation in this paper focused on evaluating how three critical parameters in the CMT-WAAM process, i.e., I, WS and CTWD, on the mechanical properties of SS 316L austenitic stainless-steel samples. The study applied Taguchi's L9 orthogonal array for a structured analysis. The research adopted the Taguchi gray relational analysis approach to optimize three mechanical properties, i.e., UTS, MH, and RS. The goal was to determine the most effective configuration of process parameters. The outcomes of the study are outlined as follows:

1. Gray relational analysis was employed to optimize input process parameters UTS, MH, and RS. The optimal levels identified for I, WS, and CTWD are 110 A, 0.7 m/min, and 3 mm, respectively. These values are hence recommended as the optimal input parameters for achieving simultaneously higher UTS, MH, and RS.
2. The ANOVA results for the gray relational grade across multiple responses emphasize that current is the most

Table 9 Confirmation test & comparison between initial level & optimum level

		Optimal input parameters	
	Initial parameters in S6 sample	Prediction	Experiment
Setting level	I_2 WS_3 Z_1	I_2 WS_3 Z_2	I_2 WS_3 Z_2
UTS, MPa	257.75		596.2
MH, HV	527		249.6
RS, MPa	144		120
Gray Relational Grade	0.711	0.745	0.747
Improvement in Grade = 0.036			

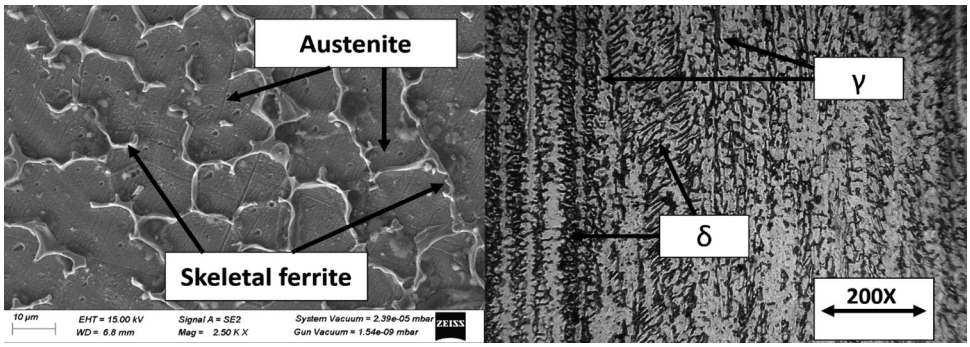


Fig. 7 (a) SEM image (b) Optical Microscope image of WAAM sample

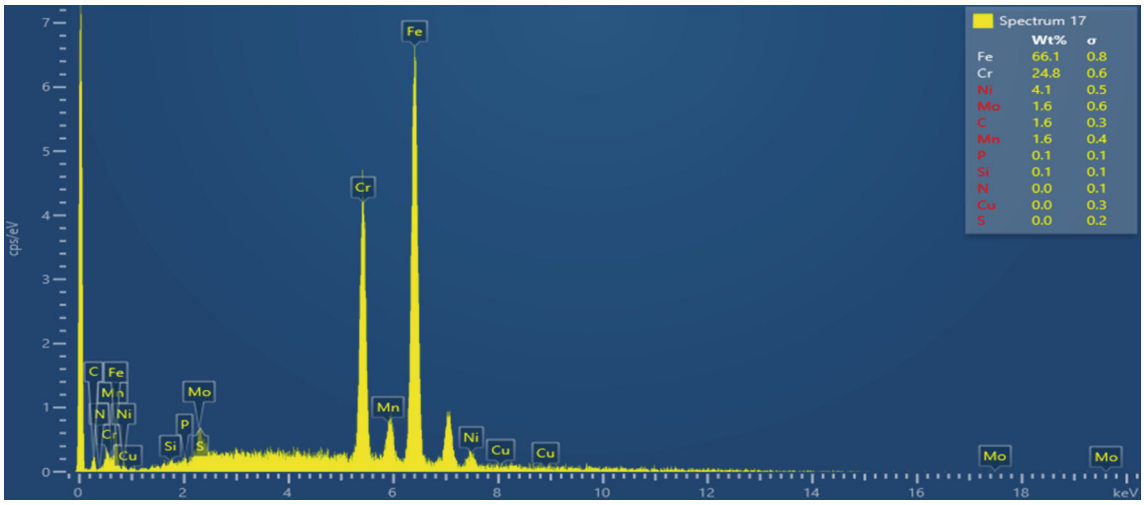


Fig. 8 EDX of 316L WAAM sample

influential parameter, trailed by welding speed and CTWD among the considered parameters.

3. At optimum input parameters UTS, MH shows an increment and slight decrease in RS; also, in confirmation results, there is an increment in gray relational grade by 0.036.
4. In accordance with the residual stress investigation, it is found that compressive residual stress is present at the base and tensile residual stress is present at the top.
5. The sample fabricated under optimal input parameters exhibits a microstructure and EDX plot resembling stainless steel for industrial purposes.

Authors Contribution

All authors contributed to the study conceptualization and design. Material preparation, experimentation, analysis, and the first draft of manuscript writing were performed by all three authors, contributed through study supervision and manuscript writing. All authors read and approved the final manuscript.

Funding

No funding was received for conducting this study.

Conflict of interest

The authors declare no competing interests.

References

1. S.W. Williams, F. Martina, A.C. Addison, J. Ding, G. Pardal, and P. Colegrove, Wire + Arc Additive Manufacturing, *Mater. Sci. Technol. (U.K.)*, 2016, **32**(7), p 641–647. https://doi.org/10.1179/1743284715Y.0000000073
2. B. Wu et al., A Review of the Wire Arc Additive Manufacturing of Metals: Properties, Defects and Quality Improvement, *J. Manuf. Process.*, 2018, **35**(August), p 127–139. https://doi.org/10.1016/j.jmapro.2018.08.001
3. S. Selvi, A. Vishvaksenan, and E. Rajasekar, Cold Metal Transfer (CMT) Technology-An Overview, *Defence Technol.*, 2018, **14**(1), p 28–44. https://doi.org/10.1016/j.dt.2017.08.002
4. A. Taşdemir and S. Nohut, An Overview of Wire Arc Additive Manufacturing (WAAM) in Shipbuilding Industry, *Ships Offshore Struct.*, 2020, **16**(7), p 797–814. https://doi.org/10.1080/17445302.2020.1786232
5. H. Pant, A. Arora, G.S. Gopakumar, U. Chadha, A. Saeidi, and A.E. Patterson, Applications of Wire Arc Additive Manufacturing (WAAM) for Aerospace Component Manufacturing, *Int. J. Adv. Manuf. Technol.*, 2023, **127**(11–12), p 4995–5011. https://doi.org/10.1007/s00170-023-11623-7
6. V. Laghi and G. Gasparini, Explorations of Efficient Design Solutions for Wire-and-Arc Additive Manufacturing in Construction, *Structures*, 2023, **56**, p 104883. https://doi.org/10.1016/j.istruc.2023.104883
7. A. Karpagaraj, S. Baskaran, T. Arunnellaiappan, and N. R. Kumar, "A Review on the Suitability of Wire Arc Additive Manufacturing (WAAM) for Stainless Steel 316," *AIP Conf Proc*, vol. 2247, no. July, 2020, https://doi.org/10.1063/5.0004148
8. C. Zhang, C. Shen, X. Hua, F. Li, Y. Zhang, and Y. Zhu, Influence of Wire-Arc Additive Manufacturing Path Planning Strategy on the Residual Stress Status in One Single Buildup Layer, *Int. J. Adv. Manuf. Technol.*, 2020, **111**(3–4), p 797–806. https://doi.org/10.1007/s00170-020-06178-w
9. V. Finishing, "Separate Print: Vol. 22 March Issue, Year 2021 A New Approach to Fast and Automated Residual Stress Measurements (p. 22-25)," vol. 22, 2021
10. K. Tanaka, The cosα Method for X-Ray Residual Stress Measurement Using Two-Dimensional Detector, *Mech. Eng. Rev.*, 2019, **6**(1), p 18-00378–18-00378. https://doi.org/10.1299/mer.18-00378
11. Y. Koli, S. Arora, S. Ahmad, Priya, N. Yuvaraj, and Z. A. Khan, Investigations and Multi-response Optimization of Wire Arc Additive Manufacturing Cold Metal Transfer Process Parameters for Fabrication of SS308L Samples, *J Mater Eng Perform*, 2023, **32**(5), p 2463–2475, https://doi.org/10.1007/s11665-022-07282-6
12. J. P. Kumar, R. J. Raman, N. J. Festus, M. Kanmani, and S. P. Nath, Optimization Of Wire Arc Additive, *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, p 108–113, 2023, https://doi.org/10.1109/ICISCoIS56541.2023.10100337
13. V. T. Le, D. Si, T. Khoa, and H. Paris, Engineering Science and Technology, An International Journal Wire and Arc Additive Manufacturing of 308L Stainless Steel Components Optimization of Processing Parameters and Material Properties, 24, p. 1015–1026, 2021
14. S.W. Han et al., Parameter Optimization of WAAM with Pulsed GMAW for Manufacturing Propeller-Shaped Blade, *Int. J. Precis. Eng. Manuf.*, 2023, **24**(7), p 1103–1110. https://doi.org/10.1007/s12541-023-00797-5

15. J.S. de Lima, J.F. da Silva Neto, T.M. Maciel, E.A.T. López, R.A.C. de Santana, and T.F. de Abreu Santos, Effect of Wire Arc Additive Manufacturing Parameters on Geometric, Hardness, and Microstructure of 316LSi Stainless Steel Preforms, *Int. J. Adv. Manuf. Technol.*, 2024, **131**(12), p 5981–5996. https://doi.org/10.1007/s00170-024-13240-4
16. V. Mishra, N. Yuvaraj, and Vipin, Tribological Behaviour of Austenitic Stainless Steel-Clad Surface Over Low Carbon Steel Produced by Cold Metal Transfer Welding Process," *Trans. Indian Inst. Metals*, vol. 77, no. 6, pp. 1639–1650, 2024, https://doi.org/10.1007/s12666-024-03261-8
17. V. Mishra, N. Yuvaraj, and V. Vipin, A Comparative Study on Weld-Bead Geometry of Austenitic Stainless-Steel Produced by Cold Metal Transfer Welding and Pulse Metal Inert Gas Welding Process, *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.*, 2023 https://doi.org/10.1177/09544089231208549
18. P.S. Gowthaman, S. Jeyakumar, and D. Sarathchandra, Effect of Heat Input on Microstructure and Mechanical Properties of 316L Stainless Steel Fabricated by Wire Arc Additive Manufacturing, *J. Mater. Eng. Perform.*, 2024, **33**(11), p 5536–5546. https://doi.org/10.1007/s11665-023-08312-7
19. N.D. Ghetiya, K.M. Patel, and A.J. Kavar, Multi-objective Optimization of FSW Process Parameters of Aluminium Alloy Using Taguchi-Based Grey Relational Analysis, *Trans. Indian Inst. Met.*, 2016, **69**, p 917–923
20. A.N. Haq, P. Marimuthu, and R. Jeyapaul, Multi response Optimization of Machining Parameters of Drilling Al/SiC Metal Matrix Composite Using Grey Relational Analysis in the Taguchi Method, *The Int. J. Adv. Manuf. Technol.*, 2008, **37**(3–4), p 250–255. https://doi.org/10.1007/s00170-007-0981-4
21. J. Vora, H. Parmar, R. Chaudhari, S. Khanna, M. Doshi, and V. Patel, Experimental Investigations on Mechanical Properties of Multi-layered Structure Fabricated by GMAW-Based WAAM of SS316L, *J. Mater. Res. Technol.*, 2022, **20**, p 2748–2757
22. J. Sun, J. Hensel, M. Köhler, and K. Dilger, Residual Stress in Wire and Arc Additively Manufactured Aluminum Components, *J. Manuf. Process.*, 2021, **65**, p 97–111. https://doi.org/10.1016/j.jmapro.2021.02.021
23. Y. Koli, S. Arora, S. Ahmad, Priya, N. Yuvaraj, and Z. A. Khan, Investigations and Multi-response Optimization of Wire Arc Additive Manufacturing Cold Metal Transfer Process Parameters for Fabrication of SS308L Samples, *J Mater Eng Perform*, **32**(5), p. 2463–2475, Mar. 2023, https://doi.org/10.1007/s11665-022-07282-6

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.

Pattern Analysis and Applications (2024) 27:119
https://doi.org/10.1007/s10044-024-01336-2

ORIGINAL ARTICLE

K-BEST subspace clustering: kernel-friendly block-diagonal embedded and similarity-preserving transformed subspace clustering

Jyoti Maggu[1] · **Anurag Goel**[2]

Received: 13 June 2024 / Accepted: 2 September 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract
Subspace clustering methods, employing sparse and low-rank models, have demonstrated efficacy in clustering high-dimensional data. These approaches typically assume the separability of input data into distinct subspaces, a premise that does not hold true in general. Furthermore, prevalent low-rank and sparse methods relying on self-expression exhibit effectiveness primarily with linear structure data, facing limitations in processing datasets with intricate nonlinear structures. While kernel subspace clustering methods excel in handling nonlinear structures, they may compromise similarity information during the reconstruction of original data in kernel space. Additionally, these methods may fall short of attaining an affinity matrix with an optimal block-diagonal property. In response to these challenges, this paper introduces a novel subspace clustering approach named Similarity Preserving Kernel Block Diagonal Representation based Transformed Subspace Clustering (KBD-TSC). KBD-TSC contributes in three key aspects: (1) integration of a kernelized version of transform learning within a subspace clustering framework, introducing a block diagonal representation term to generate an affinity matrix with a block-diagonal structure. (2) Construction and integration of a similarity preserving regularizer into the model by minimizing the discrepancy between inner products of the original data and those of the reconstructed data in kernel space. This facilitates enhanced preservation of similarity information between the original data points. (3) Proposal of KBD-TSC by integrating the block diagonal representation term and similarity preserving regularizer into a kernel self-expressing model. The optimization of the proposed model is efficiently addressed through the alternating direction method of multipliers. This study validates the effectiveness of the proposed KBD-TSC method through experimental results obtained from nine datasets, showcasing its potential in addressing the limitations of existing subspace clustering techniques.

Keywords Block diagonal representation · Kernel representations · Non-linear subspace clustering · Similarity preserving · Transformed subspace clustering

1 Introduction

In recent research endeavors, considerable attention has been directed towards the development of subspace clustering methods, addressing the efficient processing of high-dimensional data. Subspace clustering serves as a valuable technique for grouping high-dimensional data distributed across a union of subspaces. The fundamental premise of subspace clustering posits that data samples within the same cluster should reside in a common subspace. Widely applied across various domains such as computer vision, face clustering [24, 55, 75], video analysis [66], image representation and compression [14], hyperspectral image processing [73], saliency detection [6], motion segmentation [65], and domain adaptation task [28], subspace clustering has become a versatile tool.

Existing subspace clustering methods fall into five broad categories [60]: iterative models [13], algebraic models [62], statistical models [8], spectral clustering-based models [6, 23, 30, 66], and deep learning-based models [15, 17, 18, 42, 78]. Among these, spectral clustering-based subspace clustering stands out as the most renowned due to its extensive exploration and practical applications [6, 54, 71]. The

✉ Anurag Goel
anurag@dtu.ac.in

Jyoti Maggu
jyotim@iiitd.ac.in

[1] CSED, Thapar Institute of Engineering and Technology, Patiala, Punjab 147004, India

[2] Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India

Published online: 19 September 2024

spectral clustering-based subspace clustering primarily consists of two components: (1) the construction of an affinity matrix and (2) spectral clustering [53]. Recently, diverse forms of affinity matrices have surfaced, depending on the regularization term incorporated. The most notable models include sparse subspace clustering (SSC) [4, 5, 59] and clustering based on low-rank representation (LRR) [3, 12, 26, 63, 72, 77]. These approaches, rooted in sparse or low-rank representation techniques, acquire a coefficient matrix through a self-expression model. SSC enforces an l_1 norm constraint on coefficients, while LRR employs l_* norm or nuclear norm constraints on coefficients. Several works focused on enhancing the robustness of subspace clustering algorithms by employing a multi-objective framework and subspace optimization techniques [41], integrating subspace fuzzy clustering techniques [64] or adjusting the threshold based on the distribution of data points within subspaces [22].

Various techniques for subspace clustering have been developed, each employing distinct constraints to derive an optimal affinity matrix [1, 2, 30, 44, 56, 61, 71]. However, these methods are limited to processing linear subspaces, proving inadequate for handling the prevalent nonlinear structures found in real-world data [21, 52]. Another constraint is the assumption that data can be separated into distinct subspaces. Recently, techniques based on multiview clustering [11, 45–47, 57, 58, 74, 76] have also gained attention.

To address the limitations associated with linear subspace clustering on nonlinear data, kernel self-expression methods have been introduced [16, 27, 38–40, 66]. Kernelized SSC (KSSC) [40] and kernelized LRR (KLRR) [37, 66] are two prominent methods that capture nonlinear structure information in the input space, exhibiting significant advancements in various applications. Despite their efficiency in processing nonlinear data, these kernel subspace clustering methods may lose some similarity information between samples during the reconstruction of the original data in kernel space.

Recently, transform learning-based approaches for subspace clustering [9, 10, 33–35] have garnered attention. These methods operate on data originally deemed inseparable into subspaces, transforming it into a high-dimensional feature space where linear separability into subspaces is achievable. Notably, these methods eliminate the need to manually choose a mapping function, as they autonomously learn the mapping from the data itself.

In practical scenarios, real-world data exhibiting manifold structures often entails complexities beyond mere sparsity or low-rank characteristics. Consequently, it becomes crucial to formulate a representation that can adeptly capture the intricate structural information inherent in the original data. Numerous methodologies have been devised to uncover underlying structures by delving into data relationships [33–35]. Recently introduced subspace clustering methodologies grounded in structure learning comprise Similarity Learning via Kernel-Preserving Embedding (SLKE) [20] and Structure Learning with Similarity Preserving (SLSP) [19]. SLKE constructs a model that preserves similarity information among data, resulting in improved performance. In contrast, SLSP establishes a structure learning framework that integrates the similarity information of the original data, addressing potential drawbacks associated with the SLKE algorithm, which could lead to the loss of certain low-order information. Despite these methods exhibiting commendable performance, their effectiveness is dependent on a learned similarity matrix that might lack an optimal block diagonal structure for spectral clustering.

In the literature, diverse norm regularization terms have been utilized in self-expressive models to acquire a block diagonal coefficient matrix. These terms include the 1-norm, l2-norm, and nuclear norm. However, these regularization techniques exhibit two shortcomings: an inability to control the number of blocks in the coefficient matrix and the potential suboptimality of the learned coefficient matrix due to data noise. Addressing these limitations, Block Diagonal Representation (BDR) subspace clustering algorithms [30, 70] have been introduced, aiming to attain a good block diagonal structure in the coefficient matrix. For example, Implicit Block Diagonal Low-Rank Representation (IBDLR) integrates block diagonal priors and implicit feature representation into the low-rank representation model, progressively enhancing clustering performance [67]. Notably, these BDR-based subspace clustering methods, while effective, have not been integrated into similarity-preserving mechanisms. The kernelized version of transform learning was introduced in [32].

This work proposes "Similarity Preserving Kernel Block Diagonal Representation based Transformed Subspace Clustering (KBD-TSC)" that leverages the kernel self-expressing framework. Kernelized transformed subspace clustering accounts for the data that is not originally separable into subspaces by leveraging kernel self-expression-based transform learning. The proposed method transforms the data into high-dimensional feature spaces where they are linearly separable into subspaces. It doesn't suffer from choosing the mapping function, as it learns the mapping from the data itself. Although the kernel subspace clustering methods based on kernel self-expression can efficiently process the nonlinear structure data, some similarity information between samples may be lost when reconstructing

the original data in kernel space. The integration of similarity preserving regularizer and block diagonal regularizer into the proposed model facilitates enhanced preservation of similarity information between the original data points. Experimental results on nine datasets validate the effectiveness and robustness of the proposed KBD-TSC method.

The principal contributions of this paper are as follows:

- A novel subspace clustering approach is proposed which accounts for the data that is not originally separable into subspaces by leveraging kernel self-expression-based transform learning.
- A similarity preserving regularizer is incorporated in the proposed model to facilitate enhanced preservation of similarity information between the original data points.
- The integration of a block diagonal representation into the proposed model aims to derive a similarity matrix characterized by an optimal block diagonal structure. This helps in achieving the desired optimal block diagonal matrix.
- The proposed KBD-TSC model is evaluated on nine datasets featuring different types of manifolds, including handwritten digits clustering, face image clustering, object clustering, and text clustering. The experiments involved comparing the proposed model with several state-of-the-art approaches. The results strongly support the effectiveness of our proposed model.

The structure of the remaining paper is organized as follows: Sect. 2 provides a review of basic concepts in related work. Section 3 elaborates on the proposed algorithms and their solutions. Subsequently, Sect. 4 discusses experimental results, and Sect. 5 concludes the paper.

2 Background

2.1 Subspace clustering

The standard subspace clustering techniques are self-expression-oriented, where the aim is to express every single data point in terms of the linear combination of the rest of other data points that lie under the same subspace. The data must hold the basic pre-requisite of being separable into various subspaces. Let $\boldsymbol{X} = [x_1, x_2, \cdots, x_N] \in \mathfrak{R}^{d \times N}$ be the matrix of data points where every column vector $\boldsymbol{x_i}$ is drawn from a union of lower-dimensional subspaces $\{\boldsymbol{S_1} \bigcup \boldsymbol{S_2} \bigcup \cdots \boldsymbol{S_n}\}$ that have dimensions $\{d_k\}_{k=1}^{n}$ where n is the total number of manifolds. The subspace clustering technique targets to segments each set X_k of N_k points that basically belong to the same subspace S_k of dimension $d - k$.

$$\underset{\mathbf{Z}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{X} - \mathbf{XZ}\|_F^2 + \lambda(\Omega(\mathbf{Z})),\ \text{s.t.}\ \ \text{diag}(\mathbf{Z}) = 0,\ \mathbf{Z} \geq 0. \tag{1}$$

where $\Omega(\boldsymbol{Z})$ is the regularization term and $\lambda > 0$ denotes the hyperparameter. $\|\mathbf{Z}\|_*$, $\|\mathbf{Z}\|_1$, $\|\mathbf{Z}\|_F^2$ are three common regularizers. Then an affinity matrix is constructed using $\boldsymbol{Z}$, which applies any graph-cut technique to compute clusters.

2.2 Kernelized subspace clustering

The kernelized version for the subspace clustering is backed by mapping a kernel function as follows:

$$\begin{aligned} &\underset{\mathbf{Z}}{\text{minimize}}\ \frac{1}{2}\|\ker(\mathbf{X}) - \ker(\mathbf{X})\mathbf{Z}\|_F^2 + \lambda(\Omega(\mathbf{Z})) \\ &\equiv \underset{\mathbf{Z}}{\text{minimize}}\ Tr(\mathbf{I} - 2\mathbf{Z} + \mathbf{Z}^\top\mathbf{Z})K + \lambda(\Omega(\mathbf{Z})), \\ &\quad \text{s.t.}\quad \text{diag}(\mathbf{Z}) = 0,\ \mathbf{Z} \geq 0. \end{aligned} \tag{2}$$

where $\ker(\boldsymbol{X})$ is the kernel mapping function, $\boldsymbol{K}$ is the kernel matrix where each matrix element is calculated as $K_{i,j} = \ker(X_i)^\top \ker(X_j)$.

2.3 Kernelized transform learning

A more recent unsupervised representation learning methodology known as transform learning that functions as a dictionary learning method with analysis capabilities. It seeks to train the transform matrix and coefficient matrix from the input data in such a way that the learned transform is capable of analysing the data and ultimately generating the coefficients matrix [48–51]. Since we anticipate that the input data would be able to be divided into many groups, we typically use raw data pixels as input to the clustering method. However, transform learning can be applied to any high-dimensional data, resulting in the effective representations of the transform matrix and the coefficients' matrix being learned in latent space. The data's nonlinearity can be handled by the kernel approach in an effective and efficient manner. Kernelization can be used if the learned transform coefficients are not linearly divided into distinct subspaces. The kernel transform learning [32] formulation can be given as Eq. 3.

$$\underset{\mathbf{A},\mathbf{Z}}{\text{minimize}}\ \|\mathbf{AK} - \mathbf{Z}\|_F^2 + \epsilon(\|\mathbf{A}\|_F^2 - logdet(\mathbf{A})) + \mu\|\mathbf{Z}\|_1. \tag{3}$$

where $\boldsymbol{K} = \ker(\mathbf{X})^\top \ker(\mathbf{X})$ is the kernel, ϵ and μ are the hyperparameters. Constraints introduced in Eq. 3 assist prevent trivial solutions. The first constraint − logdet[1]$(\boldsymbol{A})$

[1] logdet($\mathbf{T}$)= log(singular values). If some singular value ≤ 0, then the log takes $+\infty$ as output. For the case when $\boldsymbol{T}$ is not square, the algorithm solves -logdet($\boldsymbol{T'T}$) + $||\boldsymbol{T}||_F^2$.

guarantees that $\boldsymbol{A}$ is the complete matrix. To prevent it from rising to the other extreme, the second restriction $\|\boldsymbol{A}\|_F^2$ is used. The other restriction, $\|\boldsymbol{Z}\|_1$, is used to make the coefficients sparser.

Equation 3. is solved using an alternating minimization method. Equation 4. describes how the iterative updating procedures for $\boldsymbol{A}$ and $\boldsymbol{Z}$ are carried out alternately.

$$\begin{aligned}\mathbf{A} &\leftarrow \underset{\mathbf{A}}{\text{minimize}}\ \|\mathbf{AK}-\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A}));\\ \mathbf{Z} &\leftarrow \underset{\mathbf{Z}}{\text{minimize}}\ \|\mathbf{AK}-\mathbf{Z}\|_F^2+\mu\|\mathbf{Z}\|_1.\end{aligned} \tag{4}$$

The update of $\boldsymbol{A}$ is pretty straightforward as it is directly differentiable. There are some linear algebraic tricks to solve it too which are given in [31]. Equation 5 describes how to perform one-step soft thresholding to update $\boldsymbol{Z}$.[2]

$$\mathbf{Z} \leftarrow \mathit{signum}(\mathbf{AK}).\max(0,\mathit{abs}(\mathbf{AK})-\mu). \tag{5}$$

2.4 Transformed subspace clustering

Subspace clustering operates on the underlying assumption that the given data can be segmented into numerous sub-spaces. Yet, in the context of high-dimensional data, this assumption often does not hold true. To address this challenge, we employ clustering algorithms on the features derived from transform learning. This approach involves projecting the data onto a latent space, where we acquire separable coefficients that define distinct sub-spaces [34]. The modified subspace clustering can be articulated as follows, focusing on acquiring sparse representations based on the transformed features.

$$\begin{aligned}&\underset{\mathbf{A},\mathbf{Z},\mathbf{C}}{\text{minimize}}\ \|\mathbf{AX}-\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A}))\\ &+\gamma\|\mathbf{Z}-\mathbf{ZC}\|_F^2+\lambda(\Omega(\mathbf{C})).\end{aligned} \tag{6}$$

Alternate minimization, as described in Eq. 7, is utilised to solve Eq. 6. In each cycle, we alternately update $\boldsymbol{A}$, $\boldsymbol{Z}$, and $\boldsymbol{C}$.

$$\begin{aligned}\mathbf{A} &\leftarrow \underset{\mathbf{A}}{\text{minimize}}\ \|\mathbf{AX}-\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A}));\\ \mathbf{Z} &\leftarrow \underset{\mathbf{Z}}{\text{minimize}}\ \|\mathbf{AX}-\mathbf{Z}\|_F^2+\gamma(\|\mathbf{Z}-\mathbf{ZC}\|_F^2);\\ \mathbf{C} &\leftarrow \underset{\mathbf{C}}{\text{minimize}}\ \|\mathbf{Z}-\mathbf{ZC}\|_F^2+\lambda(\Omega(\mathbf{C})).\end{aligned} \tag{7}$$

After calculating the affinity matrix and applying spectral clustering to it to discover clusters, we obtain $\boldsymbol{C}$. Additionally, to create the kernelized transform subspace clustering formulation, the subspace clustering loss is included in equation 6.

2.5 Block diagonal representation

By adding a block-diagonal regularisation term, the BDR method [29] completes the block diagonal matrix and achieves higher clustering performance. The BDR algorithm's optimization model is expressed as follows:

$$\begin{aligned}&\underset{\mathbf{Z}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{Z}\|_{\boxed{\text{m}}}+\|\ker(\mathbf{X})-\ker(\mathbf{X})\mathbf{Z}\|_F^2,\\ &s.t.\mathbf{Z}\geq 0,\mathbf{Z}^\top=\mathbf{Z},\mathit{diag}(\mathbf{Z})=0.\end{aligned} \tag{8}$$

Here, $\boldsymbol{X}$ and $\boldsymbol{Z}$ represent data matrix and coefficient matrix respectively, while $\|\boldsymbol{Z}\|_{\boxed{\text{m}}}$ denotes m-block diagonal regularizer.

3 Proposed method: KBD-TSC

The fundamental assumption underlying self-expression-based subspace clustering is the requirement for data to be segregable into distinct subspaces. Traditional methods also rely on the assumption of an inherent linear structure. However, when these assumptions are not met, particularly when data samples are non-separable into subspaces and linear subspace clustering methods struggle with non-linear structures, a more adaptable model is needed. In response to this challenge, we present a model designed to generalize effectively on non-linear structured data, even when they are not easily separable into subspaces.

The proposed Kernel Block Diagonal-based Transformed Subspace Clustering (KBD-TSC), is adept at preserving similarity information among samples, concurrently achieving an optimal block diagonal structure in the obtained similarity matrix. This approach involves embedding a non-linear model that integrates kernelized transformed subspace clustering with a kernel self-expression framework to achieve the desired objective. Incorporating a block diagonal regularization term into the kernel self-expression framework is pivotal for obtaining a similarity matrix characterized by a block diagonal structure. Furthermore, the preservation of similarity information is secured by minimizing the distinction between two inner products: one encompassing the inner products among original data in kernel space and the other involving the inner products of reconstructed data in

[2] The signum function returns -1 if the argument is negative, 0 if the argument is zero, and 1 if the argument is positive.

kernel space. The resolution of the entire optimization problem employs alternate minimization techniques.

3.1 Similarity preserving model

To uphold similarity information among samples, our objective is to minimize the difference between two inner products. One corresponds to the inner product among the original data in kernel space, and the other corresponds to the inner product of the reconstructed data in kernel space, drawing inspiration from the work of Kang et al. [20].

$$\underset{\mathbf{Z}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{K}-\mathbf{Z}^{\top}\mathbf{K}\mathbf{Z}\|_F^2. \tag{9}$$

where $\boldsymbol{K} = \ker(\mathbf{X})^{\top}\ker(\mathbf{X})$ is a positive semi-definite matrix.

3.2 Proposed algorithm

We introduce a model designed for effective generalization on non-linear manifolds. This model integrates kernelized self-expression transformed subspace clustering with a similarity-preserving kernel block diagonal representation. The kernelized component of transform learning characterizes non-linear data as a linear combination of itself in the transform domain. The incorporation of transform learning with subspace clustering loss facilitates the separation of data into subspaces. To enhance this process, a block diagonal regularization term is introduced, aiming to achieve a similarity matrix between samples with a block diagonal structure. Consequently, our proposed model not only preserves the similarity information among non-linear samples in the transform domain but also simultaneously acquires a similarity matrix with an optimal block diagonal structure.

The complete joint formulation for the proposed model is expressed as Eq. 10.

$$\begin{aligned}\underset{\mathbf{A},\mathbf{Z}}{\text{minimize}}\ \Bigg(& \underbrace{\|\mathbf{AK}\text{-}\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A}))}_{\text{Kernelized transform learning}} \\ & + \underbrace{\frac{1}{2}Tr(\mathbf{Z}-2\mathbf{K}\mathbf{Z}+\mathbf{Z}^{T}\mathbf{K}\mathbf{Z})}_{\text{Kernel self expression subspace clustering}} + \underbrace{\alpha\|\mathbf{K}-\mathbf{Z}^{T}\mathbf{K}\mathbf{Z}\|_F^2+\gamma\|\mathbf{Z}\|_{\boxed{\text{k}}}}_{\text{Similarity preserving with block diagonal}}\Bigg) \\ & s.t.\ \mathbf{Z}\geq 0, diag(\mathbf{Z})=0, \mathbf{Z}^{\top}=\mathbf{Z}.\end{aligned} \tag{10}$$

where α, ϵ, γ are positive hyperparameters.

To simplify and separate out the variables, let us introduce an auxiliary matrix $\boldsymbol{B}$ and a regularization term $\|\boldsymbol{Z}-\boldsymbol{B}\|_F^2$ into our proposed model. Thus, the optimization problem in equation 10 can be translated to

$$\begin{aligned}\underset{\mathbf{A},\mathbf{Z},\mathbf{B}}{\text{minimize}}\ & (\|\mathbf{AK}-\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A}))+\tfrac{1}{2}Tr(\mathbf{K}-2\mathbf{K}\mathbf{Z}+\mathbf{Z}^{\top}\mathbf{K}\mathbf{Z}) \\ & +\alpha\left\|\mathbf{K}-\mathbf{Z}^{\top}\mathbf{K}\mathbf{Z}\right\|_F^2+\tfrac{\beta}{2}\|\mathbf{Z}\text{-}\mathbf{B}\|_F^2+\gamma\|\mathbf{B}\|_{\boxed{\text{k}}}) \\ & s.t.\ \mathbf{Z}\geq 0, diag(\mathbf{Z})=0, \mathbf{Z}^{\top}=\mathbf{Z}.\end{aligned} \tag{11}$$

3.3 Optimization of the proposed KBD-TSC model

To facilitate the solution of the problem in equation 11, three new auxiliary variables are introduced that leads to the following equivalent problem:

$$\begin{aligned}\underset{\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{J},\mathbf{G},\mathbf{H}}{\text{minimize}}\ & (\|\mathbf{AK}-\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A}))+\tfrac{1}{2}Tr(\mathbf{K}-2\mathbf{K}\mathbf{J}+\mathbf{Z}^{\top}\mathbf{K}\,\mathbf{J}) \\ & +\alpha\left\|\mathbf{K}-\mathbf{G}^{\top}\mathbf{K}\,\mathbf{H}\right\|_F^2+\tfrac{\beta}{2}\|\mathbf{J}\text{-}\mathbf{B}\|_F^2+\gamma\|\mathbf{B}\|_{\boxed{\text{k}}}) \\ & s.t.\ \mathbf{B}\geq 0, diag(\mathbf{B})=0, \mathbf{B}^{\top}=\mathbf{Z}, \mathbf{G}=\mathbf{Z}, \mathbf{H}=\mathbf{Z}.\end{aligned} \tag{12}$$

We use ADMM for solving equation 12, and its corresponding Lagrangian [25] is given as follows:

$$\begin{aligned}\check{}\ & (\mathbf{Z},\mathbf{J},\mathbf{G},\mathbf{H},\mathbf{B},\lambda_1,\lambda_2,\lambda_3) \\ & =\|\mathbf{AK}\text{-}\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{A}\|_F^2-\log\det\mathbf{A})+\tfrac{1}{2}Tr(\mathbf{K}-2\mathbf{K}\mathbf{J}+\mathbf{Z}^{\top}\mathbf{K}\,\mathbf{J}) \\ & \quad +\alpha\left\|\mathbf{K}-\mathbf{G}^{\top}\mathbf{K}\,\mathbf{H}\right\|_F^2+\tfrac{\beta}{2}\|\mathbf{J}\text{-}\mathbf{B}\|_F^2+\gamma\|\mathbf{B}\|_{\boxed{\text{k}}} \\ & \quad +\tfrac{\mu}{2}\left[\left\|\mathbf{J}\text{-}\mathbf{Z}+\tfrac{\lambda_1}{\mu}\right\|_F^2+\left\|\mathbf{G}\text{-}\mathbf{Z}+\tfrac{\lambda_2}{\mu}\right\|_F^2+\left\|\mathbf{H}\text{-}\mathbf{Z}+\tfrac{\lambda_3}{\mu}\right\|_F^2\right].\end{aligned} \tag{13}$$

where $\lambda_1, \lambda_2, \lambda_3$ are Lagrangian multipliers and $\mu > 0$ is a penalty parameter. Now, these variables can be updated alternately. The updates for all variables are given as follows:

- **Update A**: After keeping other variables fixed, $\boldsymbol{A}$ can be updated as follows:

$$\underset{\mathbf{A}}{\text{minimize}}\ \|\mathbf{AK}\text{-}\mathbf{Z}\|_F^2+\epsilon\big(\|\mathbf{A}\|_F^2-\log\det(\mathbf{A})\big). \tag{14}$$

For updating transform, given the original data as in equation 15, we can use equation 16.

$$\underset{\mathbf{T}}{\text{minimize}}\ \|\mathbf{TX}\text{-}\mathbf{Z}\|_F^2+\epsilon(\|\mathbf{T}\|_F^2-\log\det(\mathbf{T})). \tag{15}$$

The update of the transform matrix $\boldsymbol{T}$ is straightforward, as each term is directly differentiable. But, there are better ways of solving the update of $\boldsymbol{T}$ by using some linear algebraic tricks [51].

$$\begin{aligned}\mathbf{XX}^{\top}+\epsilon\mathbf{I}&=\mathbf{LL}^{\top}, \\ \mathbf{L}^{-1}\mathbf{XZ}^{\top}&=\mathbf{USV}^{\top}, \\ \mathbf{T}&\leftarrow\frac{1}{2}\mathbf{U}(\mathbf{S}+(\mathbf{S}^2+\epsilon\mathbf{I})^{1/2})\mathbf{V}^{\top}\mathbf{L}^{-1}.\end{aligned} \tag{16}$$

Now, the solution to equation 14 is similar to the update of $\boldsymbol{T}$ in equation 16. Here $\boldsymbol{A}$ acts as transform matrix $\boldsymbol{T}$, and instead of passing original samples $\boldsymbol{X}$, we input kernel matrix $\boldsymbol{K}$.

- **Update J**: After keeping other variables fixed, $\boldsymbol{J}$ can be updated as follows:

$$\begin{aligned} &\underset{\mathbf{J}}{\text{minimize}}\ \tfrac{1}{2}Tr\big(\mathbf{K}-2\mathbf{KJ}+\mathbf{Z}^\top\mathbf{KJ}\big)\\ &+\tfrac{\beta}{2}\|\mathbf{J}\text{-}\mathbf{B}\|_F^2+\tfrac{\mu}{2}\|\mathbf{J}\text{-}\mathbf{Z}+\tfrac{\lambda_1}{\mu}\|_F^2\\ &s.t.\ \mathbf{B}\geq 0, diag(\mathbf{B})=0, \mathbf{B}^\top=\mathbf{B}. \end{aligned} \tag{17}$$

Taking its first derivative and equating it to 0 gives:

$$\mathbf{J}=(\mathbf{K}+\mathbf{B}+\mu\mathbf{I})^{-1}\big(\mathbf{K}+\mu\mathbf{Z}-\lambda_1+\beta\mathbf{B}\big). \tag{18}$$

- **Update G**: After keeping other variables fixed, $\boldsymbol{G}$ can be updated as follows:

$$\underset{\mathbf{G}}{\text{minimize}}\ \alpha\|\mathbf{K}\text{-}\mathbf{G}^\top\mathbf{KH}\|_F^2+\frac{\mu}{2}\|\mathbf{G}\text{-}\mathbf{Z}+\frac{\lambda_2}{\mu}\|_F^2. \tag{19}$$

Taking its first derivative and equating to 0 gives:

$$\mathbf{G}=\big(2\alpha\mathbf{KHH}^\top\mathbf{K}^\top+\mu\mathbf{I}\big)^{-1}\big(2\alpha\mathbf{KHK}^\top+\mu\mathbf{Z}-\lambda_2\big). \tag{20}$$

- **Update H**: After keeping other variables fixed, $\boldsymbol{H}$ can be updated as follows:

$$\underset{\mathbf{H}}{\text{minimize}}\ \alpha\|\mathbf{K}-\mathbf{G}^\top\mathbf{KH}\|_F^2+\frac{\mu}{2}\left\|\mathbf{H}\text{-}\mathbf{Z}+\frac{\lambda_3}{\mu}\right\|_F^2. \tag{21}$$

Taking its first derivative and equating to 0 gives:

$$\mathbf{H}=\big(2\alpha\mathbf{K}^\top\mathbf{GG}^\top\mathbf{K}+\mu\mathbf{I}\big)^{-1}\big(2\alpha\mathbf{K}^\top\,\mathbf{G}\,\mathbf{K}+\mu\mathbf{Z}-\lambda_3\big). \tag{22}$$

- **Update Z**: After keeping other variables fixed, the sub-problem becomes:

$$\underset{\mathbf{Z}}{\text{minimize}}\ \frac{3\mu}{2}\left\|\mathbf{Z}-\frac{\mathbf{J}+\mathbf{G}+\mathbf{H}+\mathbf{AK}+\big(\lambda_1+\lambda_2+\lambda_3\big)/\mu}{3}\right\|_F^2. \tag{23}$$

Taking its first derivative and equating to 0 gives:

$$\mathbf{Z}=\frac{\mathbf{J}+\mathbf{G}+\mathbf{H}+\mathbf{AK}+\big(\lambda_1+\lambda_2+\lambda_3\big)/\mu}{3}. \tag{24}$$

- **Update B**: After keeping other variables fixed, $\boldsymbol{B}$ can be updated as follows:

$$\underset{\mathbf{B}}{\text{minimize}}\ \frac{\beta}{2}\|\mathbf{J}-\mathbf{B}\|_F^2+\gamma\|\mathbf{B}\|_{\boxed{\text{k}}}\ s.t.\ \mathbf{B}\geq 0, diag(\mathbf{B})=0, \mathbf{B}^\top=\mathbf{B}. \tag{25}$$

Using K.Fan theorem [7], equation 25 can be rewritten as follows:

$$\begin{aligned} &\underset{\mathbf{B}}{\text{minimize}}\ \frac{\beta}{2}\|\mathbf{J}-\mathbf{B}\|_F^2+\gamma\langle diag(\mathbf{B})-\mathbf{B},\mathbf{S}\rangle\\ &s.t.\ \mathbf{B}\geq\mathbf{0}, diag(\mathbf{B})=0, \mathbf{B}^\top=\mathbf{B}, \mathbf{0}<\mathbf{S}<\mathbf{I}, Tr(\mathbf{S})=k. \end{aligned} \tag{26}$$

where $\boldsymbol{S}=\boldsymbol{U}\boldsymbol{U}^\top$, $\boldsymbol{U}$ consists of k eigenvectors that correspond to k smallest eigenvalues of $diag(\boldsymbol{B})-\boldsymbol{B}$. Now, equation 26 can be translated to:

$$\underset{\mathbf{B}}{\text{minimize}}\ \frac{1}{2}\left\|\mathbf{B}-J+\frac{\gamma}{\beta}\big(diag(\mathbf{S})1^\top-\mathbf{S}\big)\right\|_F^2. \tag{27}$$

Let us define

$$\begin{aligned} &\boldsymbol{Q}=\boldsymbol{J}-\frac{\gamma}{\beta}\big(diag(\boldsymbol{S})1^\top-\boldsymbol{S}\big),\\ &\tilde{\boldsymbol{Q}}=\boldsymbol{Q}-Diag(diag(\boldsymbol{Q})),\\ &\text{then } \boldsymbol{B}=max(0,(\boldsymbol{Q}+\tilde{\boldsymbol{Q}})/2). \end{aligned}$$

Once we obtain the matrix $\boldsymbol{B}$, the similarity matrix can be computed by $(\boldsymbol{B}+\boldsymbol{B}^\top)/2$. After this, the clustering results can be obtained by applying spectral clustering on the similarity matrix. The step-by-step algorithm is given as algorithm 1 to understand the entire model better.

Algorithm 1 The proposed algorithm: KBD-TSC

Require: $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_N}], \epsilon, \mu, \alpha, \beta, \gamma$, no. of clusters k

Compute kernel matrix $\boldsymbol{ker(X, X)}$.

Initialize random matrix $\boldsymbol{Z, G, A, H, B}$, $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 0$.

while The convergence is not reached **do**

Compute transform matrix $\boldsymbol{A}$ using equation 16.

Update matrix $\boldsymbol{J}$ using equation 18.

Update matrix $\boldsymbol{G}$ using equation 20.

Update matrix $\boldsymbol{H}$ using equation 22.

Update matrix $\boldsymbol{Z}$ using equation 24.

Update matrix $\boldsymbol{B}$ using equation 27.

Update Lagrangian multipliers $\lambda_1, \lambda_2, \lambda_3$.

end while

Compute the similarity matrix by $(\boldsymbol{B} + \boldsymbol{B}^\top)/2$.

Applying spectral clustering on the similarity matrix to obtain resulting clusters.

4 Experimental results and analysis

4.1 Dataset description

The proposed KBD-TSC algorithm is evaluated on nine images datasets which are as follows:

1. Yale[3]: It consists of 165 facial images of 15 individuals in grayscale mode. The images are resized to 32×32 pixels.
2. Jaffe[4]: This dataset consists of 213 facial images corresponding to 7 facial expressions. The images are resized to 26×26 pixels.
3. ORL[5]: It consists of 400 facial images of 40 subjects. Each image is of size 26×26 pixels.
4. ARFaces[6]: This dataset comprises of 4000 facial images of 126 people.
5. COIL20[7]: It consists of 1440 images of 20 objects. Each image size is 32×32 pixels.
6. BA[8]: This dataset contains 1404 images of handwritten digits and uppercase alphabets. Each image size is 20 ×16 pixels.
7. tr11[9]: This is a text dataset consists of 414 samples, 6429 features and 9 classes.
8. tr41[10]: This is a text dataset consists of 878 samples, 7454 features and 10 classes.
9. tr45[11]: This is a text dataset consists of 690 samples, 8261 features and 10 classes.

4.2 Baseline methods

We compare our proposed KBD-TSC method with several state-of-the-art methods, including spectral clustering (SC)

Table 1 Details of parameter values w.r.t. different kernel functions

Kernel	Function	Parameter values
Linear	$k(x_i, x_j) = x_i^\top x_j$	–
Polynomial	$k(x_i, x_j) = \left(x_i^\top x_j + a\right)^b$	$a \in \{0, 1\}, b \in \{2, 4\}$
Gaussian	$k(x_i, x_j) = exp\left(-\left\|x_i - x_j\right\|_2^2 / 2t\sigma^2\right)$	$t \in \{0.01, 0.05, 0.1, \}$ $\{1, 10, 50, 100\}$

[3] http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html.

[4] https://zenodo.org/record/3451524.

[5] https://cam-orl.co.uk/facedatabase.html.

[6] https://www2.ece.ohio-state.edu/~aleix/ARdatabase.html.

[7] https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[8] https://cs.nyu.edu/~roweis/data/.

[9] https://trec.nist.gov/.

[10] https://trec.nist.gov/.

[11] https://trec.nist.gov/.

Table 2 Hyperparameter settings

Hyperparameter	Value
ϵ	$\{1e-2, \mathbf{1e\text{-}1}, 0.5, 1\}$
α	$\{1e-5, 1e-4, 1e-3, \mathbf{1e\text{-}2}, 0.1, 1\}$
β	$\{1e-5, \mathbf{1e\text{-}3}, 0.1, 1\}$
γ	$\{1e-2, \mathbf{1e\text{-}1}, 1, 10, 30, 50\}$

[36], kernelized sparse subspace clustering (KSSC) [40], Kernel low-rank representation (KLRR) [37], Implicit block diagonal low-rank representation (IBDLR) [67], Similarity Learning via Kernel Preserving Embedding sparse (SLKEs) [20], Similarity Learning via Kernel Preserving Embedding low rank (SLKEr) [20], Structure learning with similarity preserving sparse (SLSPs) [19], Structure learning with similarity preserving low-rank (SLSPr) [19] and Kernel block

Table 3 Comparison of clustering results based upon accuracy

Dataset	SC	KSSC	KLRR	SKLEs	SKLEr	SLSPs	SLSPr	IBDLR	KBDSP	KBD-TSC
YALE	0.4425	0.4715	0.4569	0.3404	0.4768	0.3768	0.5286	0.4727	0.5626	**0.5819**
JAFFE	0.5563	0.7691	0.7358	0.6625	0.7409	0.7514	0.9031	0.9081	0.9091	**0.9192**
ORL	0.4948	0.4573	0.4971	0.3696	0.5865	0.3858	0.6075	0.6071	0.6340	**0.6542**
AR	0.2205	0.4116	0.4888	0.2866	0.2479	0.4247	0.5056	0.5010	0.5083	**0.5383**
COIL-20	0.4205	0.5744	0.5776	0.4537	0.5236	0.5341	0.5362	0.6922	0.7119	**0.7318**
BA	0.2785	0.3711	0.3280	0.3533	0.4088	0.4103	0.4128	0.3643	0.4252	**0.4452**
TR11	0.4335	0.4443	0.3364	0.4443	0.4986	0.4016	0.5135	0.5183	0.5234	**0.5537**
TR41	0.4452	0.5179	0.5344	0.3885	0.4588	0.4192	0.5555	0.4964	0.5558	**0.5688**
TR45	0.4561	0.4771	0.4953	0.5594	0.5632	0.5319	0.6025	0.5284	**0.6492**	0.6277
Avg	0.4164	0.4994	0.4945	0.4243	0.5006	0.4706	0.5739	0.5654	0.6088	**0.6245**

Table 4 Comparison of clustering results based upon NMI score

Dataset	SC	KSSC	KLRR	SKLEs	SKLEr	SLSPs	SLSPr	IBDLR	KBDSP	KBD-TSC
YALE	0.4859	0.4761	0.4882	0.3799	0.5047	0.4030	0.5425	0.5122	0.5662	**0.5771**
JAFFE	0.6069	0.7145	0.7414	0.7847	0.7691	0.7656	0.8991	0.8971	0.9178	**0.9191**
ORL	0.6981	0.6094	0.5461	0.5614	0.7580	0.5304	0.7758	0.7696	0.7964	**0.8015**
AR	0.5797	0.6687	0.7131	0.6022	0.5616	0.6750	0.7466	**0.7484**	0.6954	0.7477
COIL-20	0.5381	0.7050	0.6888	0.6047	0.6657	0.6628	0.6649	0.7877	0.7977	**0.8015**
BA	0.4387	0.5249	0.4501	0.5278	0.5361	0.5441	0.5461	0.4997	0.5704	**0.5789**
TR11	0.3142	0.3863	0.2375	0.2831	0.4408	0.2744	0.4576	0.4372	0.4752	**0.4856**
TR41	0.3657	0.4438	0.4642	0.3635	0.3691	0.3170	0.5345	0.4462	0.5453	**0.5644**
TR45	0.3310	0.3909	0.4084	0.4773	0.4995	0.4430	0.5394	0.4570	0.6027	**0.6215**
Avg	0.4843	0.5466	0.5264	0.5094	0.5672	0.5128	0.6341	0.6172	0.6630	**0.6775**

Table 5 Comparison of clustering results based upon purity

Dataset	SC	KSSC	KLRR	SKLEs	SKLEr	SLSPs	SLSPr	IBDLR	KBDSP	KBD-TSC
YALE	0.4665	0.5379	0.5862	0.6323	0.5747	0.6313	0.6138	0.5212	0.6394	**0.6514**
JAFFE	0.5798	0.8389	0.8955	0.8108	0.8187	0.9345	0.9253	**0.9369**	0.9362	0.9368
ORL	0.5387	0.6372	0.6421	0.6302	0.6812	0.7044	0.7013	0.6765	0.7110	**0.7232**
AR	0.2558	0.5090	0.5659	0.3992	0.4545	0.5749	**0.5870**	0.5502	0.5833	**0.5870**
COIL-20	0.4559	0.6751	0.6611	0.7190	0.7420	0.7241	0.6106	0.7442	0.7578	**0.7626**
BA	0.3095	0.4219	0.4383	0.5093	0.5163	0.4573	0.4749	0.5102	0.5386	**0.5512**
TR11	0.5043	0.6417	0.5064	0.6254	0.6860	0.6409	**0.6886**	0.6749	0.6747	0.6881
TR41	0.5672	0.5934	0.6359	0.6472	0.6361	0.6412	**0.6724**	0.5966	0.6501	0.6711
TR45	0.5059	0.6221	0.6781	0.6449	0.6617	0.6989	0.7014	0.6563	0.7060	**0.7218**
Avg	0.4648	0.6086	0.6233	0.6243	0.6412	0.6675	0.6639	0.6519	0.6886	**0.6992**

diagonal representation subspace clustering with similarity preservation (KBDSP) [68].

4.3 Evaluation metrics

During experiments, it is typically presumed that the quantity of clusters is already established. Under these circumstances, various metrics, including accuracy, Normalized Mutual Information (NMI) and Purity, are commonly employed [42, 43]. The metrics are described below:

- Accuracy: The accuracy is defined as the ratio of the number of data instances that are assigned the same cluster as in the ground truth to the total number of data instances.
- Normalized Mutual Information (NMI): This metric computes the normalized measure of similarity between the labels of same data instances. The range of NMI is [0,1] where 0 signifies no correlation and 1 signifies the perfect correlation.
- Purity: Purity measures the extent to which data points within each cluster are assigned to the same true class [69]. A larger purity value indicates better clustering performance.

4.4 Kernel design

We have designed 12 different kernels in this work, including one linear kernel, four polynomial kernels, and seven Gaussian kernels (Table 1).

In the case of the Gaussian kernel, σ is the maximum distance between x_i, x_j.

4.5 Computational complexity

In the proposed algorithm, the first part is the construction of kernel matrix which is bounded by $O(n^2)$. The second part is updating step of different variables, each one of them is bounded by $O(n^3)$. Thus, the proposed algorithm has overall time complexity of $O(tn^3)$ where t and n represents the number of iterations and number of data samples respectively.

4.6 Parameter sensitivity analysis

There are four hyper-parameters in the proposed KBD-TSC algorithm 11, i.e., $\epsilon, \alpha, \beta, \gamma$. The parameter ϵ controls the good conditioning of the transform. The parameter α balances the similarity-preserving term $\left\|\boldsymbol{K} - \boldsymbol{Z}^T\boldsymbol{K}\boldsymbol{Z}\right\|_F^2$, the parameter β is used to control the term $\|\boldsymbol{Z} - \boldsymbol{B}\|_F^2$, the parameter γ is used to control the block-diagonal structure term $\|\boldsymbol{B}\|_{\boxed{\text{k}}}$.The YALE and JAFFE datasets are used for parameter evaluation using NMI. The parameters $\epsilon, \alpha, \beta, \gamma$ take values from the sets $\{1e-2, 1e-1, 0.5, 1\}$, $\{1e-5, 1e-4, 1e-3, 1e-2, 0.1, 1\}$, $\{1e-5, 1e-3, 0.1, 1\}$, and $\{1e-2, 1e-1, 1, 10, 30, 50\}$ respectively. The best values of the parameters for the optimal clustering performance is 0.1, 0.01, 0.001, 0.1 for $\epsilon, \alpha, \beta, \gamma$, respectively. Therefore, we keep $\epsilon = 0.1, \alpha = 0.01, \beta = 0.001, \gamma = 0.1$ in all the experiments of this paper. The parameter tuning is done using grid searching. The parameter settings of all experiments have been given in the Table 2, in which recommended parameters are indicated in bold. We performed parameter sensitivity analysis on JAFFE Dataset for wide range of α, β and γ values against NMI score. It is observed that the proposed technique is not very sensitive to these hyperparameters. Moreover, when these parameters are set to $\alpha = 1e-2, \beta = 1e-3, and \gamma = 1e-1$, the best clustering performance is achieved. Hence, we fixed these values for all the experiments.

4.7 Results and discussion

The experimental results for all the nine datasets are shown in terms of Accuracy, NMI, and Purity in Tables 3, 4, and 5 respectively. We average out the results of all experiments on 12 kernels. The experimental results are reported by averaging the results of ten iterations in the Tables 3, 4 and 5. From the results, it can be observed that the proposed KBD-TSC approach outperforms the state-of-the-art methods.

To be more specific in analysis, the results from Tables 3, 4, 5 are discussed below:

1. Compared to SC algorithm, the proposed KBD-TSC method obtains better results regarding all evaluation metrics: accuracy, NMI, and purity. From Tables 3, 4, 5, it can be easily observed that the average value of accuracy, NMI, and purity of the proposed method are 20.81, 19.32, and 23.44 % higher than SC, respectively. The reason for the same is that the input to spectral clustering is the learned $\boldsymbol{Z}$ instead of the raw kernel matrix.
2. Our proposed KBD-TSC method also outperforms the kernel-based methods KSSC and KLLR. This is because of the similarity-preserving trick in the transform domain.
3. In comparison to SLKEs and SLKEr, our proposed algorithm exhibits superior performance. This improvement can be attributed to two key aspects: firstly, the proposed framework for kernel self-expression has the capability to retain specific low-order details from the input data; secondly, the introduction of the term representing block diagonal structures in our model, within the latent transform space, facilitates the acquisition of a similarity matrix characterized by a block diagonal arrangement.

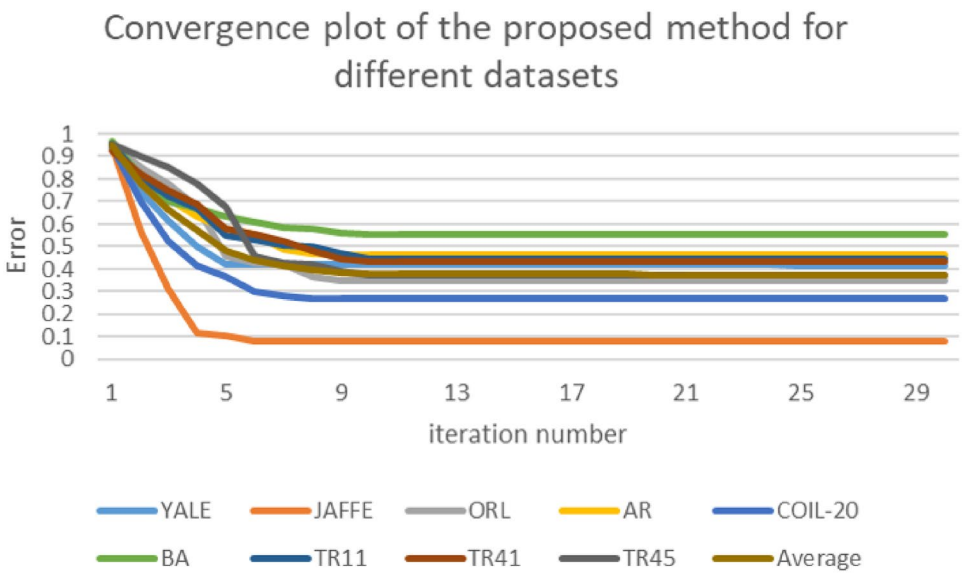


Fig. 1 Convergence graph of the proposed method with nine different datasets in 30 iterations

Table 6 Runtime comparison (in seconds)

Dataset	SSC	KSSC	KLRR	SKLEs	SKLEr	SLSPs	SLSPr	IBDLR	KBD-TSC
YALE	77	88	93	92	94	80	82	79	**68**
JAFFE	62	84	89	87	90	71	74	70	**55**
ORL	86	97	101	99	103	89	93	88	**78**
AR	91	113	119	117	122	95	97	94	**83**
COIL-20	98	117	125	121	128	102	105	100	**88**
BA	77	89	93	90	94	82	85	80	**67**
TR11	110	124	130	125	131	115	118	110	**96**
TR41	121	137	141	139	145	127	131	124	**105**
TR45	117	129	136	133	137	123	128	120	**101**

Table 7 Comparison of NMI score of the proposed objective function against the objective function without similarity preserving term and the objective function without block diagonal term for all the datasets

Dataset	Similarity-	Block-	Proposed
YALE	0.4665	0.5542	**0.5771**
JAFFE	0.8678	0.8832	**0.9191**
ORL	0.7798	0.7814	**0.8015**
AR	0.6872	0.7185	**0.7477**
COIL-20	0.7112	0.7665	**0.8015**
BA	0.4887	0.5443	**0.5789**
TR11	0.4661	0.4727	**0.4856**
TR41	0.5311	0.5392	**0.5644**
TR45	0.5991	0.6003	**0.6215**

4. SLSPs and SLSPr, which are capable of handling nonlinear datasets and preserving similarity information, gives better performance as compared to SC, KSSC, KLRR, SLKEs, and SLKEr. However, the proposed KBD-TSC algorithm consistently outperforms them in most instances. Specifically, average values of accuracy, NMI, and purity in Tables 3, 4, and 5 indicate that the proposed method surpasses SLSPs by 15.39%, 16.47%, and 3.17%, respectively. These findings confirm that the introduced term representing block diagonal structures significantly contributes to improving performance.
5. Both IBDLR and the proposed KBD-TSC mthod facilitate the acquisition of a desired affinity matrix with an optimal block diagonal structure by integrating the block diagonal representation term. Tables 3, 4, and 5 demonstrate that the proposed KBD-TSC method and IBDLR outperform other compared algorithms on all datasets. This underscores the effectiveness of methods incorporating the block diagonal representation term, particularly for datasets with multiple classes. Remarkably, in the case of COIL20 and BA datasets characterized by a larger number of instances, the proposed method demonstrates a performance improvement of almost 15% compared to alternative methods, with the exception of IBDLR on the COIL20 dataset. Additionally, the suggested KBD-TSC method surpasses the performance of

IBDLR specifically on the COIL20 dataset, underscoring the advantages of incorporating a similarity-preserving strategy in the transform domain.

6. In datasets with high-dimensional features such as TR11, TR41, and TR45, SLSPr demonstrates superior performance to IBDLR, attributed to its integration of a similarity-preserving mechanism. Capitalizing on both the similarity-preserving strategy and the block diagonal representation term, the suggested KBD-TSC consistently outshines IBDLR and even surpasses SLSPr in most instances across TR11, TR41, and TR45 datasets. These outcomes underscore the effectiveness of the proposed KBD-TSC method in effectively managing datasets with intricate features, enabling the extraction of inherent data structures.

In a nutshell, the experimental results demonstrate the effectiveness of our proposed KBD-TSC method combined with similarity preserving regularizer, transform learning-based kernel self-expressing model, and block diagonal representation term.

4.8 Convergence analysis

The convergence plots of the proposed method is shown in Fig. 1. For all the datasets, the proposed method converges within 10 iterations.

4.9 Computational time

The experiments are conducted on a 64-bit Windows system with Intel i7 processor and 32GB RAM. The running time of the proposed method and the various state-of-the-art methods for all the datasets are shown in Table 6. From Table 6, it can be observed that the proposed KBD-TSC method is the fastest among all kernel-based techniques.

4.10 Ablation experiments

For the ablation experiments, we have compared the NMI score of the proposed technique against the objective function without the similarity preserving term and the objective function without the block diagonal term. It is observed that these terms are important for making clustering efficient. The NMI score gives the best value for all the datasets when these terms are included in the objective function. The ablation experiment results are discussed in Table 7. The objective function without the similarity preserving term is labelled as "similarity-" and the objective function without the block diagonal term is labelled as "block-."

5 Conclusion

This paper presents a novel subspace clustering approach that integrates transform learning-based kernel block diagonal representation and a similarity-preserving strategy. The method exhibits effective performance even when the raw data lacks inherent separability into subspaces and demonstrates robust generalization capabilities for non-linear manifolds. The proposed KBD-TSC operates through a three-step process. Initially, it captures the non-linear structure of the input data by incorporating the kernel self-expressing framework into the transform learning-based framework. The second step introduces the block diagonal representation term to create a similarity matrix with a block diagonal structure. In the final step, the similarity-preserving term is introduced to capture pairwise similarity information between various data points. The effectiveness of the proposed approach is evaluated on nine benchmark datasets, showcasing its superiority over several state-of-the-art methods. In future work, we aim to extend the proposed method to multiple kernel learning.

References

1. Bai L, Liang J (2020) Sparse subspace clustering with entropy-norm. ICML'20. JMLR.org
2. Chen H, Wang W, Feng X, He R (2018) Discriminative and coherent subspace clustering. Neurocomputing 284:177–186. https://doi.org/10.1016/j.neucom.2018.01.006
3. Chen J, Zhang H, Mao H, Sang Y, Yi Z (2016) Symmetric low-rank representation for subspace clustering. Neurocomputing 173:1192–1202. https://doi.org/10.1016/j.neucom.2015.08.077
4. Chen Y, Gu Y (2017) Active orthogonal matching pursuit for sparse subspace clustering. IEEE Signal Process Lett. https://doi.org/10.1109/LSP.2017.2741509
5. Cheng W, Chow TW, Zhao M (2016) Locality constrained-lp sparse subspace clustering for image clustering. Neurocomputing 205:22–31. https://doi.org/10.1016/j.neucom.2016.04.010
6. Elhamifar E, Vidal R (2013) Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans Pattern Anal Mach Intell 35:2765–2781. https://doi.org/10.1109/TPAMI.2013.57
7. Fan K (1949) On a theorem of Weyl concerning eigenvalues of linear transformations I. Proc Nat Acad Sci 35:652–655. https://doi.org/10.1073/pnas.35.11.652
8. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24:381–395. https://doi.org/10.1145/358669.358692
9. Goel A, Majumdar A (2021) Sparse subspace clustering friendly deep dictionary learning for hyperspectral image classification. IEEE Geosci Remote Sens Lett 19:1–5. https://doi.org/10.1109/LGRS.2021.3112603
10. Goel A, Majumdar A (2024) Sparse subspace clustering incorporated deep convolutional transform learning for hyperspectral band selection. Earth Sci Inf 17:2727–2735. https://doi.org/10.1007/s12145-024-01312-8
11. Gupta P, Goel A, Majumdar A, Chouzenoux E, Chierchia G (2024) Deconfcluster: Deep convolutional transform learning based

multiview clustering fusion framework. Sig Process 224:109597. https://doi.org/10.1016/j.sigpro.2024.109597
12. He R, Wang L, Sun Z, Zhang Y, Li B (2015) Information theoretic subspace clustering. IEEE Trans Neural Netw Learn Syst 27:2643–2655. https://doi.org/10.1109/TNNLS.2015.2500600
13. Ho J, Yang M-H, Lim J, Lee K-C, Kriegman D (2003) Clustering appearances of objects under varying illumination conditions. In: Proceedings of the 2003 IEEE computer society conference on computer vision and pattern recognition, 2003. IEEE, vol 1, p I. https://doi.org/10.1109/CVPR.2003.1211332
14. Hong W, Wright J, Huang K, Ma Y (2006) Multiscale hybrid linear models for lossy image representation. IEEE Trans Image Process 15:3655–3671. https://doi.org/10.1109/TIP.2006.882016
15. Huang Q, Zhang Y, Peng H, Dan T, Weng W, Cai H (2020) Deep subspace clustering to achieve jointly latent feature extraction and discriminative learning. Neurocomputing 404:340–350. https://doi.org/10.1016/j.neucom.2020.04.120
16. Ji P, Reid I, Garg R, Li H, Salzmann M (2017) Adaptive low-rank kernel subspace clustering. arXiv preprint arXiv:1707.04974. https://doi.org/10.48550/arXiv.1707.04974
17. Ji P, Zhang T, Li H, Salzmann M, Reid I (2017) Deep subspace clustering networks. In: Proceedings of the 31st international conference on neural information processing systems NIPS'17. Curran Associates Inc, Red Hook, pp 23–32
18. Ji P, Zhang T, Li H, Salzmann M, Reid I (2017) Deep subspace clustering networks. Adv Neural Inf Process Syst 30:5509–5521
19. Kang Z, Lu X, Lu Y, Peng C, Chen W, Xu Z (2020) Structure learning with similarity preserving. Neural Netw 129:138–148. https://doi.org/10.1016/j.neunet.2020.05.030
20. Kang Z, Lu Y, Su Y, Li C, Xu Z (2019) Similarity learning via kernel preserving embedding. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 4057–4064. https://doi.org/10.1609/aaai.v33i01.33014057
21. Kang Z, Peng C, Cheng Q (2017) Kernel-driven similarity learning. Neurocomputing 267:210–219. https://doi.org/10.1016/j.neucom.2017.06.005
22. Kelkar BA, Rodd SF, Kulkarni UP (2019) Estimating distance threshold for greedy subspace clustering. Expert Syst Appl 135:219–236. https://doi.org/10.1016/j.eswa.2019.06.011
23. Li C-G, You C, Vidal R (2017) Structured sparse subspace clustering: a joint affinity learning and subspace clustering framework. IEEE Trans Image Process 26:2988–3001. https://doi.org/10.1109/TIP.2017.2691557
24. Liao M, Li Y, Gao M (2022) Graph-based adaptive and discriminative subspace learning for face image clustering. Expert Syst Appl 192:116359. https://doi.org/10.1016/j.eswa.2021.116359
25. Lin Z, Chen M, Ma Y (2010) The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055. https://doi.org/10.48550/arXiv.1009.5055
26. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y (2012) Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell 35:171–184. https://doi.org/10.1109/TPAMI.2012.88
27. Liu M, Wang Y, Sun J, Ji Z (2022) Adaptive low-rank kernel block diagonal representation subspace clustering. Appl Intell 52:2301–2316. https://doi.org/10.1007/s10489-021-02396-1
28. Liu Z, Ou W, Zhang K, Xiong H (2024) Robust manifold discriminative distribution adaptation for transfer subspace learning. Expert Syst Appl 238:122117. https://doi.org/10.1016/j.eswa.2023.122117
29. Lu C, Feng J, Lin Z, Mei T, Yan S (2018) Subspace clustering by block diagonal representation. IEEE Trans Pattern Anal Mach Intell 41:487–501. https://doi.org/10.1109/TPAMI.2018.2794348
30. Lu C, Feng J, Lin Z, Mei T, Yan S (2019) Subspace clustering by block diagonal representation. IEEE Trans Pattern Anal Mach Intell 41:487–501. https://doi.org/10.1109/TPAMI.2018.2794348
31. Maggu J, Majumdar A (2016) Alternate formulation for transform learning. In: Proceedings of the tenth Indian conference on computer vision, graphics and image processing, pp 1–8. https://doi.org/10.1145/3009977.3010069
32. Maggu J, Majumdar A (2017) Kernel transform learning. Pattern Recogn Lett 98:117–122. https://doi.org/10.1016/j.patrec.2017.09.002
33. Maggu J, Majumdar A, Chouzenoux E (2018) Transformed locally linear manifold clustering. In: 2018 26th European signal processing conference (EUSIPCO), pp 1057–1061. https://doi.org/10.23919/EUSIPCO.2018.8553061
34. Maggu J, Majumdar A, Chouzenoux E (2021) Transformed subspace clustering. IEEE Trans Knowl Data Eng 33:1796–1801. https://doi.org/10.1109/TKDE.2020.2969354
35. Maggu J, Majumdar A, Chouzenoux E, Chierchia G (2020) Deeply transformed subspace clustering. Signal Process 174:107628. https://doi.org/10.1016/j.sigpro.2020.107628
36. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 14:256
37. Nguyen H, Yang W, Shen F, Sun C (2015) Kernel low-rank representation for face recognition. Neurocomputing 155:32–42. https://doi.org/10.1016/j.neucom.2014.12.051
38. Patel VM, Van Nguyen H, Vidal R (2013) Latent space sparse subspace clustering. In: Proceedings of the IEEE international conference on computer vision, pp 225–232
39. Patel VM, Van Nguyen H, Vidal R (2015) Latent space sparse and low-rank subspace clustering. IEEE J Sel Top Signal Process 9:691–701. https://doi.org/10.1109/JSTSP.2015.2402643
40. Patel VM, Vidal R (2014) Kernel sparse subspace clustering. In: 2014 IEEE international conference on image processing (ICIP), pp 2849–2853. IEEE. https://doi.org/10.1109/ICIP.2014.7025576
41. Paul D, Saha S, Mathew J (2020) Improved subspace clustering algorithm using multi-objective framework and subspace optimization. Expert Syst Appl 158:113487. https://doi.org/10.1016/j.eswa.2020.113487
42. Peng X, Feng J, Zhou JT, Lei Y, Yan S (2020) Deep subspace clustering. IEEE Trans Neural Netw Learn Syst 31:5509–5521. https://doi.org/10.1109/TNNLS.2020.2968848
43. Peng X, Xiao S, Feng J, Yau W-Y, Yi Z (2016) Deep subspace clustering with sparsity prior. In: IJCAI, pp 1925–1931
44. Pham D-S, Budhaditya S, Phung D, Venkatesh S (2012) Improved subspace clustering via exploitation of spatial constraints. In: 2012 IEEE conference on computer vision and pattern recognition, pp 550–557. IEEE. https://doi.org/10.1109/CVPR.2012.6247720
45. Qin Y, Pu N, Wu H (2023) EDMC: efficient multi-view clustering via cluster and instance space learning. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2023.3331197
46. Qin Y, Pu N, Wu H (2023) Elastic multi-view subspace clustering with pairwise and high-order correlations. IEEE Trans Knowl Data Eng
47. Qin Y, Tang Z, Wu H, Feng G (2023) Flexible tensor learning for multi-view clustering with Markov chain. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2023.3305624
48. Ravishankar S, Bresler Y (2013) Closed-form solutions within sparsifying transform learning. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp 5378–5382. IEEE. https://doi.org/10.1109/ICASSP.2013.6638690
49. Ravishankar S, Bresler Y (2013) Learning sparsifying transforms. IEEE Trans Signal Process 61:1072–1086. https://doi.org/10.1109/TSP.2012.2226449
50. Ravishankar S, Bresler Y (2015) Online sparsifying transform learning—Part II. IEEE J Sel Top Signal Process 9:637–646. https://doi.org/10.1109/JSTSP.2015.2407860

51. Ravishankar S, Wen B, Bresler Y (2015) Online sparsifying transform learning—Part I. IEEE J Sel Top Signal Process 9:625–636. https://doi.org/10.1109/JSTSP.2015.2417131
52. Seung HS, Lee DD (2000) The manifold ways of perception. Science 290:2268–2269. https://doi.org/10.1126/science.290.5500.2268
53. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22:888–905. https://doi.org/10.1109/34.868688
54. Soltanolkotabi M, Candés EJ (2012) A geometric analysis of subspace clustering with outliers. Ann Stat 40:2195–2238. https://doi.org/10.1214/12-AOS1034
55. Somandepalli K, Narayanan S (2019) Reinforcing self-expressive representation with constraint propagation for face clustering in movies. In: ICASSP 2019—2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4065–4069. https://doi.org/10.1109/ICASSP.2019.8682314
56. Song J, Yoon G, Hahn K, Yoon SM (2019) Subspace clustering via structure-enforced dictionary learning. Neurocomputing 362:1–10. https://doi.org/10.1016/j.neucom.2019.07.025
57. Tang K, Xu K, Jiang W, Su Z, Sun X, Luo X (2022) Selecting the best part from multiple Laplacian autoencoders for multi-view subspace clustering. IEEE Trans Knowl Data Eng 35:7457–7469. https://doi.org/10.1109/TKDE.2022.3178145
58. Tang K, Xu K, Su Z, Zhang N (2023) Multi-view subspace clustering via consistent and diverse deep latent representations. Inf Sci 651:119719. https://doi.org/10.1016/J.INS.2023.119719
59. Tschannen M, Bölcskei H (2018) Noisy subspace clustering via matching pursuits. IEEE Trans Inf Theory 64:4081–4104. https://doi.org/10.1109/TIT.2018.2812824
60. Vidal R (2011) Subspace clustering. Sig Process Mag IEEE 28:52–68. https://doi.org/10.1109/MSP.2010.939739
61. Vidal R, Favaro P (2014) Low rank subspace clustering (LRSC). Pattern Recogn Lett 43:47–61. https://doi.org/10.1016/j.patrec.2013.08.006
62. Vidal R, Ma Y, Sastry S (2005) Generalized principal component analysis (GPCA). IEEE Trans Pattern Anal Mach Intell 27:1945–1959. https://doi.org/10.1109/TPAMI.2005.244
63. Wang J, Shi D, Cheng D, Zhang Y, Gao J (2016) LRSR: low-rank-sparse representation for subspace clustering. Neurocomputing 214:1026–1037. https://doi.org/10.1016/j.neucom.2016.07.015
64. Waqas M, Tahir MA, Khan SA (2023) Robust bag classification approach for multi-instance learning via subspace fuzzy clustering. Expert Syst Appl 214:119113. https://doi.org/10.1016/j.eswa.2022.119113
65. Xia G, Sun H, Feng L, Zhang G, Liu Y (2017) Human motion segmentation via robust kernel sparse subspace clustering. IEEE Trans Image Process 27:135–150. https://doi.org/10.1109/TIP.2017.2738562
66. Xiao S, Tan M, Xu D, Dong ZY (2015) Robust kernel low-rank representation. IEEE Trans Neural Netw Learn Syst 27:2268–2281. https://doi.org/10.1109/TNNLS.2015.2472284
67. Xie X, Guo X, Liu G, Wang J (2017) Implicit block diagonal low-rank representation. IEEE Trans Image Process 27:477–489. https://doi.org/10.1109/TIP.2017.2764262
68. Yang Y, Li F (2023) Kernel block diagonal representation subspace clustering with similarity preservation. Appl Sci 13:9345. https://doi.org/10.3390/app13169345
69. Yang Z, Oja E (2010) Linear and nonlinear projective nonnegative matrix factorization. IEEE Trans Neural Netw 21:734–749. https://doi.org/10.1109/TNN.2010.2041361
70. Yin M, Liu W, Li M, Jin T, Ji R (2021) Cauchy loss induced block diagonal representation for robust multi-view subspace clustering. Neurocomputing 427:84–95. https://doi.org/10.1016/j.neucom.2020.11.017
71. You C, Robinson D, Vidal R (2016) Scalable sparse subspace clustering by orthogonal matching pursuit, pp 3918–3927. https://doi.org/10.1109/CVPR.2016.425
72. Yu S, Yiquan W (2018) Subspace clustering based on latent low rank representation with frobenius norm minimization. Neurocomputing 275:2479–2489. https://doi.org/10.1016/j.neucom.2017.11.021
73. Zhai H, Zhang H, Zhang L, Li P, Plaza A (2016) A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery. IEEE Geosci Remote Sens Lett 14:43–47. https://doi.org/10.1109/LGRS.2016.2625200
74. Zhang C, Li H, Chen C, Jia X, Chen C (2022) Low-rank tensor regularized views recovery for incomplete multiview clustering. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3232538
75. Zhang C, Li H, Chen C, Qian Y, Zhou X (2020) Enhanced group sparse regularized nonconvex regression for face recognition. IEEE Trans Pattern Anal Mach Intell 44:2438–2452. https://doi.org/10.1109/TPAMI.2020.3033994
76. Zhang C, Li H, Lv W, Huang Z, Gao Y, Chen C (2023) Enhanced tensor low-rank and sparse representation recovery for incomplete multi-view clustering. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 11174–11182. https://doi.org/10.1609/aaai.v37i9.26323
77. Zhang H, Lin Z, Zhang C, Gao J (2014) Robust latent low rank representation for subspace clustering. Neurocomputing 145:369–373. https://doi.org/10.1016/j.neucom.2014.05.022
78. Zhou L, Xiao B, Liu X, Zhou J, Hancock ER et al (2019) Latent distribution preserving deep subspace clustering. In: 28th international joint conference on artificial intelligence. York, pp 4440–4446. https://doi.org/10.24963/ijcai.2019/617

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.

Natural Hazards
https://doi.org/10.1007/s11069-024-06934-1

ORIGINAL PAPER

Machine learning approach for evaluating soil liquefaction probability based on reliability method

Sophia Khatoon[1] · Kishan Kumar[1] · Pijush Samui[1] · Laith Sadik[2] · Sanjay Kumar Shukla[3,4,5]

Received: 19 March 2024 / Accepted: 22 September 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract

Reliability analysis is necessary to address the many uncertainties, including both model and parametric uncertainties. This study systematically assesses the reliability index (β) and probability of occurrence of liquefaction (P_L) using the first-order reliability method (FORM) approach on the cone penetration test (CPT) dataset, taking into account parametric uncertainties. Acknowledging the recent advancements in machine learning models and their ability to capture complex, non-linear relationships and interactions within the data, a deep learning model, namely a deep neural network (DNN), is developed and suggested based on its performance in predicting P_L. We use eight statistical performance metrics to evaluate the DNN model's performance across three distinct dataset split ratios. Additional charts, such as regression plots like Taylor's diagrams, rank analysis, regression error characteristics curves, and loss and epoch curves, is provided to comprehensively assess the DNN model's performance. The current investigation demonstrates that the DNN model is promising for predicting P_L on CPT datasets. Additionally, we conduct a reliability-sensitivity analysis to determine the contribution of each variable in evaluating P_L. According to the sensitivity analysis, the most important parameter is the equivalent clean sand penetration resistance (q_{c1Ncs}). It is followed by the magnitude scaling factor (MSF) and the stress reduction factor (r_d). This study contributes valuable risk assessments for geotechnical engineering design and advocates for the broader integration of FORM-based ML models in liquefaction evaluation.

Keywords Liquefaction · Reliability · Machine learning · Deep neural network · Cone penetration test

List of symbols

CPT	Cone penetration test
β	Reliability index
P_L	Probability of occurrence of liquefaction
FORM	First order reliability method
ML	Machine learning
DNN	Deep neural network
CSR	Cyclic stress ratio

Extended author information available on the last page of the article

CRR	Cyclic resistance ratio
σ_v	Vertical total stress
σ'_v	Vertical total effective stress
a_{max}	Maximum horizontal ground acceleration
q_{c1Ncs}	Equivalent clean sand penetration resistance
MSF	Magnitude scaling factor
r_d	Stress reduction factor
COV	Coefficient of variation
TR	Training dataset
TS	Testing dataset
VS	Validation dataset
R^2	Coefficient of determination
U_{95}	Expanded uncertainty
GPI	Global performance indicator
$RMSE$	Root mean square error
MAE	Mean absolute error
MBE	Mean bias error
MAD	Mean absolute deviation
$WMAPE$	Weighted mean absolute percentage error
AUC	Area under the curve
REC	Regression error characteristic

1 Introduction

Soil liquefaction due to seismic events is one of the significant geotechnical phenomena investigating the intricate mechanisms during seismic events, particularly emphasizing the transient transformation of granular, saturated soils into a quasi-liquid state due to elevated pore water pressure. As per Terzaghi and Peck, soil liquefaction is a phenomenon in soil mechanics in which the saturated soil loses strength quickly (Nova 2013). Youd and Idriss (2001) state that when soil liquefies, it has major impacts on the foundation, leading to significant damage to people and property. Therefore, it is essential to quantify soil liquefaction and its associated risks. One widely used method for liquefaction evaluation is the simplified procedure proposed by Seed and Idriss (1971). This method is based on the concepts of the cyclic stress ratio (CSR), which is the ratio of earthquake-induced cyclic stress to the effective overburden stress, and the cyclic resistance ratio (CRR), which measures the resistance of the soil. Robertson and Wride (1998) explained soil liquefaction conditions to evaluate cyclic resistance loads using the CPT test. Juang et al. (2002) developed a Bayesian mapping function based on simplified procedures using CPT, SPT, and SWVT data. Using the CPT as the in-situ index test, Moss (2003) developed a correlation for assessing the probability of soil liquefaction. Cetin et al. (2004a) proposed probabilistic and deterministic correlations to better evaluate the in-situ cyclic shear stress ratio for calculating the probability of liquefaction. Idriss and Boulanger (2006) evaluated the probability of cyclic softening in fine-grained soils like clay using CPT test data.

Although these simplified procedures provide a clear and distinct methodology in terms of obtaining the CRR parameter, there is no assurance that liquefaction will not occur when the factor of safety (FOS) is greater than one in some situations, as deterministic analysis does not account for uncertainty in the input parameters. Therefore, carrying out

a probabilistic analysis that considers these uncertainties in addition to a deterministic approach is an efficient technique. The literature presents several deterministic and probabilistic techniques for evaluating liquefaction. Several methods, such as the first-order reliability method (FORM), the first-order second moment (FOSM), and Monte Carlo simulation (MCS), among others, calculate the reliability index to estimate the probability of liquefaction. Juang and Chen (1999) used MCS to analyse the reliability index (β), considering uncertainties in the input parameters and using the specified limit state function. Juang et al. (2006) used the FORM method to quantify the reliability index, considering model uncertainty. Low and Tang (2007) used the FORM method, which has been suggested and demonstrated for situations involving correlated non-normal and performance functions that are both explicit and implicit. Jha and Suzuki (2009a) presented a comparative study of reliability approaches, including the FOSM method, advanced FOSM (Hasofer–Lind), point estimation method (PEM), and MCS. The jointly distributed random variables (JDRV) method was used for modelling probability of liquefaction (Johari et al. 2012; Johari and Khodaparast 2013). Vivek and Raychowdhury (2014) performed reliability analysis using an integral approach to estimate P_L of soil using CPT data. A reliability analysis for estimating P_L was conducted by Kasebzadeh et al. (2015) using MCS method. Zou et al. (2017) used FORM with Box–Cox transformation to evaluate probability of liquefaction. Jha et al. (2020) applied the FOSM method and Ghani and Kumari (2022) applied the FORM for evaluating P_L based on the SPT data. Kumar et al. (2023b) and Sett et al. (2023) used the FOSM method to quantify the probability of liquefaction values. Golmoghani Ebrahimi et al. (2024) used an improved hypercube sampling (IHS) method to evaluate liquefaction potential. However, the FOSM method assumes that the input variables follow a normal distribution which may not be valid for all datasets, leading to potential inaccuracies in reliability estimates. In this study, the reliability index is calculated using the FORM approach under the simplified procedure as suggested by Boulanger and Idriss (2014). The advantage of using the FORM is that it requires only a small number of iterations to achieve convergence, making it more computationally efficient compared to MCS. Therefore, FORM approach is considered in this study to evaluate the P_L and β values. Further, these values are used to estimate the reliability sensitivity analysis for different parameters used in the FORM method.

In recent years, machine learning (ML) techniques have advanced significantly, enabling more efficient handling of complex data, reducing processing time, and delivering highly accurate predictions (Kumar et al. 2023a). Machine learning and artificial intelligence (AI) have proven to be effective in solving several engineering research problems. Recent works in the development and applications of ML models in civil engineering include predicting material properties (Asteris et al. 2017) and strength of concrete materials (Asteris and Kolovos 2019) using artificial neural networks (ANN), the development of an optimized Naïve Bayes model for predicting rock-burst failures in underground projects (Ke et al. 2021), predictions of anchors shear breakout capacity in thin concrete members (Tarawneh et al. 2021), slope stability analysis (Sabri et al. 2023), prediction of liquefaction potential index (LPI) (Ghani et al. 2024), and prediction of uniaxial compressive strength in Indian rocks (Sabri et al. 2024). Tarawneh et al. (2023) presented a hybrid data-driven machine learning methodology that combines Bayesian regularization-based ANN and genetic expression programming for determining prestressed concrete losses. Alkayem et al. (2024) presented a detailed and in-depth state-of-the-art review highlighting the superior merits of machine learning models over traditional analytical models, emphasizing their accuracy, generalization abilities, flexible frameworks, and cost-effectiveness. The review underscores the remarkable performance of deep learning models in predicting the

properties of concrete and fiber-reinforced concrete at high temperatures, due to their capability to handle complex nonlinear correlations and challenging regression problems. Various machine learning models, such as ANN (Hanna et al. 2007), support vector machine (SVM) (Samui et al. 2011; Lee and Chern 2013), random forests (RF) (Kohestani et al. 2015), extreme learning machines (ELM) (Zhang et al. 2021), hybrid adaptive neuro-based fuzzy inference system (ANFIS) (Ghani et al. 2022), extreme gradient boosting (XGBoost) (Kumar et al. 2022; Kumar and Kumari 2023), explainable machine learning models such as XGBoost (Jas and Dodagoudar 2023a) and LightGBM (Jas et al. 2024b) with SHapley Additive exPlanations (SHAP) and nature inspired hybrid ANN (Kumari and Ghani 2024) models have been used to evaluate the P_L, which are found to be more efficient compared to statistical methods. Jas and Dodagoudar (2023b) presented a state-of-the-art assessment on the application of machine learning algorithms for estimating the liquefaction potential of soils, spanning the literature from 1994 to 2021. They critically reviewed and summarized numerous ML techniques applied to liquefaction analysis, stressing the limits of conventional methodologies and available ML models. The work underlines the need for a more comprehensive database and the refinement of ML algorithms to improve computational efficiency and reliability in predicting soil liquefaction potential. Furthermore, ML models can handle large datasets and a large number of variables efficiently and demonstrate outstanding performance in evaluating the probability of liquefaction, as evident from the literature. The literature presents several ML models, however, the application of deep learning models in liquefaction evaluation still needs to be improved. While it is true that neural network-based models can have higher computational costs than conventional ML algorithms, several factors must be considered when choosing a model. Since this study involves tabular data and requires the ability to capture complex patterns, the benefits of using neural networks might outweigh the computational cost. Thus, this study proposes a deep learning model, deep neural networks (DNN), to predict the P_L.

Hence, the aim of this study is (a) to evaluate the reliability index and probability of occurrence of liquefaction using the FORM method. (b) To evaluate the reliability sensitivity analysis by using the P_L and β values to determine the input variables that are most sensitive in reliability analysis. (c) To develop a deep neural networks (DNN) model, to predict probability of occurrence of liquefaction on three different dataset split ratio. (d) To evaluate the performance of the DNN model using several statistical parameters and visual plots such as the actual vs. predicted scatter plot, Taylor diagram, rank analysis, regression error characteristic curves, and loss versus epoch curves.

2 Research significance

Accurate predictions of soil liquefaction potential are essential for avoiding geological disasters. The investigation of probabilistic and deterministic methods for assessing liquefaction indicates the necessity for a comprehensive approach that considers many sources of uncertainty in liquefaction studies. Different models give varying factors of safety and/or probabilities of liquefaction for identical input values. Therefore, reliability analysis is important to consider various uncertainties, such as model and parametric uncertainties (Jha and Suzuki 2009a). Furthermore, reliability estimates offer a method for assessing the overall impact of uncertainties and offer a rational framework for selecting suitable safety measures based on the level of uncertainty and the potential implications of failure. Therefore, a reliability assessment of liquefaction potential can be a valuable tool for making

more informed engineering decisions, either as an alternative or in addition to deterministic assessments (Jha and Suzuki 2009b). In this regard, a spreadsheet implementation of FORM using CPT datasets is used to calculate P_L. To determine the most important parameters in the reliability analysis, a reliability sensitivity analysis is conducted. However, it is well known that the problem of liquefaction is highly nonlinear. From literature, it is evident that machine learning models have emerged as a highly valuable tool in modelling complex, nonlinear problems. To this end, a DNN model has been developed and suggested based on its performance to accurately predict P_L. The results suggest that the proposed DNN model is highly generalizable, which broadens its applicability.

3 Data and statistical analysis

The database includes records referring to a varied range of soil and seismic parameters. The collection consists of 473 records from the case histories of worldwide earthquakes, including liquefaction and non-liquefaction events. Table 1 below presents the details of the dataset, soil type, and corresponding references.

The CPT-based field dataset contains six parameters such as vertical total stress (σ_v), vertical total effective stress (σ'_v), a_{max}, which represents the maximum horizontal ground acceleration (as a fraction of gravity) at the ground surface, equivalent clean sand penetration resistance (q_{c1Ncs}), the magnitude scaling factor (*MSF*), and the stress reduction factor (r_d). Reliable data is crucial for developing robust computational models. This requires not just a large number of data points but a dataset that covers the full range of parameter values, ensuring a comprehensive simulation of the problem (Asteris et al. 2021b). This requires exploratory and statistical analysis of the dataset before model training. Table 2 shows the statistical descriptions of the data, including the minimum, maximum, mean, median, and standard deviation (SD) of each input variable for the entire dataset. These statistical values define the limits to which the proposed computational model applies. It is important to note that for parameter values outside these limits, the model cannot provide reliable predictions (Le et al. 2022). Additionally, Fig. 1 depicts a frequency histogram for input variables. These frequency histograms are fundamental in exploratory data analysis because they provide crucial details on the characteristics of the data, including its distribution, range, central tendency, variability, skewness, and outliers. Figure 2 displays the Pearson correlation coefficient heatmap matrix, which provides valuable information about the interrelationships among all parameters. These correlation coefficients are crucial for evaluating the reliability of the input data. A high correlation between certain parameters can indicate redundancy, while a low or no correlation may highlight the independence of variables. Understanding these relationships helps in assessing whether the data adequately captures the variability and interactions necessary for reliable model training. The correlation matrix reveals significant correlations among certain variables. For instance, total stress showed a high correlation with other variables such as effective stress (0.92) and stress reduction factor (− 0.90). The correlation coefficient among variables σ'_v and r_d is also slightly high (− 0.85). Jas and Dodagoudar (2023a) successfully employed these two important input parameters (MSF and r_d) for the first time in ML model development and analysis within a CPT-based framework. This decision was made based on a correlation analysis of 16 input parameters. Also, the results of the sensitivity analysis showed r_d as an important parameter. Thus, r_d is kept as an input parameter in ML model development. Furthermore, in order to avoid redundancy and multicollinearity issues, the total stress

Table 1 Description of datasets from the case histories of earthquake

Earthquake	No. of datasets	Soil type	References
Loma Prieta	186	Alluvial deposits, sand, and clay	Boulanger et al. (1997), Juang et al. (2003)
Kocaeli	89	Clayey and alluvial soils	Youd et al. (2009)
Tangshan	34	Alluvial soils and loess deposits	Moss et al. (2009)
Darfield	25	Loess and alluvial deposits	Green et al. (2014)
Christchurch	25	Alluvial deposits, loess, and volcanic ash	Green et al. (2014)
Hyogoken Nambu	23	Alluvial deposits and volcanic ash	Suzuki et al. (2003)
Edgecumbe	22	Volcanic activity and sedimentary deposits	Bastin et al. (2020)
Superstition Hills	20	Sandy and gravelly soils	Bennett (1984), Cetin et al. (2004b)
Chi-Chi	15	Alluvial deposits, clay, and volcanic ash	Moss (2003), Chu et al. (2006)
Victoria (Mexicali)	9	Sandy and gravelly soils	Bennett (1984)
Tohoku	8	Volcanic ash, sedimentary deposits, and alluvial soils	Cox et al. (2013)
Borah Peak	7	Alluvial deposits and bedrock	Moss (2003)
Imperial Valley	6	Sedimentary deposits and alluvial soils	Bennett (1984)
North Ridge	4	Sedimentary and alluvial soils	Moss (2003)

Table 2 Descriptive statistics of the input parameters

Particulars	a_{max}	q_{c1Ncs} (MPa)	σ'_v (kPa)	σ_v (kPa)	*MSF*	r_d
Mean	0.31	97.90	65.13	91.99	1.05	0.94
Median	0.26	90.60	57.70	83.00	1.04	0.95
Mode	0.25	76.56	50.00	90.30	1.01	0.94
Standard deviation	0.15	33.62	30.35	46.97	0.06	0.04
Kurtosis	0.57	1.35	1.87	1.65	4.63	5.01
Skewness	1.03	1.05	1.24	1.26	1.41	− 1.87
Minimum	0.08	24.37	19.00	11.30	0.79	0.73
Maximum	0.84	216.24	215.20	274.00	1.40	1.01

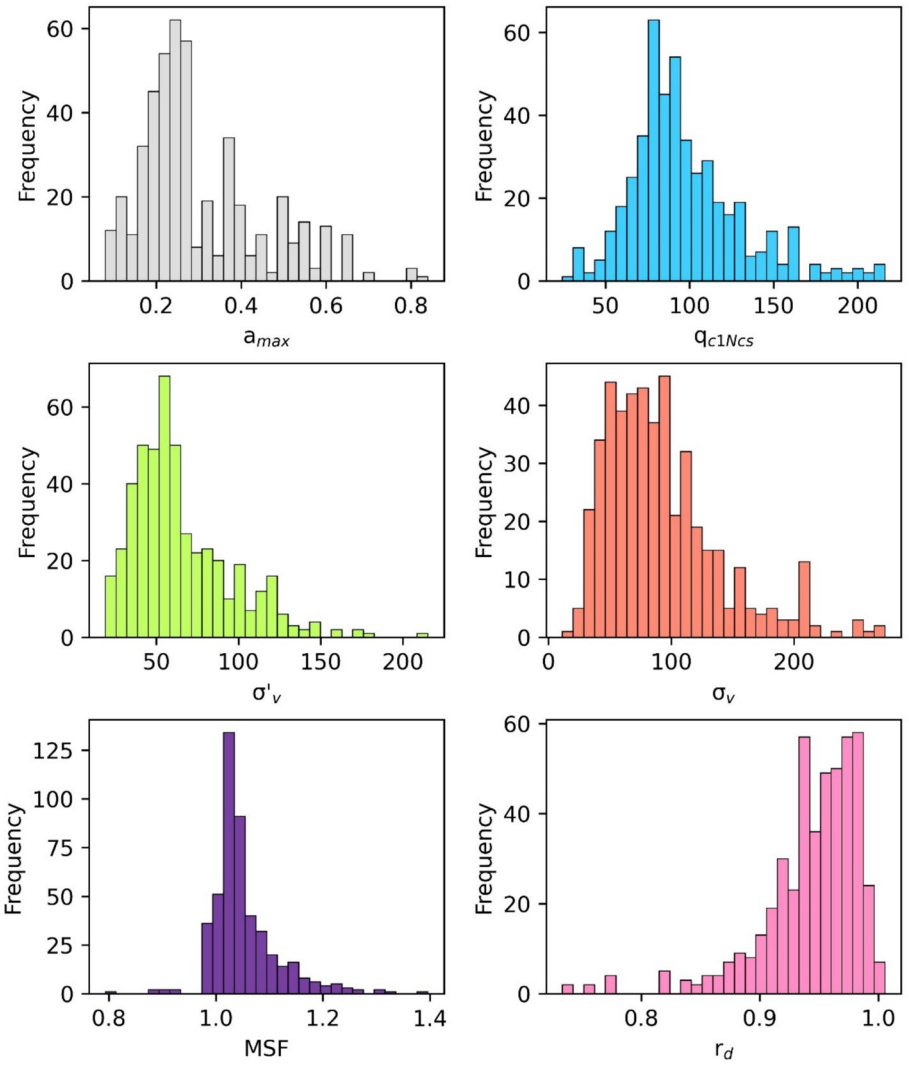


Fig. 1 Frequency histogram of the six input variables

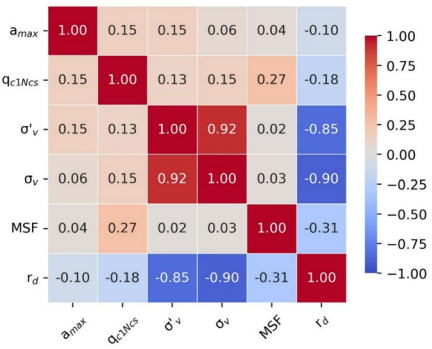


Fig. 2 Correlation heatmap matrix of the six input variables

parameter is removed from input features in the machine learning model. This allowed for the retention of essential information without including both correlated variables, and addressed independence and interdependence among input variables.

4 Methodology

4.1 Deterministic approach

In this paper, the potential for soil liquefaction has been first calculated by using CSR and CRR, employing the method described by Boulanger and Idriss (2014). CSR is calculated as follows:

$$(CSR)_{M=7.5,\sigma'_v=1} = 0.65\left(\frac{\sigma_v}{\sigma'_v}\right)\left(\frac{a_{max}}{g}\right)\frac{r_d}{MSF}\left(\frac{1}{K_\sigma}\right) \tag{1}$$

where the vertical total and effective stresses at depth h are denoted by σ_v and σ'_v, respectively. The stress reduction factor (r_d) that explains the soil profile's dynamic response, the magnitude scaling factor (MSF), and the effective overburden correction factor (K_σ) are shown, respectively. Here, r_d is calculated by Eq. (2).

$$r_d = e^{[\alpha(z)+\beta(z)\bullet M_w]} \tag{2}$$

$$\alpha(\mathrm{h}) = -1.012 - 1.126\sin\left(\frac{\mathrm{h}}{11.73} + 5.133\right) \tag{3}$$

$$\beta(\mathrm{h}) = 0.106 + 0.118\sin\left(\frac{h}{11.28} + 5.142\right) \tag{4}$$

where h is the depth below the surface of the ground in meters. The MSF parameter, which includes a dependence on soil parameters, is calculated as follows:

$$MSF = 1 + \left(MSF_{\max} - 1\right)\left[8.64\exp\left(\frac{-M_w}{4}\right) - 1.325\right] \tag{5}$$

where MSF_{max} is related to q_{c1Ncs} values, and it is calculated as follows:

$$MSF_{\max} = 1.09 + \left(\frac{q_{c1Ncs}}{180}\right)^3 \tag{6}$$

The K_σ, and the coefficient C_σ are calculated as follows:

$$K_\sigma = 1 - C_\sigma \ln\left(\frac{\sigma_v'}{P_a}\right) \tag{7}$$

$$C_\sigma = \frac{1}{37.3 - 8.27\left(q_{c1Ncs}\right)^{0.264}} \tag{8}$$

where P_a is air pressure (101.3 kPa, 1 atm). For $M_w = 7.5$ and $\sigma'_v = 1$ atm (101.3 kPa), *CRR* is determined in terms of q_{c1Ncs} as below:

$$CRR_{M=7.5,\sigma'_v=1} = exp\left(\frac{qc1Ncs}{113} + \left(\frac{qc1Ncs}{1000}\right)^2 - \left(\frac{qc1Ncs}{140}\right)^3 + \left(\frac{qc1Ncs}{137}\right)^4 - C_o\right) \tag{9}$$

where q_{c1Ncs} denotes the equivalent clean sand penetration resistance, and C_o represents the unknown fitting parameter. For this Eq. (9), consider $C_o = 2.8$. The fine content (FC) of the soil affects the relationship between CRR and q_{c1Ncs} in cohesionless soils as well. The relationship is expressed mathematically by the following expressions:

$$q_{c1Ncs} = q_{c1N} + \Delta q_{c1N} \tag{10}$$

$$\Delta q_{c1N} = \left(11.9 + \frac{q_{c1N}}{14.6}\right) exp\left[1.63 - \frac{9.7}{FC + 2} - \left(\frac{15.7}{FC + 2}\right)^2\right] \tag{11}$$

The relationship used for estimating FC is as below:

$$FC = 80\left(I_c + C_{FC}\right) - 137 \tag{12}$$

A soil behaviour type index (I_c) that depends on sleeve friction (f_s) and q_c is frequently connected with the FC and soil categorization. The I_c term is calculated by Eq. (13) Robertson and Wride (1998).

$$I_c = \left\{\left[3.47 - \log(Q)\right]^2 + \left[1.22 + \log(F)\right]^2\right\}^{0.5} \tag{13}$$

where Q is the normalized tip ratio and F is the sleeve friction ratio expressed as Eqs. (14) and (15), respectively.

$$Q = \left(\frac{q_c - \sigma_{vc}}{P_a}\right)\left(\frac{P_a}{\sigma_{vc}'}\right)^n \tag{14}$$

$$F = \left(\frac{f_s}{q_c - \sigma_{vc}}\right) \bullet 100\% \tag{15}$$

The CPT penetration resistance for overburden stress effects (q_{c1N}) is calculated using Eq. (16).

$$q_{c1N} = C_N \frac{q_c}{P_a} \tag{16}$$

The overburden correction factor (C_N) relationship by Boulanger (2003) based on cone penetration numerical modelling and calibration test data is calculated using the following expression:

$$C_N = \left(\frac{P_a}{\sigma_v'}\right)^m \tag{17}$$

where the exponent m is calculated as follows:

$$m = 1.338 - 0.249\left(q_{c1Ncs}\right)^{0.264} \tag{18}$$

4.2 Reliability analysis using FORM

The first-order reliability method (FORM) is a widely used engineering and probabilistic analytical technique for estimating the reliability index and the probability of the occurrence of liquefaction in complex systems. FORM is a closed-form method for reliability estimation where the reliability level is estimated as the minimum distance between the limit state surface and the point that corresponds to the mean values of all design parameters (Saleh et al. 2022). FORM reliability analysis approximates the system's performance limit state function using a first-order Taylor series expansion. The limit state function is defined as:

$$g(x) = CRR - CSR \tag{19}$$

If $g(x) < 0$, the performance function fails and liquefaction occurs; if $g(x) > 0$, liquefaction will not occur; and if $g(x) = 0$, the performance function is limiting. Using a CPT-based dataset, six parameters are used in the first-order reliability method to determine P_L. The parameter in this analysis is a lognormal distribution, which is the precise distribution of the variables in liquefaction (Hsein et al. 2012). The range of coefficient of variation (COV) and COV values used in this study is given in Table 3. Based on this range, the COV of the parameters has been extracted and used in this study for reliability analysis.

Figure 3 depicts the flowchart of the FORM methodology used in this study. The FORM method uses the Microsoft Excel sheet, where different equations are used by the brief program code EqvN (made within Microsoft Excel). The EqvN function is

Table 3 COV of parameters

Parameters	COV range	Reference	COV (this study)
a_{max}	0.1–0.2	Juang and Chen (1999)	0.1
q_{c1Ncs}	0.2–0.6	Phoon and Kulhawy (1999)	0.2
σ_v	0.05–0.2	Juang and Chen (1999)	0.1
σ'_v	0.05–0.2	Juang and Chen (1999)	0.15
r_d	0.05–0.1	Juang and Chen (1999)	0.05
MSF	0.05–0.15	Jha and Suzuki (2009b)	0.05

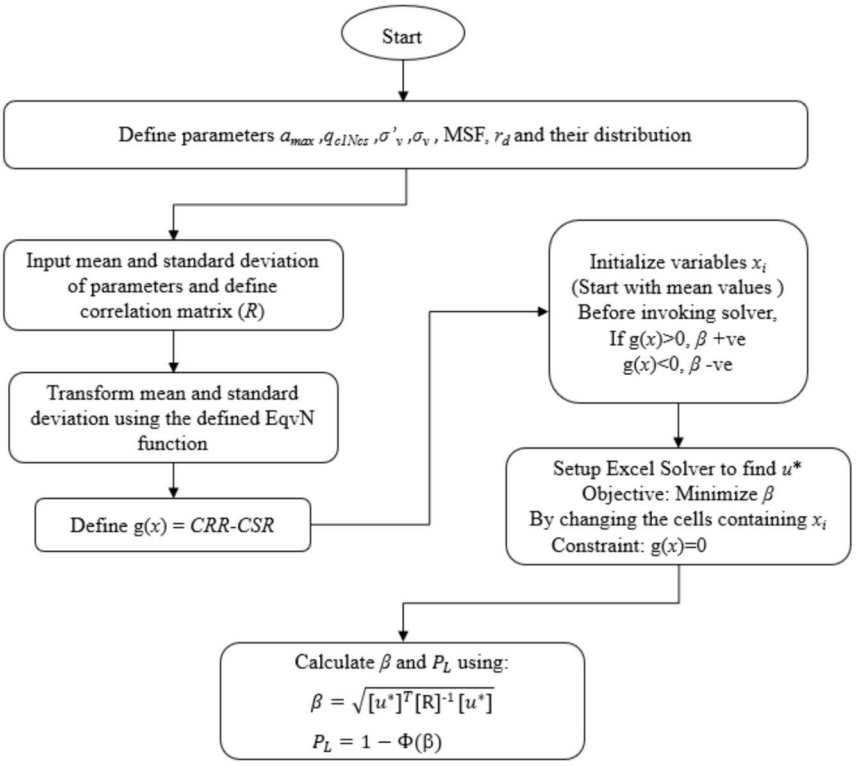


Fig. 3 Flow chart of the first-order reliability method as per Low and Tang (2007)

used to transform the mean and standard deviation of the lognormal distribution. Equations (20) to (22) define the lognormal transformation defined by the EqvN function.

$$\mu_i = exp\left(\lambda + \left(\frac{\sigma_i^2}{2}\right)\right) \tag{20}$$

$$\sigma_i = \sqrt{\log\left(1 + \left(\frac{\sigma}{\mu}\right)^2\right)} \tag{21}$$

$$\lambda = \log(\mu) - 0.5\log\left(1 + \left(\frac{\sigma}{\mu}\right)^2\right) \tag{22}$$

Here μ is the mean, σ is the standard deviation of the lognormal distribution, μ_i is the equivalent normal mean, and σ_i is the equivalent normal standard deviation respectively. The reliability index is estimated using Eq. (23) below:

$$\beta = \min_{x \in F} \sqrt{\left[\frac{x_i - \mu_i}{\sigma_i}\right]^T [R]^{-1} \left[\frac{x_i - \mu_i}{\sigma_i}\right]} \tag{23}$$

where x is the vector representing the set of random variables x_i, R is the correlation matrix. On the other hand, the P_L is determined with the following mathematical expression presented in Eq. (24) below:

$$P_L = 1 - \Phi(\beta) \tag{24}$$

where P_L is assumed to be the standard normal cumulative distribution function (Φ) of the reliability index.

4.3 Sensitivity analysis

Sensitivity analysis (SA) is a technique used to measure and evaluate the level of uncertainty in mathematical and computational models. The goal of sensitivity analysis is to identify the key input variables of a model and accurately measure the extent to which uncertainties in these inputs affect the model's outcome(s). Sensitivity studies may be carried out using various methods, which can be categorized into gradient, variance, and regression-based methods (Gan and Duan 2019). Gradient-based approaches calculate the sensitivity analysis by analysing how the output gradient changes in response to variations in the input parameter. These methods include the one-at-a-time (OAT) method (Daniel 1973) and the Morris one-at-a-time (MOAT) method (Morris 1991). Variance-based methods include quantitatively decomposing the variance of a model's output as the contributions from each parameter and their interactions. Although computationally costly, they are accurate and independent of models. Sobol's sensitivity indices (Sobol′ 2001), the Fourier amplitude sensitivity test (FAST) (Cukier et al. 1973), and the Extended version (eFAST) (Saltelli et al. 1999) are some examples of variance-based methods. Regression-based methods are computationally less costly as they use surrogate models to approximate the results of sensitivity analysis. Regression-based methods include linear regression (Galton 1886), multivariate adaptive regression splines (MARS) (Friedman 1991), and Gaussian process (GP) (MacKay 1998) models.

The selection of an appropriate sensitivity analysis method is based upon multiple factors, including (1) the number of variables being investigated; (2) the computational costs linked to the model and the SA method; (3) the ability to handle nonlinear and non-monotonic relationships between variables and output; and (4) the ability to take variable relationships into account. It should be noted that different analysis may yield different answers to this problem (Iman and Helton 1988; Hamby 1995). In this study, using a gradient-based approach, reliability sensitivity analysis is calculated to find the rate of change in the output due to the changes in the input parameters. This method given by Xiaoping (2005) uses computed design points and reliability index in the determination of sensitivity of input variables used in the FORM approach. The sensitivity of the mean of the input parameters is calculated using Eq. (25).

$$s_{\mu_i} = \Phi(-\beta)\frac{u_i^*}{\beta\sigma} \tag{25}$$

where $\Phi(-\beta)$ = probability of failure, u_i^* = design points, β = reliability index, and σ = standard deviation. In this study, a constant 10% change in standard deviation has been

considered to calculate sensitivity values using Eq. (25) for each of the six parameters for 473 datasets. The average of the sensitivity values is then calculated, and the results are compared with previous studies.

4.4 Description of machine learning

A key aspect of machine learning is the selection of suitable ML models, which is greatly affected by the data's features and the desired output. Time-series data requires models that can accurately capture sequential patterns. For example, shear stress and pore water pressure ratio time history data for soils (Jas et al. 2024a). RNN-based models such as long short-term memory (LSTM), bidirectional LSTM (BiLSTM), and gated recurrent unit (GRU) can handle this type of data efficiently (Pirani et al. 2022). CNNs are well-suited for handling spatial data, such as images or grid-like data. They excel at tasks like pattern identification and feature extraction in spatially organized data (Hacıefendioğlu et al. 2021). Tabular data, which is structured in tables with attributes and categories, requires models that can efficiently handle intricate, non-sequential connections. Neural networks and ensemble models like XGBoost and RF are appropriate for tabular data, as stated by Borisov et al. (2024). Furthermore, the choice between a classification or regression model is contingent upon the characteristics of the desired output. The CPT datasets utilized in this study are organized tabular data that is non-sequential in nature. A suitable choice for this type of dataset, where the relationships between features are complex and not necessarily spatial or sequential, is the DNN model. In this context, there is a popular theorem referred to as 'no free lunch', meaning no one algorithm works best for all problems. The description of the DNN model and its parameters have been discussed in the subsections that follow. Additionally, the flowchart of the methodology used in this study is presented in Fig. 4 below.

4.4.1 Data normalization and statistical distribution analysis

In data analysis, statistical normalization is essential for aligning variables for equitable comparisons across diverse datasets. It modifies numerical distributions, frequently bringing scales together or adhering to statistical norms through the use of techniques like

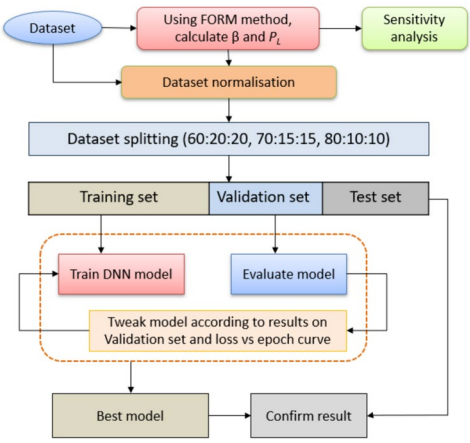


Fig. 4 Illustration of DNN model implementation and validation

min–max scaling. This procedure leads to more dependable statistical results by improving interpretability, reducing the impact of outliers, and guaranteeing that each variable contributes proportionately to the study (Lee and Chern 2013; Asteris et al. 2021b). Data normalization is defined as:

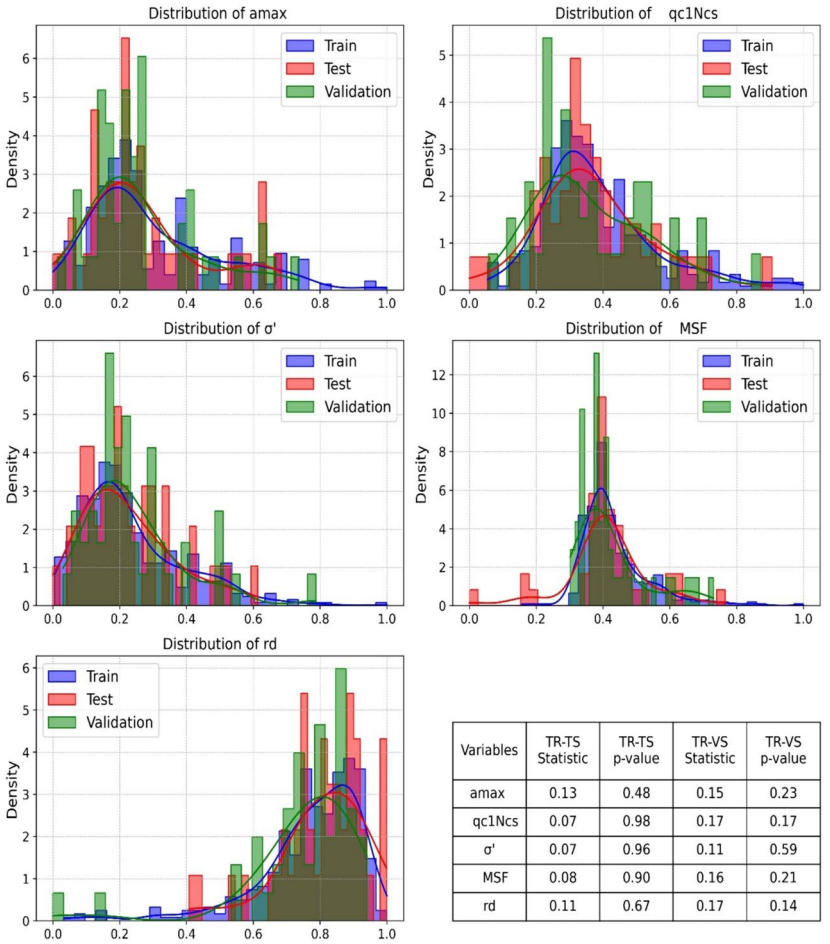
$$X_{\text{norm}} = \frac{\mathrm{X}_{\text{act}} - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{26}$$

where X_{max} and X_{min} represent the parameter's highest and lowest values, respectively. The X_{act} and X_{norm} stand for the actual and normalized values of the parameters, respectively. After normalization, the complete dataset is randomly divided into three subgroups: training, testing, and validation. The training, testing, and validation datasets are indicated, respectively, by the symbols TR, TS, and VS. In this study, three split ratios of the dataset have been carried out. These include 60:20:20, 70:15:15, and 80:10:10 between the training, testing, and validation datasets, respectively. The training set is used for model training, the validation set is used to validate the model during training, and the testing set is used to evaluate model performance after model training. The test set helps us ensure that the trained and validated model performs well and generalizes to new, unseen data. This unbiased assessment is essential because it helps determine if the model is overfitting. The most important difference between testing and validation sets is that we use the test data only once, at the very end of the model development process.

After data normalization and data splitting, the distribution of all three sets was analysed using density plots. Assessing and quantifying the distribution of these sets before model training is crucial to ensuring that the datasets across all sets are comparable and representative. This is essential because significantly different distributions of the training, testing, and validation sets can lead to biased estimates and poor generalization performance of the ML model. For this analysis, we used the Kolmogorov–Smirnov (KS) test (Massey Jr. 1951), which is a non-parametric statistical test that evaluates whether two sample distributions are significantly different. The KS test compares the empirical cumulative distribution functions of the datasets and provides a statistic and p value that quantify the extent of any differences. A high p value (greater than the level of significance of 0.05) suggests that there is no significant difference between the distributions. Additionally, density plots for each variable to visually assess the distribution of the data are provided. These plots use histograms combined with kernel density estimation (KDE). KDE is a non-parametric method to estimate the probability density function of a random variable based on kernels as weights. KDE plots are particularly useful as they provide a continuous approximation of the data distribution, facilitating a more intuitive understanding of the dataset's characteristics across different sets. The results of the KS test and density plots for the split ratio of 80:10:10, shown in Fig. 5 below, suggest that the distribution of the TR, TS, and VS sets is statistically similar. The density plots for the other two split ratios are provided in the supplementary appendix (Fig. SA1 and Fig. SA2).

4.4.2 Deep neural network (DNN)

DNN is a type of neural network in which the input and output layers are separated by several hidden layers. DNN models are being used more often because of their improved ability to understand nonlinear input–output mapping and the fundamental structure of input data vectors (Sze et al. 2017). These are built into a complicated network with



Variables	TR-TS Statistic	TR-TS p-value	TR-VS Statistic	TR-VS p-value
amax	0.13	0.48	0.15	0.23
qc1Ncs	0.07	0.98	0.17	0.17
σ'	0.07	0.96	0.11	0.59
MSF	0.08	0.90	0.16	0.21
rd	0.11	0.67	0.17	0.14

Fig. 5 Density plot of the normalized data distribution for training, testing, and validation datasets (80:10:10 split ratio) of five input variables along with the results of the KS test

some hidden layers when they receive information and transmit it to the other neurons in their network. The basic building blocks of this include neural networks, an activation function, input/output, weights, layers, and an optimizer (learning technique). In neural networks, weights are used to characterize the correlation between neurons in subsequent layers. These weights regulate how an input influences the subsequent neuron's output. Initially, these weights are assigned at random. However, to help the network learn and make accurate predictions of the intended output, they are progressively changed during the iterative process of model training. The weights are updated step-by-step using an optimizer algorithm called Adam, which converts the weights' original random values into more appropriate weights that improve the network's prediction power. Figure 6 shows the DNN architecture, where X_1, X_2, and X_k are input variables, H_{11} up to H_{mn} are hidden layers, and O represents the output layers.

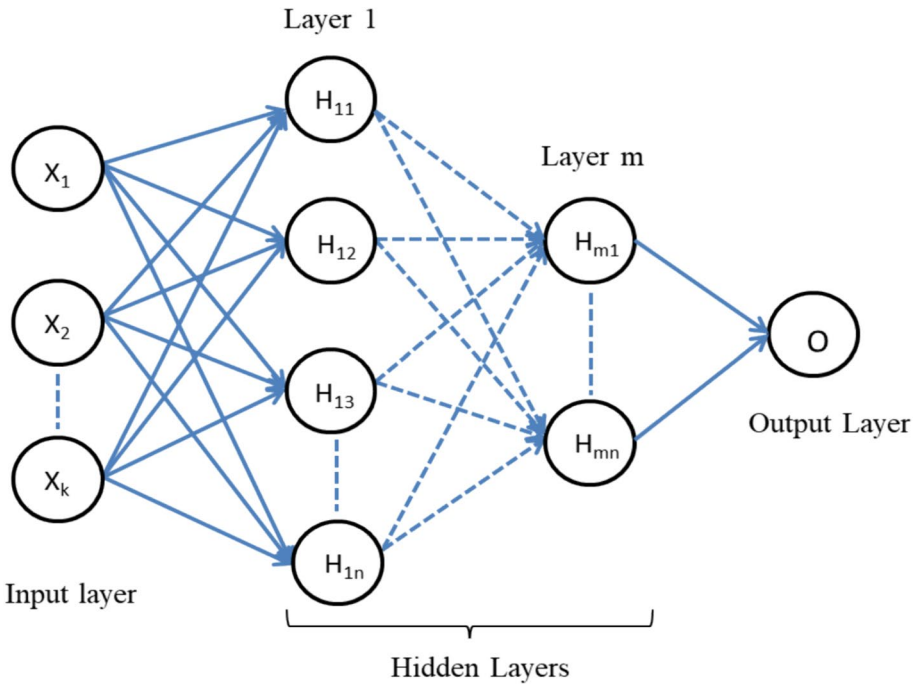


Fig. 6 Deep neural network (DNN) architecture

4.4.3 Model building

In neural network models, several key parameters influence performance and training efficiency. Thus, careful consideration must be given to each of these parameters. Hidden layers are crucial for capturing complex patterns and abstractions from the input data; their number and size affect the model's capacity. The number of neurons in each layer affects the model's learning capacity; while more neurons can improve performance, they also increase the risk of overfitting. The batch size determines how often the model's weights are updated. Smaller batch sizes can make the model more accurate but slower to train, while larger batch sizes can speed up training but may lead to less accurate models. Larger batch sizes speed up training but may reduce model accuracy, while smaller batch sizes enhance accuracy. The optimizer updates model weights to minimize the loss function, with popular alternatives like stochastic gradient descent (SGD) and adaptive moment estimation (Adam) providing trade-offs in learning speed and stability. The loss function, such as mean squared error (MSE), provides a quantitative metric for the accuracy of the model's predictions, which is used to guide the optimizer in adjusting weights and biases during the model's training process. The learning rate controls the magnitude of weight updates, influencing convergence speed and stability. Activation functions are fundamental elements of the neural network architecture. They introduce nonlinearity into the network, allowing the model to learn more complex interactions. Finally, the output layer delivers the final predictions, which are configured based on the task, such as classification or regression. The model parameter of the DNN model is described in Table 4.

Table 4 Model parameters

Model parameters	DNN
Hidden layers	4
No. of neurons	(48,32,16,16)
Batch size	16
Optimizer	Adam
Loss	MSE
Learning rate	0.001
Activation functions	
(1) Hidden layers	Relu
(2) Output layer	Sigmoid

The relationship between epochs and loss in machine learning is a crucial aspect of the training process. Each epoch represents a single run throughout the training dataset. Due to the model's repeated random setting of parameters, the loss is often more significant at the start of training or in the first epoch. As training progresses through multiple epochs, the loss tends to decrease. This loss reduction signifies that the model is improving its ability to make predictions that align more closely with the actual target values. The ultimate goal is for the loss to decrease gradually with each epoch until it converges to a minimum. However, it's essential to be cautious because overtraining, or overfitting, can occur if training continues for too many epochs (Kumar et al. 2024). Overfitting in machine learning models occurs when a model learns the details and noise in the training data to such an extent that it negatively impacts the model's performance on new data. Essentially, an overfitted model performs very well on the training data but poorly on the testing or validation data (Armaghani and Asteris 2021; Asteris et al. 2021a). A regularization strategy called 'early stopping' stops training if it observes no appreciable improvement over several successive epochs. This is done by assigning a 'patience' parameter. If the metric does not improve for a number of epochs equal to the patience value, training is stopped. In this study, a patience value equal to ten has been considered. This involves monitoring the model's performance on a validation set and stopping training when performance stops improving.

4.4.4 Performance evaluation

The last and very crucial step in machine learning modelling is performance evaluation. Statistical indices are commonly used to assess the model's performance. In this study, three of the eight statistical parameters—coefficient of determination (R^2), expanded uncertainty (U_{95}), and global performance indicator (*GPI*)—are used to calculate the trend value, while root mean square error (*RMSE*), mean absolute error (*MAE*), mean bias error (*MBE*), mean absolute deviation (*MAD*), and weighted mean absolute percentage error (*WMAPE*) are used to calculate the error values of the suggested models (Chai and Draxler 2014; Kumar and Samui 2022; Bardhan et al. 2023). Other than these statistical metrics, several other model assessment techniques, such as the actual vs. predicted plot, rank analysis, Taylor diagram, error matrix, REC curves, loss vs. epoch curves, residual plots, Williams plot, etc., can be used in the context of regression analysis (Kumar and Samui 2024). The performance evaluation of ML models should be comprehensive and not restricted to a few statistical indices, as each index and plot has its own advantages and disadvantages. For instance, R^2, which is a widely recognized and used metric to measure goodness of fit, is highly sensitive to outliers and does

not provide information regarding the bias of coefficient predictions. Similarly, *MBE* indicates the average bias in the predictions but does not provide information regarding accuracy. Each metric provides different insights into model performance, and selecting the appropriate metrics depends on the specific goals of the analysis and the nature of data. Combining multiple metrics can offer a comprehensive evaluation of model performance. Table 5 below shows the ideal value of the statistical indices. A value that is closer to the ideal value indicates a better model.

These parameters are described from Eqs. (27) to (35) as follows:

$$R^2 = \frac{\sum_{i=1}^{n}\left(d_i - d_{mean}\right)^2 - \sum_{i=1}^{n}\left(d_i - y_i\right)^2}{\sum_{i=1}^{n}\left(d_i - d_{mean}\right)^2} \tag{27}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(d_i - y_i\right)^2} \tag{28}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|\left(y_i - d_i\right)\right| \tag{29}$$

$$MBE = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - d_i\right) \tag{30}$$

$$MAD = \frac{1}{N}\sum_{i=1}^{n}\left|d_i - d_{mean}\right| \tag{31}$$

$$WMAPE = \frac{\sum_{i=1}^{n}\left|\frac{d_i - y_i}{d_i}\right| \times d_i}{\sum_{i=1}^{n} d_i} \tag{32}$$

$$U_{95} = 1.96\left(SD^2 + RMSE^2\right)^{\frac{1}{2}} \tag{33}$$

$$GPI = MBE \times RMSE \times U_{95} \times t_{stat} \times \left(1 - R^2\right) \tag{34}$$

Table 5 Ideal value of statistical indices

Indices	Ideal value	Indices	Ideal value
R^2	1	*MAD*	0
RMSE	0	*WMAPE*	0
MAE	0	U_{95}	0
MBE	0	*GPI*	1

$$t_{stat} = \sqrt{\frac{(N-1)MBE^2}{RMSE^2 - MBE^2}} \tag{35}$$

where d_i = observed *ith* value, d_{mean} = average of the observed values, y_i = predicted *ith* value, N = the number of sample data points, SD = standard deviation, and t_{stat} = t-statistic, which is calculated as per Eq. (35).

5 Results

5.1 Probability of failure and sensitivity analysis

The results of the P_L and β calculations using the FORM approach are displayed in Fig. 7 which depicts that the value of reliability index varies from − 8 to 6. The probability of occurrence of liquefaction and the reliability index value obtained for 473 CPT datasets are further used in the sensitivity analysis and development of the DNN model.

The sensitivity analysis results for different parameters are displayed in Fig. 8. According to sensitivity analysis results, the sequence of sensitive variables is q_{c1Ncs}, MSF, r_d, σ_v, σ'_v, and a_{max}. For regression models, Muduli and Das (2014) concluded in their findings that q_c and a_{max} are the most important parameters. Similar observations were made by Sebaaly and Rahhal (2019), and Alzahamie and Abdul-Husain (2021) where q_c had been established as the most sensitive parameter.

5.2 Actual versus predicted plot

A scatter plot shows the result of actual versus predicted values for both the training and testing datasets for the ML model. This provides an intuitive visual assessment of the model's performance and can be used in conjunction with other diagnostic tools and metrics to evaluate and improve the regression model. Figures 9, 10, and 11 display the scatter plot separately for the TR and TS datasets for each data split. Table 6 shows the summary of the DNN model's performance using eight statistical parameters for different split ratios.

5.3 Rank analysis

The ML models may not perform equally on training, testing and validation datasets. For comparing models based on overall performance of ML models across all statistical parameters, simple rank analysis is a useful tool. A simple rank-analysis approach is employed to

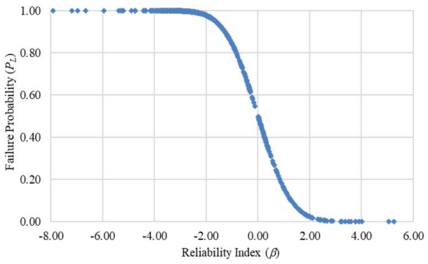


Fig. 7 Plot between P_L and β for 473 data obtained from reliability analysis considering parametric uncertainty

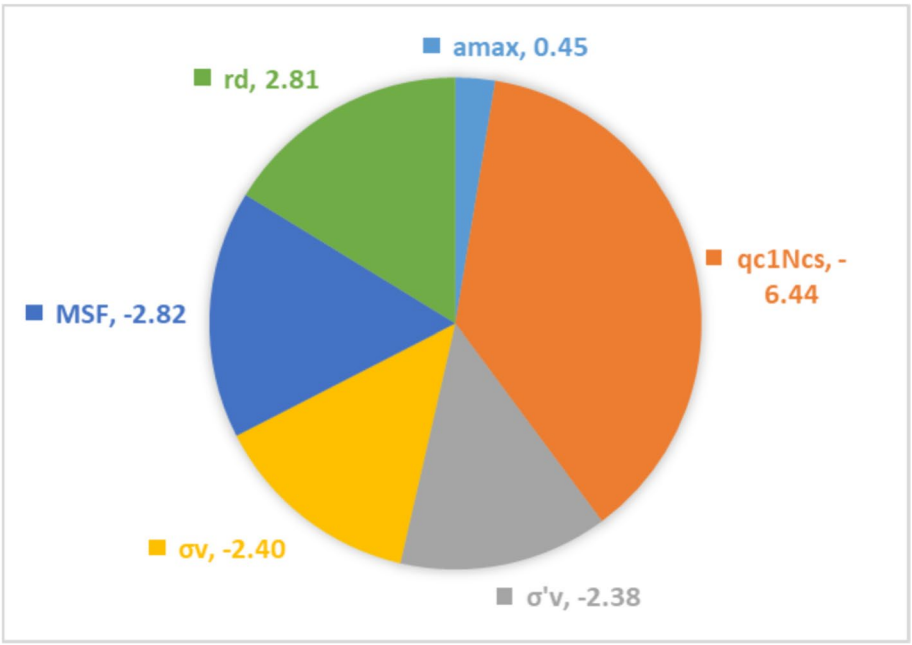


Fig. 8 Comparison of reliability sensitivity analysis of six input variables

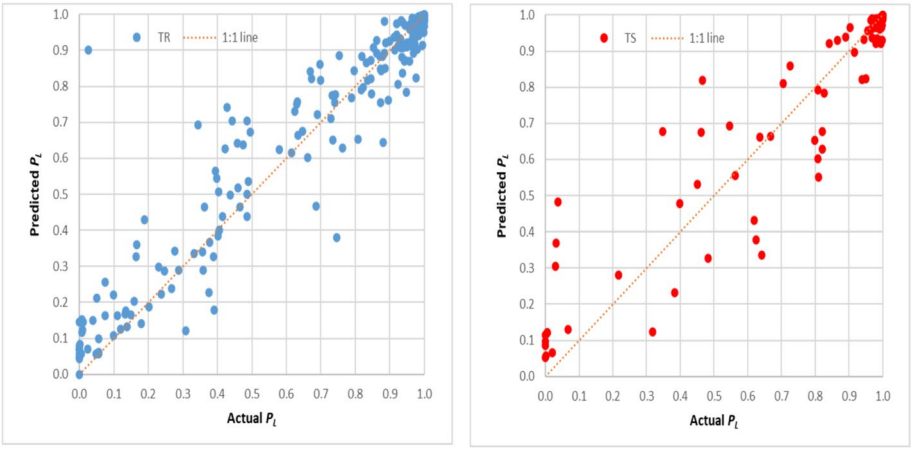


Fig. 9 Scatter plot for training and testing datasets for a data split ratio of 60:20:20

compare the performance of the DNN model across three data splits. The score value of each prediction model has been determined separately for the training, testing, and validation datasets. During the TR, TS, and VS datasets of this study, the datasets that exhibit the lowest error values are ranked as 1, those that exhibit the greatest error values are ranked as 2, 3, and so on, and the split ratio for which the DNN model demonstrates the best accuracy is ranked as 1. The rank of each dataset is then added to determine the overall

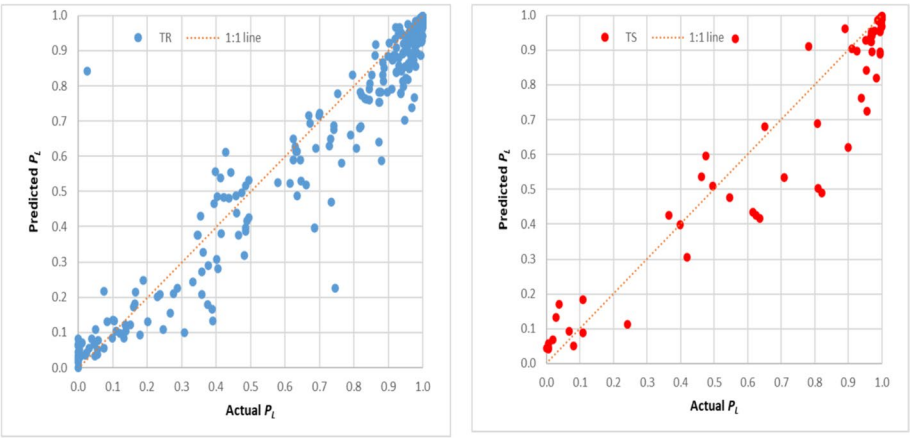


Fig. 10 Scatter plot for training and testing datasets for a data split ratio of 70:15:15

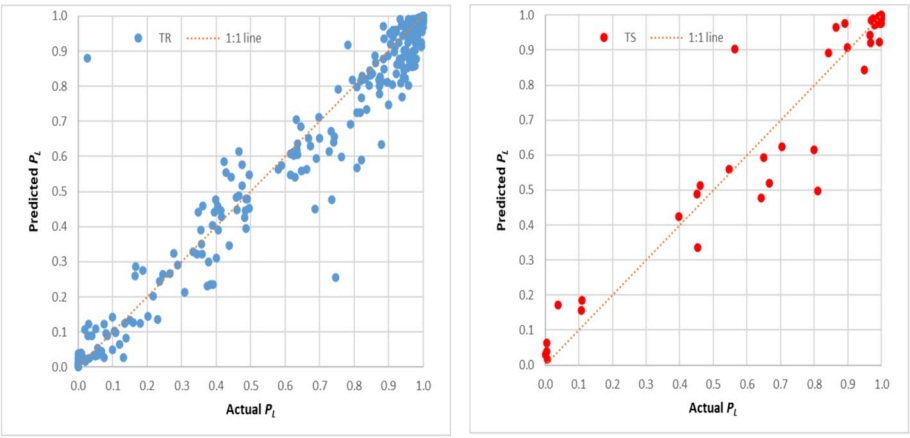


Fig. 11 Scatter plot for training and testing datasets for a data split ratio of 80:10:10

performance of all the models mentioned above. Finally, the sum of the scores from all sets is used to get the final score for each model. The results of the rank analysis, are shown in Fig. 12 below.

5.4 Taylor diagram

A Taylor diagram has been implemented to evaluate the accuracy of the DNN model during training and testing datasets across different data splits. The Taylor diagram was first introduced by Taylor (2001) and is used to display the accuracy of different models. This diagram gives a comprehensive and graphical overview of multiple models within a single plot. The parameters such as *R, RMSE*, and the ratio of *SD* values are used to measure and describe the degree of agreement between observed and predicted actions as a singular data point. The ability to present ML models' performance across

Table 6 Performance assessment summary of eight statistical indices

	R^2	*RMSE*	*MAE*	*MBE*	*MAD*	*WMAPE*	U_{95}	*GPI*
Split ratio (60:20:20)								
TR	0.925	0.094	0.049	0.012	0.018	0.065	0.258	0.000
TS	0.879	0.119	0.072	0.005	0.031	0.095	0.330	0.000
VS	0.919	0.106	0.063	0.029	0.025	0.084	0.288	0.000
Split ratio (70:15:15)								
TR	0.938	0.090	0.052	− 0.030	0.027	0.069	0.243	0.000
TS	0.908	0.111	0.071	− 0.031	0.035	0.098	0.303	0.000
VS	0.897	0.103	0.069	− 0.031	0.043	0.087	0.281	0.000
Split ratio (80:10:10)								
TR	0.952	0.076	0.038	− 0.014	0.016	0.050	0.208	0.000
TS	0.928	0.092	0.054	− 0.006	0.023	0.074	0.255	0.000
VS	0.938	0.081	0.046	− 0.020	0.018	0.059	0.222	0.000

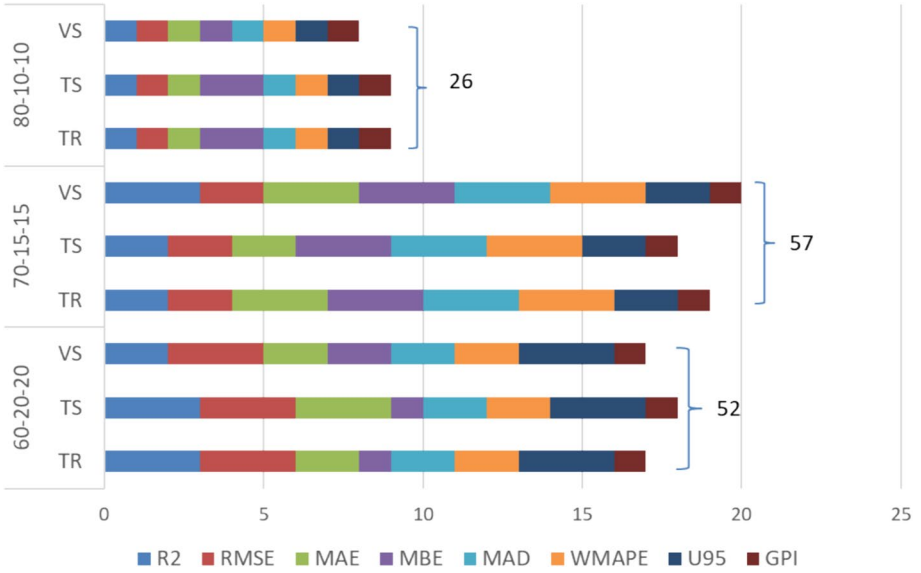


Fig. 12 Rank score of all data split ratios across training, testing, and validation sets for the eight statistical parameters and the total score of each split ratio

three parameters at a single point makes the Taylor diagram a valuable visualization tool in ML performance assessment. In an ideal model, the point would be positioned at a closer distance from the reference point. The Taylor diagram for training and testing datasets is shown in Fig. 13 below.

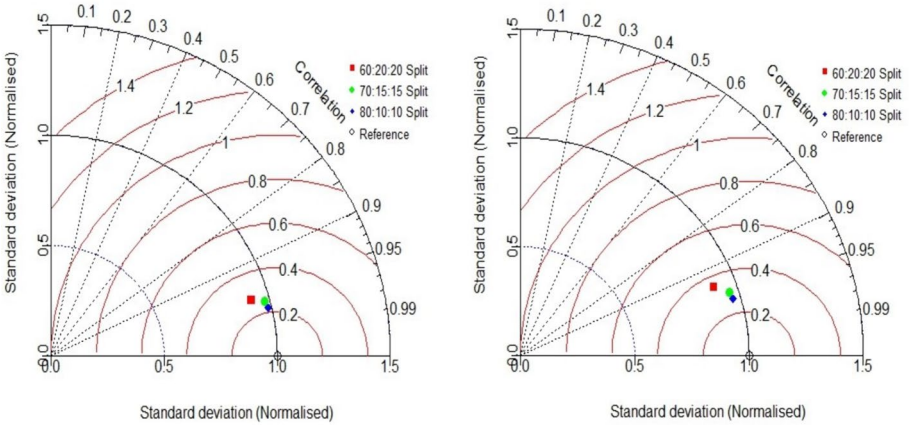


Fig. 13 Taylor diagram for training and testing dataset

5.5 Regression error characteristic (REC) and loss versus epochs curve

REC curves generalize receiver operating characteristic (ROC) curves to regression (Bi and Bennett 2003). ROC curves provide a powerful tool for visualizing and comparing classification results (Fawcett 2004). REC curves are a useful visual tool for communicating results to non-experts. The REC curves depict the error tolerance on the x-axis and the percentage of predicted data within the tolerance on the y-axis. The REC curve displays the error tolerance as either a squared residual or an absolute deviation. As a result, there is a clear trade-off between error tolerance and regression function accuracy. The generated curve computes the error in the cumulative distribution function between the experimental and observed values. The regression models are easily distinguishable when they are similar and when they are significantly different. An accurate measure of a regression model's performance is the area under the curve (AUC), often known as the curve area. The AUC value should be as high as practicable, which indicates better model performance. The REC and loss vs epochs curves are shown in Figs. 14, 15, and 16.

6 Discussion

The previous subsections comprehensively assessed the DNN model for predicting the probability of the occurrence of soil liquefaction. Table 6 displays the performance summary for the eight statistical indices used in this study. Notably, the R^2, *RMSE,* and *MAE* values of the training dataset for the split ratio 80:10:10 are 0.952, 0.076, and 0.032, and for the testing set, 0.928, 0.092, and 0.054, respectively, indicate superior performance of the DNN model. The R^2 values for training, testing, and validation sets are close to each other and are relatively high, which suggests that the proposed model performs well across different datasets and is not overfitting or underfitting.

The results of reliability analysis using the FORM approach shown in Fig. 7 indicate that the spreadsheet solution is an effective tool for reliability analysis of a single case using CPT datasets. A more efficient and time-saving solution for handling numerous

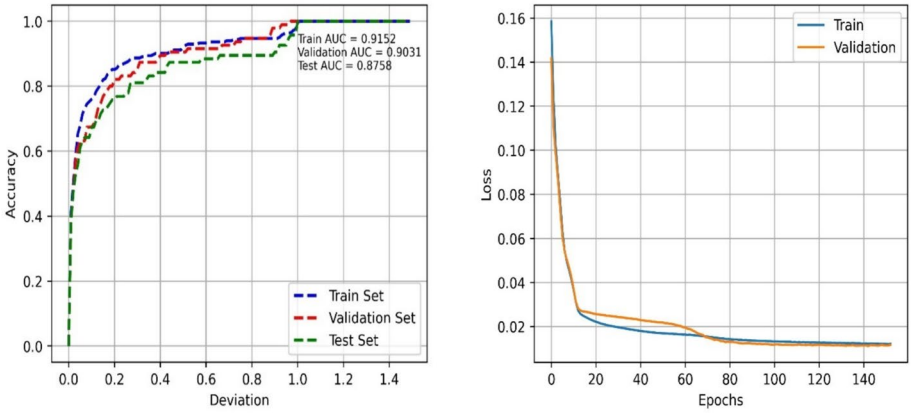


Fig. 14 REC curve and loss versus epoch curve for 60:20:20 dataset split ratio

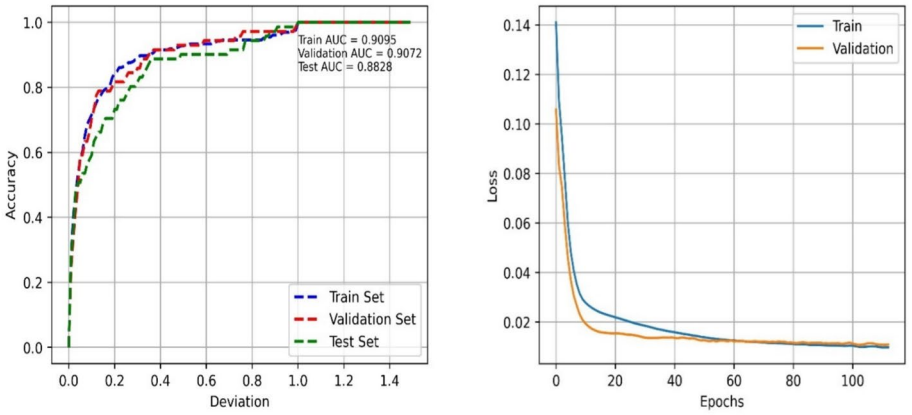


Fig. 15 REC curve and loss versus epoch curve for 70:15:15 dataset split ratio

datasets would be to implement this approach in MATLAB. The sensitivity analysis results establish q_{c1Ncs} as the most sensitive parameter in FORM-based liquefaction analysis using CPT datasets. Furthermore, we found MSF and r_d, which were not previously considered in the CPT-based sensitivity assessment, to be the second and third most sensitive parameters, respectively. This could be because MSF is a function of q_{c1Ncs} and M_w whereas r_d is derived from M_w and the depth parameter (Eqs. 2 to 6). Although various sensitivity studies on SPT datasets with several parameters are available in the literature, a comprehensive reliability-based sensitivity analysis assessment using the CPT dataset is still missing. Future studies can include advanced sensitivity analysis techniques discussed previously (Sect. 4.3). The results of the rank analysis show that the dataset with a split ratio of 80:10:10 had the lowest score (26) and performed best. The point closest to the reference point in the Taylor diagram indicates that this split ratio performs optimally for both training and testing datasets. Furthermore, a high AUC value for training and testing validates the DNN model's superior performance.

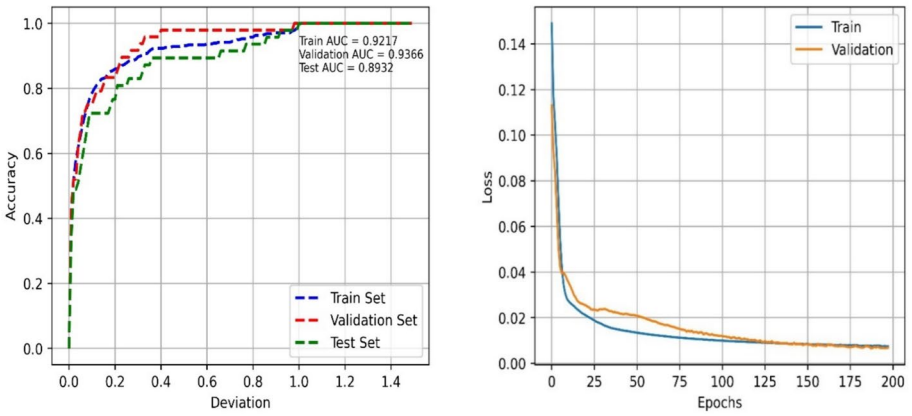


Fig. 16 REC curve and loss versus epoch curve for 80:10:10 dataset split ratio

Future research may address some of this study's limitations. First, the linear function approximation of limit state function in FORM can be addressed by using the second order reliability method (SORM). Second is the assumption that the COV remains constant in all situations for each parameter. Site-specific COV should be used if available. Furthermore, additional data points may be added to improve the DNN model's performance. In terms of generalizability to different soil types, this methodology should be refined by incorporating appropriate equations for the Cyclic Resistance Ratio (CRR) specific to gravelly soils. Also, exploring alternative machine learning architectures, incorporating additional soil and seismic parameters, or investigating ensemble modelling techniques such as XGBoost and hybrid machine learning models can significantly improve the performance.

7 Conclusions

In this study, using CPT datasets and the FORM technique, the reliability index and probability of occurrence of liquefaction have been calculated. The most sensitive input parameters for soil liquefaction have then been identified by calculating the reliability sensitivity analysis. A DNN model has been developed and validated to predict the probability of occurrence of liquefaction. The dataset has been first divided into training, testing, and validation datasets in three split ratios. The DNN model is then validated using a separate testing dataset after being built using the training and validation datasets. Based on the results and discussion presented earlier, the following conclusions can be drawn:

- The spreadsheet solution is a valuable tool for performing reliability analysis on CPT datasets. The results of the FORM analysis validate its robustness, with the computed reliability index ranging from − 8 to 6.
- The DNN model is very promising in terms of predictive accuracy. A very high R^2 value on the testing set (0.928), close to the R^2 of the training set (0.952), suggests that the proposed DNN model generalizes well.

- The rank analysis and Taylor diagram show that the DNN model performs best for the datasets with an 80:10:10 split ratio. The REC curve and corresponding area under the curve further confirm the DNN model's robustness.
- The results of sensitivity analysis show q_{c1Ncs} as the most important parameter contributing to the probability of occurrence of soil liquefaction in the FORM based approach, followed by *MSF*, r_d, σ_v, σ'_v, and a_{max}. This suggests *MSF* and r_d as important parameters that should be considered in the reliability and machine learning based evaluation of P_L.

Based on the performance of the DNN models in predicting the probability of occurrence of soil liquefaction, it can be offered as a realistic choice to help geotechnical professionals predict the soil liquefaction probability and in liquefaction hazard mapping for quick and effective risk assessment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11069-024-06934-1.

Funding No funding was received to conduct this study.

Declarations

Competing interests On behalf of all the authors, the corresponding author states that there are no competing interests.

References

Alkayem NF, Shen L, Mayya A et al (2024) Prediction of concrete and FRC properties at high temperature using machine and deep learning: a review of recent advances and future perspectives. J Build Eng 83:108369. https://doi.org/10.1016/j.jobe.2023.108369

Alzahamie ZH, Abdul-Husain HA (2021) Artificial neural network for prediction of liquefaction triggering based on CPT data. J Phys Conf Ser 1973:12197

Armaghani DJ, Asteris PG (2021) A comparative study of ANN and ANFIS models for the prediction of cement-based mortar materials compressive strength. Neural Comput Appl 33:4501–4532. https://doi.org/10.1007/s00521-020-05244-4

Asteris PG, Kolovos KG (2019) Self-compacting concrete strength prediction using surrogate models. Neural Comput Appl 31:409–424. https://doi.org/10.1007/s00521-017-3007-7

Asteris PG, Roussis PC, Douvika MG (2017) Feed-forward neural network prediction of the mechanical properties of sandcrete materials. Sensors 17:1344

Asteris PG, Lemonis ME, Le T-T, Tsavdaridis KD (2021a) Evaluation of the ultimate eccentric load of rectangular CFSTs using advanced neural network modeling. Eng Struct 248:113297. https://doi.org/10.1016/j.engstruct.2021.113297

Asteris PG, Skentou AD, Bardhan A et al (2021b) Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. Cem Concr Res 145:106449

Bardhan A, Alzo'ubi AK, Palanivelu S et al (2023) A hybrid approach of ANN and improved PSO for estimating soaked CBR of subgrade soils of heavy-haul railway corridor. Int J Pavement Eng 24:2176494. https://doi.org/10.1080/10298436.2023.2176494

Bastin S, van Ballegooy S, Mellsop N, Wotherspoon L (2020) Liquefaction case histories from the 1987 Edgecumbe earthquake, New Zealand-insights from an extensive CPT dataset and paleo-liquefaction trenching. Eng Geol 271:105404

Bennett MJ (1984) Geotechnical investigation of liquefaction sites, Imperial Valley, California. US Geological Survey

Bi J, Bennett KP (2003) Regression error characteristic curves. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 43–50

Borisov V, Leemann T, Seßler K et al (2024) Deep neural networks and tabular data: a survey. IEEE Trans Neural Netw Learn Syst 35:7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161
Boulanger RW (2003) High overburden stress effects in liquefaction analyses. J Geotech Geoenviron Eng 129:1071–1082. https://doi.org/10.1061/(asce)1090-0241(2003)129:12(1071)
Boulanger RW, Mejia LH, Idriss IM (1997) Liquefaction at moss landing during Loma Prieta Earthquake. J Geotech Geoenviron Eng 123:453–467. https://doi.org/10.1061/(ASCE)1090-0241(1997)123:5(453)
Boulanger RW, Idriss IM (2014) CPT and SPT based liquefaction triggering procedures. Rep No UCD/CGM-14 1
Cetin KO, Seed RB, Der Kiureghian A et al (2004a) Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential. J Geotech Geoenviron Eng 130:1314–1340
Cetin KO, Youd TL, Seed RB et al (2004b) Liquefaction-induced lateral spreading at Izmit Bay during the Kocaeli (Izmit)-Turkey earthquake. J Geotech Geoenviron Eng 130:1300–1313
Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250
Chu DB, Stewart JP, Youd TL, Chu BL (2006) Liquefaction-induced lateral spreading in near-fault regions during the 1999 Chi-Chi, Taiwan earthquake. J Geotech Geoenviron Eng 132:1549–1565
Cox BR, Boulanger RW, Tokimatsu K et al (2013) Liquefaction at strong motion stations and in Urayasu City during the 2011 Tohoku-Oki earthquake. Earthq Spectra 29:55–80
Cukier RI, Fortuin CM, Shuler KE et al (1973) Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I theory. J Chem Phys 59:3873–3878. https://doi.org/10.1063/1.1680571
Daniel C (1973) One-at-a-time plans. J Am Stat Assoc 68:353–360. https://doi.org/10.1080/01621459.1973.10482433
Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. Mach Learn 31:1–38
Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67
Galton F (1886) Regression towards mediocrity in hereditary stature. J Anthropol Inst Gt Britain Irel 15:246–263. https://doi.org/10.2307/2841583
Gan Y, Duan Q (2019) Sensitivity analysis methods BT. In: Duan Q, Pappenberger F, Wood A et al (eds) Handbook of hydrometeorological ensemble forecasting. Springer, Berlin, pp 637–671
Ghani S, Kumari S (2022) Reliability analysis for liquefaction risk assessment for the city of Patna, India using hybrid computational modeling. J Geol Soc India 98:1395–1406. https://doi.org/10.1007/s12594-022-2187-7
Ghani S, Kumari S, Ahmad S (2022) Prediction of the seismic effect on liquefaction behavior of fine-grained soils using artificial intelligence-based hybridized modeling. Arab J Sci Eng 47:5411–5441. https://doi.org/10.1007/s13369-022-06697-6
Ghani S, Sapkota SC, Singh RK et al (2024) Modelling and validation of liquefaction potential index of fine-grained soils using ensemble learning paradigms. Soil Dyn Earthq Eng 177:108399
Golmoghani Ebrahimi S, Noorzad A, Kupaei HJ (2024) Reliability analysis of soil liquefaction using improved hypercube sampling (IHS) method. Int J Civ Eng 22:195–207. https://doi.org/10.1007/s40999-023-00863-z
Green RA, Cubrinovski M, Cox B et al (2014) Select liquefaction case histories from the 2010–2011 Canterbury earthquake sequence. Earthq Spectra 30:131–153. https://doi.org/10.1193/030713EQS066M
Hacıefendioğlu K, Başağa HB, Demir G (2021) Automatic detection of earthquake-induced ground failure effects through faster R-CNN deep learning-based object detection using satellite images. Nat Hazards 105:383–403. https://doi.org/10.1007/s11069-020-04315-y
Hamby DM (1995) A comparison of sensitivity analysis techniques. Health Phys 68:195–204
Hanna AM, Ural D, Saygili G (2007) Evaluation of liquefaction potential of soil deposits using artificial neural networks. Eng Comput 24:5–16. https://doi.org/10.1108/02644400710718547
Hsein JC, Ye FS, Kun LD (2012) Reliability analysis of soil liquefaction potential. Earthq Eng Soil Dyn 158:1–15. https://doi.org/10.1061/40779(158)24
Idriss IM, Boulanger RW (2006) Semi-empirical procedures for evaluating liquefaction potential during earthquakes. Soil Dyn Earthq Eng 26:115–130
Iman RL, Helton JC (1988) An investigation of uncertainty and sensitivity analysis techniques for computer models. Risk Anal 8:71–90
Jas K, Dodagoudar GR (2023a) Explainable machine learning model for liquefaction potential assessment of soils using XGBoost-SHAP. Soil Dyn Earthq Eng 165:107662
Jas K, Dodagoudar GR (2023b) Liquefaction potential assessment of soils using machine learning techniques: a state-of-the-art review from 1994–2021. Int J Geomech 23:3123002
Jas K, Jana A, Dodagoudar GR (2024a) Prediction of shear strain and excess pore water pressure response in liquefiable sands under cyclic loading using deep learning model. Jpn Geotech Soc Spec Publ 10:1729–1734. https://doi.org/10.3208/jgssp.v10.OS-35-05

Jas K, Mangalathu S, Dodagoudar GR (2024b) Evaluation and analysis of liquefaction potential of gravelly soils using explainable probabilistic machine learning model. Comput Geotech 167:106051. https://doi.org/10.1016/j.compgeo.2023.106051
Jha SK, Suzuki K (2009a) Reliability analysis of soil liquefaction based on standard penetration test. Comput Geotech 36:589–596. https://doi.org/10.1016/j.compgeo.2008.10.004
Jha SK, Suzuki K (2009b) Liquefaction potential index considering parameter uncertainties. Eng Geol 107:55–60. https://doi.org/10.1016/j.enggeo.2009.03.012
Jha SK, Karki B, Bhattarai A (2020) Deterministic and probabilistic evaluation of liquefaction potential: a case study from 2015 Gorkha (Nepal) earthquake. Geotech Geol Eng 38:4369–4384. https://doi.org/10.1007/s10706-020-01277-7
Johari A, Khodaparast AR (2013) Modelling of probability liquefaction based on standard penetration tests using the jointly distributed random variables method. Eng Geol 158:1–14
Johari A, Javadi AA, Makiabadi MH, Khodaparast AR (2012) Reliability assessment of liquefaction potential using the jointly distributed random variables method. Soil Dyn Earthq Eng 38:81–87. https://doi.org/10.1016/j.soildyn.2012.01.017
Juang CH, Chen CJ (1999) Cpt-based liquefaction evaluation using artificial neural networks. Comput Civ Infrastruct Eng 14:221–229
Juang CH, Jiang T, Andrus RD (2002) Assessing probability-based methods for liquefaction potential evaluation. J Geotech Geoenviron Eng 128:580–589. https://doi.org/10.1061/(ASCE)1090-0241(2002)128:7(580)
Juang CH, Yuan H, Lee D-H, Lin P-S (2003) Simplified cone penetration test-based method for evaluating liquefaction resistance of soils. J Geotech Geoenviron Eng 129:66–80. https://doi.org/10.1061/(ASCE)1090-0241(2003)129:1(66)
Juang CH, Fang SY, Khor EH (2006) First-order reliability method for probabilistic liquefaction triggering analysis using CPT. J Geotech Geoenviron Eng 132:337–350
Kasebzadeh J, Noorzad A, Mahboubi AR (2015) Reliability analysis of liquefaction utilizing monte carlo simulation based on simplified stress method. J Seismol Earthq Eng 17:233–248
Ke B, Khandelwal M, Asteris PG et al (2021) Rock-burst occurrence prediction based on optimized naïve Bayes models. IEEE Access 9:91347–91360. https://doi.org/10.1109/ACCESS.2021.3089205
Kohestani VR, Hassanlourad M, Ardakani A (2015) Evaluation of liquefaction potential based on CPT data using random forest. Nat Hazards 79:1079–1089
Kumar P, Samui P (2022) Design of an energy pile based on CPT data using soft computing techniques. Infrastructures 7:169. https://doi.org/10.3390/infrastructures7120169
Kumar P, Samui P (2024) Reliability-based load and resistance factor design of an energy pile with CPT data using machine learning techniques. Arab J Sci Eng 49:4831–4860. https://doi.org/10.1007/s13369-023-08253-2
Kumar DR, Samui P, Burman A (2022) Prediction of probability of liquefaction using soft computing techniques. J Inst Eng Ser A 103:1195–1208. https://doi.org/10.1007/s40030-022-00683-9
Kumar DR, Samui P, Burman A, Kumar S (2023a) Seismically induced liquefaction potential assessment by different artificial intelligence procedures. Transp Infrastruct Geotechnol 11:1–22
Kumar K, Samui P, Choudhary SS (2023b) State parameter based liquefaction probability evaluation. Int J Geosynth Gr Eng 9:76. https://doi.org/10.1007/s40891-023-00495-2
Kumar K, Samui P, Choudhary SS (2024) Assessment of maximum liquefaction distance using soft computing approaches. Geomech Eng 37:395–418. https://doi.org/10.12989/gae.2024.37.4.395
Kumar D, Kumari S (2023) Assessment of liquefaction susceptibility of fine-grained soil based on soil classification. Geomech Eng 54:1366–1380. https://doi.org/10.1007/s40098-023-00812-6
Kumari S, Ghani S (2024) Prediction of liquefaction behaviour of fine-grained soil using nature-inspired optimized algorithms coupled with neural network. Indian Geotech J. https://doi.org/10.1007/s40098-024-01016-2
Le T-T, Skentou AD, Mamou A, Asteris PG (2022) Correlating the unconfined compressive strength of rock with the compressional wave velocity effective porosity and Schmidt hammer rebound number using artificial neural networks. Rock Mech Rock Eng 55:6805–6840. https://doi.org/10.1007/s00603-022-02992-8
Lee C-Y, Chern S-G (2013) Application of a support vector machine for liquefaction assessment. J Mar Sci Technol 21:10
Low BK, Tang WH (2007) Efficient spreadsheet algorithm for first-order reliability method. J Eng Mech 133:1378–1387
MacKay DJC (1998) Introduction to Gaussian processes. NATO ASI Ser F Comput Syst Sci 168:133–166
Massey FJ Jr (1951) The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 46:68–78. https://doi.org/10.1080/01621459.1951.10500769

Morris MD (1991) Factorial sampling plans for preliminary computational experiments. Technometrics 33:161–174. https://doi.org/10.1080/00401706.1991.10484804
Moss RES (2003) CPT-based probabilistic assessment of seismic soil liquefaction initiation. University of California, Berkeley
Moss R, Kayen R, Tong L, et al (2009) Reinvestigation of liquefaction and nonliquefaction case histories from the 1976 Tangshan Earthquake. Pac Earthq Eng Res Cent Rep 2009/102
Muduli PK, Das SK (2014) CPT-based seismic liquefaction potential evaluation using multi-gene genetic programming approach. Indian Geotech J 44:86–93. https://doi.org/10.1007/s40098-013-0048-4
Nova R (2013) Introduction: basic concepts. In: Soil mechanics. Wiley, pp 1–25
Phoon K-K, Kulhawy FH (1999) Characterization of geotechnical variability. Can Geotech J 36:612–624
Pirani M, Thakkar P, Jivrani P, et al (2022) A comparative analysis of ARIMA, GRU, LSTM and BiLSTM on financial time series forecasting. In: 2022 IEEE international conference on distributed computing and electrical circuits and electronics (ICDCECE). pp 1–6
Robertson PK, Wride CE (1998) Evaluating cyclic liquefaction potential using the cone penetration test. Can Geotech J 35:442–459
Sabri MS, Ahmad F, Samui P (2023) Slope stability analysis of heavy-haul freight corridor using novel machine learning approach. Model Earth Syst Environ. https://doi.org/10.1007/s40808-023-01774-7
Sabri MS, Jaiswal A, Verma AK, Singh TN (2024) Advanced machine learning approaches for uniaxial compressive strength prediction of Indian rocks using petrographic properties. Multiscale Multidiscip Model Exp Des. https://doi.org/10.1007/s41939-024-00513-4
Saleh E, Alghossoon A, Tarawneh A (2022) Optimal allocation of material and slenderness limits for the rectangular concrete-filled columns. J Constr Steel Res 193:107283. https://doi.org/10.1016/j.jcsr.2022.107283
Saltelli A, Tarantola S, Chan KP-S (1999) A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41:39–56. https://doi.org/10.1080/00401706.1999.10485594
Samui P, Kim D, Sitharam TG (2011) Support vector machine for evaluating seismic-liquefaction potential using shear wave velocity. J Appl Geophys 73:8–15
Sebaaly GT, Rahhal ME (2019) Probabilistic analysis of soil liquefaction based on CPT and SPT results. COMPDYN Proc 1:141–150
Seed HB, Idriss IM (1971) Simplified procedure for evaluating soil liquefaction potential. J Soil Mech Found Div 97:1249–1273
Sett S, Chattopadhyay KK, Ghosh A (2023) Liquefaction hazard mapping and relationship between probability of liquefaction and factor of safety for Kolkata city, India. Nat Hazards 118:347–376. https://doi.org/10.1007/s11069-023-06007-9
Sobol′ IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 55:271–280. https://doi.org/10.1016/S0378-4754(00)00270-6
Suzuki Y, Tokimatsu K, Moss R, et al (2003) CPT-based liquefaction field case histories from the 1995 Hyogoken-Nambu (Kobe) Earthquake, Japan. Geotech Eng Res Rep No UCB/GT-2003/03
Sze V, Chen Y-H, Yang T-J, Emer JS (2017) Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE 105:2295–2329
Tarawneh A, Momani Y, Alawadi R (2021) Leveraging artificial intelligence for more accurate and reliable predictions of anchors shear breakout capacity in thin concrete members. Structures 32:1005–1014. https://doi.org/10.1016/j.istruc.2021.03.074
Tarawneh A, Saleh E, Almasabha G, Alghossoon A (2023) Hybrid data-driven machine learning framework for determining prestressed concrete losses. Arab J Sci Eng 48:13179–13193. https://doi.org/10.1007/s13369-023-07714-y
Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106:7183–7192
Vivek B, Raychowdhury P (2014) Probabilistic and spatial liquefaction analysis using CPT data: a case study for Alameda County site. Nat Hazards 71:1715–1732. https://doi.org/10.1007/s11069-013-0976-4
Xiaoping D (2005) Probabilistic engineering design-first order and second reliability methods, University of Missouri–Rolla
Youd TL, Idriss IM (2001) Liquefaction resistance of soils: summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils. J Geotech Geoenviron Eng 127:297–313
Youd TL, DeDen DW, Bray JD et al (2009) Zero-displacement lateral spreads, 1999 Kocaeli, Turkey, Earthquake. J Geotech Geoenviron Eng 135:46–61. https://doi.org/10.1061/(ASCE)1090-0241(2009)135:1(46)/ASSET/74DA8F7B-24E4-4D1B-AAB8-51065A74FE75/ASSETS/IMAGES/LARGE/13.JPG
Zhang Y, Qiu J, Zhang Y, Wei Y (2021) The adoption of ELM to the prediction of soil liquefaction based on CPT. Nat Hazards 107:539–549. https://doi.org/10.1007/s11069-021-04594-z

Zou H, Liu S, Cai G et al (2017) Mapping probability of liquefaction using geostatistics and first order reliability method based on CPTU measurements. Eng Geol 218:197–212

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

Sophia Khatoon[1] · Kishan Kumar[1] · Pijush Samui[1] · Laith Sadik[2] · Sanjay Kumar Shukla[3,4,5]

✉ Kishan Kumar
kishank.phd22.ce@nitp.ac.in

Sophia Khatoon
sophiak.pg22.ce@nitp.ac.in

Pijush Samui
pijush@nitp.ac.in

Laith Sadik
sadikla@mail.uc.edu

Sanjay Kumar Shukla
s.shukla@ecu.edu.au; sanjaykshukla1@gmail.com

[1] Department of Civil Engineering, National Institute of Technology Patna, Patna, India

[2] Department of Civil and Architectural Engineering and Construction Management, University of Cincinnati, Cincinnati, USA

[3] School of Engineering, Edith Cowan University, 270 Joondalup Drive, Joondalup, Perth, WA 6027, Australia

[4] Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India

[5] Delhi Technological University, New Delhi, Delhi, India

MASSFormer: Mobility-Aware Spectrum Sensing using Transformer-Driven Tiered Structure

Dimpal Janu, *Member, IEEE,* Sandeep Mandia, *Member, IEEE,* Kuldeep Singh, *Senior Member, IEEE,* and Sandeep Kumar, *Senior Member, IEEE*

arXiv:2409.17546v1 [cs.IT] 26 Sep 2024

***Abstract*—In this paper, we develop a novel mobility-aware transformer-driven tiered structure (MASSFormer) based cooperative spectrum sensing method that effectively models the spatio-temporal dynamics of user movements. Unlike existing methods, our method considers a dynamic scenario involving mobile primary users (PUs) and secondary users (SUs)and addresses the complexities introduced by user mobility. The transformer architecture utilizes an attention mechanism, enabling the proposed method to adeptly model the temporal dynamics of user mobility by effectively capturing long-range dependencies within the input data. The proposed method first computes tokens from the sequence of covariance matrices (CMs) for each SU and processes them in parallel using the SU-transformer network to learn the spatio-temporal features at SU-level. Subsequently, the collaborative transformer network learns the group-level PU state from all SU-level feature representations. The attention-based sequence pooling method followed by the transformer encoder adjusts the contributions of all tokens. The main goal of predicting the PU states at each SU-level and group-level is to improve detection performance even more. We conducted a sufficient amount of simulations and compared the detection performance of different SS methods. The proposed method is tested under imperfect reporting channel scenarios to show robustness. The efficacy of our method is validated with the simulation results demonstrating its higher performance compared with existing methods in terms of detection probability, sensing error, and classification accuracy.**

***Index Terms*—Spectrum Sensing, Transformer, Self-attention, Mobility, Fading Channels.**

I. Introduction

THE fifth generation (5G) mobile communication systems confront new requirements and challenges in terms of ultra-large range, massive connection, and ultra-low latency due to the exponential growth of mobile data traffic and the excessive connecting devices. The increasing use of wireless networks worldwide has caused a shortage of radio spectrum. Cognitive Radio (CR) networks offer a potential solution by allowing secondary users (SUs) to opportunistically access the spectrum allocated to primary users (PUs) when it is not in use. Therefore, spectrum sensing (SS) has gained intense attention from academia in recent decades. Numerous SS approaches have been proposed, with energy detection (ED) being the most commonly used method due to its low structure and minimal processing complexity [1].

D. Janu, and K. Singh are with the Department of Electronics and Communication Engineering, Malaviya National Institute of Technology Jaipur, 302017,India (e-mail: 2019rec9146@mnit.ac.in;kuldeep.ece@mnit.ac.in)

S. Mandia is with Thapar Institute of Engineering & Technology, Patiala, 147004, India (e-mail: sandeep.mandia@thapar.edu)

S. Kumar is with Central Research Lab, Bharat Electronics Ltd., Ghaziabad, 201010, India (e-mail: sann.kaushik@gmail.com)

Various deep/machine learning (DL/ML) methods have been used recently to enhance cooperative spectrum sensing (CSS) detection performance. In [2], several ML-based SS methods have been proposed. A detailed analysis of ML-based approaches applied to CSS has been provided in [3]. This paper examines the strengths and limitations associated with these ML-based approaches, focusing on aspects such as performance improvement, variety of features used, applicable scenarios and complexity. There are certain limitations associated with ML-based approaches, including the necessity for manually crafted features during the ML model training process, as these features may not effectively capture the complexities of the real environment. Nowadays, DL-based methods have been applied to wireless communication demonstrating remarkable performance. Specifically, DL-based spectrum sensing has started to make a notable impact. Convolutional neural networks (CNNs) possess a robust ability to extract spatial features from data in matrix format. The covariance matrices (CMs) are regarded as a versatile statistical measure, encompassing correlation features of sensing signals of SUs. Consequently, CNNs have been widely applied in SS to acquire correlation features from CMs. Liu *et. al.* have designed CNN-based approaches in [4] and [5], where CMs were fed as input to derive the test statistics, and these two works achieved performance improvement in CSS. The SS methodology discussed in [6] integrates the CNNs and short Time Fourier transform (STFT) to predict PU states leveraging time-frequency domain information. Graph convolutional neural networks (GCN) possess a robust capability to capture relationships among graph nodes by encoding the structural information of non-grid data. Additionally, it can concurrently incorporate graphs of variable sizes. Leveraging the distinctive ability of GCN, an SS method has been proposed in [7] to solve the hidden node problem. In [8], the authors used a CNN to provide a suitable alternative solution to the likelihood ratio test (LRT) under various general noises, which include Middleton class A, isometric complex symmetric α -stable, and isometric complex generalized Gaussian noise.

The methodologies mentioned above leverage sensing data produced during the current sensing duration for the prediction of PU states while refraining from incorporating sensing data from preceding durations. These methods utilize CNN architecture to learn the graphical features from CMs of sensed signals. Concurrently, researchers have also concentrated on capturing temporal information by leveraging historical data. An approach based on activity pattern aware spectrum sensing (APASS) has been developed in [9] in order

to enhance the sensing performance even further. This method involves the simultaneous deployment of two parallel CNN structures, utilizing CMs generated from both current and past sensing data to learn both graphical and temporal features. A limited number of existing works in literature have applied the Long Short-Term Memory (LSTM) network, demonstrating its ability to learn temporal features through the utilization of historical sensing data across numerous sensing durations. In [10], an SS model that applies the LSTM network to capture temporal correlation features from sequential data of current and past sensing events has been proposed. Additionally, the model employed statistical insights of PU activity such as PU ON and OFF periods' durations as well as duty cycle to improve the sensing performance. Further, an LSTM-based SS approach has been developed in [11] to learn temporal information from the CMs. Authors of [12] have developed a spectrum prediction method that utilizes the Taguchi method to optimize the hyperparameter of the LSTM network. Considering the single antenna SU scenarios, authors of [13] and [14] have developed an SS method employing a hybrid structure of CNN-LSTM to extract the spatio-temporal information from observations of single sensing intervals. A combination of 1-D CNN and LSTM network has been employed to learn the time and frequency domain features, to find the presence of PU specifically in low signal-to-noise ratio (SNR) environments [15]. In order to enhance the SS performance further, xie *et. al.* have introduced a CNN-LSTM method in [16]. Initially, the detector employs CNN to capture spatial information from CMs derived from sensing signals. Subsequently, the features extracted from different sensing durations are inputted into the LSTM network to determine the PU activity pattern. An SS model utilizing CNN and LSTM network has been proposed in [17], where a 1-D signal vector is inputted to CNN to learn graphical features, and then learned features are provided as input to the LSTM network to capture the temporal features.

The methodologies discussed above have utilized an LSTM network and a hybrid CNN-LSTM network to learn temporal information for predicting the PU states. While some LSTM-based CSS methods learn temporal features from observations collected during one sensing duration. However, such features are comparatively less precise than those obtained by extracting temporal features across several sensing periods to find the activity pattern of PU. The two activities, predicting PU states at individual SUs and at the fusion centre, occur simultaneously over time. However, the methods discussed above treat all sensing outcomes from participating SUs with equal importance, which can misguide the PU activity prediction results by overstating the importance of irrelevant features. In general, these methods face challenges to effectively model long-range dependence in the input sequence and capture comprehensive spatio-temporal features among cooperating SUs. Moreover, all the above SS methods have assumed static SUs, and none of these methods have considered mobility scenarios. However, in real-world scenarios, PUs and SUs are mobile, and their mobility has a significant impact on detection performance. In dynamic wireless environments, accurately detecting PU activity patterns is challenging due to severe path loss and channel fading resulting from user mobility.

Based on the above observation, we consider a mobile scenario having multiple SUs equipped with multiple antenna and propose a MASSFormer based method to effectively model the spatio-temporal dynamics of users by utilizing historical sensing data and exploiting user mobility patterns. Since transformer [18], an attention mechanism based architecture can address the challenges that LSTM and RNN networks struggle with, particularly in effectively modeling long-range dependencies in input sequences. Vision transformer (ViT) [19], a transformer based model, was first proposed for image classification and sparked considerable interest within the research community, leading to subsequent works and its extension to video vision transformer (ViViT) in [20]. Our proposed MASSFormer method is inspired by ViViT, which utilizes an SU-transformer network to learn relevant spatio-temporal features from the sensing outcomes of each SU over several sensing durations. Subsequently, a collaborative transformer network is utilized to model the SU-level spatio-temporal features of all participating SUs to learn group-level features to predict PU states. Attention across multiple levels of the transformer layers can effectively capture more relevant features from the individual SUs in a progressive manner. With the considered scenario, the CNN-LSTM method [16] captures the temporal information from all SUs concurrently using a single LSTM, thereby neglecting temporal information at individual SU-level. Similarly, the 3-D CNN [21] method fails to capture the spatio-temporal information at individual SU-level.

We summarize the main contributions of the paper as follows:

- We develop a novel MASSFormer method that effectively models the spatio-temporal dynamics of user movements. The developed method first uses an SU-transformer network which uses the transformer encoder with attention mechanism to learn the spatio-temporal features of each SU to predict the PU states at SU-level and a collaborative transformer network to model the learned representations from SUs to predict the PU states at group-level.
- We adopt a practical system model having multiple SUs with multiple antennas. We address the challenges posed by the mobile users in real-world scenarios by accounting for varying levels of path loss and fading severity at each SU. We consider the impact of imperfect reporting channels between SUs and fusion centre to show the robustness of our approach.
- With extensive simulations, we compared and analyzed the detection performance of the proposed method with state-of-the-art methods. We have validated that the proposed method outperforms the state-of-the-art methods in considered mobile scenarios in terms of detection probability, sensing error, and classification accuracy.

II. System model

We consider that a single PU with a single antenna and S SUs with multiple antenna are distributed randomly in a predefined area of the CR network. In the considered mobile scenarios, it is assumed that users are moving with random

velocity, resulting in dynamic changes to their locations over time, and their movements are independent of each other. At the beginning of each frame, each SU conducts SS and gathers N observations at each antenna during u-th sensing intervals. We formulate two hypotheses regarding the states of PU, where H_1 and H_0 denote active state and inactive states, respectively. The signal $y_s^m(n)$ from m-th antenna s-th SU is received at fusion centre represented as

$$y_s^m(n) = \begin{cases} h_{s,r}^m(n)(h_s^m(n)w(n) + \eta_s^m(n)) + \eta_c(n), & \mathcal{H}_1 \\ h_{s,r}^m(n)\eta_s^m(n) + \eta_c(n), & \mathcal{H}_0 \end{cases} \quad (1)$$

Where, PU signal is denoted as $w(n)$, $m = 1, 2, ..., M$, $s = 1, 2, ..., S$ and $n = 1, 2, 3, ..., N$. $\eta_s^m(n)$ and $h_s^m(n)$ denotes the noise signal and channel gain between PU and m-th antenna of s-th SU, respectively. The noise signal and channel gain received at fusion centre from m-th antenna of s-th SU is represented as $\eta_c(n)$ and $h_{s,r}^m(n)$, respectively. It is assumed that signals traveling through sensing as well as reporting channel may undergo different amounts of fading, which is quantified by varying fading parameter values. Channel gains of sensing channel and reporting channel are assumed to follow fading distribution over multiple sensing periods and during a particular sensing period it remains constant as sensing period is shorter than channel coherence time.

A. Mobility scenario

Mobility scenarios involve the movement of both PUs and SUs within the network environment. PUs mobility affects the prediction of PU states due to changes in PU activity patterns with their movements. The PU signals reaching SUs encounter varying levels of path loss and different channel conditions due to SUs' mobility. In this work, we consider the Random Waypoint (RW) Mobility Model [22] to find the movement patterns, trajectories, speed, direction, and location information of PUs and SUs. According to the RW mobility model, PU and SUs are initially distributed randomly in the simulation area. The users wait for pause time by staying in one location, and once this time expires, they choose a random destination in the simulation area and a speed that is uniformly distributed $[v_{min}, v_{max}]$. The users move to the new location at the selected speed and wait for a pause time after arrival. Fig. 1 shows the movement pattern of PU and SUs within a defined simulation area. Initially the position of moving user i.e. PU or SUs is represented as (X, Y), after a time-interval Δt, the position is updated using the following formulas:

$$X(t + \Delta t) = X(t) + v\Delta t * cos\theta(t) \quad (2)$$

$$Y(t + \Delta t) = Y(t) + v\Delta t * sin\theta(t) \quad (3)$$

where speed v is randomly selected between $[v_{min}, v_{max}]$, $\theta(t)$ denotes the direction at time t. The position traces of PU and SUs are calculated using equation (2) and (3).

The instantaneous signal-to-noise ratio experienced by s-th SU is denoted as $\gamma_s = \frac{|h_s|^2 Pt}{N_0 B_W} = \frac{Pr_s}{N_0 B_W}$. We assume that the transmit power of PU is fixed to P_t and PU signals are transmitted through the channel whose bandwidth is B_W. The

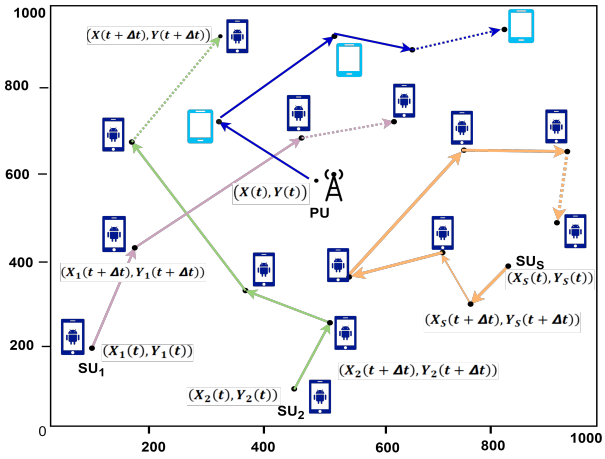


Fig. 1: Movement pattern of PUs and SUs using RW Mobility model

received PU power at SU s at a distance $d_s(u)$ from the PU at time u can be expressed in dB as

$$Pr_s(dB) = Pt(dB) - \{10log_{10}(\beta) + 10\alpha log_{10}(d_s(u))\} \quad (4)$$

where α and β represent path-loss exponent and path-loss constant respectively, and N_0 denotes noise power spectral density.

B. Data Preprocessing

The proposed model requires labeled training dataset collected in U sensing durations. The raw sensing signal $\boldsymbol{Y}_u^s$ collected during a single sensing duration is too large with dimension $M \times N$, to be directly considered as input for the neural network. Most of the existing SS methods utilized CMs as test statistics. Therefore, it is required to compress the sensing signal into CMs to construct the training set. The signal matrix $\boldsymbol{Y}_u^s$ composed of sensing signals from M antennas of SU s to fusion centre can be expressed as

$$\boldsymbol{Y}_u^s = \begin{bmatrix} y_s^1(1) & y_s^1(2) & \cdots & y_s^1(N) \\ y_s^2(1) & y_s^2(2) & \cdots & y_s^2(N) \\ \vdots & \vdots & \ddots & \vdots \\ y_s^M(1) & y_s^M(2) & \cdots & y_s^M(N) \end{bmatrix} \quad (5)$$

Let us construct the CM of the signal matrix as defined below

$$\boldsymbol{R}_u^s = \frac{1}{N}\boldsymbol{Y}_u^s \boldsymbol{Y}_u^{s\,H} \quad (6)$$

Now, construct the training set by arranging the CMs as defined as $\{(\boldsymbol{R}_1^s, b_1), (\boldsymbol{R}_2^s, b_2), (\boldsymbol{R}_3^s, b_3).....(\boldsymbol{R}_U^s, b_U)\}$, where U denotes total examples in the set. The class of u-th example is denoted as $b_u \in \{0, 1\}$, where the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$ are selected when $b_u = 1$ and $b_u = 0$, respectively. To predict the PU states at SU-level, we concurrently analyze the CMs of sensed signals gathered at $s-$th SU during numerous sensing intervals. The dataset for s-th SU is written as

$$\Phi^s = \{(\boldsymbol{\phi}_1^s, b_\lambda), (\boldsymbol{\phi}_2^s, b_{2\lambda}), ..., (\boldsymbol{\phi}_u^s, b_{u\lambda}), .., (\boldsymbol{\phi}_{U/\lambda}^s, b_U)\} \quad (7)$$

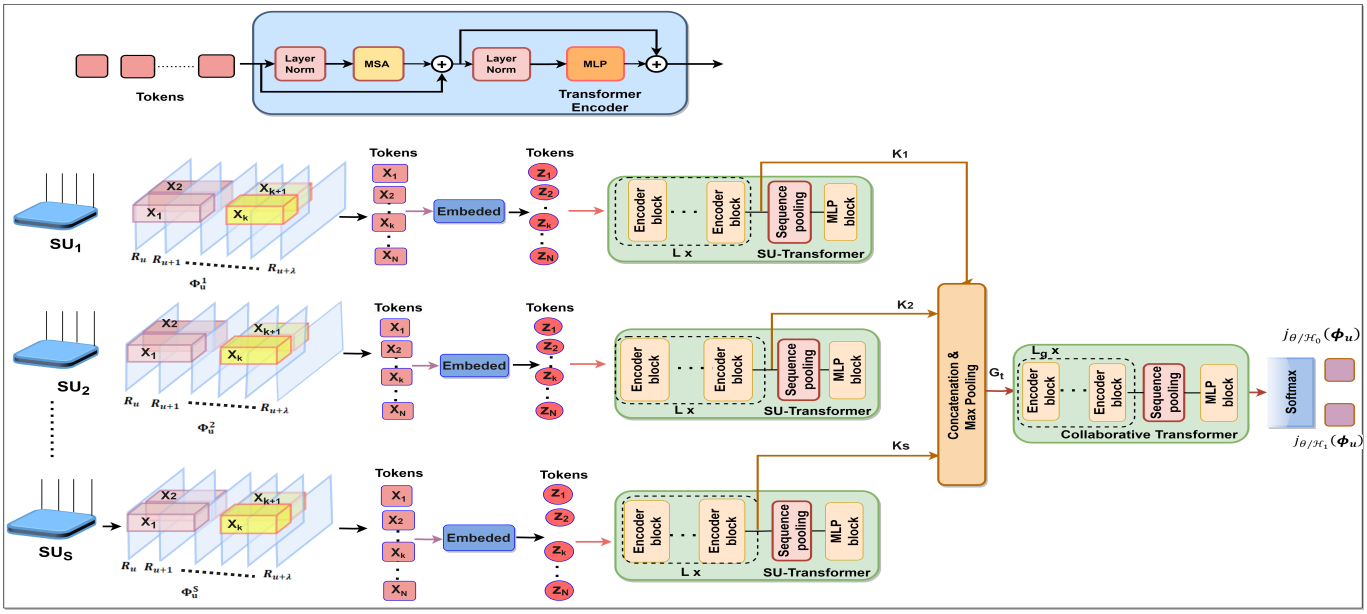


Fig. 2: Architecture of MASSFormer

where, $\boldsymbol{\phi}_u^s = [\boldsymbol{R}^s_{\lambda(u-1)+1}, \boldsymbol{R}^s_{\lambda(u-1)+2}, ..., \boldsymbol{R}^s_{\lambda(u-1)+\lambda}]$ and λ denotes the length of input's temporal sequence. The u-th example in the set of s-th SU is denoted as $\boldsymbol{\phi}_u^s$. The complete dataset for s-th SU is represented as Φ^s. To generate data for all SUs, we follow the same method as described in equation (5), (6), and (7). Since every participating SU sends its sensing data to the fusion centre, the complete dataset constructed by collecting samples from all SUs is defined as

$$\begin{aligned}\boldsymbol{\Phi} = \{&(\boldsymbol{\phi}_1^1, \boldsymbol{\phi}_1^2, .., \boldsymbol{\phi}_1^S, b_\lambda), (\boldsymbol{\phi}_2^1, \boldsymbol{\phi}_2^2, .., \boldsymbol{\phi}_2^S, b_{2\lambda}) \\ .., &(\boldsymbol{\phi}_u^1, \boldsymbol{\phi}_u^2, .., \boldsymbol{\phi}_u^S, b_{u\lambda}) .., (\boldsymbol{\phi}_{U/\lambda}^1, \boldsymbol{\phi}_{U/\lambda}^2, .., \boldsymbol{\phi}_{U/\lambda}^S, b_U)\}\end{aligned} \tag{8}$$

III. MASSFormer based framework

The equation (8) denotes the complete dataset, where each sample in the set consists of data samples from all participating SUs. The data point $\boldsymbol{\phi}_1^s$ in equation (7), consists of a sequence of CMs computed for each of SU is represented as $\mathbb{R}^{\lambda \times M \times M}$. Since, the transformer takes input in tokenized format, which is discussed in section III-A.

A. Tokenization

In this section, we discuss the tokenization process of ViViT, as we adopt a similar approach for tokenization in our methodology. A video sample is denoted as $X \in \mathbb{R}^{T \times H \times W \times C}$, where T, H, W, and C referring to the temporal length, height, width and depth of the input respectively. The ViT-based model processes the 2-D images to extract N_t non-overlapping patches. In the context of videos, a sequence of spatio-temporal "tubes" $x_1, x_2 .....x_{N_t}, x \in \mathbb{R}^{t \times h \times w}$ are extracted from the input volume as described in ViViT [20]. A linear operator E is applied to the tubes to linearly project them to 1-D tokens $z \in \mathbb{R}^{N_t \times d}$, where $N_t = n_t n_h n_w$, $n_t = \lfloor \frac{T}{t} \rfloor, n_h = \lfloor \frac{H}{h} \rfloor$, and $n_w = \lfloor \frac{W}{w} \rfloor$. To perform linear projection, 3-D convolution is used with the kernel size (t, h, w) and the strides (t, h, w) in time, height, and width dimensions, respectively. In ViViT, a learnable token z_{cls} is prepended to the token sequence. However, to incorporate the information from all the tokens, we apply attention based sequence pooling operation to compute a 1-D feature vector from 2-D encoder output. As the self-attention mechanism in the transformer encoder is order agnostic, positional embedding $P \in \mathbb{R}^{N_t \times d}$ is added to tokens to preserve the positional information. Therefore, we extract tokens from the sequence of CMs constructed for each of the SUs as depicted in Fig. 2. The sequence of tokens for s-th SU is denoted as

$$z_s = [Ex_1, Ex_2, Ex_3 .....Ex_{N_t} + P] \tag{9}$$

B. Transformer encoder

The SU-transformer network and collaborative transformer network are made up of several transformer encoder layers. The encoder layer consists of multi-head self-attention (MSA) and multi-layer perceptron (MLP) in sequence with skip connections. The sequence of tokens z corresponding to each SU is processed by the SU-transformer network which consists of L_1 encoder layers. Each encoder layer is applied sequentially which consists of the following operations,

$$c^l = MSA(LN(z^{l-1})) + z^{l-1}, \tag{10}$$

$$z^l = MLP(LN(c^l)) + c^l \tag{11}$$

where LN is layer normalization [23], MSA is multi-head self attention [18] and MLP [24] consists of two dense layer separated by GeLU non-linearity. The self-attention aims to capture interactions among all N_t token embeddings. The self-attention is computed as

$$Z_{att} = \text{Softmax}\left(\frac{Q\,(K)^T}{\sqrt{d_k}}\right) \tag{12}$$

The output of the self-attention block is obtained by multiplying Z_{att} with the value matrix V. MSA allows the model to concurrently attend to information from various representation subspaces at different positions. The output of MSA block is computed as:

$$\text{MultiHead}(Q, K, V) = Concat(head_1, head_2 ... head_{h_d})W^O, \tag{13}$$

where

$$\text{head}_i = \text{Softmax}\left(\frac{QW^{Q_i}\left(KW^{K_i}\right)^T}{\sqrt{d_k/h_d}}\right)VW^{V_i} \tag{14}$$

where $W^{Q_i} \in R^{d\times d_q}$, $W^{K_i} \in R^{d\times d_k}$ and $W^{V_i} \in R^{d\times d_v}$, and $W^O \in R^{h_d d_v \times d}$, $d_v = d_k = d/h_d$. i denotes number of heads ranging from 1 through h_d and h_d being the total number of heads in MSA block.

C. Architecture of MASSFormer

In this section, we provide a brief introduction to the proposed MASSFormer method. The main objective of the proposed architecture is to learn the collective PU states by selectively extracting the most relevant features from the sensing data of all cooperating SUs. The proposed MASSFormer model is divided into two components i.e. the SU-transformer network and collaborative transformer network to predict PU states by extracting spatio-temporal features at individual SU-level and at group-level respectively. The proposed MASSFormer architecture is depicted in Fig. 2, where SU-transformer network is used to predict the SU-level PU states by extracting spatio-temporal features from the sequence of tokens extracted from a series of CMs of a s-th SU. Further, collaborative transformer network is used to model the output representations of all SUs to predict the PU states at the group-level. Attention from various levels of a transformer can extract contributing features from the participating SUs in a progressive manner.

To predict the PU states, tokens computed from samples of each SU are fed to SU-transformer network in different parallel pipelines. SU-transformer network consists of L_1 encoder layers, followed by sequence pooling as described in section III-D and MLP block to predict the PU states at SU-level. We concatenate the output feature representations K_s of each s-th SU.

$$G_t = K_1 \diamond K_2 ... \diamond K_s ... \diamond K_S \tag{15}$$

Further, max pooling operation is performed over the feature K_s of all SUs to get output representation G_t provided in equation (15). Instead of spatial pooling (like typical max pooling in CNNs, which pools over spatial dimensions), this operation pools over the dimension representing different SU and combines their predictions by selecting the maximum value for each position across SU dimension. The output features G_t is fed to collaborative transformer network which consists of L_2 transformer encoder layer, followed by sequence pooling and MLP block, to extract the group-level features from all the participating SU. Finally, the output features are fed to Softmax layer for predicting the activity states of PU.

D. Sequence pooling

For the prediction of class label, ViViT [20] forwards a learnable class token through the encoder layers and later to the MLP for classification. In contrast, we use sequence pooling, first proposed by Hassani *et. al.* [25] to extract representation. Sequence pooling, an attention-based method, assigns weights to sequential embeddings, transforming the input sequence into a vector representation. This involves assigning importance weights to the processed data from transformer encoder. The motivation is rooted in the dispersion of information across all tokens in the output sequence, necessitating aggregation by assigning suitable weights to each token. Sequence pooling involves the transformation of sequence as $T : \mathbb{R}^{N_t\times d} \mapsto \mathbb{R}^d$. Given the output $z \in \mathbb{R}^{N_t\times d}$ of encoder is passed to linear layer $g(z) \in \mathbb{R}^{N_t\times 1}$, and softmax activation to the token sequence as follows

$$Weigths = softmax(g(z)^T) \in \mathbb{R}^{1\times N_t} \tag{16}$$

Hence, we calculate importance weights for each input token as described in equation (16), and further, these weights are assigned to each token as described in equation (17). The computed weights are employed to adjust the contributions of the tokens through a weighting operation as follows

$$z_{seq} = Weigths \times z \in \mathbb{R}^{1\times d} \tag{17}$$

z_{seq} is passed to MLP block of the SU-transformer network and collaborative transformer network to detect activity states of PU at SU-level and group-level respectively.

E. Network training

The developed model is trained in two distinct stages. In the initial stage, the SU-transformer network is trained end-to-end using a training dataset comprising sequences of CMs converted into tokens and associated labels. This process aims to extract spatio-temporal features at SU-level. After the SU-transformer network gets trained, the output features, K_s, extracted for each SU during the initial stage are concatenated and further, a max pooling operation is applied to obtain pooled feature representation denoted as P_t. In the subsequent stage, the pooled features from multiple sensing periods are inputted into the collaborative transformer network which comprises various encoder layers, followed by sequence pooling and an MLP block.

The resulting output features are fed to the Softmax layer for the prediction of probabilities associated with hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. The u-th example in the set contains data samples from each SU is denoted by symbol Φ_u, where $\Phi_u = (\phi_u^1, \phi_u^2 ... \phi_u^S)$. Normalized output of the developed model is represented as $[j_{\theta/\mathcal{H}_0}(\mathbf{\Phi}_u), j_{\theta/\mathcal{H}_1}(\mathbf{\Phi}_u)]$, where $j_{\theta/\mathcal{H}_0}(\mathbf{\Phi}_u) + j_{\theta/\mathcal{H}_1}(\mathbf{\Phi}_u) = 1$. For a new data point $\mathbf{\Phi}_u$, the trained model predicts a probability associated to either Hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$. $j_{\theta/\mathcal{H}_i}(\mathbf{\Phi}_u)$ denotes the class probability of hypothesis $\mathcal{H}_i$. Based on this, the main goal of training is to maximize the likelihood as

$$\prod_{u=1}^{U/\lambda} \left(j_{\theta|\mathcal{H}_1}(\mathbf{\Phi}_u)\right)^{b_{u\lambda}} \left(j_{\theta|\mathcal{H}_0}(\mathbf{\Phi}_u)\right)^{1-b_{u\lambda}}. \tag{18}$$

This is equivalent to minimizing the cost function, a cross-entropy loss function, which is defined as logarithm of the likelihood function and normalizing it with the number of training points.

$$L(\theta) = -\frac{1}{U/\lambda}\sum_{u=1}^{U/\lambda} b_{u\lambda}\log j_{\theta|\mathcal{H}_1}(\mathbf{\Phi}_u) + (1-b_{u\lambda})\log j_{\theta|\mathcal{H}_0}(\mathbf{\Phi}_u) \tag{19}$$

To achieve the maximum likelihood, the cross-entropy loss function is minimized to obtain the optimal parameters as

$$\theta^* = \arg\min_{\theta} L(\theta). \tag{20}$$

While training the developed model, Back propagation (BP) algorithm is used to compute gradients of the loss function, and applying Adam optimizer to optimize the model parameters to get well-trained network. The output of trained model is denoted as $j_{\theta^*|\mathcal{H}_0}(\mathbf{\Phi}_u)$ and $j_{\theta^*|\mathcal{H}_1}(\mathbf{\Phi}_u)$. According to Neyman-Pearson (N-P) theorem, the optimal test statistics is likelihood ratio is represented as

$$\beta_{MASSFormer}(\mathbf{\Phi}_u) = \frac{j_{\theta|\mathcal{H}_1}(\mathbf{\Phi}_u)}{j_{\theta|\mathcal{H}_0}(\mathbf{\Phi}_u)}. \tag{21}$$

Upon the arrival of a new data point $\mathbf{\Phi}_u$, the model output is used to determine the activity states of PU. However, this approach does not provide control over the false alarm probability p_{fa}. Therefore, we calculate the detection threshold γ using Monte-Carlo method that ensures the desired false alarm probability p_{fa} is achieved. We collect the sequence of CMs corresponding to hypothesis $\mathcal{H}_0$ to build a set $\mathbf{\Phi}_{\mathcal{H}_0} = \{\tilde{\mathbf{\Phi}}_1, ..., \tilde{\mathbf{\Phi}}_{\tilde{Q}}\}$, where total examples in set $\mathbf{\Phi}_{\mathcal{H}_0}$ denoted by $\tilde{Q}$. Now, we fed the noisy data samples to the MASSFormer and get the expression of test statistics under $\mathcal{H}_0$ i.e. $\beta_{MASSFormer|\mathcal{H}_0}$. Further, reorganize the all these $\beta_{MASSFormer|\mathcal{H}_0}$ to construct a set $\tilde{\mathbf{\Phi}}_{MASSFormer|\mathcal{H}_0}$ in ascending order such as $\forall 1 \le p \le q \le \tilde{Q}$,

$$\beta_{MASSFormer|\mathcal{H}_0}\left(\tilde{\mathbf{\Phi}}_p\right) \le \beta_{MASSFormer|\mathcal{H}_0}\left(\tilde{\mathbf{\Phi}}_q\right) \tag{22}$$

Lastly, we compute the detection threshold with constraint of p_{fa} as

$$\gamma = \tilde{\mathbf{\Phi}}_{MASSFormer|\mathcal{H}_0}\left(round\left(\tilde{Q}\left(1-p_{fa}\right)\right)\right). \tag{23}$$

The function $round(.)$ provides the nearest integer value of a given number. For a test data, a decision about the activity states of PU can be made as follows

$$\gamma \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lessgtr}} \beta_{MASSFormer}(\mathbf{\Phi}_u). \tag{24}$$

IV. Numerical Results

A. Simulation environment and Hyperparameters

We consider the simulation area of 1000m × 1000m area, and they move with a velocity randomly chosen between [20, 25] m/sec, resulting in dynamic changes in their positions over time. The pause time is chosen as 1msec. This setup allows us to assess the feasibility of the proposed model under conditions of user mobility. We assume that PU signals are independent and identically distributed (i.i.d.) Gaussian Random vector with zero mean and signal variance σ_w^2. We have chosen values of $N = 100$, $S = 3$, and $M = 15$, length of temporal sequence $\lambda = 20$ randomly. PU signals are assumed to be sent on a channel having bandwidth B_W is 10MHz. Furthermore, we assume that $P_t = 200mW$, $\beta = 10^{3.453}$ and $\alpha = 3.8$, and noise power density is $N_0 = -150dBm/Hz$. We assumed that the noise signal represented by $\eta_s^m(n)$ and $\eta_c(n)$ follow an i.i.d Gaussian random vector with zero mean and noise variance σ_η^2. To address the challenges occurring due to user mobility in real-world scenarios, we assumed that PU signals reaching from multiple SUs to fusion centre may experience different path loss and fading severity. Consequently, we assumed that channel gains $h_s^m(n)$ and $h_{s,r}^m(n)$ follow Rayleigh distributions and fading severity is modeled by adjusting the fading parameter. The received signal power is calculated using equation (4), which incorporates path-loss between PU and each SU. The hyperparameters used for our proposed architecture are provided in Table I. The number of training and testing points used for training and evaluating the proposed models are 104, 000 and 15, 000. The number of epochs and batch size are considered to be 100 and 16.

TABLE I: HYPER-PARAMETERS SETTINGS

Parameter settings		
Input shape:- Dimension (20,16,16,3)		
Patch/token size: Dimension (20,1,1)		
Parameter description	SU-Tranformer Network	Collaborative Transformer Network
3D convoultion	Kernel size:24@(20,1,1) stride: Patch size	-
Linear layer projection of input		
Projection dimention	24	24
No. of Heads	4	4
No. of encoder layers	5	4
Transformer units	[48,24]	[48,24]
MLP head units	[128,64]	[128,64]
Learning Rate	$1e^{-5}$	$1e^{-5}$
Total parameters	42667	31747

TABLE II: FLOPs for MASSFormer

Layers Description	FLOPs (SU-Transformer)	FLOPs (Collaborative Transformer)
Patch Embedding	245,760	576
MSA	1,376,256	1,376,256
MLP in Encoder	1,179,648	1,179,648
Layer Normalization	24,576	24,576
Total FLOPs for Encoder Layer	5 × (1,376,256 + 1,179,648 + 24,576)	4 × (1,376,256 + 1,179,648 + 24,576)
Sequence Pooling	12,864	12,864
MLP Head	11,392	11,392
Total FLOPs	13,172,416	10,346,752
MASSFormer	Params: 74,414 FLOPs: 23,519,168	

TABLE III: FLOPs for CNN-LSTM

Layers Description	Parameters value	FLOPs
Input:	(20,16,16,3)	
Conv. Layer	Kernel: 32 @(3,3)	8,847,360
Max-pooling layer	Kernel (2,2) stride:(1,1)	-
Global Avg. Pooling	-	-
Dense layer 1	dimension: 128	163,840
LSTM	Units:32	819,200
Dense layer 2	Unit: 2	128
Total FLOPs and Params	Total Params: 25,794	9,830,400

TABLE IV: FLOPs for 3D CNN

Layers Description	Parameters value	FLOPs
Input:	(20,16,16,3)	
Conv. Layer	Kernel: 32@(3,3,3)	26,542,080
Max-pooling layer	Kernel (2,2,2) stride:(1,1,1)	-
Conv. Layer	Kernel: 24 @(3,3,3)	177,292,800
Max-pooling layer	Kernel (2,2,2) stride:(1,1,1)	-
Global Avg. Pooling	-	-
Dense layer 1	dimension: 64	1536
Dense layer 2	Unit: 2	128
Total FLOPs	Total Params: 25,114	203,836,544

B. Simulation Results

We conducted comprehensive simulations to illustrate the efficacy of the proposed MASSFormer method. We evaluated the detection performance in terms of receiver operating characteristics (ROC) curve, probability of detection (P_d) vs noise power density (N_0), and sensing error vs noise power density. We compared and analyzed the performance of MASSFormer method with existing methods such as CNN-LSTM [16], APASS [9], and 3D CNN [21] methods. We rigorously evaluated the performance of the proposed method under various scenarios, including imperfect sensing as well as both perfect and imperfect reporting channels, to thoroughly test its robustness. We calculated the P_d values at different detection thresholds computed at different values of p_{fa} to evaluate the detection performance. Fig. 3 shows the ROC curve of the proposed MASSFormer method along with existing methods at two different noise power density values such as $N_0 = -150dBm/Hz$ and $N_0 = -145dBm/Hz$.

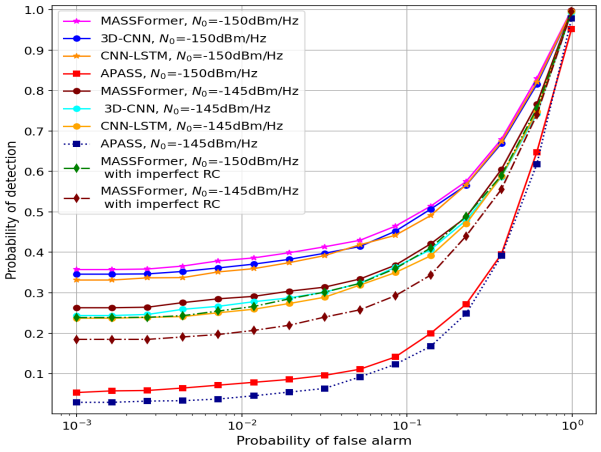


Fig. 3: ROC curve for various SS methods

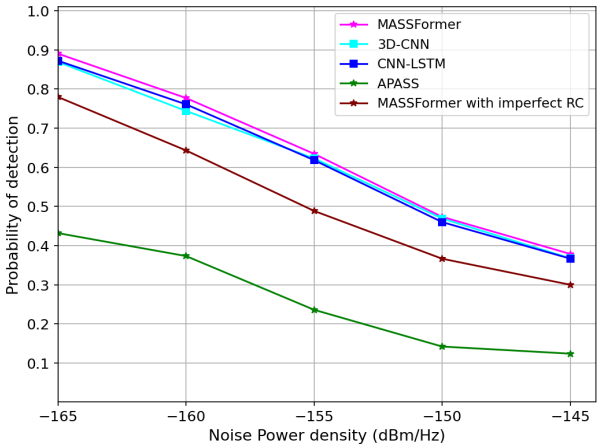


Fig. 4: Probability of detection vs noise power density (p_{fa} = 0.09)

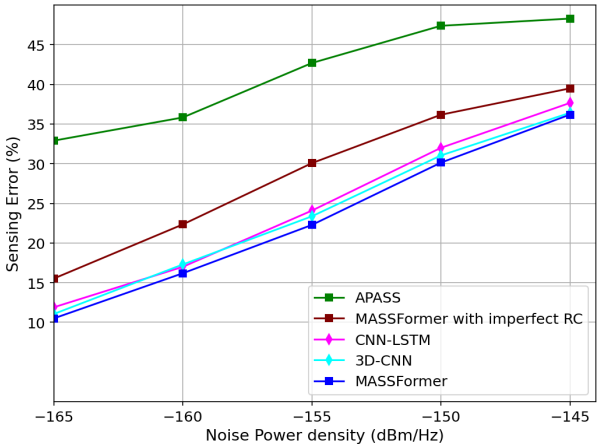


Fig. 5: Sensing error vs noise power density

The results show that the proposed MASSFormer method performs better as compared to methods CNN-LSTM, APASS, and 3D CNN in terms of detection performance. In accordance with IEEE 802.22 standard, the maximum acceptable value of false alarm probability is ($P_f \leq 0.1$). Therefore, we provide a plot of P_d versus noise power density at a fixed value of false alarm probability P_f = 0.09 in Fig. 4. From Fig. 3 and Fig. 4, we observe that the detection probability decreases as the noise power density increases. The sensing error, which is defined as the mean of the probability of miss detection and probability of false alarm, another performance metric, is used to evaluate the performance of the proposed method.

Fig. 5 depicts a plot of sensing error vs noise power density for different methods. We observe that the sensing

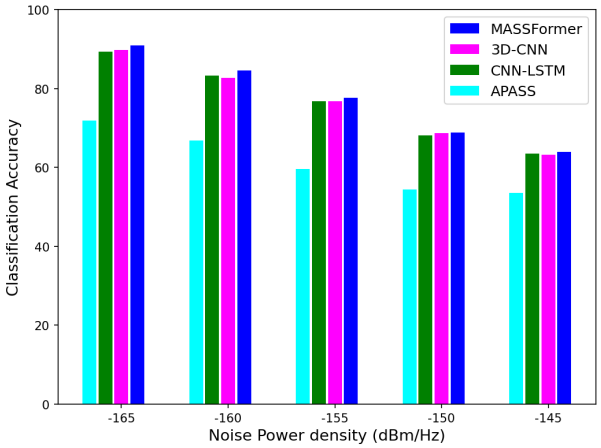


Fig. 6: Classification Accuracy vs noise power density

error increases with increasing noise power density, reflecting inaccuracies in individual SUs' sensing outcomes. After analysis, we find that the proposed method achieves lower sensing error than the existing methods, highlighting its benefits. Fig. 6 depicts the plot of classification accuracy vs noise power density for different methods. After a comprehensive analysis of the results, it is evident that the proposed MASSFormer method outperforms the existing methods, exhibiting superior detection performance in terms of detection probability and classification accuracy.

V. Computational Complexity Analysis

Although the proposed method achieves improved detection performance, the computational complexity of detection methods should be analyzed for a fair comparison. Since training of neural networks can be performed offline, the paper focused on the computational complexity of inference after the model is deployed. The computational complexity to convert the sensing signal $\boldsymbol{Y}_u^s$ into the CM $\boldsymbol{R}_u^s$ is $O(M^2N)$. Let l and L_{CNN} represent present convolutional layer's index and total number of convolutional layers. We define the computational complexity of the convolutional layer as $O\left(\sum_{l=1}^{L_{CNN}} n_{f,l-1} \cdot n_{s,l}^2 \cdot n_{f,l} \cdot m_{s,l}^2\right)$, where $n_{f,l}$ and $n_{f,l-1}$ show the convolutional kernels count of lth and $(l-1)$th layer. We define the spatial size of kernel current layer and the resulting output feature map as $n_{s,l}$ and $m_{s,l}$ respectively. With the CNN stride to be 1, we have $m_{s,1} = M$ and the total number of SUs becomes the number of input channel i.e. $n_{f,0} = S$. The complexity of single convolutional layer is increased by the temporal length λ since every sample has λ CMs in the dataset. The complexity of LSTM network is denoted as $O\left(4\lambda n_l\left(n_{f,1} + n_l\right)\right)$, where the dimension of the each gate of LSTM cell are represented by n_l. Therefore, complexty of CNN-LSTM method, which consists of Convolutional layers, Max-pooling layers, Global average pooling layer, and two dense layers, is provided below

$$O\left(\lambda M^2 n_{f,1} n_{s,1}^2 S + 4\lambda n_{l1}(n_{f,1} + n_l) + n_{f,1} n_{fc1} + n_l n_{fc2}\right) \tag{25}$$

where, n_{fc1} and n_{fc2} denotes the dimensions of the first and second dense layer respectively, The computational complexity of 3-D CNN is computed as below

$$O\left(\lambda M^2 (n_{f,1} m_{s,1}^3 S + n_{f,1} n_{f,2} m_{s,2}^3) + n_{f,2} d_{fc1} + d_{fc1} d_{fc2}\right) \tag{26}$$

where $n_{f,1}$, and $n_{f,2}$ denotes the number of filters of first and second convolutional layer respectively. We define the spatial size of the filter in the first layer and second as $m_{s,1}$ and $m_{s,2}$, respectively. d_{fc1} and d_{fc2} denote the dimensions of first and second dense layer respectively.

The computational complexity of the proposed MASSFormer method is

$$O\left(L_1 h_{att1} M^2 \lambda d_{emb1} + L_1 M d_{fc1}\right) + O\left(L_2 h_{att2} M^2 \lambda d_{emb2} + L_2 M d_{fc2}\right) \tag{27}$$

where L_1 and L_2 denote the number of transformer layers, h_{att1} and h_{att2} number of attention heads in SU-transformer and collaborative transformer network respectively. d_{emb1} and d_{emb2} denote the projection dimensions, d_{fc1} and d_{fc2} denote the dimension of the dense layer of MLP block in SU-transformer and collaborative transformer network respectively.

The computational complexity of these methods is also computed in terms of multiply accumulates (MACs) or floating point operations (FLOPs) as provided in Table II, III, and IV. From the analysis, it is concluded that the FLOPs required for the proposed MASSFormer method are less as compared to the 3D CNN method and higher than the CNN-LSTM method. We measured the time needed for data preparation and model inference. We calculated the time for preprocessing the data for 1000 samples, where each sample comprises of $\lambda = 20$ CMs. After averaging, the time required for one data sample is 0.099 msec. For model inference, the CNN-LSTM, 3-D CNN, and APASS detector had inference times of 0.17, 0.65, and 0.87 msec, respectively. Our proposed MASSFormer method requires 2.46 msec for inference. According to the IEEE 802.22 standards, SUs are required to evacuate the spectrum within 2 seconds when a PU becomes active. Therefore, our model detects the PU state within 2.46 msec satisfying the real-time latency constraint.

VI. Conclusion

In this work, we proposed a MASSFormer method to predict the PU states in mobile scenarios. Since the prediction of PU state at SU-level and group-level, both events are occurring simultaneously over time. Therefore, inspired by this, a model is developed that uses SU-transformer network to predict PU states at SU-level and collaborative transformer network to predict PU states at group-level by modeling the spatio-temoral dynamics of movements of all contributing SUs. From simulations results, it is evident that our MASSFormer method outperforms existing methods in terms of P_d demonstrating superior sensing performance.

REFERENCES

[1] S. Kumar, "Performance of ed based spectrum sensing over α–η–μ fading channel," *Wireless Personal Communications*, vol. 100, no. 4, pp. 1845–1857, 2018.

[2] K. M. Thilina, K. W. Choi, N. Saquib, and E. Hossain, "Machine learning techniques for cooperative spectrum sensing in cognitive radio networks," *IEEE Journal on selected areas in communications*, vol. 31, no. 11, pp. 2209–2221, 2013.

[3] D. Janu, K. Singh, and S. Kumar, "Machine learning for cooperative spectrum sensing and sharing: A survey," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 1, p. e4352, 2022.

[4] C. Liu, X. Liu, and Y.-C. Liang, "Deep cnn for spectrum sensing in cognitive radio," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[5] C. Liu, J. Wang, X. Liu, and Y.-C. Liang, "Deep cm-cnn for spectrum sensing in cognitive radio," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2306–2321, 2019.

[6] Z. Chen, Y.-Q. Xu, H. Wang, and D. Guo, "Deep stft-cnn for spectrum sensing in cognitive radio," *IEEE Communications Letters*, vol. 25, no. 3, pp. 864–868, 2020.

[7] D. Janu, S. Kumar, and K. Singh, "A graph convolution network based adaptive cooperative spectrum sensing in cognitive radio network," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 2269–2279, 2022.

[8] A. Mehrabian, M. Sabbaghian, and H. Yanikomeroglu, "Cnn-based detector for spectrum sensing with general noise models," *IEEE Transactions on Wireless Communications*, vol. 22, no. 2, pp. 1235–1249, 2022.

[9] J. Xie, C. Liu, Y.-C. Liang, and J. Fang, "Activity pattern aware spectrum sensing: A cnn-based deep learning approach," *IEEE Communications Letters*, vol. 23, no. 6, pp. 1025–1028, 2019.

[10] B. Soni, D. K. Patel, and M. Lopez-Benitez, "Long short-term memory based spectrum sensing scheme for cognitive radio using primary activity statistics," *IEEE Access*, vol. 8, pp. 97 437–97 451, 2020.

[11] W. Chen, H. Wu, and S. Ren, "Cm-lstm based spectrum sensing," *Sensors*, vol. 22, no. 6, p. 2286, 2022.

[12] L. Yu, J. Chen, G. Ding, Y. Tu, J. Yang, and J. Sun, "Spectrum prediction based on taguchi method in deep learning with long short-term memory," *IEEE Access*, vol. 6, pp. 45 923–45 933, 2018.

[13] J. Gao, X. Yi, C. Zhong, X. Chen, and Z. Zhang, "Deep learning for spectrum sensing," *IEEE Wireless Communications Letters*, vol. 8, no. 6, pp. 1727–1730, 2019.

[14] K. Yang, Z. Huang, X. Wang, and X. Li, "A blind spectrum sensing method based on deep learning," *Sensors*, vol. 19, no. 10, p. 2270, 2019.

[15] D. Ke, Z. Huang, X. Wang, and X. Li, "Blind detection techniques for non-cooperative communication signals based on deep learning," *IEEE Access*, vol. 7, pp. 89 218–89 225, 2019.

[16] J. Xie, J. Fang, C. Liu, and X. Li, "Deep learning-based spectrum sensing in cognitive radio: A cnn-lstm approach," *IEEE Communications Letters*, vol. 24, no. 10, pp. 2196–2200, 2020.

[17] S. Solanki, V. Dehalwar, and J. Choudhary, "Deep learning for spectrum sensing in cognitive radio," *Symmetry*, vol. 13, no. 1, p. 147, 2021.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[22] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, no. 5, pp. 483–502, 2002.

[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[24] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[25] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.



Dimpal Janu Dimpal Janu received her B.Tech. degree in Electronics and Communication Engineering from Govt. Mahila engineering College, Ajmer, India in 2014, and M.Tech. degree in Electronics and Communication Engineering from Malaviya national institute of technology (MNIT), Jaipur, India in 2018. She is currently pursuing a Ph.D. degree from MNIT Jaipur. Her research interests include wireless communication, cognitive radio network, machine learning, and deep learning. She is a member of the IEEE student branch.



Sandeep Mandia received the MTech degree from National Institute of Technology, Silchar, India in 2014 and the PhD degree in computer vision from Malaviya National Institute of Technology Jaipur, India in 2023. He is currently serving as an Assistant Professor at Thapar Institute of Engineering and Technology, Patiala, India. His research interests are machine/ deep learning applications in student engagement analysis, medical diagnostics, and beyond.



Kuldeep Singh received his MTech degree in Signal Processing from Delhi University in 2006 and a Ph.D. degree in Computer Vision from Delhi Technological University, India, in 2016. He is currently an Associate Professor with the Department of Electronics & Communication Engineering, Malaviya National Institute of Technology, Jaipur, India. Previously, he was a Senior Scientist with the Central Research Lab, Bharat Electronics Ltd., India. He also worked as a postdoctoral fellow at the University of Alberta, Canada from October 2017 to April 2018. His research interest includes Computer Vision, Machine/ Deep Learning, Biometrics, and Cyber Security. He is a reviewer of various IEEE transactions, Elsevier, and Springer journals.



Sandeep Kumar received his B. Tech. in electronics and communication from Kurukshetra University, India in 2004 and Master of Engineering in Electronics and Communication from Thapar University, Patiala, India in 2007. He received his Ph.D. from Delhi Technological University, Delhi, India in 2018. He is currently working as Member (Senior Research Staff) at Central Research Laboratory, Bharat Electronics Limited Ghaziabad, India. He has received various awards and certificates of appreciation for his research activities. His research interests include the study of wireless channels, performance modeling of fading channels, and cognitive radio networks. He is also serving as a reviewer for IEEE, Elsevier, and Springer journals.

MĀPAN-Journal of Metrology Society of India
https://doi.org/10.1007/s12647-024-00771-0

ORIGINAL PAPER

Measurement of Heavy Metals Content in Suspended Sediment of Ganges River Using Atomic Absorption Spectrometry

M. Aggarwal[1,2]*, S. Anbukumar[1] and T. Vijaya Kumar[1]

[1]Department of Civil Engineering, Delhi Technological University, Delhi, India

[2]Galgotias College of Engineering and Technology, Greater Noida, Uttar Pradesh, India

Received: 05 June 2024 / Accepted: 25 August 2024

© Metrology Society of India 2024

Abstract: This investigation aimed to evaluate the severity of heavy metal concentrations in suspended sediments along a 225 km section of the Ganges River in the eastern part of Uttar Pradesh, India. Metal concentrations were measured via Atomic Absorption Spectrometry, with averages recorded as Fe (50230 µg/g), Mn (1283 µg/g), Cr (201 µg/g), Zn (183 µg/g), Pb (55.1 µg/g), Ni (50.9 µg/g), Cu (42.8 µg/g), and Cd (1.01 µg/g). Zn, Fe, Cr and Pb showed anthropogenic contribution at all selected locations, while anthropogenic inputs of Mn and Ni were also observed at few locations. Risk evaluation was conducted using various pollution indicators. The Contamination Factor (CF) revealed moderate contamination (CF between 1 and 3) by Cr. The Geoaccumulation Index (GI) indicated a stage of uncontaminated to moderate contamination (GI between 0 and 1) for Cr at upstream locations. The values of Enrichment Ratios (ER) indicated minor enrichment (ER between 1 and 3) of Cr at all locations. The Pollution Load Index values ranged from 0.72 to 0.86, suggesting that the selected section is overall uncontaminated. The study highlights significant but localized anthropogenic impacts on heavy metal concentrations, particularly for chromium. While the overall assessment suggests that the suspended sediment in the studied stretch are not heavily contaminated, the elevated concentrations of iron, manganese, chromium, and nickel pose significant ecological risks. The study underscores the need for targeted pollution control measures, especially for metals with high anthropogenic contributions, to ensure the long-term health and safety of this vital water resource.

Keywords: Atomic absorption spectrometry; Heavy metals measurement; Ganges River; Risk evaluation

1. Introduction

Though various investigations are been conducted to identify the heavy metal concentrations in the water and riverbed sediment of Ganges River but this study focuses on suspended sediment of the Ganges River. The "Ganges" is India's biggest river, stretching 2525 km from the Gangotri Glacier till the Bay of Bengal supplying nearly 400 million individuals. Industrial discharge is a basic cause of pollution in the Ganges [1, 2].

The Ganges River basin, encompassing over 220 cities, generates a sewage volume of 8.25 billion liters per day (BLD). However, the treatment facilities are inadequate, with only 3.5 BLD of sewage being treated. Consequently, 2.46 BLD of untreated sewage is directly released in the Ganges River, and 4.51 BLD is released in its tributaries. The leftover sewage is disposed of on the ground surface, where it percolates and contaminates surface and groundwater downstream [2]. Since 1985, the Indian government has implemented numerous measures to address sewage discharge into the Ganges River. However, In the past few years, the amount of untreated wastewater entering the river has gone up [3]. In addition to sewage originating from towns and cities, industries such as distilleries, power plants, textiles, electroplating, paper and pulp and sugar mills generate an additional 2.5 BLD of sewage. This industrial waste is thrown straight into the Ganges [4]. Another research [5] indicates that the central stretch of the Ganges, particularly across Kannauj to Varanasi, receives a significant portion of industrial discharge, predominantly from tanneries. The Ganges River traverses densely populated regions with intense human activities, leading to metal contamination in its sediment. These

*Corresponding author, E-mail: mohitjvg@gmail.com

Published online: 11 September 2024

metals within the sediments of the river are subsequently transferred into the ocean by the flowing water in the river. Heavy metals, known for their high toxicity, persistence, and propensity to accumulate in aquatic environments, pose a severe threat to marine life when rivers become contaminated by them[6–8]. The primary causes of metal pollution in aquatic elements are atmospheric deposition [9], dumping of untreated sewage, erosion of soil [10], usage of fertilizers and pesticides comprising metals [11], surface runoff [12] and various chemical substances from industries, urban areas and farming activities [13, 14].

The Ganga River geomorphology in the study area is defined by the Upland Terrace Surface (T2). This surface, also known as Older Alluvium or Bangar, has a gentle southeast slope (5–10 cm/km) and consists of oxidized sediments showing early soil development like calcretization and ferrugenization. Formed 74 to 128 thousand years ago, T2 features varied micro-geomorphology and centripetal drainage. It acts as a base for newer geomorphic surfaces, with areas of both erosion and sedimentation and remains unaffected by major floods. T2 is mainly composed of fine sand-silt and clay with distinct mottling and calcrete development. Sediment transfer during monsoon occurs through sheet flow, attributed to interfluve sedimentation processes [15, 16]. The geology of the study area reflects a complex history of sedimentation, tectonic activity, and resource development. While the region benefits from fertile alluvial soils and valuable mineral resources, it also faces challenges related to groundwater management and environmental sustainability in the face of rapid urbanization and industrial growth. Understanding these geological aspects is crucial for effective land-use planning, resource management, and environmental conservation efforts in the Kanpur region. The study area faces significant environmental challenges, primarily related to air and water pollution, waste management, and the impact of rapid urbanization. While efforts are being made to mitigate these issues, ongoing and enhanced measures are necessary to improve the environmental quality in these cities.

In aquatic environments, heavy metals exist in dissolved, suspended and settled states [17–21]. While heavy metals during their dissolved state are highly poisonous to both human beings and water creatures, however, their concentration is often lower relative to that in suspension [22]. Metals in dissolved state readily bind to suspended sediment because suspended sediment are extremely volatile and has a huge contact area [21, 23]. Therefore, metals in suspension demand more focus than those in the dissolved state. Previous investigations suggest that riverbed sediment constitutes just ten percent or less of the entire sediments that reach river-sea junctions, while more than 90% of the sediment reaches in the suspended state [24, 25]. Metals present in the water of rivers are prone to adsorption onto suspended sediment throughout the river because of their immediate contact with the water. Subsequently, the deposition of these sediments, which contain metals, represents the primary process for the accumulation of metals in riverbed sediment [23]. These contaminated riverbed sediment are capable of being resuspended due to disturbances in the water flow [21, 23]. This mechanism is crucial for the interaction between water and sediment, particularly regarding the environmental concerns pertaining to heavy metals. Consequently, numerous investigations have been performed for the analysis of heavy metals in suspended sediment, impacts on fluvial environments and subsequent implications for oceanic systems [17, 19, 21, 23, 25–35]. A research investigation of the Gulf of Tunis demonstrates contamination of Fe, Zn, Pb and Cu, is predominantly caused by industrial and fishing activities [17]. The findings from the Tigris River investigation demonstrated that numerous metals dominated the dissolved state, although Ti, Fe, Al, Th and Pb exhibited larger amounts in suspended because of increased concentrations of suspended sediment [19]. In a study [27], a fresh database on the chemical investigations comprising heavy metals is presented for the suspended sediment of rivers worldwide. This dataset captured changes in concentrations of heavy metals and provided a “snapshot” of the fluctuations in heavy metals across several continents. This comprehensive assessment aimed to assess the impact of human operations on the natural isotopic heavy metal cycle in diverse environments. Over two years of continuous monitoring, data on granular including Pb, Cd, Cu, Ni, Cr, Zn, Fe and Mn exhibited no discernible variations between seasons or correlations with tides. Instead, the data suggested a significant influx of metals stemming from intensive farming operations in the river’s catchment [36]. A research was undertaken on the Lerma River [37], and it was found that particulate Mn and Fe primarily originated from sources of nature. Conversely, Cr, Pb, Cu and Zn present in suspended sediment were mostly obtained from lithogenic processes. The supervision of catchments and restoration efforts necessitate an understanding of levels of pollution by heavy metals as well as accompanying environmental risks. To assess the level of pollution, various methods have been developed, including the pollution load index, enrichment ratio, geoaccumulation index and contamination factor. These approaches help estimate metal enrichments, contamination levels, bioavailability, anthropogenic contributions and the toxicity of heavy metals in river sediments [38].

Several research’s [3–5, 39–46] have performed heavy metals investigations for the Ganges and its associated streams, focusing primarily on river water and riverbed sediment. However, these studies did not determine the

pollution levels by heavy metals in the suspended sediment of Ganges. This gap in research leaves a significant area unexplored, as suspended sediments have a significant part in transporting pollutants downstream & affecting water quality throughout the river basin. This inspires us to focus on the suspended sediments of the Ganges River, as illustrated by a concept picture in Fig. 1. Numerous researchers assessed the pollution of heavy metal in the Ganges water [5, 43–45] and riverbed sediments [40, 43, 44, 46] and revealed significant pollution, particularly in the river's middle stretch. Investigating heavy metals in suspended sediments could be highly beneficial for evaluating downstream contamination caused by transported sediments. This investigation focuses on a 225-km central section of the Ganges River from Kanpur to Prayagraj in the eastern part of U.P., India. This area is significant due to its industrial, agricultural, social and religious activities. The Jajmau area, located downstream of Kanpur, is an industrial hub with facilities like tanneries, electroplating, paint factories and leather processing, all of which use heavy metals in their operations. Consequently, from these industries untreated sewage containing heavy metals is often discharged into the Ganges. Such metals after getting adsorbed with suspended sediment transport downwards and reach Prayagraj, where individuals engage in religious bathing and other spiritual activities on the banks of the Ganges. Given the contamination risks, this 225 km stretch was selected for investigation to develop the required corrective steps to protect Ganges.

This study is intended to investigate the concentrations of Pb, Cd, Cu, Ni, Cr, Zn, Fe and Mn metals and to assess their geographical spread in the suspended sediment of Ganges River. Primary benefits of this research include:

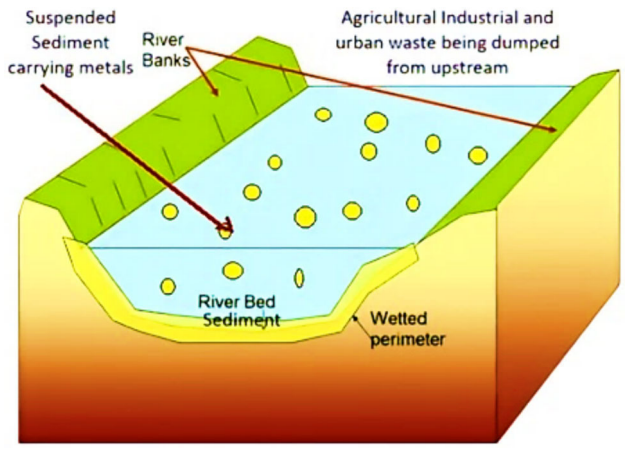


Fig. 1 Concept picture demonstrating the downward movement and settling of suspended sediment as riverbed sediment with the flow of the river

i. Determination of concentrations of metals in suspended sediment collected along the chosen stretch of Ganges.
ii. Classification of sources that resulted in metal concentrations in the chosen stretch.
iii. Study of the geographical spread of concentrations of metals along the chosen stretch of Ganges.
iv. Risk evaluation of contamination by heavy metal using various pollution indices like pollution load index (PLI), enrichment ratio (ER), geoaccumulation index (GI) and contamination factor (CF).

2. Methodology

The overall methodology for the present study is outlined with the help of a flow diagram (Fig. 2).

2.1. Sampling Points and Procedure for Collection & Analysis of Specimen

As river water contains maximum suspended sediment load during the rainy season due to high water flow and turbulence in the river, so the sampling was carried out during peak monsoon in July 2022 to obtain enough suspended sediment for heavy metal analysis. Since the river's discharge fluctuates throughout the day due to varying water consumption at different times, all samples were collected between 1:00 PM and 4:00 PM from the same depth of water to minimize these variations. Samples were taken from a 225-km section of the Ganges, commencing from Prayagraj till Kanpur in the eastern part of U.P., India. This complete chosen section lies within the central part of the Ganges River Basin, with a uniform lithography composed of alluvium [16]. The Upper Ganges River (UGR) extends from Haridwar in Uttarakhand to Prayagraj in Uttar Pradesh. This span of the river is characterized by sediment and water originating from theHimalayas. Onsite photographs of sampling points are displayed in Fig. 3. Sampling points have been chosen across the main course of the Ganges and the suspended sediment samples were retrieved by collecting ten water samples from ten distinct sites, mentioned in Table 1 and located in Fig. 4.

To obtain sufficient amount of samples of suspended sediment, twenty liters of river water samples were drawn from the mid-section. The water was initially collected in sterilized plastic vessels to avoid contamination and ensure accurate measurements. Each vessel was properly labelled each with the date, time, location**.** Over some time, the water samples were filtered through a pre-washed 0.2 µm MCE membrane disc filter. The suspended sediment retained on the filter was rinsed off with demineralized water. The moisture in the sediment was then removed

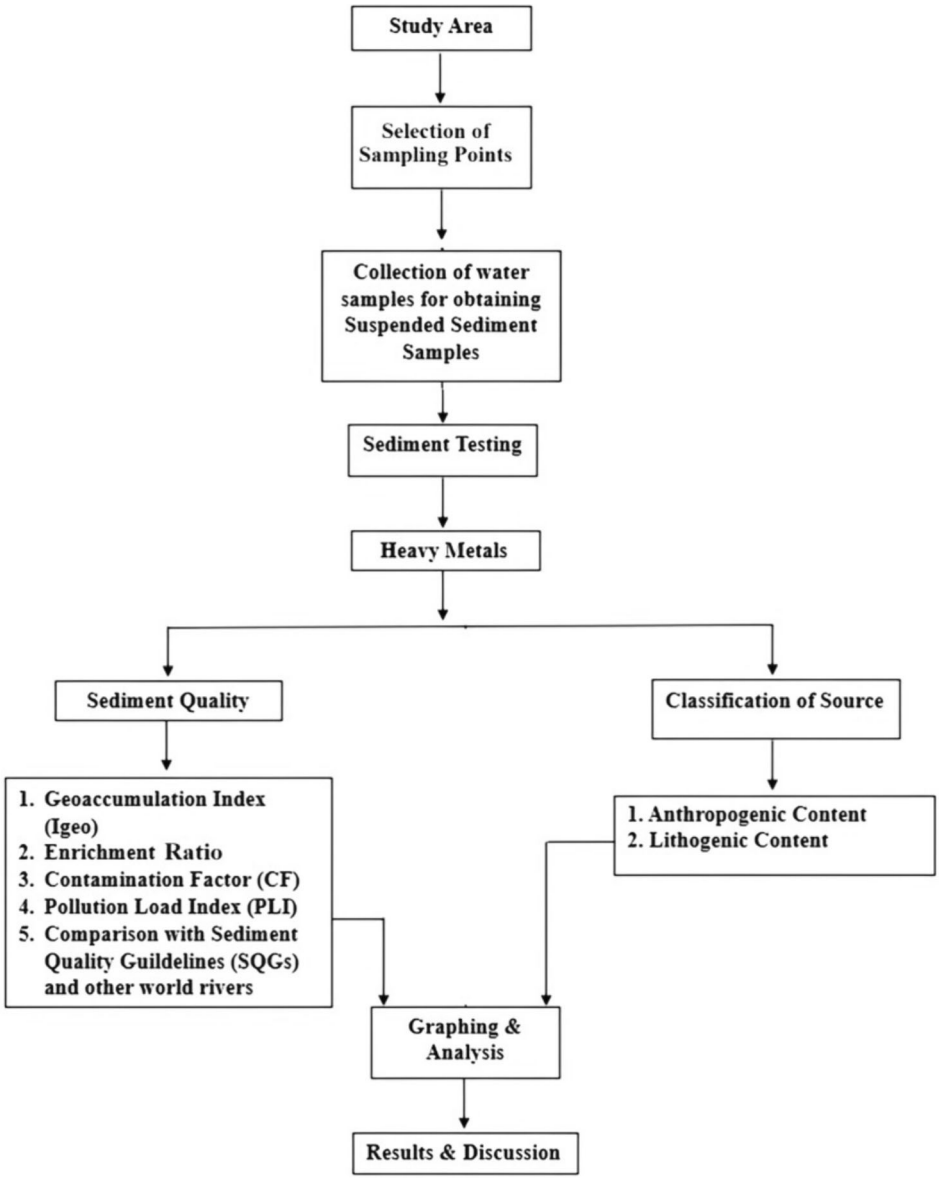


Fig. 2 Flow Diagram showing overall methodology

Table 1 Selected sampling points, their codes, GPS coordinates and distance on the selected stretch

Sampling point	Code	GIS location	Distance (km)
Jajmau Ganga Bridge	SP1	26°26′17′′N 80°27′43′′E	0
Village Maharajpur	SP2	26°22′42′′N 80°29′26′′E	24
Village Sarsaul	SP3	26°18′13′′N 80°32′35′′E	44
Dalmau Ganges Bridge	SP4	26°03′18′′N 81°01′56′′E	89
Unchahar Town	SP5	25°53′22′′N 81°12′59′′E	112
Ganges Ghat, Nawabganj Ghat	SP6	25°49′06′′N 81°19′46′′E	134
Ganges Ghat, Manikpur	SP7	25°46′03′′N 81°23′58′′E	143
Ganges Ghaat, Shringverpur Ghaat	SP8	25°35′20′′N 81°38′01′′E	181
Ganges Ghat, Phaphamau	SP9	25°30′24′′N 81°51′59′′E	214
Sangam, Prayagraj	SP10	25°25′41′′N 81°53′20′′E	225



Fig. 3 Onsite Field Photographs of the various selected sampling points

with the help of an oven set at 40 degrees Celsius. The mass of the resulting solids was measured, and the heavy metals concentrations were subsequently determined.

There are various methods to measure metal concentrations in river sediment like Atomic Absorption Spectroscopy (AAS), Inductively Coupled Plasma Mass Spectrometry (ICP-MS) and X-ray Fluorescence (XRF). AAS has several advantages over ICP-MS and XRF in specific applications. AAS is generally more cost-effective, with lower instrument and operational costs compared to ICP-MS, making it accessible for laboratories with limited budgets. It offers high selectivity, using element-specific lamps that reduce spectral interferences, which is beneficial when analyzing complex samples. AAS is also simpler to operate, requiring less technical expertise, and involves straightforward sample preparation, making it ideal for routine analysis.

Compared to XRF, AAS provides higher sensitivity for trace metal detection, especially in liquid samples. While XRF is effective for solid samples and can measure multiple elements simultaneously, AAS excels in detecting very low concentrations of specific metals with high accuracy. Additionally, AAS typically requires smaller sample volumes and experiences minimal matrix effects, which can be challenging in both ICP-MS and XRF.

Overall, AAS is particularly advantageous for laboratories focused on the precise and selective analysis of a few specific metals, offering a balance of sensitivity, simplicity, and cost-efficiency that makes it a preferred choice for routine and regulatory testing.

Nearly two grams of oven-dried suspended sediment specimen was ground and sieved via Whatman filter media 42. For digesting the sediment, a mixture of HCl, HF, $HClO_4$ and HNO_3 was introduced at extreme pressures and temperatures. The concentrations of heavy metals namely Ni, Cd, Cu, Fe, Cr, Zn, Pb, and Mn were measured in the digested samples using atomic absorption spectroscopy (AAS) (Make: Shimadzu, Model: AA-6300) (Refer to Fig. 5). Technical details of AAS AA-6300 are given in Table 2. AA-6300 is an atomic absorption spectrophotometer (AAS) commonly used in laboratories for analyzing the concentration of elements in samples. In this technique concentration of metals is determined by measuring the absorption characteristics of radiation by the atomic vapour of the element. The major source of radiation is a hollow cathode lamp. Standard solutions in optimum range were prepared for each heavy metal (Cd, Cr,

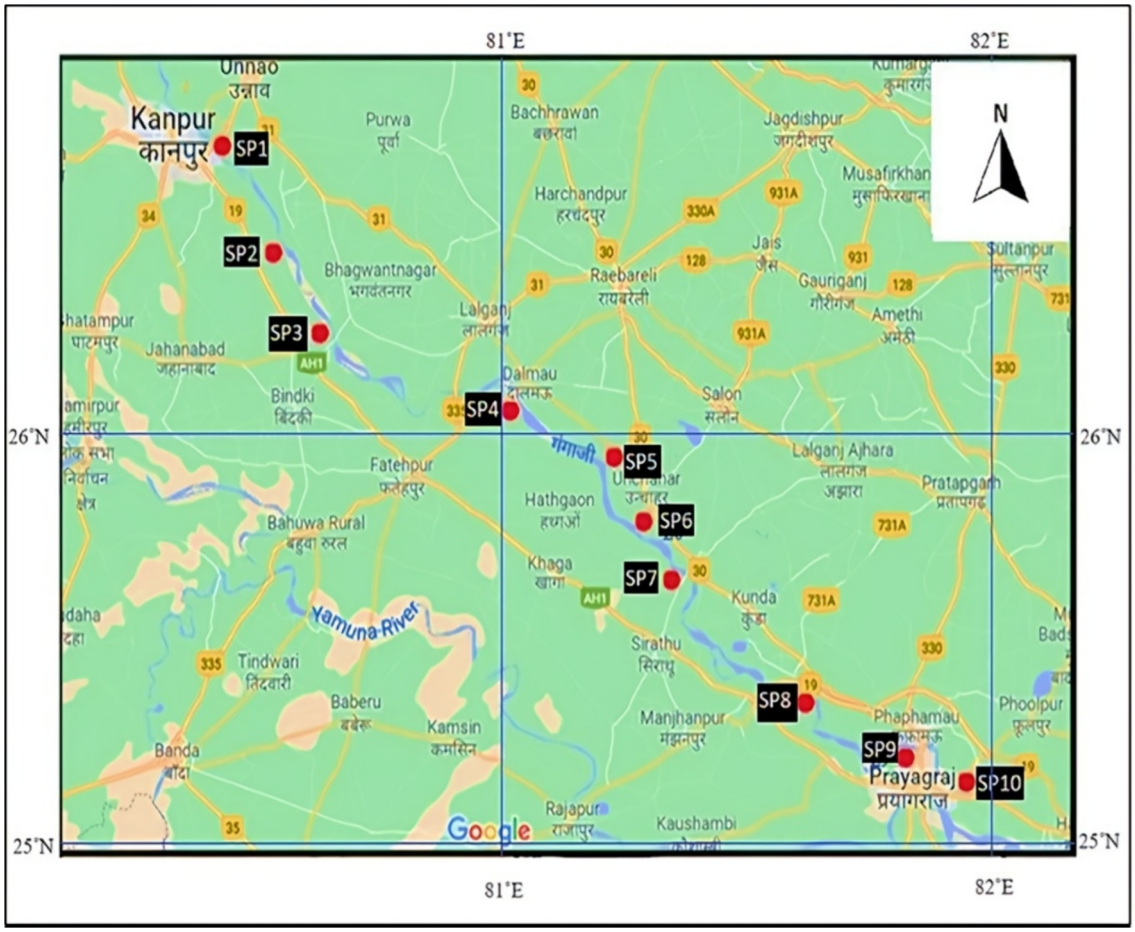


Fig. 4 Sampling Points along the selected stretch of Ganges River in eastern part of U.P., India

Cu, Fe, Ni, Pb, Zn) separately by their stock solutions. After that each standard was aspirated in turn into flame and absorbance was recorded. Calibration curve between absorbance and concentrations of standards were plotted and concentration of heavy metals in sediment samples was determined by using at respective wavelengths at 326.1 nm, 425.4 nm, 327.4 nm, 372 nm, 403.1 nm, 341.5 nm, 405.8 nm and 213.9 nm for cadmium, chromium, copper, iron, manganese, nickel, lead and zinc respectively.

When measuring metal concentrations using AAS, several sources of uncertainty can affect the accuracy and reliability of the results. Understanding and managing these uncertainties is crucial for obtaining precise and reliable data. Below are the primary sources of uncertainties in AAS measurements:

1 Instrumental Uncertainties

- Wavelength Calibration: Small errors in the calibration of the spectrometer's wavelength can lead to inaccuracies in the absorbance measurement.
- Lamp Intensity Fluctuations: Variations in the intensity of the hollow cathode lamp or other light sources can cause fluctuations in the absorbance signal.
- Detector Noise: Electronic noise in the detector can introduce random errors in the measurement of absorbance.
- Baseline Drift: Changes in the baseline absorbance due to temperature variations or other environmental factors can lead to inaccuracies.

2 Sample Preparation Uncertainties

- Sample Digestion and Dilution: Incomplete digestion or inaccurate dilution of samples can lead to systematic errors in the measured metal concentration.
- Homogeneity of Samples: Inconsistent mixing or inhomogeneity in the sample can cause variations in the measured concentration.
- Contamination: Contamination from glassware, reagents, or the environment during sample preparation can introduce additional metals, leading to overestimation of concentrations.

Table 2 Technical details of Atomic Absorption Spectrophotometer, Make- Shimadzu, Model- AA6300

Wavelength range	185 to 900 nm (Automated wavelength selection)
Mounting	Aberration-corrected Czerny-Turner mounting, Number of grating grooves:1800 lines/mm, Focal length: 298 mm
Bandwidth	0.2 0.7 0.7 2.0(Low) nm (4 step automatic switching)
Detector	Photomultiplier (short wavelengths), semiconductor (long wavelength)
No. of lamps mounted	6 lamps, 2 user-selected lamps can be illuminated at the same time
	(1 for measurement, 1 warming up for next measurement)
Photometric mode	Flame: Optical double beam. Furnace: Double beam
Background correction	High-speed Self-Reversal method (BGC-SR), high-speed deuterium lamp method (BGC-D2)
Lamp mode	Emission, Non-BGC, BGC-D2, BGC-SR
Burner type	Air-cooled premix type
Burner head	Made of titanium, 10 cm slot (the 5 cm slot for high temperature flames is optional)
Nebulizer	Pt-lr capillary with Teflon orifice and ceramic impact bead (Hydrofluoric acids can be used)
Chamber	Polypropylene
Positioning control	Automatic search for the optimal burner height, automatic adjustment of vertical position, manual adjustment of back and forth position
Gas control—flow control unit	Fuel gas, support gas, automatic flow rate setting (0.1L/min step), automatic search for the optimal gas flow rate
Safety measures	Safe ignition/extinction sequence
	Automatic extinction in a power failure
	Flashback prevention through gas pressuring monitoring
	Burner misuse prevention
	Prevention of raw gas release when the flame dies out
Ignition	Push ignite button
Software	Software based on Microsoft Windows ™
Dimensions & weight	835 (W) × 460 (H) × 545 (D) mm, 70 kg (not including the flue, protruding parts)
Power requirements	AC100V ± 10% (no sudden voltage fluctuations)
	230VA, 50/60HZ
Ambient temperature and humidity range	Temperature: 10 ~ 35 °C
	Humidity: 20% ~ 80% (less than 70% when temperature is greater than 30 °C

3 Analytical Method Uncertainties

- Calibration Curve Accuracy: Errors in preparing or measuring standards for the calibration curve can propagate through to the sample measurements.
- Matrix Effects: The presence of other substances in the sample matrix can cause interferences that affect the absorption of the target metal, leading to inaccurate results.
- Background Absorption: Non-specific absorption by other components in the sample can interfere with the measurement, causing errors in the metal concentration.

4 Measurement Process Uncertainties

- Repeatability and Reproducibility: Variations in measurements when the same sample is analyzed multiple times (repeatability) or by different operators or instruments (reproducibility) can introduce uncertainty.
- Operator Skill: Differences in operator technique, such as pipetting accuracy or flame adjustment, can affect the consistency and accuracy of the measurements.

5 Environmental Factors

- Temperature and Humidity: Changes in environmental conditions can affect both the instrument and the sample, leading to fluctuations in measurement results.
- Air Quality: Contaminants or variations in the air quality (e.g., dust, aerosols) in the laboratory can affect the absorption measurements, especially in flame AAS.

6 Data Processing Uncertainties

- Signal Processing: Errors in data processing, such as incorrect baseline correction or improper handling of noise, can lead to inaccurate concentration calculations.
- Interpolation Errors: When concentrations fall between standard values on a calibration curve, interpolation is



Fig. 5 Atomic Absorption Spectrophotometer

used, which can introduce uncertainty if the curve is not perfectly linear.

7 Standards and Reference Material Uncertainties

- Purity of Standards: Impurities in calibration standards or reference materials can introduce bias into the measurement results.
- Traceability: Uncertainties in the traceability of reference standards can lead to errors in the final metal concentration reported.

Factors influencing measurement accuracy with AAS include:

Calibration standards: The accuracy depends heavily on the quality and preparation of calibration standards. If the standards are not accurately prepared or if they degrade over time, the accuracy of the measurement will be compromised.

Matrix effects: The presence of other substances in the sample matrix can affect the absorption of light by the analyte, leading to inaccuracies. This is often mitigated by using matrix-matched standards or applying correction techniques.

Instrumental factors: Instrumental factors such as the stability of the light source (hollow cathode lamp), the accuracy of the wavelength setting, and the efficiency of the monochromator play crucial roles in ensuring accurate measurements.

Chemical interferences: Chemical interactions within the sample can lead to either suppression or enhancement of the signal, thus affecting the accuracy. These interferences can often be managed using chemical modifiers or by optimizing the method.

Detection limits: The accuracy near the detection limit of the instrument can be lower, as the signal-to-noise ratio diminishes.

Factors affecting precision in measurements with AAS include:

Sample homogeneity: For consistent results, the sample must be homogeneous. Any variation in the sample composition can lead to variations in the results.

Instrumental stability: The stability of the instrument over time is critical for precision. Factors such as lamp drift, fluctuations in flame temperature (for flame AAS), and electronic noise can affect precision.

Operator skill: The skill and experience of the operator in preparing samples and setting up the instrument can significantly influence precision.

Measurement conditions: Consistency in the measurement conditions, such as flame or furnace temperature, gas flow rates, and aspiration rates, is crucial for maintaining precision.

Number of measurements: Precision generally improves with the number of measurements taken, as random errors can be averaged out.

2.2. Risk Evaluation of Metal Concentrations

The measured concentrations of metals were contrasted with the relevant levels specified in Sediment Quality Guidelines (SQGs) as specified in Table 2. Sediment Quality Guidelines (SQGs) are instrumental in assessing the level at which a particular metal concentration in suspended sediment could negatively impact aquatic flora and wildlife. They serve to estimate the quality of suspended sediments [47]. SQGs are established in two categories for the freshwater ecosystem[48]. Threshold Effect Concentration (TEC) values indicate concentrations below which harmful effects are predicted to be observed seldom., while Probable Effect Concentration (PEC) values indicate concentrations that are likely to have harmful consequences.

To assess the degree of pollution, various pollution indices were used such as pollution load index (PLI), enrichment ratio (ER), geoaccumulation index (GI) and contamination factor (CF). Their values were computed by making use of averages of world rivers mentioned in a study [27].

2.2.1. Calculation of Contamination Factor

Contamination factor values are obtained by dividing the present concentrations of metals by their corresponding background concentrations [49]. The levels of pollution are categorized as follows: low (CF is smaller than one), moderate (CF ranging between one to three), considerable (CF ranging between three to six), and very high contamination (CF is more than 6). CF is obtained employing Eq. 1;

$$\mathrm{CF} = \frac{C}{B} \tag{1}$$

2.2.2. Calculation of Pollution Load Index

Pollution load index serves to determine the entire metal contamination at an individual site. It is obtained employing Eq. 2[50];

$$\mathrm{PLI} = (\mathrm{CF1}^{*}\ \mathrm{CF2}^{*}\ \mathrm{CF3}\ldots\ldots\ {}^{*}\mathrm{CFn})^{\frac{1}{n}} \tag{2}$$

where CF signifies the contamination factor of the 'n' number of metals. A PLI of lesser than or equal to one indicates pollution at baseline levels only, whereas a value greater than one indicates rapid deterioration.

2.2.3. Calculation of Geoaccumulation Index

The geoaccumulation index (GI) is employed for assessing the pollution in sediments[51]. GI has been commonly utilized in previous investigations for contamination assessment [21, 35]. It is obtained using the following Eq. 3;

$$\mathrm{GI} = \mathrm{Log}_2 \frac{C}{1.5 * B} \tag{3}$$

where 'C' is the metal's current computed concentration, while 'B' is the metal's background concentration. A factor of correction of 1.5 is applied to decrease fluctuations in background concentrations caused by lithogenic fluctuations. The classification of GI is termed as: uncontaminated (GI lesser than zero); uncontaminated to moderately contaminated (GI ranging between zero to one); moderately contaminated (GI ranging between one to two); moderately to heavily contaminated (GI ranging between two to three); heavily contaminated (GI ranging between three to four); heavy to extremely contaminated (GI ranging between four to five); and extremely contaminated (GI more than five).

2.2.4. Calculations for Enrichment Ratio

The enrichment ratio is employed to assess the level of metal pollution originating through anthropogenic activities [52]. Equation 4 is used to derive CF for various metals;

$$\mathrm{ER} = \frac{\left[\frac{X_X}{X_{Fe}}\right] s}{\left[\frac{X_X}{X_{Fe}}\right] b} \tag{4}$$

where $(X_x/X_{Fe})s$ sample represents the proportion of the current metal's 'x' concentration and current iron's concentration in the suspended sediment specimen, whereas $(X_x/X_{Fe})b$ shows the proportion of the metal's 'x" background concentration and iron's background concentration. An ER below 1 signifies the origination of metal's concentration from a crystalline source and exhibits no enrichment due to anthropogenic activities[53], specimens having ER values from 1 to 3 show slight enrichment due to anthropogenic operations, while those from 3 to 5 are categorized as moderately enriched, from 5 to 10 represent moderately severe enriched, from 10 to 25 indicate severely enriched, from 25 to 50 are classified as very severely enriched and those above 50 are considered extremely severe enriched [38].

2.3. Classification of Metal Source

The source of metals are categorized as natural or anthropogenic. Anthropogenic influence on metal concentrations is determined using Eq. 5;

$$\text{Anthropogenic Input, AI } (\%) = \frac{C - BC}{C} \times 100 \tag{5}$$

where C signifies metal's current concentration at a certain site and BC signifies metal's baseline concentration. In this

study, metals baseline concentrations are taken for the sediment of the Ganges River at Hastinapur as presented in a previous investigation [46]. The values of BC for Fe, Cr, Mn, Cu, Zn, Ni, Cd and Pb were taken as 39,200, 136, 1250, 53, 104, 50, 1.31 and 19 μg/g respectively. Negative readings obtained for AI have been considered as zero which signifies zero anthropogenic matter in the specimen. Equation 6 has been applied to determine the influence of lithogenic contributions on metal concentrations;

$$\text{Lithogenic Input, LI } (\%) = (100 - \text{AI}) \quad (6)$$

3. Results and Discussion

3.1. Concentrations of metals, variations along chosen stretch and comparison with SQGs

Table 3 displays the quantities of all eight metals found in collected suspended sediment samples with their SQG values.

The decreasing order of averages of metals concentrations (in μg/g) in suspended sediment of the Ganges were recorded as follows: Fe (50,230) > Mn (1283) > Cr (201) > Zn (183) > Pb (55.1) > Ni (50.9) > Cu (42.8) > Cd (1.01). The three most abundant metals within suspended sediment include Fe, Mn, and Cr, having peak readings of 57,977, 1614, and 228 μg/g witnessed at SP4 and Cd had the highest readings of 1.19 μg/g at SP1, possessed the least concentration amid all metals.

3.1.1. Iron (Fe)

Fe concentrations were observed to be the highest amongst all elements in the Ganges River sediment, primarily due to natural processes such as weathering and erosion, as well as human activities including municipal solid waste, urban discharge, construction and demolition wastes, agricultural activities and industrial effluent. Fe values in the specimens varied within 42,716 to 57,977 μg/g, exhibiting the peak at SP4 subsequent to SP6, SP7, SP5, SP3, SP8, SP1, SP2, SP9 and lowest at SP10. The concentrations of iron at all locations crossed the SQGs value for PEC of 40,000 μg/g, indicating that the Fe element in suspended sediment has increased beyond the safe level which will negatively impact water species. The Fe concentrations in all sampling points were observed lesser than that recorded for Ganges river sediment in 2018 but greater than those measured in 2003 (refer to Table 2).

Table 3 Suspended sediment load (SSL) (mg/l), concentrations of metals (in μg/g) at chosen locations & SQGs

Code		SSL	Fe	Cr	Mn	Cu	Zn	Ni	Cd	Pb
SP1		145	49,081	206	1190	43.8	192	52.4	1.19	56.6
SP2		139	48,701	202	1335	41.7	172	50.3	1.06	58.6
SP3		156	50,026	219	1415	36.9	163	48.6	1.09	54.9
SP4		171	57,977	228	1614	36.5	186	50.6	0.98	50.1
SP5		159	52,156	216	1402	45.3	201	52.0	1.01	49.2
SP6		166	55,540	219	1376	48.5	196	51.5	0.97	55.3
SP7		160	53,024	198	1140	46.9	202	54.3	0.93	63.7
SP8		149	49,577	181	1098	45.1	198	53.8	0.98	56.8
SP9		129	43,502	172	1131	43.8	169	48.8	0.99	54.9
SP10		127	42,716	167	1128	39.2	151	46.9	0.94	50.6
Average		150	50,230	201	1283	42.8	183	50.9	1.01	55.1
Standard Deviation		*15.02*	*4779*	*21*	*171*	*4.09*	*18*	*2.36*	*0.08*	*4.37*
Background Conc. (Bn)[a]			58,100	130	1679	75.9	208	74.5	1.55	61.1
Sediment Quality Guidelines (SQG)[b]	TEC		20,000	43.4	460	31.6	121	22.7	0.99	35.8
	PEC		40,000	111	1100	149	456	48.6	4.98	128
Ganges Sediment, Kanpur 2018[c]			58,821	135.9	–	19.17	66.35	22.74	3.89	6.55
Ganges Sediment, Hastinapur 2003[d]			39,200	136	1250	53	104	50	1.31	19
Ganges Sediment, Kanpur 2017[e]			–	83	–	35.6	–	–	9.7	12.5

– Not mentioned

[a]Averages of world rivers from Viers et al. [27]

[b]Given by MacDonald et al. [48]

[c]Data obtained from Kumar et al. [43]

[d]Data obtained from Singh et al. [46]

[e]Data obtained from Kumar et al. [44]

3.1.2. Manganese (Mn)

Mn values in the specimens varied within 1098–1614 μg/g, exhibiting the peak at SP4 subsequent to SP3, SP5, SP6, SP2, SP1, SP7, SP9, SP10 and lowest at SP8. The concentrations of manganese at all locations crossed the SQGs value for TEC of 460 μg/g. At all locations except SP8, Mn concentrations are also observed greater than the PEC of 1100 μg/g indicating that Mn element in suspended sediment is increasing at a higher rate which is very likely to negatively impact water species. Sources of manganese in sediments include emissions from alloy production units, sewage sludge, mining waste, waste from various metal processing units, combustion of fossil fuels and municipal wastewater. The average of Mn concentrations was observed higher than that recorded for Ganges River sediment in 2003 (refer to Table 2).

3.1.3. Chromium (Cr)

Chromium compounds are utilized in various industries such as dyes, paint and leather industries. These compounds can be transported into the groundwater and soil of industrial spaces. Paints contain chromium as a base and are commonly used for refinishing automotive, contributing to Cr pollution in the Ganges River. Cr values in the specimens varied within 167 to 228 μg/g, exhibiting the peak at SP4 subsequent to SP3, SP6, SP5, SP1, SP2, SP7, SP8, SP9 and least at SP10. The concentrations of chromium at all locations crossed the SQGs value for TEC and PEC of 43.4 μg/g and 111 μg/g respectively indicating that Cr element in suspended sediment has drastically increased beyond the safe levels and will affect water creatures. The Cr concentrations in all sampling points were observed higher than that recorded for Ganges river sediment in 2017, 2018 & 2003 (refer Table 2).

3.1.4. Zinc (Zn)

Zn values in the specimens varied within 151 to 202 μg/g, exhibiting the peak at SP7 subsequent to SP5, SP8, SP6, SP1, SP4, SP2, SP9, SP3, and least at SP10. The concentrations of zinc at all locations crossed the SQGs value for TEC of 121 μg/g but were detected lower compared to the PEC level of 456 μg/g indicating that Zn element in suspended sediment is still under safe levels and has a rare probability to negatively impact water species. Zinc sources come from both human activities and natural processes. The use of fertilizers for agricultural reasons in the Ganges River basin increases zinc concentrations in river sediments. The Zn concentrations in all sampling points were observed higher than those recorded for Ganges river sediment in 2018 & 2003 (refer to Table 2).

3.1.5. Lead (Pb)

Pb values in the specimens varied within 49.2–63.7 μg/g, exhibiting the peak at SP7 subsequent to SP2, SP8, SP1, SP6, SP3, SP9, SP10, SP4 and least at SP5. The concentrations of lead at all locations crossed the SQGs value for TEC of 35.8 μg/g but detected lower compared to the PEC level of 128 μg/g indicating that Pb element in suspended sediment is still under the safe levels and have a rare probability to negatively impact water species. The primary sources of lead include airborne particles, forest fires, effluents from leather factories, vehicle emissions, paints, volcanic activity, waste incineration and pesticides. The Earth's crust typically contains lead concentrations ranging from 15 to 20 ppm. The Pb concentrations at all sampling points were observed much higher than those recorded for Ganges river sediment in 2017, 2018 & 2003 (refer to Table 2).

3.1.6. Nickel (Ni)

Ni is commonly used in alloys with zinc, chromium, copper and iron, and finds widespread applications across various industries including food processing, household appliances, fuel production, electroplating, ceramics, pigments, jewellery manufacturing, magnets, heat exchangers coins, batteries, medical prostheses and others. **N**i values in the specimens varied within 46.9–54.3 μg/g, exhibiting the peak at SP7 subsequent to SP8, SP1, SP5, SP6, SP4, SP2, SP9, SP3 and least at SP10. The concentrations of Nickel at all locations except at SP10 crossed the SQGs value for PEC of 48.6 μg/g indicating that Ni element in suspended sediment has increased beyond the safe level which will negatively impact water species. The Ni concentrations in all sampling points were observed higher than that recorded for Ganges river sediment in 2018 & the average of Ni is similar to that recorded for 2003 (refer Table 2).

3.1.7. Copper (Cu)

Cu values in the specimens varied within 36.5 to 48.5 μg/g, exhibiting the peak at SP6 subsequent to SP7, SP5, SP8, SP1, SP9, SP2, SP10, SP3 and minimum at SP4. The concentrations of copper at all locations crossed the SQGs value for TEC of 31.6 μg/g but were detected lower compared to the PEC level of 149 μg/g indicating that Cu element in suspended sediment is still under safe levels and has a rare probability to negatively impact water species. Copper is introduced into the environment through natural processes such as decaying vegetation, volcanic eruptions, sea spray, and forest fires, as well as anthropogenic activities including wastewater from industries and municipal corporations**.** The Cu concentrations in all sampling points

were observed higher than those recorded for Ganges river sediment in 2017 & 2018 (refer to Table 2).

3.1.8. Cadmium (Cd)

Cd values in the specimens varied within 0.93 to 1.19 µg/g, exhibiting the peak at SP6 subsequent to SP3, SP2, SP5, SP9, SP4, SP8, SP6, SP10 and least at SP7. The concentrations of cadmium at locations SP1, SP3, SP2, SP5 and SP9 crossed the SQGs value for TEC of 0.99 µg/g but were detected lower compared to the PEC level of 4.98 µg/g at all locations indicating that Cd element in suspended sediment is still under the safe levels and have a rare probability to negatively impact water species. Cd is typically an after-product of zinc and lead mining and smelting processes and exhibits higher mobility in aquatic environments compared to most other metals. Sources of cadmium release into the environment include metallurgical industries, atmospheric deposition, power plants, fertilizers, natural atmospheric processes, municipal solid waste (MSW) incineration, and the discharge of toxic effluents from industries and wastewater treatment facilities. The Cd concentrations in all sampling points were observed lesser than those recorded for Ganges river sediment in 2017, 2018 & 2003 (refer to Table 2).

3.2. Comparison with Averages of Rivers of Several Continents

Averages of metal concentrations obtained in the current research have been contrasted to the averages of rivers of the various continents [27] and are reported in Table 4. Cr concentration is significantly elevated than the global river average, while Zn and Pb concentrations are very near to global averages. In contrast, Cd, Ni, Fe, Mn and Cu concentrations are lesser as compared to the global river average. When compared to the river averages of neighboring continents, specifically Asia (China), the averages of Cr, Zn, Mn and Fe found in the current investigation are larger, suggesting that the sediments of the Ganges are impacted by anthropogenic operations. However, apart from Pb, all other metal averages are lower than those reported in Asia (Russia).

3.3. Risk Evaluation of Heavy Metals Concentrations

3.3.1. Pollution Load Index (PLI) & Contamination Factor (CF)

Table 5 displays the calculated CF and PLI values. The PLI for all locations varied within 0.72 & 0.86. Maximum value has been identified at SP6 subsequent to SP5, SP4, SP1, SP2, SP3, SP8, SP9 & SP10. PLI values at all sites stayed below 1 which indicates all sampling locations in the selected stretch are overall uncontaminated.

The averages of CF for metals varied within 0.56 to 1.54 which follows the ascending sequence as: Cu (0.56) < Cd (0.65) < Ni (0.68) < Mn (0.76) < Fe (0.86) < Zn (0.88) < Pb (0.90) < Cr (1.54). The values of CF for Cr metal lie between 1 & 3 at all sampling locations indicating sediment to be moderately contaminated by Cr. At site SP4, the CF for Fe metal is 1 and at site SP7, the CF value for Pb metal is 1.04 indicating sediment to be moderately contaminated by Fe and Pb only at these locations. The values of CF for all other metals are below 1 at all sites which shows that contamination of other metals is at a low level.

3.3.2. Geoaccumulation Index (GI)

The GI measurements can be seen in Table 6. The GI measurements for Cr for locations SP1, SP2, SP3, SP4, SP5, SP6 and SP7 lie in Class 1 ($0 < GI < 1$) which indicates an unpolluted to moderate level of pollution of Cr metal in these sampling points. GI values for all other metals for all sampling locations lie in Class 0 ($GI < 0$)

Table 4 Comparison with global river average and averages of rivers of the various continents

Metal	Ganges River (This Study)	Global River Average	Asia (China) River Average	Asia (Russia) River Average	Europe River Average	North America River Average	Africa River Average	South America River Average
Cd	1.01	1.55	–	–	–	–	–	–
Cr	201	130	117	260	164	79	115	164
Cu	42.8	75.9	53	145	172	59	34	172
Fe	50,230	58,100	46,000	78,800	43,000	52,900	45,000	43,000
Mn	1283	1679	970	5767	1884	700	1430	1884
Ni	50.9	74.5	68	123	66	46	50	66
Pb	55.1	61.1	64	35	71	76	22	71
Zn	183	208	145	300	346	184	137	346

– Not mentioned

Table 5 Contamination Factor (CF) and Pollution Load Index (PLI) for the observed metal concentrations

Code	Fe	Mn	Cr	Zn	Pb	Ni	Cu	Cd	PLI
SP1	0.84	0.71	1.58	0.92	0.93	0.70	0.58	0.77	0.84
SP2	0.84	0.80	1.55	0.83	0.96	0.67	0.55	0.69	0.82
SP3	0.86	0.84	1.68	0.78	0.90	0.65	0.49	0.70	0.81
SP4	1.00	0.96	1.75	0.90	0.82	0.68	0.48	0.63	0.84
SP5	0.90	0.84	1.66	0.97	0.81	0.70	0.60	0.65	0.85
SP6	0.96	0.82	1.68	0.94	0.91	0.69	0.64	0.62	0.86
SP7	0.91	0.68	1.52	0.97	1.04	0.73	0.62	0.60	0.84
SP8	0.85	0.65	1.39	0.95	0.93	0.72	0.59	0.63	0.81
SP9	0.75	0.67	1.32	0.81	0.90	0.66	0.58	0.64	0.77
SP10	0.74	0.67	1.28	0.73	0.83	0.63	0.52	0.61	0.72
Average	0.86	0.76	1.54	0.88	0.90	0.68	0.56	0.65	0.82

Table 6 Geoaccumulation index (GI) for the observed metal concentrations

Code	Fe	Mn	Cr	Zn	Pb	Ni	Cu	Cd
SP1	− 0.83	− 1.08	0.08	− 0.70	− 0.70	− 1.09	− 1.38	− 0.97
SP2	− 0.84	− 0.92	0.05	− 0.86	− 0.65	− 1.15	− 1.45	− 1.13
SP3	− 0.80	− 0.83	0.17	− 0.94	− 0.74	− 1.20	− 1.62	− 1.09
SP4	− 0.59	− 0.64	0.23	− 0.74	− 0.87	− 1.14	− 1.64	− 1.25
SP5	− 0.74	− 0.85	0.15	− 0.63	− 0.90	− 1.10	− 1.33	− 1.20
SP6	− 0.65	− 0.87	0.17	− 0.67	− 0.73	− 1.12	− 1.23	− 1.27
SP7	− 0.72	− 1.14	0.02	− 0.63	− 0.52	− 1.04	− 1.28	− 1.32
SP8	− 0.81	− 1.20	− 0.11	− 0.66	− 0.69	− 1.05	− 1.33	− 1.25
SP9	− 1.00	− 1.15	− 0.18	− 0.89	− 0.74	− 1.20	− 1.38	− 1.23
SP10	− 1.03	− 1.16	− 0.22	− 1.04	− 0.86	− 1.25	− 1.54	− 1.31

which revealed that sediment were uncontaminated by all other metals.

3.3.3. Enrichment Ratio (ER)

The ER is computed and displayed in Table 7. The averages of ER for the examined metals in decreasing order are as follows: Cu (0.66) > Cd (0.76) > Ni (0.80) > Mn (0.88) > Zn (1.02) > Pb (1.05) > Cr (1.79) revealing slight enrichment ($1 < ER < 3$) of Zn, Pb and Cr in suspended sediment of Ganges and no enrichment by Cu, Cd, Ni, Mn at anthropogenic levels. The measurements of ER for Cr metal ranged between 1.63 to 1.96 indicating slight enrichment of Cr across each sampling point. In summary,

Table 7 Enrichment ratio (ER) for the observed metal concentrations

Code	Mn	Cr	Zn	Pb	Ni	Cu	Cd
SP1	0.84	1.88	1.09	1.10	0.83	0.68	0.91
SP2	0.95	1.85	0.99	1.14	0.80	0.66	0.82
SP3	0.98	1.96	0.91	1.04	0.76	0.57	0.82
SP4	0.96	1.76	0.90	0.82	0.68	0.48	0.63
SP5	0.93	1.85	1.08	0.90	0.78	0.66	0.73
SP6	0.86	1.76	0.99	0.95	0.72	0.67	0.65
SP7	0.74	1.67	1.06	1.14	0.80	0.68	0.66
SP8	0.77	1.63	1.11	1.09	0.85	0.70	0.74
SP9	0.90	1.77	1.08	1.20	0.88	0.77	0.85
SP10	0.91	1.75	0.99	1.13	0.86	0.70	0.82
Average	0.88	1.79	1.02	1.05	0.80	0.66	0.76

ER data show that anthropogenic activities have a minor impact on elevated concentrations of metals in suspended sediment along the chosen segment of the Ganges.

3.4. Source of Metal Classification

Figure 6 highlights the fraction of heavy metals originating from anthropogenic and lithogenic processes. Strong anthropogenic input has been identified for Pb, Zn, Fe & Cr at all sampling points (SP1 – SP10). On average 65.30% of Pb, 42.66% of Zn, 21.31% of Fe & 31.55% of Cr concentrations were observed to have originated due to anthropogenic factors and the rest seem to be originating due to geogenic factors. The highest anthropogenic contribution was detected for Pb ranging between 61.38% to 70.17% among selected sampling locations. In the case of Zn, anthropogenic sources were responsible for 31.33% to 48.51% of the concentrations observed at the chosen sample stations. A small amount of anthropogenic input of Mn and Ni was also observed at few sampling points with an average of 6.06% for Mn and 2.81% for Ni, while Cu & Cd seem to be originating only due to geogenic factors. Anthropogenic input of Mn metal was recorded at sampling locations SP2, SP3, SP4, SP5 & SP6 whereas anthropogenic input of less than 8% was reported for Ni metal at sampling locations SP1, SP4, SP5, SP6, SP7 & SP8.

4. Conclusion

This study identifies the concentrations and risk evaluation of heavy metals (Fe, Cr, Mn, Cu, Zn, Ni, Cd and Pb) in the suspended sediment of 225 km segment of the Ganges River using various pollution indicators, background concentrations, river averages of various continents and SQGs. The analysis of metal concentrations in the suspended sediments of the Ganges reveals a significant presence of various metals, with iron, manganese and chromium being the most abundant.

The concentrations of metals in the suspended sediments of the Ganges River, measured in μg/g, exhibit the following descending order: Fe (50,230) > Mn (1283) > Cr (201) > Zn (183) > Pb (55.1) > Ni (50.9) > Cu (42.8) > Cd (1.01). The peak readings for the most abundant metals, Fe, Mn, and Cr, were recorded at SP4, with values of 57,977, 1614, and 228 μg/g respectively. Cd had the highest reading of 1.19 μg/g at SP1, indicating the lowest concentration among all metals.

- Iron (Fe): Concentrations ranged from 42,716 to 57,977 μg/g, surpassing the SQGs value for PEC of 40,000 μg/g at all sites, indicating a potential negative impact on aquatic life.
- Manganese (Mn): Concentrations ranged from 1098 to 1614 μg/g, exceeding the SQGs value for TEC of 460 μg/g and PEC of 1100 μg/g at all sites except SP8, suggesting a likely adverse effect on water species.
- Chromium (Cr): Concentrations ranged from 167 to 228 μg/g, well above the SQGs values for TEC (43.4 μg/g) and PEC (111 μg/g), indicating significant contamination and potential harm to aquatic life.
- Zinc (Zn): Concentrations ranged from 151 to 202 μg/g, above the SQGs value for TEC of 121 μg/g but below the PEC level of 456 μg/g, implying a rare probability of negative impact.
- Lead (Pb): Concentrations ranged from 49.2 to 63.7 μg/g, above the SQGs value for TEC of 35.8 μg/g but below the PEC level of 128 μg/g, indicating a rare probability of harm.
- Nickel (Ni): Concentrations ranged from 46.9 to 54.3 μg/g, surpassing the SQGs value for PEC of 48.6 μg/g at most sites except SP10, indicating potential negative effects.
- Copper (Cu): Concentrations ranged from 36.5 to 48.5 μg/g, above the SQGs value for TEC of 31.6 μg/g but below the PEC level of 149 μg/g, suggesting a rare probability of harm.
- Cadmium (Cd): Concentrations ranged from 0.93 to 1.19 μg/g, with some sites exceeding the SQGs value for TEC of 0.99 μg/g but all below the PEC level of 4.98 μg/g, indicating a rare probability of negative impact.

The calculated contamination factor and pollution load index values further support the assessment of the Ganges' sediment quality. With PLI values below 1 at all sampling sites ranging from 0.72 to 0.86, the overall contamination level is classified as low, indicating that despite the high concentrations of certain metals, the stretch of the Ganges studied is not severely contaminated. The contamination factor analysis revealed varying levels of metal contamination, with CF values indicating moderate contamination by chromium ranging between 0.56 to 1.54 across all sites and by iron and lead at specific locations (SP4 and SP7, respectively). The Geoaccumulation Index further confirmed that chromium poses a moderate pollution risk (Class 1: $0 < GI < 1$) at several sites, while all other metals were classified as uncontaminated ($GI < 0$). The Enrichment Ratio analysis indicated slight enrichment of zinc, lead and chromium, pointing to minor anthropogenic influences on these metals' concentrations.

Anthropogenic contributions were notably high for lead, zinc, iron, and chromium, with the highest impact observed for lead, where anthropogenic sources accounted for up to

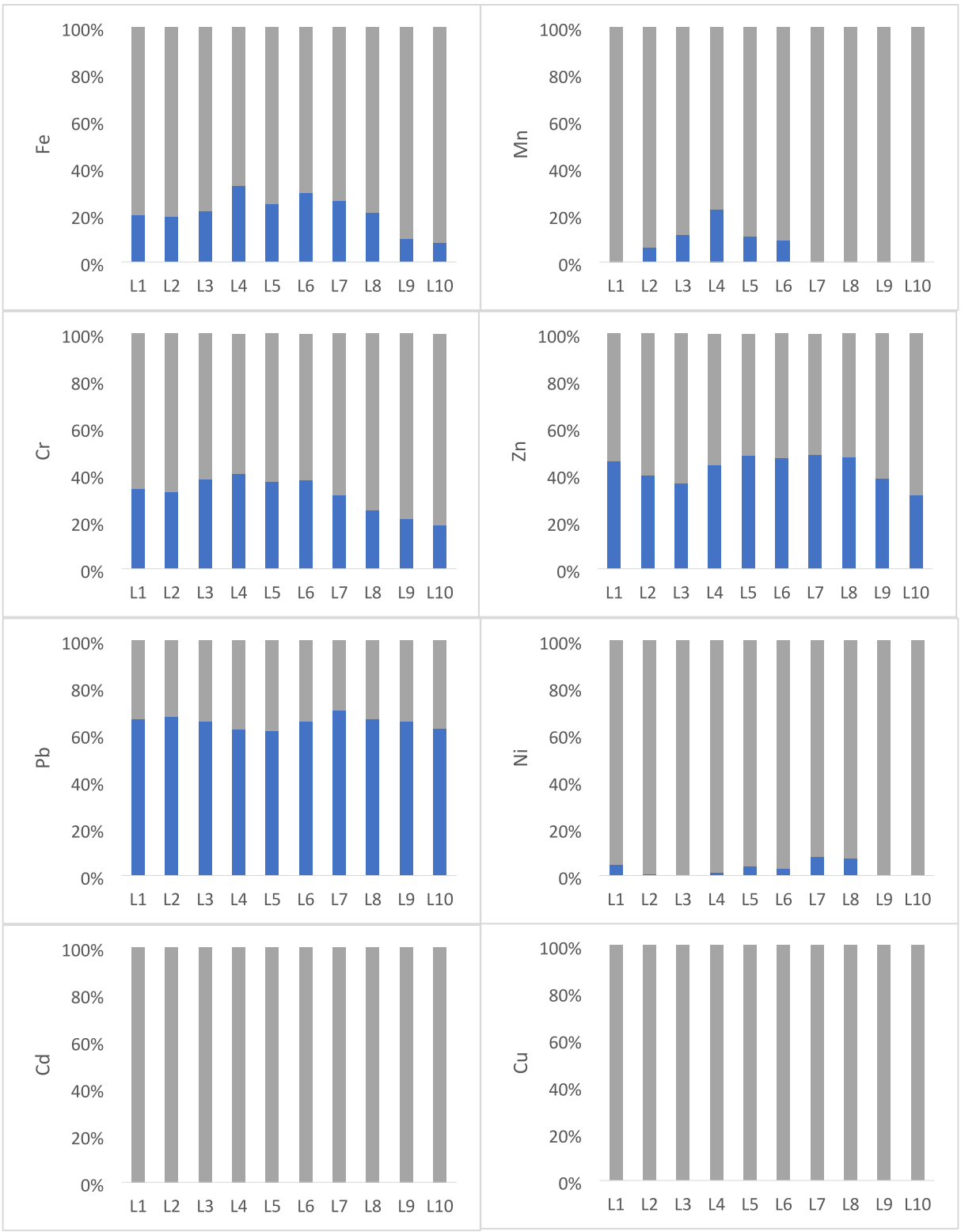


Fig. 6 Anthropogenic and Lithogenic content in suspended sediment of Ganges River

70.16% of its concentration and the rest from geogenic factors. Conversely, copper and cadmium showed no significant anthropogenic input and all contributions were observed to be from geogenic factors, manganese and nickel had minimal anthropogenic influence at a few sampling points.

To conclude, the study highlights the significant but localized anthropogenic impact on heavy metal concentrations, particularly for chromium in the Ganges River while the overall assessment suggests that the Ganges' suspended sediment in the studied stretch is not heavily contaminated, the elevated concentrations of iron, manganese, chromium, and nickel pose significant ecological risks. Continuous monitoring and effective management strategies are essential to mitigate the adverse effects of metal pollution and protect the river's aquatic ecosystem. The study underscores the need for targeted pollution control measures, especially for metals with high anthropogenic contributions like lead and zinc, to ensure the long-term health and safety of this vital water resource. Sustainable practices and comprehensive environmental monitoring are critical to preventing further contamination and preserving the ecological balance of the Ganges River. The metal concentrations averages determined in the present study will serve as the standards for assessing the further contamination of Ganges River and will be used as back ground concentrations for the risk assessment in various pollution indices.

References

[1] A. Yadav and J. Pandey, Contribution of point sources and non-point sources to nutrient and carbon loads and their influence on the trophic status of the Ganga River at Varanasi, India. Environ. Monit. Assess., **189** (2017) 1–19. https://doi.org/10.1007/s10661-017-6188-8.

[2] Central Pollution Control Board (CPCB), Pollution assessment: River Ganga, (2013), https://cpcb.nic.in/wqm/pollution-assessment-ganga-2013.pdf

[3] S. Dwivedi, S. Mishra and R.D. Tripathi, Ganga water pollution: a potential health threat to inhabitants of Ganga basin. Environ. Int., **117** (2018) 327–338. https://doi.org/10.1016/j.envint.2018.05.015.

[4] R.C. Trivedi, Water quality of the Ganga River-an overview. Aquat. Ecosyst. Heal. Manag., **13** (2010) 347–351. https://doi.org/10.1080/14634988.2010.528740.

[5] J. Pandey, K. Shubhashish and R. Pandey, Heavy metal contamination of Ganga River at Varanasi in relation to atmospheric deposition. Trop. Ecol., **51** (2010) 365–373.

[6] A. Qadir and R.N. Malik, Heavy metals in eight edible fish species from two polluted tributaries (Aik and Palkhu) of the River Chenab, Pakistan. Biol. Trace Elem. Res., **143** (2011) 1524–1540. https://doi.org/10.1007/s12011-011-9011-3.

[7] J. Fu, C. Zhao, Y. Luo, C. Liu, G.Z. Kyzas, Y. Luo, D. Zhao, S. An and H. Zhu, Heavy metals in surface sediments of the Jialu River, China: their relations to environmental factors. J. Hazard. Mater., **270** (2014) 102–109. https://doi.org/10.1016/j.jhazmat.2014.01.044.

[8] S. Madhav, S. Kanhaiya, A. Srivastav, V. Singh and P. Singh, Ecological significance of riparian ecosystems: challenges and management strategies. Elsevier, 2022

[9] A. Demirak, F. Yilmaz, A.L. Tuna and N. Ozdemir, Heavy metals in water, sediment and tissues of *Leuciscus cephalus* from a stream in southwestern Turkey. Chemosphere, **63** (2006) 1451–1458. https://doi.org/10.1016/j.chemosphere.2005.09.033.

[10] L. Zhang, X. Ye, H. Feng, Y. Jing, T. Ouyang, X. Yu, R. Liang, C. Gao and W. Chen, Heavy metal contamination in western Xiamen Bay sediments and its vicinity, China. Mar. Pollut. Bull., **54** (2007) 974–982. https://doi.org/10.1016/j.marpolbul.2007.02.010.

[11] J. Iqbal and M.H. Shah, Occurrence, risk assessment, and source apportionment of heavy metals in surface sediments from Khanpur Lake, Pakistan. J. Anal Sci. Technol., **5** (2014) 1–12. https://doi.org/10.1186/s40543-014-0028-z.

[12] A. Zahra, M.Z. Hashmi, R.N. Malik and Z. Ahmed, Enrichment and geo-accumulation of heavy metals and risk assessment of sediments of the Kurang Nallah—feeding tributary of the Rawal Lake Reservoir, Pakistan. Sci. Total Environ., **470** (2014) 925–933. https://doi.org/10.1016/j.scitotenv.2013.10.017.

[13] J. Park and B.J. Presley, Trace metal contamination of sediments and organisms from the Swan Lake area of Galveston Bay. Environ. Pollut., **98** (1997) 209–221. https://doi.org/10.1016/S0269-7491(97)00137-1.

[14] R. Xiao, J. Bai, L. Huang, H. Zhang, B. Cui and X. Liu, Distribution and pollution, toxicity and risk assessment of heavy metals in sediments from urban and rural rivers of the Pearl River delta in southern China. Ecotoxicology, **22** (2013) 1564–1575. https://doi.org/10.1007/s10646-013-1142-1.

[15] S. Kanhaiya, B.P. Singh, M. Tripathi, S. Sahu and V. Tiwari, Lithofacies and particle-size characteristics of late quaternary floodplain deposits along the middle reaches of the Ganga River, Central Ganga Plain, India. Geomorphology, **284** (2017) 220–228.

[16] U.K. Shukla, Geomorphology of Ganga Basin, Varanasi and Ganga. In: Geomorphology Archaeology Investigation Ramnagar, 1st edn. Aryab Books International, New Delhi (2016), pp. 3–16.

[17] L. Chouba and N. Mzoughi, Assessment of heavy metals in sediment and in suspended particles affected by multiple anthropogenic contributions in harbours. Int. J. Environ. Sci. Technol., **10** (2013) 779–788. https://doi.org/10.1007/s13762-013-0233-0.

[18] L. Gao, Z. Wang, S. Li and J. Chen, Bioavailability and toxicity of trace metals (Cd, Cr, Cu, Ni, and Zn) in sediment cores from the Shima River, South China. Chemosphere, **192** (2018) 31–42. https://doi.org/10.1016/j.chemosphere.2017.10.110.

[19] S.H. Hamad, J.J. Schauer, M.M. Shafer, E. Abed Al-Raheem and H. Satar, The distribution between the dissolved and the particulate forms of 49 metals across the Tigris River, Baghdad, Iraq. Sci. World J. (2012). https://doi.org/10.1100/2012/246059.

[20] M.S. Islam, M.K. Ahmed, M. Raknuzzaman, M. Habibullah-Al-Mamun and M.K. Islam, Heavy metal pollution in surface water and sediment: a preliminary assessment of an urban river in a developing country. Ecol. Indic., **48** (2015) 282–291. https://doi.org/10.1016/j.ecolind.2014.08.016.

[21] R. Li, C. Tang, Y. Cao, T. Jiang and J. Chen, The distribution and partitioning of trace metals (Pb, Cd, Cu, and Zn) and metalloid (As) in the Beijiang River. Environ. Monit. Assess., **190** (2018) 1–15. https://doi.org/10.1007/s10661-018-6789-x.

[22] N. Zhang, S. Zang and Q. Sun, Health risk assessment of heavy metals in the water environment of Zhalong Wetland, China.

Ecotoxicology, **23** (2014) 518–526. https://doi.org/10.1007/s10646-014-1183-0.

[23] C. Liu, C. Fan, Q. Shen, S. Shao, L. Zhang and Q. Zhou, Effects of riverine suspended particulate matter on post-dredging metal re-contamination across the sediment-water interface. Chemosphere, **144** (2016) 2329–2335. https://doi.org/10.1016/j.chemosphere.2015.11.010.

[24] N.E.M. Asselman, Fitting and interpretation of sediment rating curves. J. Hydrol., **234** (2000) 228–248. https://doi.org/10.1016/S0022-1694(00)00253-5.

[25] W. Zhang, X. Wei, Z. Jinhai, Z. Yuliang and Y. Zhang, Estimating suspended sediment loads in the Pearl River Delta region using sediment rating curves. Cont. Shelf Res., **38** (2012) 35–46. https://doi.org/10.1016/j.csr.2012.02.017.

[26] F. Nie, T. Li, H. Yao, M. Feng and G. Zhang, Characterization of suspended solids and particle-bound heavy metals in a first flush of highway runoff. J. Zhejiang Univ. A, **9** (2008) 1567–1575. https://doi.org/10.1631/jzus.A0820271.

[27] J. Viers, B. Dupré and J. Gaillardet, Chemical composition of suspended sediments in World Rivers: new insights from a new database. Sci. Total Environ., **407** (2009) 853–868. https://doi.org/10.1016/j.scitotenv.2008.09.053.

[28] Q. Yao, X. Wang, H. Jian, H. Chen and Z. Yu, Characterization of the particle size fraction associated with heavy metals in suspended sediments of the Yellow River. Int. J. Environ. Res. Public Health, **12** (2015) 6725–6744. https://doi.org/10.3390/ijerph120606725.

[29] M. Aggarwal, S. Anbukumar and T.Vijaya Kumar, Analysis and pollution assessment of heavy metals in suspended solids of the middle stretch of river Ganga between Kanpur to Prayagraj, U.P., India. In: Sadhana - Academy Proceedings Engineering Sciences, **48** (2023) 257. https://doi.org/10.1007/s12046-023-02325-7.

[30] S. Kanhaiya, S. Singh, A. Dixit and A.K. Singh, Rivers of India: past, present and future. Springer, Cham (2024).

[31] U. Bhosale and K.C. Sahu, Heavy metal pollution around the island city of Bombay, India Part II: distribution of heavy metals between water, suspended particles and sediments in a polluted aquatic regime. Chem. Geol., **90** (1991) 285–305. https://doi.org/10.1016/0009-2541(91)90105-Z.

[32] T.I. Kassim, H.A. Al-Saadi, A.A. Al-Lami and H.H. Al-Jaberi, Heavy metals in water, suspended particles, sediments and aquatic plants of the upper region of Euphrates river, Iraq. J. Environ. Sci. Heal. Part A, **32** (1997) 2497–2506. https://doi.org/10.1080/10934529709376698.

[33] J. Liu, S.-L. Li, J.-B. Chen, J. Zhong, F.-J. Yue, Y. Lang and H. Ding, Temporal transport of major and trace elements in the upper reaches of the Xijiang River SW China. Environ. Earth Sci., **76** (2017) 1–17. https://doi.org/10.1007/s12665-017-6625-6.

[34] T. Matsunaga, K. Tsuduki, N. Yanase, R. Kritsananuwat, T. Ueno, Y. Hanzawa and H. Naganawa, Temporal variations in metal enrichment in suspended particulate matter during rainfall events in a rural stream. Limnology, **15** (2014) 13–25. https://doi.org/10.1007/s10201-013-0409-9.

[35] S. Nazeer, M.Z. Hashmi and R.N. Malik, Heavy metals distribution, risk assessment and water quality characterization by water quality index of the River Soan, Pakistan. Ecol. Indic., **43** (2014) 262–270. https://doi.org/10.1016/j.ecolind.2014.03.010.

[36] M.O. Beltrame, S.G. De Marco and J.E. Marcovecchio, Dissolved and particulate heavy metals distribution in coastal lagoons a case study from Mar Chiquita Lagoon, Argentina. Estuar. Coast. Shelf Sci., **85** (2009) 45–56. https://doi.org/10.1016/j.ecss.2009.04.027.

[37] P. Ávila-Pérez, G. Zarazúa, L. Carapia-Morales, S. Tejeda, C. Diaz-Delgado and I. Barceló-Quintal, Evaluation of heavy metal and elemental composition of particles in suspended matter of the upper course of the Lerma River. J. Radioanal. Nucl. Chem., **273** (2007) 625–633. https://doi.org/10.1007/s10967-007-0922-6.

[38] R. Li, C. Tang, X. Li, T. Jiang, Y. Shi and Y. Cao, Reconstructing the historical pollution levels and ecological risks over the past sixty years in sediments of the Beijiang River, South China. Sci. Total Environ., **649** (2019) 448–460. https://doi.org/10.1016/j.scitotenv.2018.08.283.

[39] M. Aggarwal, S. Anbukumar and T.Vijaya Kumar, Heavy metals concentrations and risk assessment in the sediment of Ganga River between Kanpur and Prayagraj, UP, India. In: Sadhana - Academy Proceedings Engineering Sciences, **47** (2022), 195. https://doi.org/10.1007/s12046-022-01972-6.

[40] H. Singh, S. Yadav, B.K. Singh, B. Dubey, K. Tripathi, V. Srivastava and D.N. Shukla, Assessment of geochemical environment from study of river sediments in the middle stretch of River Ganga at Ghazipur, Buxar and Ballia area. In: Academy Proceedings Engineering Sciences, India Section B Biology Sciences, 83 (2013) 371–384. https://doi.org/10.1007/s40011-012-0134-x.

[41] V.K. Singh, K.P. Singh and D. Mohan, Status of heavy metals in water and bed sediments of river Gomti–a tributary of the Ganga river, India. Environ. Monit. Assess., **105** (2005) 43–67. https://doi.org/10.1007/s10661-005-2816-9.

[42] J. Pandey and R. Singh, Heavy metals in sediments of Ganga River: up-and downstream urban influences. Appl. Water Sci., **7** (2017) 1669–1678. https://doi.org/10.1007/s13201-015-0334-7.

[43] M. Kumar, N. Gupta, A. Ratn, Y. Awasthi, R. Prasad, A. Trivedi and S.P. Trivedi, Biomonitoring of heavy metals in river ganga water, sediments, plant, and fishes of different trophic levels. Biol. Trace Elem. Res., **193** (2020) 536–547. https://doi.org/10.1007/s12011-019-01736-0.

[44] D. Kumar, D.S. Malik, N. Kumar, N. Gupta and V. Gupta, Spatial changes in water and heavy metal contamination in water and sediment of river Ganga in the river belt Haridwar to Kanpur. Environ. Geochem. Health, **42** (2020) 2059–2079. https://doi.org/10.1007/s10653-019-00471-8.

[45] W. Aktar, M. Paramasivam, M. Ganguly, S. Purkait and D. Sengupta, Assessment and occurrence of various heavy metals in surface water of Ganga River around Kolkata: a study for toxicity and ecological impact. Environ. Monit. Assess., **160** (2010) 207–213. https://doi.org/10.1007/s10661-008-0688-5.

[46] M. Singh, G. Müller and I.B. Singh, Geogenic distribution and baseline concentration of heavy metals in sediments of the Ganges River India. J. Geochem. Explor., **80** (2003) 1–17. https://doi.org/10.1016/S0375-6742(03)00016-5.

[47] R.J. Wenning, Use of sediment quality guidelines and related tools for the assessment of contaminated sediments. SETAC, (2005). https://www.cerc.usgs.gov/pubs/sedtox/wg0_setac_sqg_summary.pdf.

[48] D.D. MacDonald, C.G. Ingersoll and T.A. Berger, Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. Arch. Environ. Contam. Toxicol., **39** (2000) 20–31. https://doi.org/10.1007/s002440010075.

[49] K. Loska, J. Cebula, J. Pelczar, D. Wiechuła and J. Kwapuliński, Use of enrichment, and contamination factors together with geoaccumulation indexes to evaluate the content of Cd, Cu, and Ni in the Rybnik water reservoir in Poland. Water. Air. Soil Pollut., **93** (1997) 347–365. https://doi.org/10.1023/A:1022121615949.

[50] D.L. Tomlinson, J.G. Wilson, C.R. Harris and D.W. Jeffrey, Problems in the assessment of heavy-metal levels in estuaries and the formation of a pollution index. Helgoländer Meeresunters., **33** (1980) 566–575.

[51] G. Muller, Index of geoaccumulation in sediments of the Rhine River. GeoJournal, **2** (1969) 108–118.

[52] K. Selvaraj, V.R. Mohan and P. Szefer, Evaluation of metal contamination in coastal sediments of the Bay of Bengal, India: geochemical and statistical approaches. Mar. Pollut. Bull., **49** (2004) 174–185. https://doi.org/10.1016/j.marpolbul.2004.02.006.

[53] J. Zhang and C.L. Liu, Riverine composition and estuarine geochemistry of particulate metals in China—weathering features, anthropogenic impact and chemical fluxes. Estuar. Coast. Shelf Sci., **54** (2002) 1051–1070. https://doi.org/10.1006/ecss.2001.0879.

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.

Mental Health and Stress Prediction Using NLP and Transformer-based Techniques

Abhishek Pandey
Dept. of Computer Science and Engineering
Delhi Technological University
Main Bawana Road, New Delhi-110042, India
abhiabhishekpandey11@gmail.com

Sanjay Kumar
Dept. of Computer Science and Engineering
Delhi Technological University
Main Bawana Road, New Delhi-110042, India
sanjay.kumar@dtu.ac.in

***Abstract*— The integration of Natural Language Processing (NLP) and machine learning technologies into the field of mental health and stress prediction represents a significant advancement in the early detection and intervention of mental health issues. This paper explores the application of various NLP and transformer-based techniques, such as TF-IDF, n-grams, BERT, RoBERTa, and ALBERT, to analyze textual data from diverse sources such as social media platforms towards mental health and stress prediction. The developed methodology involves comprehensive data preprocessing, innovative feature extraction, and the application of deep learning models to understand the refinement of language indicative of mental health states. By comparing the effectiveness of different models and approaches, this paper aims to establish a robust framework for accurately predicting mental health conditions, thereby contributing to the broader goal of improving mental health care through technology. The simulation performed on a real-life dataset obtained from Reddit reveals that RoBERTa-based techniques perform better than the other methods.**

Keywords—A Lite BERT (ALBERT), Bidirectional Encoder Representation from Transformers (BERT), Machine learning, Mental health, Natural language processing, Robustly Optimized BERT Pre-training Approach (RoBERTa), Term frequency-inverse document frequency (TF-IDF)

I. INTRODUCTION

The rapid increase in the field of Natural Language Processing (NLP) has significantly advanced our capabilities in understanding and processing human language, opening up new avenues for analyzing and addressing complex societal issues. Among these, mental health stands out as a critical area where NLP techniques can have a profound impact. The early prediction and identification of stress and mental health issues with the help of NLP not only promises to revolutionize the approach towards mental healthcare but also offers a pathway in the prompt intervention which potentially saves lives and hence helps to improve the quality of life for the individuals across the globe [1].

The use of the NLP methods in mental health and stress prediction influenced the large amounts of textual data generated from daily communications such as social media posts, clinical notes, and personal diaries. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) can quantify the importance of words within documents while also considering their frequency across documents and provide a foundational approach for extracting meaningful patterns from the text. This technique, among others, helps in the identification of linguistic features that are indicative of stress and mental health issues [2].

Deep learning architectures that go far beyond these traditional models have enabled new levels of NLP capabilities. Powerful models like BERT (Bidirectional Encoder Representations from Transformers), RoBERTa(a robustly optimized BERT approach), and ALBERT(A Lite BERT) have come a long way in helping the field to really recognize what words mean based on textual form considerably better. Things like deep contextual embeddings are what make these models so good at capturing the subtlety of mental health language that in turn increases the precision of stress and mental disorders.

Moreover, the use of n-grams which are continuous sequences of n items from a given sample of text has been instrumental in capturing the context and then identifying patterns in language uses that correlate with the mental health issues. By analyzing these patterns researchers can easily identify the potential indicators of the stress and the mental health conditions from the textual data which then also offers the various insights that were previously difficult to identify. The integration of these NLP techniques presented a multidimensional approach to mental health and stress prediction and each of them is contributing their unique strengths to the challenge of understanding and then predicting the mental health issues from text data. This research paper also aims to explore the application of these varied NLP techniques—TF-IDF, BERT [4], RoBERTa, ALBERT, and n-grams—in the context of mental health and stress prediction. By examining their methodologies and effectiveness and also the potential abilities this research aims to demonstrate the transformative potential of the NLP techniques in mental health care by offering a glimpse into the future where mental health interventions are more accessible and timelier and are also informed by the given complexities of human language. In this research, we have considered social media posts extracted from Reddit [3-5]. To predict the mental status we have trained our model on the extracted dataset from Reddit and then we have analyzed the comparison of various Machine Learning algorithms and the Natural Language Processing techniques.

Embedding a novel method of text analysis has emerged as a result of recent developments in language models that are derived from the large corpus. In this method, each word is positioned in a multidimensional vector space according to its neighboring words in the underlying corpus. Word embedding can be created using pre-trained word embedding as those provided by Global Vectors for Word Representation (GloVe) [16], and Word2Vec [17], or they can be trained using the input corpus itself. Each of these two models has pros and cons that vary based on the particular use case. GloVe can handle large corpus efficiently because of its co-occurrence matrix-based methodology, which enables quick and scalable training [6].

2024 IEEE Symposium on Wireless Technology & Applications (ISWTA) | 979-8-3503-5135-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISWTA62130.2024.10651843

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:34:53 UTC from IEEE Xplore. Restrictions apply.

The remainder of the paper is organized as follows. Section II presents related work on mental health and stress prediction using Natural Language Processing. Section III presents the proposed methodology for this study including data collection, data processing, and classification models which further explain techniques used like TF-IDF, n-gram, and BERT. Section IV presents experimental evaluations which are used in this study providing all the details which are used for implementation. Finally, section V presents the conclusion of this work and future work.

II. RELATED WORK

Natural Language Processing has emerged as a powerful technique in the domain of mental health research for early diagnosis and monitoring of mental health conditions. Machine learning and NLP techniques have advanced mental health care facilities by providing new insights into the complicated relationship between language use, behavioral patterns, and mental health.

ML Methods for Mental Health: The use of machine learning methods in mental health research is complex, ranging from analyzing social media content to identifying the early signs of mental distress to developing predictive models for various mental health conditions. For instance, different studies have shown that ML algorithms can effectively classify text data such as social media posts and therapy session transcripts which can be used to identify the markers of mental health issues like depression, anxiety, and PTSD [2].

Transformer Methods for Mental Health: NLP methods particularly the advanced models such as TF-IDF, BERT, and RoBERTa, played a crucial role in this procedure. They have enabled the extraction of meaningful information from unstructured text data by understanding the context and semantics of the language.

Zanwar et al. [7] emphasized the potential of NLP and ML techniques while detecting mental health issues from social media data. It critiques the predominance of binary classification in the field, proposing a multiclass approach to discern subtle linguistic patterns linked to specific mental health conditions. The study utilizes advanced NLP models like BERT and RoBERTa, alongside BiLSTM networks, to analyze linguistic features from Reddit posts, aiming to predict six mental health conditions. The above research innovatively combined the hybrid and ensemble models in return demonstrating their superiority in capturing the complexity in mental health discourse over the traditional method and hence helped in broadening the scope for early and nuanced mental health diagnosis through social media analysis.

Dristy et al. [8] examined the effectiveness of the various supervised machine learning classifiers while predicting the mental health status from text messages. Their study showed that the performance of Logistic Regression, Bernoulli Naive Bayes, K-Nearest Neighbors, Decision Trees, and Random Forest finding Logistic Regression algorithms are performing better with the highest accuracy. The research highlighted the importance of selecting relevant algorithms for mental health predictions and then contributed to the evolving field of mental health diagnostics using NLP and different machine learning techniques.

Deshmukh et al. [9] examined the application of Natural Language Processing (NLP) and deep learning (DL) methods to analyze social media content and also to identify mental health conditions. This study highlighted the use of various DL models like CNNs and RNNs that helped into process and classify the user-generated content on the platforms like Reddit. The study focused on the potential of these technologies to enhance the understanding and detection of mental health issues by showcasing the importance of machine learning in mental health research [2]. Villa-Pérez et al. [10] demonstrated the use of traditional and deep learning algorithms to identify the mental health conditions of English and Spanish-speaking Twitter users. The study makes the two datasets which are then labeled as the "diagnosed" and "control" to differentiate between the users with and without self-reported diagnoses for the specific mental health conditions. Various text features like and that includes n-grams, q-grams, Part-of-Speech tags, topic modeling, Linguistic Inquiry and Word Count (LIWC), and word embeddings were extracted and then used to train classifiers for the binary and the multiclass classification tasks. The paper highlighted the effectiveness of the XGBoost and the convolutional neural networks (CNNs) in these tasks while underscoring the potential of lexical attributes like n-grams and q-grams in mental health prediction from social media content.

Nouman et al. [11] examined the application of the various deep learning models to classify mental health conditions from text chats. They not only utilized the novel dataset but also the study compared the performance of different models which helped in finding the bidirectional gated recurrent unit (BiGRU) model to achieve an accuracy of over 83%. The research underscores the potential of deep learning in identifying mental health conditions from digital communication by highlighting the BiGRU model's effectiveness in processing and then classifying complex conversational data.

III. METHODOLOGY

Our methodology focuses on enhancing an NLP Model that is initially trained on the Reddit dataset for the specific use case in the mental health domain. For mental health and stress detection using Natural Language Processing (NLP) techniques, we proposed a multi-phased approach integrating various state-of-the-art NLP methods, including TF-IDF, n-grams, BERT, ALBERT, and RoBERTa. Fig. 1 shows the overall workflow of the implementation which has been carried out in this paper explaining every step and the process involved in each step.

A. Data Collection

Collecting data for mental health and stress detection from Reddit poses a unique set of challenges and opportunities. Reddit is a social media platform rich in diverse user-generated content for extracting insights into mental health conditions and stress indicators through various subreddits. However, the enormity and unstructured nature of Reddit data entails sophisticated Natural Language Processing (NLP) techniques to delve into and analyze the information effectively.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:34:53 UTC from IEEE Xplore. Restrictions apply.

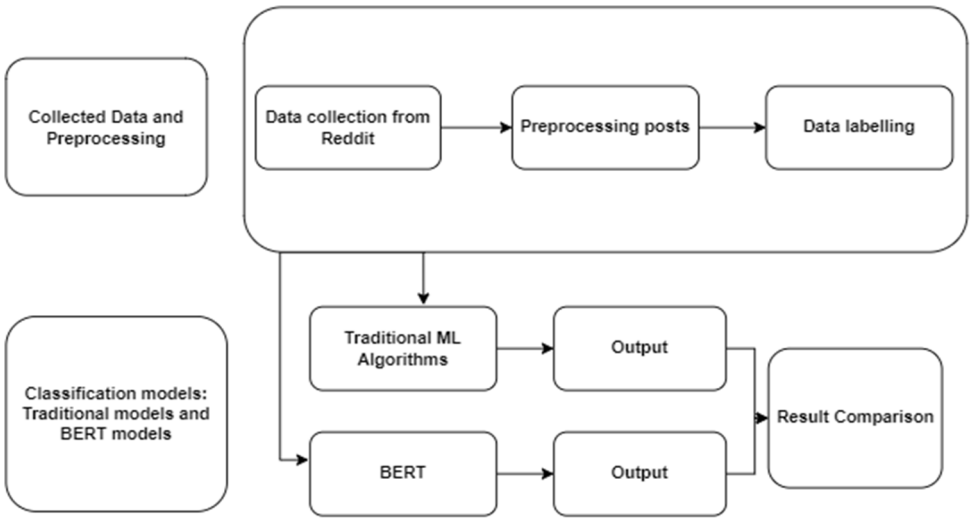


Fig. 1. Workflow of the proposed model for mental health and stress prediction: data collection and preprocessing, model training and classification, result evaluation

B. Data Processing

Data processing is a crucial stage in handling textual data for mental health and stress detection, primarily given that the dataset is usually complex and unstructured for those collected on social media platforms. The phase incorporates a series of major steps meant to convert raw data into an analytically valuable format. First, the data is cleaned to eliminate noise, comprising irrelevant posts, spam, and non-textual components, making sure a dataset is concentrated on the topic and useful for study purposes. Data cleaning removes text normalization, comprising transformation of the text into a unique case, cleaning from punctuation, stop words, stemming, and lemmatization are then applied to reduce words to their root form, facilitating the grouping of similar expressions of concepts. Fig. 2, and 3 show the word cloud of non-stress weights and stress weights respectively from the utilized dataset with a glance of the whole data.

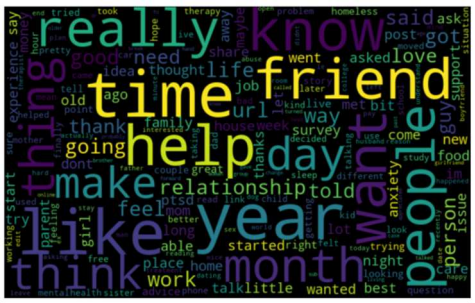


Fig. 2. Word cloud: non-stress weights

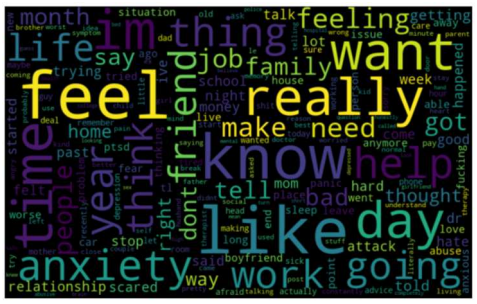


Fig. 3. Word cloud: stress weights

C. Classification Models

After data collection and pre-processing phase, classification models are applied to classify the post into different types of mental health. In this phase, we used various Natural Language Processing (NLP) and transformer-based techniques like TF-IDF, n-gram BERT, ALBERT, and RoBERTa to evaluate the models. BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Approach) were chosen as they give the more promising results. In addition to these various traditional machine learning techniques like AdaBoost, linear support vector machine, Naïve Bayes, and random forest were used for comparison of results.

i. TF-IDF

Words are considered an effective means of communication between two people and for expressing ideas. It can also be used to express thoughts, feelings, and ideas. A statistical metric called TF-IDF is applied to measure the significance of a word to a document inside a collection or corpus. TF-IDF increases corresponding to the number of times a word appears in the document but it is balanced out by the frequency of the word in the corpus which helps us to adjust for the fact that some words can appear more frequently in general. This approach is very useful for extracting the important information from the texts and selecting the features in machine learning models for the Natural Language Processing tasks. Fig. 4 shows us the top 20 words of TF-IDF for the stress posts and the non-stress posts. This technique works well for the smaller dataset but it is computationally very slow for the larger datasets than this.

$$TF\text{–}IDF\ (t, d, D) = TF\ (t, d) \cdot IDF\ (t, D),\ where:$$

$$TF\ (t) = \frac{\text{Number of times term t appears in a document}}{Total\ no\ of\ terms\ in\ a\ document\ d}$$

$$IDF\ (t, D) = log_e\left(\frac{\text{Total number of documents in the corpus}}{\text{Number of documents with term t in them}}\right)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:34:53 UTC from IEEE Xplore. Restrictions apply.

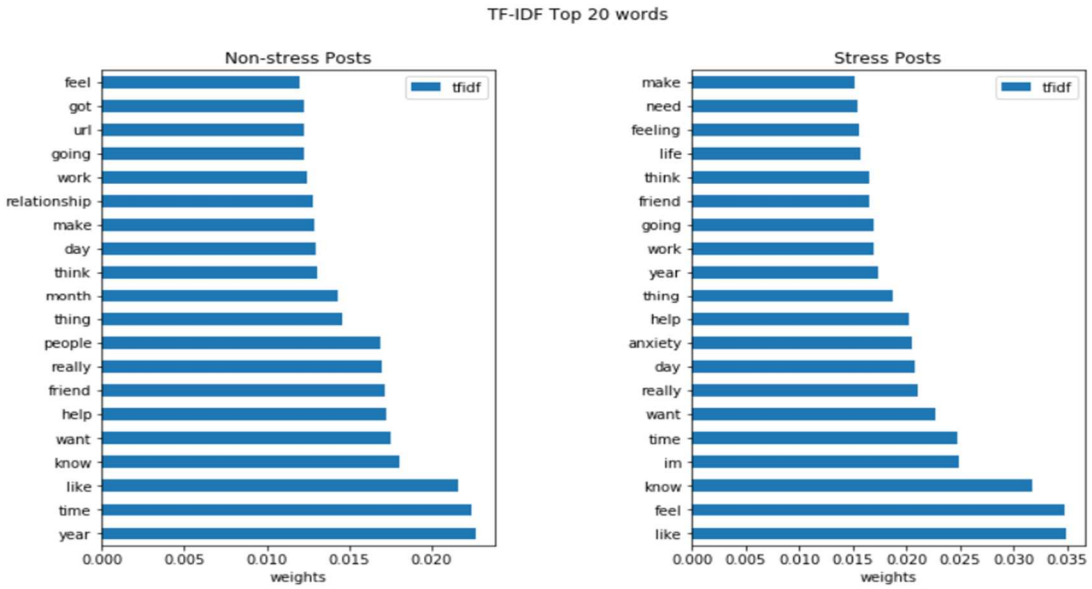


Fig. 4. TF-IDF Top 20 words

ii. N-Grams

N-grams are sequences of 'n' items from a given text or speech sample, where (n = 1, 2, 3). In the context of Natural Language Processing (NLP), these items are typically words. N-grams are used to model the probability of a word (or sequence of words) following a given sequence of words, aiding in tasks such as text prediction, classification, and language modeling. The choice of 'n' affects the balance between the detail of the model and its computational complexity. Although implementing N-grams is computationally less expensive than other advanced NLP techniques, it presents the challenge of data sparsity while employing it. This can be due to various reasons such as limited training dataset, attempt of overfitting, and higher memory requirements. Addressing the data sparsity challenge while employing n-grams requires employing various techniques such as smoothening, pruning, and data augmentation.

D. Transformer Models

i. Bidirectional Encoder Representation from Transformers

BERT applies the bidirectional training of Transformer from a popular attention mechanism that learns the contextual relation between the words in a text to language modeling. According to Devlin et al. [12], BERT is a "language understanding" model. Similar to transfer learning in computer vision neural networks, a neural network is trained to learn a language in the bidirectional encoder representations from the transformer-based method. The neural network then follows the linguistic representation in a bidirectional manner, examining the words that come before and after each other. Its success in Natural Language Processing can be attributed to the combination of these techniques. In this study, we follow the process by choosing a model for classification tasks from the 'transformers' library and then configuring the model for the no of classes. Converted the tokenized data into a format for PyTorch models. Since, the model is computationally expensive, therefore, various fine-tuning and training are done to achieve the best performance. Table I presents the parameters used in BERT for the classification of different mental states.

TABLE I. BERT MODEL PARAMETERS

Models	***BERT Base-uncased***
Parameters	110M
Batch size	32, 64
Epochs	50
Optimizers	AdamW
Learning rate	2×10^{-5}
Length Embeddings	27

ii. Robustly Optimized BERT Pretraining Approach

The RoBERTa model, an optimization of BERT, is designed for more robust performance in NLP tasks. It differs in several key aspects, such as training on larger datasets, longer sequences, and with more iterations. RoBERTa does not employ the next sentence prediction objective used in BERT and dynamically changes the masking pattern applied to the training data. While specific configurations may vary, RoBERTa models can have parameters ranging from 125 million in its base version to over 355 million in larger configurations. The base version of RoBERTa typically consists of 12 layers of transformer blocks. Each transformer block contains a set of self-attention mechanisms and feed-forward neural networks. The dimensionality of the model's hidden states was set to 768 in the model. This means that each token in the input sequence is represented by a vector of size 768 in the model's hidden layers. The maximum input sequence length accepted by RoBERTa was set to 512 tokens. RoBERTa requires high computational resources, typically involving GPUs or TPUs, to manage its extensive training requirements effectively. Fig. 5 shows how BERT and RoBERTa are differentiated based on static and dynamic masking [19] [20].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:34:53 UTC from IEEE Xplore. Restrictions apply.

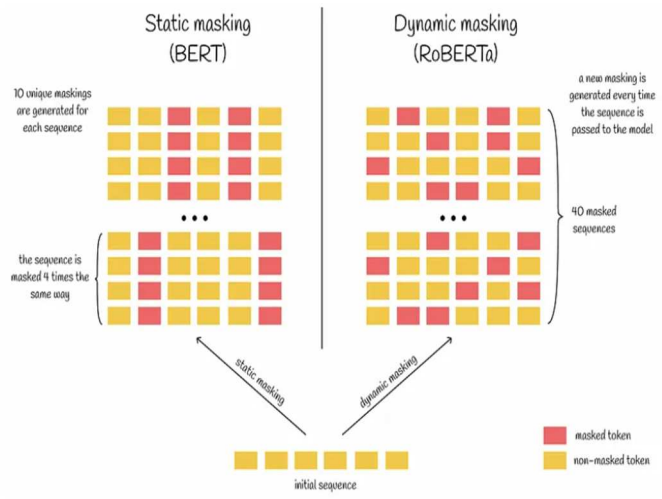


Fig. 5. Static masking vs. dynamic masking [19]

iii. A Lite BERT

ALBERT was created especially to balance speed and accuracy by lowering the overall size of the model and the amount of computing power needed, while still retaining or raising its performance than BERT. ALBERT outperforms BERT in terms of speed by utilizing fewer parameters. This size reduction means that inference and training periods are significantly shorter than before something that makes it appropriate for systems with low computer power or are very sensitive to response times for example in web searches. Parameters shared between layers and factorized embedding are utilized in the ALBERT model. It does this by having a considerable drop in the count of parameters. Specifically, the base model has only 12 million parameters while BERT has over 110 million parameters [21, 22]. Its use is effective in terms of both training duration and amount of memory consumed allowing it to be used in diverse natural language processing tasks with various resource intensity levels.

IV. EXPERIMENTAL EVALUATIONS

The data used in this analysis was sourced from Reddit to predict mental health and stress. The dataset was retrieved from various subreddits which are communities where people talk about their experiences, problems, and how to cope with mental health and stress. The Reddit platform allows people to remain anonymous which means they share everything without fear of revealing their identities. This allows for extensive data, with its rich and detailed information which could be used to find the subtle patterns of language and semantics that are specific to mental health conversations on the web. The dataset represents valuable insight for early identification and prediction of mental health using traditional ML algorithms, NLP, and transformer-based models. The dataset was divided into training and testing data in the ratio 80:20.

We train our models using the following methods. We start by carrying out data cleaning (handling null values, categorical values, and removing outliers) and preprocessing (tokenization, removing stop words, stemming, and lemmatization). For the classification task 'BertForSequenceClassification' model from the transformers library has been chosen and the model for the no of classes. We conducted fine-tuning and training where AdamW was used as an optimizer in the training and cross-entropy loss function was used. The training was set to 50 epochs with batch sizes of 32 and 64 respectively. For the evaluation of the model performance metrics such as accuracy, precision, recall, F1-score, classification report, and confusion matrix were used on the testing dataset to evaluate the performance of the models.

Despite raw model performance, the process of choosing the best classification model in practice will frequently take into account a wide range of factors, such as computational overhead and the degree of explainability of the model. We use the F1 score as the main criterion for model evaluation to emphasize the best-performing model.

TABLE II. COMPARISON OF F1-SCORES IN THE MENTAL HEALTH CLASSIFICATION TASK

Category	***Classifier***	***F1-score***
Traditional ML Algorithms	Logistic Regression Naïve Bayes Linear SVC Random Forest	0.94 0.93 0.93 0.86
NLP techniques	TF-IDF N-gram	0.85 0.84
Transformer based models	BERT RoBERTa ALBERT	0.95 0.97 0.94

The findings from TABLE II, shown above suggest that the transformer-based model performs better than the traditional ML algorithms and NLP techniques such as TF-IDF. Among the transformer models, RoBERTa shows a better F1-score from BERT as well as ALBERT. In addition to the traditional ML algorithms, logistic regression stays on the top with an F1-score of 0.92, in comparison to Naïve Bayes, Linear SVC, and random forest.

V. CONCLUSION AND FUTURE WORK

This research paper conclusively demonstrates that advanced NLP techniques like TF-IDF, n-gram, BERT, ALBERT, and RoBERTa, significantly outperform traditional machine learning methods in the context of mental health and stress prediction. Our findings underscore the potential of transformer-based models to significantly enhance prediction accuracy by capturing the context of language used in expressing mental health conditions. The efficacy of transformer-based models in understanding context-rich expressions enables a deeper comprehension of mental states, facilitating early intervention and personalized care strategies. Our work lays the groundwork for future studies to build upon, highlighting the need for continued innovation and interdisciplinary collaboration to harness the full potential of NLP in mental health and stress prediction.

REFERENCES

[1] M. Danner, B. Hadzic, S. Gerhardt, S. Ludwig, I. Uslu, P. Shao, T. Weber, Y. Shiban, and M. Ratsch, "Advancing Mental Health

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:34:53 UTC from IEEE Xplore. Restrictions apply.

Diagnostics: GPT-Based Method for Depression Detection". In *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, September 2023, (pp. 1290-1296). IEEE.

[2] V. Vajre, M. Naylor, U. Kamath, and A. Shehu. "PsychBERT: a mental health language model for social media mental health behavioral analysis". In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* December 2021, (pp. 1077-1082). IEEE.

[3] N. Proferes, N. Jones, S. Gilbert, C. Fiesle and M. Zimmer "Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics". *Social Media+ Society*, *7*(2), 2021, p.20563051211019004.

[4] Dhelim, Sahraoui, et al. "Detecting Mental Distresses Using Social Behavior Analysis in the Context of COVID-19: A Survey." *ACM Computing Surveys* 2023.

[5] K. Yang, T. Zhang, H. Alhuzali and S. Ananiadou, "Cluster-level contrastive learning for emotion recognition in conversations". *IEEE Transactions on Affective Computing*, 2023.

[6] J. Devlin, M. Chang, W., K. Lee and K. Toutanova "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805,* 2018.

[7] S. Zanwar, D. Wiechmann, Y. Qiao, and E. Kerz, "Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health Status on Social Media". *arXiv preprint arXiv:2212.09839,* 2022.

[8] I.J. Dristy, A.M. Saad, and A.A. Rasel, "Mental Health Status Prediction Using ML Classifiers with NLP-Based Approaches". In *2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET),* December 2022, (pp. 1-6). IEEE.

[9] V.M. Deshmukh, B. Rajalakshmi, S. Dash, P. Kulkarni and Gupta, S.K. "Analysis and characterization of mental health conditions based on user content on social media". In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI),* January 2022, (pp. 1-5). IEEE.

[10] M.E. Villa-Pérez, L.A. Trejo, M.B. Moin, and E. Stroulia "Extracting Mental Health Indicators From English and Spanish Social Media: A Machine Learning Approach". *IEEE Access*, *11*, 2023, pp.128135-128152.

[11] M. Nouman, H. Sara, S.Y. Khoo, M.P. Mahmud, and A.Z. Kouzani, "Mental Health Prediction through Text Chat Conversations". In *2023 International Joint Conference on Neural Networks (IJCNN),* June 2023, (pp. 1-6). IEEE.

[12] B.L. Cook, A.M. Progovac, P. Chen, B. Mullin, S. Hou and E. Baca-Garcia "Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid". *Computational and mathematical* methods in medicine, 2016.

[13] J.M. Liu, D. Li, H. Cao, T. Ren, Z. Liao and J. Wu "Chatcounselor: A large language models for mental health support". *arXiv preprint arXiv:2309.15461,* 2023.

[14] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, "Towards interpretable mental health analysis with ChatGPT". *arXiv,* 2023.

[15] D. Van Le, J. Montgomery, K.C. Kirkby, and J. Scanlan. "Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting". *Journal of Biomedical Informatics*, *86*, pp.49-58, 2018.

[16] R. Stewart and S. Velupillai "Applied natural language processing in mental health big data". *Neuropsychopharmacology*, *46*(1), 2021, p.252.

[17] J. Pennington, R. Socher and C.D. Manning "Glove: Global vectors for word representation". In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* October 2014, (pp. 1532-1543).

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781,* 2013.

[19] V. Efimov "Large language models: ROBERTA — a robustly optimized BERT approach". *Medium*. https://towardsdatascience.com/roberta-1ef07226c8d8 ,2023.

[20] *RoBERTa* .(n.d.).https://huggingface.co/docs/transformers/ model_doc /roberta, (2024)

[21] *ALBERT*.(n.d.). https://huggingface.co/docs/transformers/model_doc/ albert, (2024)

[22] *BERT*.(n.d.).https://huggingface.co/docs/transformers/model_doc/ber t, (2024)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2024 at 04:34:53 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384240855

Volatility Clustering and Long-Memory -a FIGARCH Analysis of Selected Currency Pairs

Preprint · September 2024

CITATIONS
0

READS
3

2 authors, including:



Sanjay Kumar
Dyal Singh College
6 PUBLICATIONS **2** CITATIONS

SEE PROFILE

All content following this page was uploaded by Sanjay Kumar on 22 September 2024.

The user has requested enhancement of the downloaded file.

Modelling Persistence and Asymmetry in Currency Volatility: A Multi-Country Application of FIGARCH Models

Sanjay Kumar*[1] and Dr. Nand Kumar†[2]

[1,2]Department of Humanities, Delhi Technological University, New Delhi.

September 22, 2024

*M. Phil in Economics. Research Scholar, Department of Humanities, Delhi Technological University, New Delhi. Dyal Singh College, Delhi University, New Delhi.sanjaykumar.eco@dsc.du.ac.in Orcid: http://orcid.org/0000-0003-4613-076X

†Ph.D. in Economics.Professor, Department of Humanities, Delhi Technological University, New Delhi.nandkumar@dce.ac.in Orcid: http://orcid.org/0000-0002-5684-6078

Volatility Clustering and Long- Memory -a FIGARCH Analysis of Selected Currency Pairs

Abstract

This paper employs the Fractionally Integrated GARCH (FIGARCH) model to analyze the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD, using data from January 4, 1999, to August 6, 2021. Before model fitting, the squared-innovation series is examined for the presence of long memory through informal and formal methods. While the Hurst Exponent (H) does not indicate long memory, the Local Whittle Estimator and GPH estimator suggest evidence of long memory for all pairs except JPY-USD.

The findings of the FIGARCH study reveal significant short-term volatility clustering and persistence in volatility, with negligible average returns for the currency pairs. These results contribute to the existing literature on financial time series analysis and have practical implications for market participants. Understanding the presence of long memory in volatility can help traders, risk managers, and policymakers anticipate and manage risks associated with extreme market events and volatility clustering, leading to more informed decision-making in currency markets. The study highlights the importance of considering long memory in volatility when analysing and modelling exchange rate dynamics.

Keywords: FIGARCH, long memory, volatility clustering, exchange rates, risk management.

JEL Classification: C22, C58, F31, G15

1 Introduction

The complex dynamics of financial time series, particularly in foreign exchange markets, have long fascinated researchers and practitioners alike. Three interrelated and stylized facts about financial time series resulting from autocorrelation-induced dependency are long memory, volatility clusters, and fat tails (Cont, 2001). Accurately modeling and forecasting the volatility of exchange rates is crucial for effective risk management, asset allocation, and derivative pricing (Andersen et al., 2000).

The seminal works of Engle (1982) and Bollerslev (1986) introduced the ARCH and GARCH models, respectively, which have become the standard tools for capturing volatility clustering in financial time series. However, these models fail to account for the long memory property often observed in the volatility of asset returns (Baillie et al., 1996). Long memory, also known as long-range dependence or persistence, refers to the slow decay of autocorrelations at long lags, implying that past volatility has a long-lasting impact on future volatility (McElroy and Politis, 2020).

To address this shortcoming, Baillie et al. (1996) introduced the Fractionally Integrated GARCH (FIGARCH) model, which extends the GARCH framework by allowing for fractional integration in the volatility process. The FIGARCH model

has gained popularity due to its ability to capture both short-term volatility clustering and long memory in financial time series (Davidson, 2004; Belkhouja and Boutahary, 2011).

Despite the growing body of literature on FIGARCH models, the understanding of long memory in the volatility of foreign exchange rates remains limited, particularly for emerging market currencies. Moreover, the practical implications of long memory for market participants, such as traders, risk managers, and policymakers, have not been fully explored.

This paper aims to fill these gaps by employing the FIGARCH model to analyse the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD. By examining the presence of long memory in the volatility of these currency pairs, we contribute to the existing literature on financial time series analysis and provide valuable insights for market participants.

Our study is motivated by the need for a better understanding of the long-term dynamics of exchange rate volatility and its potential implications for risk management and decision-making in currency markets. The inclusion of an emerging market currency (INR) alongside major developed market currencies (EUR, GBP, JPY) allows for a comparative analysis and sheds light on the differences in volatility characteristics across markets.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature on FIGARCH models and their applications in finance. Section 3 describes the data and methodology employed in our analysis. Section 4 presents the empirical results and discusses their implications. Finally, Section 5 concludes the paper and offers suggestions for future research.

2 Literature Review

The application of fractionally integrated models to capture long memory in financial time series has gained significant attention in the literature. Granger (1980) introduced the concept of long memory and its implications for economic and financial time series. Hosking (1981) and Geweke and Porter-Hudak (1983) further developed the theory and estimation methods for fractionally integrated processes.

In the context of volatility modeling, Ding et al. (1993) found evidence of long memory in the absolute and squared returns of the SP 500 index. Andersen and Bollerslev (1997) and Baillie et al. (1996) confirmed the presence of long memory in the volatility of foreign exchange rates using high-frequency data and the FIGARCH model, respectively.

Since the introduction of the FIGARCH model by Baillie et al. (1996), numerous

studies have applied the model to various financial assets. Bollerslev and Mikkelsen (1996) used the FIGARCH model to analyze the volatility of exchange rates and found evidence of long memory in the Deutsche Mark-US Dollar and Japanese Yen-US Dollar rates. Tse (1998) applied the FIGARCH model to stock market indices and demonstrated its superiority over the GARCH model in capturing long memory.

Beine et al. (2002) investigated the impact of central bank interventions on the volatility of exchange rates using the FIGARCH model. They found that accounting for long memory and structural breaks improves the fit of the model. Conrad and Haag (2006) examined the persistence of volatility in the FIGARCH model and derived the impulse response function for the conditional variance.

More recently, Kiliç (2011) proposed a smooth transition FIGARCH (ST-FIGARCH) model to capture both long memory and nonlinearity in the volatility of exchange rates and stock market indices. Belkhouja and Boutahary (2011) employed a time-varying FIGARCH model to study the volatility of crude oil prices and found evidence of long memory and structural changes.

In the context of emerging markets, Goudarzi and Ramanarayanan (2010) applied the FIGARCH model to the Indian stock market and found strong evidence of long memory in volatility. Oberholzer and Venter (2015) used the FIGARCH model to

analyze the volatility of the South African Rand against major currencies and found that long memory is more pronounced in the Rand-US Dollar exchange rate compared to other currency pairs.

Several studies have compared the performance of FIGARCH with other long memory volatility models. Davidson (2004) introduced the hyperbolic GARCH (HYGARCH) model and showed that it outperforms the FIGARCH model in terms of goodness-of-fit and forecasting ability. Yalama and Sevil (2008) compared the forecasting performance of FIGARCH, HYGARCH, and fractionally integrated APARCH (FIAPARCH) models for the Turkish stock market and found that the FIAPARCH model provides the best out-of-sample forecasts.

3 Research Objectives

The primary objective of this study is to analyze the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD, using the Fractionally Integrated GARCH (FIGARCH) model. Specifically, we aim to:

1. Examine the presence of long memory in the volatility of the selected currency pairs using informal methods, such as the autocorrelation function and

periodogram, and formal tests, including the Hurst Exponent, Local Whittle Estimator, and GPH estimator.

2. Estimate the FIGARCH model for each currency pair to capture short-term volatility clustering and long-term persistence in volatility.

3. Assess the significance of the FIGARCH model parameters, particularly the fractional differencing parameter (d), which indicates the degree of long memory in volatility.

4. Investigate the practical implications of the findings for market participants, such as traders, risk managers, and policymakers, in terms of understanding and managing risks associated with extreme market events and volatility clustering.

The relevance of this study lies in its focus on four currency pairs that play a significant role in global financial markets. By examining the presence of long memory in the volatility of these pairs, we contribute to a deeper understanding of how volatility dynamics manifest in these specific markets. The inclusion of an emerging market currency (INR) alongside major developed market currencies (EUR, GBP, JPY) allows for a comparative analysis and provides insights into potential differences in volatility characteristics across markets.

Furthermore, this research aims to bridge the gap between theoretical models and practical applications by discussing the implications of long memory in volatility for market participants. Understanding the persistence of volatility can help investors, risk managers, and policymakers make more informed decisions when dealing with currency risk, portfolio optimisation, and market stability.

Overall, this study contributes to the existing literature on financial time series analysis by providing a comprehensive examination of long memory in the volatility of key currency pairs using the FIGARCH model, while also highlighting the practical significance of the findings for stakeholders in the foreign exchange market.

4 Methodology

4.1 The FIGARCH Process

Let t be an index of time and $P_t, t \in T$ be the observed price of an asset. Let $r_t = \ln P_t - \ln P_{t-1}$ be its return r_t such that its parameters θ are the maximum likelihood ($L(\theta)$) estimates of an appropriate mean model with pre-chosen density function (f_θ). Then we have:

$$L(\theta; r_1, r_2, \ldots, r_T) = \prod_{t=1}^{T} \log(f_\theta(r_t)) \tag{4.1}$$

The log-likelihood will be given by:

$$l(\theta; r_1, r_2, \ldots, r_T) = \sum_{t=1}^{T} \log(f_\theta(r_t)) \tag{4.2}$$

To get the estimator of θ, we maximize $l(\theta)$. Then, the innovations ϵ_t are:

$$\epsilon_t = r_t - \theta(B) r_t \tag{4.3}$$

where B is a backshift operator, $\theta(B) = \sum_{i=1}^{n} \theta_i B^i$, and ϵ_t is white noise with

$\mathrm{E}(\epsilon_t) = 0$ and $\mathrm{E}(\epsilon_t^2) = \sigma_t^2$.

The GARCH(p, q) model (Bollerslev, 1986) can be given as:

$$\sigma_t^2 = \omega + \alpha(B)\epsilon_t^2 + \beta(B)\sigma_t^2 \tag{4.4}$$

where $\omega, \alpha, \beta > 0$, and conditional variance is stationary for $\alpha + \beta < 1$.

$\alpha(B) = \sum_{i=1}^{q} \alpha_i B^i$ and $\beta(B) = \sum_{j=1}^{p} \beta_j B^j$. For $q = 0$, it is an ARCH(p) process,

and for $p = q = 0$, it becomes a white noise process. After rearranging, we get:

$$(1 - \alpha(B) - \beta(B))\epsilon_t^2 = \omega + (1 - \beta(B))(\sigma_t^2 - \epsilon_t^2) \tag{4.5}$$

Hence,

$$(1 - B)\phi(B)\epsilon_t^2 = \omega + (1 - \beta(B))(\sigma_t^2 - \epsilon_t^2) \tag{4.6}$$

Here, $\phi(B) = \sum_{i=1}^{m-1} \phi_i B^i$. In the fractionally integrated model, $(1 - B)$ is replaced by $(1 - B)^d$ for $0 < d < 1$. This is done to capture the slow hyperbolic decay of memory. Thus, we can write the FIGARCH(p, d, q) model equation as:

$$(1 - B)^d\phi(B)\epsilon_t^2 = \omega + (1 - \beta(B))(\sigma_t^2 - \epsilon_t^2) \tag{4.7}$$

We have:

$$(1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \tag{4.8}$$

Solving Equation (7) and truncating the lag polynomial at 1000 for operational

purposes (Baillie et al., 1996), we have the FIGARCH(p, d, q) model with $0 < d < 1$ (allowing for slow decline of volatility after volatility shocks, a phenomenon called long memory) as:

$$\sigma_t^2 = (\omega - \bar{\epsilon}^2) + \sum_{j=1}^{q} \alpha_j(\epsilon_{t-j}^2 - \bar{\epsilon}t - j^2) + \sum j = 1^p \beta_j(\sigma_{t-j}^2 - \epsilon_{t-j}^2) \tag{4.9}$$

Thus, GARCH(1, 1) and GARCH(1, 1, 1) can be taken as special cases of the FIGARCH(1, d, 1) model with $d = 0$ and $d = 1$, respectively.

4.2 Data and Model Estimation

The data set consists of daily exchange rates for four currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD. The sample period spans from January 4, 1999, to August 6, 2021, resulting in 5,669 observations for each currency pair.

We first calculate the logarithmic returns for each currency pair as $r_t = \ln P_t - \ln P_{t-1}$, where P_t is the exchange rate at time t. Following Baillie et al. (1996), we fit the mean model $r_t = \mu + \epsilon_t$ to all four return series.

Before estimating the FIGARCH model, we examine the squared returns for the presence of long memory using informal methods, such as the autocorrelation

function (ACF) and periodogram, and formal tests, including the Hurst Exponent (H), Local Whittle Estimator, and GPH estimator.

Finally, we estimate the FIGARCH(1, d, 1) model for each currency pair using the maximum likelihood method, assuming normal and Student's t distributions for the standardized innovations. The model parameters, including the fractional differencing parameter (d), are assessed for statistical significance, and the implications of the findings are discussed.

5 Research Findings

In this section, we present the results of the FIGARCH analysis of the squared returns of four selected currency pairs: Euro-USD, GBP-USD, JPY-USD, and INR-USD. The FIGARCH model, an extension of the GARCH model, is used to capture volatility and long memory, specifically to model volatility clustering. The model is specified in terms of a univariate conditional mean model, a variance model, a distribution model, and a fractional integration parameter. Before fitting the model to the data, we examine the squared-innovation data for the presence of long memory using the autocorrelation function and periodogram. Additionally, we analyze the data using the Hurst Exponent (H) (Hurst, 1956), Local Whittle

Estimator (Robinson, 1995), and GPH estimator (Geweke and Porter-Hudak, 1983).

The ADF test, tau statistics value (with p-values indicated in parentheses) for error terms of the four mean models with constant is -15.4397 (1.858e-036), -17.5335 (1.858e-036), -27.3391 (1.539e-051), and -76.5754 (0.0001), respectively. Very low p-values indicate that we can reject the null that the four series have unit root. Table 1 shows the summary statistics of the error terms of the mean models of all the four daily return series. All the four series are fat-tailed as indicated by excess kurtosis values. All the popular tests of normality point to the return series being non-normally distributed.

Table 1: Summary Statistics, Returns, 1999-21

Statistics/Currency Pairs	**EUROUSD**	**GBPUSD**	**INRUSD**	**JPYUSD**
Mean	-1.1643e-019	-1.1781e-020	-9.8762e-020	5.1043e-018
Median	-7.6303e-007	8.5727e-005	-9.7861e-005	-0.0012322
Maximum	0.044380	0.044380	0.039285	6.9988
Minimum	-0.081662	-0.081662	- - -0.037658	-0.053389
Std. Dev.	0.0058991	0.0058991	0.0042961	0.093181
Skewness	-0.099739	-0.65764	0.20986	74.763
Ex-Kurtosis	2.4824	10.688	10.810	5612.9
Jarque-Bera Test	1465	27390.6	27643.1	7.44703e+009
P Value	0	0	0	0
Doornik-Hansen Test	808.679	5119.98	6182.23	3.58164e+007
P Value	2.49787e-176	0	0	0
Shapiro-Wilk Test	0.978187	0.946218	0.879669	0.0138291
P Value	1.24335e-028	22.44678e-041	9.7656e-055	6.37173e-099
Lilliefors Test	0.048886	0.0446517	0.109872	0.4295
P Value	0	0	0	0

Source: Author's calculation.

Bai and Ng (2005), with the help of their test specially designed for time series,

tested the normality of 21 time series out of these six-time series are found to be non-symmetric and all of them are financial time series. In Table 2, we have presented the result of the Bai & Ng Normality Test for error terms of our mean models for all the four series as most popular tests for normality may not be applicable to financial time series. All return series seem to be non-normally distributed.

Table 2: Bai & Ng Long Run Normality Test, Returns 1999-21

Currency Pairs	π_3	$\pi_3/\sqrt{5}$	π_4	π_{34}
EuroUSD	1.187239	2.213913	3.498400	19.98223
	(0.117567)	(0.013417)	(0.000234)	(4,58E-05)
GBPUSD	0.878083	0.952704	1.343738	3.781008
	(0.189949)	(0.170370)	(0.089517)	(0.150996)
INRUSD	0.306107	0.458222	3.031943	10.32863
	(0.379762)	(0.323397)	(0.001215)	(0.005717)
JPYUSD	1.440571	2.335931	2.854371	17.10611
	(0.074853)	(0.009747)	(0.002156)	(0.000193)

Source: Author's calculation.

Note: p-values are in parentheses. As per Bai and Ng (2005), $\pi_3, \pi_4 \rightarrow N$, $\pi_{34} \rightarrow \chi^2$, 5% critical value is 1.96 for π_3, π_4 and 5.99 for π_{34} at 2df.

In Figure 2, we have plotted error square (ϵ^2) of the mean model of the FIGARCH process of Baillie et al. (1996).

An examination of Figure 2 reveals that all (ϵ^2) series, except JPY-USD series, have GARCH effect as well as long memory. In the case of Euro-USD, GBP-USD, and

INR-USD currency pairs, we have volatility and clustering around the 2008 global financial crisis. For GBP-USD, we have high volatility and clustering around the Brexit period, and in the case of Euro-USD, we also have high volatility clustering during the Brexit period. For autocorrelated series with k lag, we have $\epsilon^2 \propto \sum_k^T \rho_k$, i.e., clusters in the return series will be caused if lagged observations are significantly correlated, individually or jointly. Individual dependencies can be gauged by autocorrelation function at different lags and joint dependencies can be gauged by portmanteau tests. Long memory is indicated by a slow or hyperbolic decline of autocorrelation as opposed to fast or geometric decline in case of short-term memory. Figure 3 presents ACF of (ϵ^2) of all the four currency pairs. The autocorrelation function (ACF) at lag k is given by the power rule $\rho_k = c_\rho k^{2d-1}$ with $k \to \infty$, where c_ρ is a constant and d is a memory parameter. We have a sample size of 5669 for our series, hence 95% confidence interval for autocorrelations, as shown by the horizontal bars in Figure 3, using Bartlett standard error will be given by $\pm\frac{1.96}{\sqrt{5669}} = \pm 0.026$. The volatility clustering phenomenon seen in Figure 1 is confirmed by ACFs given in Figure 2 . The (ϵ^2) autocorrelations for very long lags frequently exceed the two 95% Bartlett (1946) confidence bands except for JPY-USD series. Also, the Ljung and Box (1978) portmanteau test for the joint significance of autocorrelations is quite significant for all lags with p-value equal to zero and this is for all the (ϵ^2) series except for

JPY-USD series for which we do not have any significant Q statistics]. This apparent persistence in the autocorrelation function is substantially reduced if we take the first difference of the (ϵ^2) series. This is indicated by reduced values for ACF though these ACF values are still significant at quite distant lags and the overall significance shown by Q statistics is still very high. The frequency domain definition of long memory states that the spectral density function is unbounded at some frequency in the interval [0,). Most of the empirical literature has concentrated on the case where the singularity, or pole, in the spectrum takes place at the zero frequency (Gil-Alana and Hualde, 2009). We present periodogram, with Bartlett window length of 150, of (ϵ^2) for all the four series in Figure 4. The figure reveals long memory in the case of all the four return series.

Following Engle and Bollerslev (1986), we will use fractionally differenced return series to capture the long memory in the series. Taylor (2008) studied power transformation of the return from a number of time series and found that squared returns are more likely to be autocorrelated than returns. For a formal examination of long memory, if any, in ϵ^2, we use Hurst Exponent (H) (Hurst et al., 1965), Local Whittle Estimator (Robinson, 1995), and GPH estimator (Geweke and Porter-Hudak, 1983) of memory parameter d. The results are reported in Table 3 and Table4, respectively. We do not find evidence of long memory in any of the series on basis of Hurst Exponent (H), but on the basis of Local Whittle Estimator

and GPH estimator, we find evidence of long memory. So we don't have a conclusive outcome here](Conclusion). Following Weron and Weron (2002), we estimate Hurst Exponent (H) = 0.5 + Empirical Hurst Exponent – Anis-Lloyd corrected Hurst Exponent = 0.5 + slope of (R/S)n – E(R/S)n. Here (R/S)n is empirical Hurst Exponent estimated by (R/S)n = CHn and E(R/S)n is the expected or theoretical Hurst Exponent at size n (Anis and Lloyd, 1976). There is a lot of debate in literature on the asymptomatic property of Hurst Exponent, original R/S analysis gave only the true value of the parameter H but it did not provide its probability distribution. Some researchers have shown that the original Hurst Exponent is normally distributed. Couillard and Davison (2005) studied 10,000 Hurst exponents from independent time series of 10,000 observations each and found them to be normally distributed. In this paper, we use 95% confidence interval suggested by Weron and Weron (2002) for Anis-Lloyd corrected H (R/S-AL). The estimated value of Empirical Hurst Exponent, Theoretical Hurst Exponent, and Anis-Lloyd corrected H (R/S-AL) Exponent, along with 95% confidence interval for Anis-Lloyd corrected H (R/S-AL), is reported in Table 3 .

GPH test, $H_0 : d \geq 0.5$, no presence of long memory; $H_1 : d < 0.5$ presence of long memory. LW test, $H_0 : d \geq 0.5$, no presence of long memory; $H_1 : d < 0.5$ presence of long memory, Level of Significance 5%.

Table 3: Hurst Exponent Test for Long Memory in Square Returns

	1.INR USD, Sample Size (T) = 5669	**2.EUROUSD, Sample Size (T) = 5669**	**3.GBPUSD, Sample Size (T) = 5669**	**4.JPYUSD, Sample Size (T) = 5669**
Empirical Hurst Exponent H(R/S)n	0.7771236	0.8956424	0.8956424	0.8229069
Theoretical Hurst Exponent H(E(R/S)n)	0.5287461	0.5287461	0.5287461	0.5287461
Anis-Lloyd Corrected Hurst Exponent H(R/S-AL) (= 0.5 + H(R/S)n - H(E(R/S)n))	0.7483775	0.8668963	0.8604084	0.7908433
95% Confidence Interval for Anis-Lloyd Hurst Exponent	-0.49931 <R/S-AL< 1.408454	-0.49931 <R/S-AL< 1.408454	-0.49931 <R/S-AL< 1.408454	-0.49931 <R/S-AL< 1.408454
H_0	Not rejected	Not rejected	Not rejected	Not rejected
Conclusion	No Long memory	No Long memory	No Long memory	No Long memory

Source: Author's calculation.

Table 4: Frequency Domain Long Memory Estimators

	1.INR USD, Sample Size (T) = 5669	**2.EUROUSD, Sample Size (T) = 5669**	**3.GBPUSD, Sample Size (T) = 5669**	**4.JPYUSD, Sample Size (T) = 5669**
Local Whittle Estimator (Robinson, 1995)				
m	177	177	177	177
d	0.291183	0.452571	0.298797	0.344853
Standard Error	0.0375823	0.0375823	0.0375823	0.0375823
Test statistic: z =	7.74786	12.0421	7.95048	9.17595
p-value	0.0000	0.0000	0.0000	0.0000
Conclusion	Reject the null. Long memory	Reject null. Long memory	Reject null. Long memory	Reject null. Long memory
GPH Estimator				
m	177	177	177	177
d	0.293947	0.518973	0.297206	0.40157
Standard Error	0.0465883	0.0502571	0.0393567	0.0489204
Test statistic: t(175) =	6.30947	0.0502571	7.55159	8.20865
p-value	0.0000	0.0000	0.0000	0.0000
Conclusion	Reject the null. Long memory	Reject the null. Long memory	Reject the null. Long memory	Reject the null. Long memory

Source: Author's calculation.

The estimated values of the FIGARCH(1, d, 1) model with normal and Student's t distributions for the four currency pairs are presented in Table 5 and Table 6, respectively. The mean return (μ) is estimated to be very close to zero for all currency pairs, suggesting negligible average returns. The constant term (ω) in the volatility equation is not statistically significant for any of the pairs, indicating that the models do not require a constant mean or variance component.

However, both past squared shocks (α_1) and long memory (δ) have significant impacts on current volatility for all currency pairs, suggesting the presence of short-term volatility clustering and strong persistence in volatility. The Student's t distribution parameter (ν) is also significant for all currency pairs, indicating the presence of heavy tails in the standardized residuals.

Table 5: FIGARCH(1, d, 1) Mean Model: ARFIMA(0, 0, 0) Distribution Model: Normal. Optimal Parameters

Currency Pairs	μ	ω	α_1	β_1	δ
EuroUSD	-0.000036	0.000000	0.051781	0.937948	0.887184
	(0.57868)	(0.37205)	(0.00000)	(0.00000)	(0.00000)
GBPUSD	0.000031	0.000001	0.242715	0.654901	0.446753
	(0.641479)	(0.007183)	(0.000000)	(0.000000)	(0.000000)
INRUSD	0.000016	0.000000	0.151041	0.914590	0.956082
	(0.626094)	(0.012456)	(0.000000)	(0.000000)	(0.000000)
JPYUSD	0.000016	0.000000	0.151041	0.914590	0.956082
	(0.626094)	(0.012456)	(0.000000)	(0.000000)	(0.000000)

Source: Author's calculation.

These findings have several important implications for market participants. The presence of long memory in volatility suggests that extreme market events and

Table 6: FIGARCH(1, d, 1) Mean Model: ARFIMA(0, 0, 0) Distribution Model: Student's t distribution. Optimal Parameters

Currency Pairs	μ	ω	α_1	β_1	δ	ν
EuroUSD	-0.000030	0.000000	0.052874	0.937948	0.883961	8.385051
	(0.634173)	(0.144420)	(0.000005)	(0.00000)	(0.00000)	(0.000000)
GBPUSD	0.000055	0.000000	0.004340	0.951721	0.984915	8.313940
	(0.39247)	(0.19537)	(0.82997)	(0.000000)	(0.000000)	(0.00000-)
INRUSD	-0.000037	0.000000	0.110330	0.888670	0.956082	4.250676
	(0.080559)	(0.934861)	(0.000000)	(0.000000)	(0.000000)	(0.00000)
JPYUSD	0.000016	0.000000	0.151041	0.914590	0.956082	8.310940
	(0.626094)	(0.012456)	(0.000000)	(0.000000)	(0.000000)	(0.00000)

Source: Author's calculation.

volatility clustering are more likely to occur than in a standard GARCH setting, and market participants should be prepared for prolonged periods of high or low volatility. The heavy tails of the standardized residuals also highlight the importance of accounting for extreme events in risk models and decision-making processes.

6 Summary and Conclusion

In this study, we employed the Fractionally Integrated GARCH (FIGARCH) model to analyze the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD, using data from January 4, 1999, to August 6, 2021. Our primary objectives were to examine the presence of long memory in volatility, estimate the FIGARCH model to capture short-term volatility clustering and long-term persistence, and investigate the practical implications of

our findings for market participants.

In summary, our FIGARCH analysis of the daily USD returns for four major currency pairs (Euro-USD, GBP-USD, INR-USD, and JPY-USD) reveals several important findings. The preliminary analysis shows that the return series are stationary but exhibit fat tails and non-normality. The squared returns display volatility clustering and long memory, particularly for the Euro-USD, GBP-USD, and INR-USD pairs, as evidenced by the autocorrelation function and periodogram.

The formal tests for long memory provide mixed results, with the Hurst Exponent not indicating long memory for any of the series, while the Local Whittle Estimator and GPH estimator suggest evidence of long memory for all pairs except JPY-USD. The FIGARCH(1, d, 1) model estimates confirm the presence of short-term volatility clustering and strong persistence in volatility for all currency pairs, as indicated by the significant past squared shocks (α_1) and long memory (δ) parameters. The heavy tails of the standardized residuals are also evident from the significant Student's t distribution parameter (ν).

Our study contributes to the existing literature on financial time series analysis by providing a comprehensive examination of long memory in the volatility of key currency pairs using the FIGARCH model. The inclusion of an emerging market currency (INR) alongside major developed market currencies (EUR, GBP, JPY)

allows for a comparative analysis and provides insights into potential differences in volatility characteristics across markets.

7 Policy Prescription

The findings of our study have several important implications for policymakers, particularly central banks and financial regulators. Understanding the presence of long memory in the volatility of major currency pairs can help inform policy decisions related to exchange rate stability, financial market regulation, and risk management.

First, central banks should take into account the persistence of volatility when designing and implementing exchange rate policies. The presence of long memory in volatility suggests that extreme market events and prolonged periods of high or low volatility are more likely to occur than in a standard GARCH setting. This implies that central banks may need to be prepared for more frequent and longer-lasting interventions in foreign exchange markets to maintain stability and prevent excessive fluctuations in exchange rates.

Second, financial regulators should consider the implications of long memory in volatility when setting capital requirements and other risk management standards

for financial institutions. The heavy tails of the standardized residuals in our FIGARCH model estimates highlight the importance of accounting for extreme events in risk models. Regulators may need to require financial institutions to hold more capital or maintain higher liquidity buffers to cushion against potential losses arising from prolonged periods of high volatility.

Third, policymakers should promote the development and use of more sophisticated risk management tools that take into account the presence of long memory in volatility. This could include encouraging the adoption of long memory volatility models, such as FIGARCH, HYGARCH, or FIAPARCH, by financial institutions and market participants. Policymakers could also support research and education initiatives aimed at improving the understanding and application of these models in practice.

Fourth, given the evidence of long memory in volatility, policymakers should be cautious about relying too heavily on short-term measures of volatility, such as implied volatility indices or short-term historical volatility, when making policy decisions. These measures may not fully capture the long-term persistence of volatility and could lead to suboptimal policy choices. Instead, policymakers should consider using longer-term measures of volatility and incorporating the insights from long memory models when formulating policies.

Finally, our findings highlight the importance of international cooperation and information sharing among policymakers and regulators. The presence of long memory in volatility in major currency pairs suggests that volatility shocks in one market can have long-lasting effects on other markets. This underscores the need for policymakers to work together to monitor and manage risks arising from cross-border financial market linkages.

8 Limitations and Future Research

While our study provides valuable insights into the presence of long memory in the volatility of major currency pairs and its implications for market participants and policymakers, it is important to acknowledge the limitations of our analysis and identify areas for future research.

One limitation of our study is the mixed results from the formal tests for long memory. While the Local Whittle Estimator and GPH estimator suggested evidence of long memory for all currency pairs except JPY-USD, the Hurst Exponent did not indicate long memory for any of the series. These mixed results suggest that further research may be needed to fully understand the nature of long memory in these currency pairs and to develop more robust tests for detecting long memory in

financial time series.

Another limitation is that our analysis focuses on a specific time period (1999-2021) and may not capture potential changes in volatility dynamics over time. Future research could extend our analysis by considering longer time periods or by using rolling window estimation techniques to examine how the degree of long memory in volatility evolves over time. This could provide insights into how changes in economic conditions, financial market structure, or policy regimes affect the persistence of volatility in foreign exchange markets.

Our study also focuses on a limited set of currency pairs (Euro-USD, GBP-USD, INR-USD, and JPY-USD). While these pairs represent a mix of major developed and emerging market currencies, future research could extend the analysis to a broader set of currency pairs to provide a more comprehensive understanding of long memory in foreign exchange market volatility. This could include examining the presence of long memory in the volatility of other major currency pairs, such as USD-CHF or USD-CAD, as well as in the volatility of currency pairs involving other emerging market currencies, such as USD-BRL or USD-ZAR.

Another area for future research is to explore the use of alternative long memory volatility models, such as the Hyperbolic GARCH (HYGARCH) or the Fractionally Integrated Asymmetric Power ARCH (FIAPARCH) models, to assess the robustness

of our findings. These models may provide additional insights into the nature of long memory in volatility and its implications for risk management and policy design.

Future research could also investigate the economic factors driving the differences in long memory across currency pairs. This could involve examining how differences in macroeconomic fundamentals, financial market development, or institutional factors affect the persistence of volatility in different currency markets. Understanding the underlying drivers of long memory in volatility could help market participants and policymakers better anticipate and manage risks arising from volatility persistence.

Finally, future research could explore the implications of long memory in volatility for other financial markets, such as commodities or derivatives. Long memory in volatility may have important implications for the pricing and risk management of financial derivatives, such as options or futures contracts, as well as for the hedging strategies used by market participants. Examining the presence and implications of long memory in volatility across different asset classes could provide a more comprehensive understanding of the role of long memory in financial market dynamics.

References

T. G. Andersen and T. Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2-3):115–158, 1997.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Exchange rate returns standardized by realized volatility are (nearly) gaussian. *Multinational Finance Journal*, 4(3/4):159–179, 2000.

A. A. Anis and E. H. Lloyd. Expected value of the adjusted rescaled hurst range of independent normal summands. *Biometrika*, 63(1):111–116, 1976.

J. Bai and S. Ng. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics*, 23(1):49–60, 2005.

R. T. Baillie, T. Bollerslev, and H. O. Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1):3–30, 1996.

M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41, 1946.

M. Beine, A. B'enassy-Qu'er'e, and C. Lecourt. Central bank intervention and

foreign exchange rates: new evidence from figarch estimations. *Journal of International Money and Finance*, 21(1):115–144, 2002.

M. Belkhouja and M. Boutahary. Modeling volatility with time-varying figarch models. *Economic Modelling*, 28(3):1106–1116, 2011.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

T. Bollerslev and H. O. Mikkelsen. Modeling and pricing long memory in stock market volatility. *Journal of Econometrics*, 73(1):151–184, 1996.

C. Conrad and B. R. Haag. Inequality constraints in the fractionally integrated garch model. *Journal of Financial Econometrics*, 4(3):413–449, 2006.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

M. Couillard and M. Davison. Hurst exponents and the prediction of stock market crashes. *Studies in Economics and Finance*, 2005.

J. Davidson. Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *Journal of Business Economic Statistics*, 22(1):16–29, 2004.

Z. Ding, C. W. Granger, and R. F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106, 1993.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, pages 987–1007, 1982.

R. F. Engle and T. Bollerslev. Modelling the persistence of conditional variances. *Econometric reviews*, 5(1):1–50, 1986.

J. Geweke and S. Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4):221–238, 1983.

L. A. Gil-Alana and J. Hualde. Fractional integration and cointegration: An overview and an empirical application. *Palgrave Handbook of Econometrics*, pages 434–469, 2009.

H. Goudarzi and C. Ramanarayanan. Modeling and estimation of volatility in the indian stock market. *International Journal of Business and Management*, 5(2): 85–98, 2010.

C. W. Granger. Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14(2):227–238, 1980.

J. R. Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 1981.

H. E. Hurst. Methods of using long-term storage in reservoirs. *Proceedings of the Institution of Civil Engineers*, 5(5):519–543, 1956.

H. E. Hurst, R. Black, and Y. Simaika. Long-term storage: an experimental study. *London*, 1965.

R. Kiliç. Long memory and nonlinearity in conditional variances: A smooth transition figarch model. *Journal of Empirical Finance*, 18(2):368–378, 2011.

G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

T. McElroy and D. N. Politis. Properties of the figarch and related processes. *Econometric Theory*, 36(6):1266–1289, 2020.

N. Oberholzer and P. Venter. Univariate garch models applied to the jse/ftse stock indices. *Procedia Economics and Finance*, 24:491–500, 2015.

P. M. Robinson. Log-periodogram regression of time series with long range dependence. *The annals of Statistics*, 23(3):1048–1072, 1995.

S. J. Taylor. *Modelling financial time series*. world scientific, 2008.

Y. K. Tse. The conditional heteroscedasticity of the yen–dollar exchange rate. *Journal of Applied Econometrics*, 13(1):49–55, 1998.

R. Weron and A. Weron. Hurst analysis in matlab. *HSC Research Report, Wrocław University of Technology*, 2002.

A. Yalama and G. Sevil. Forecasting world stock markets volatility. *International Research Journal of Finance and Economics*, 15:159–174, 2008.

Appendix: Figures

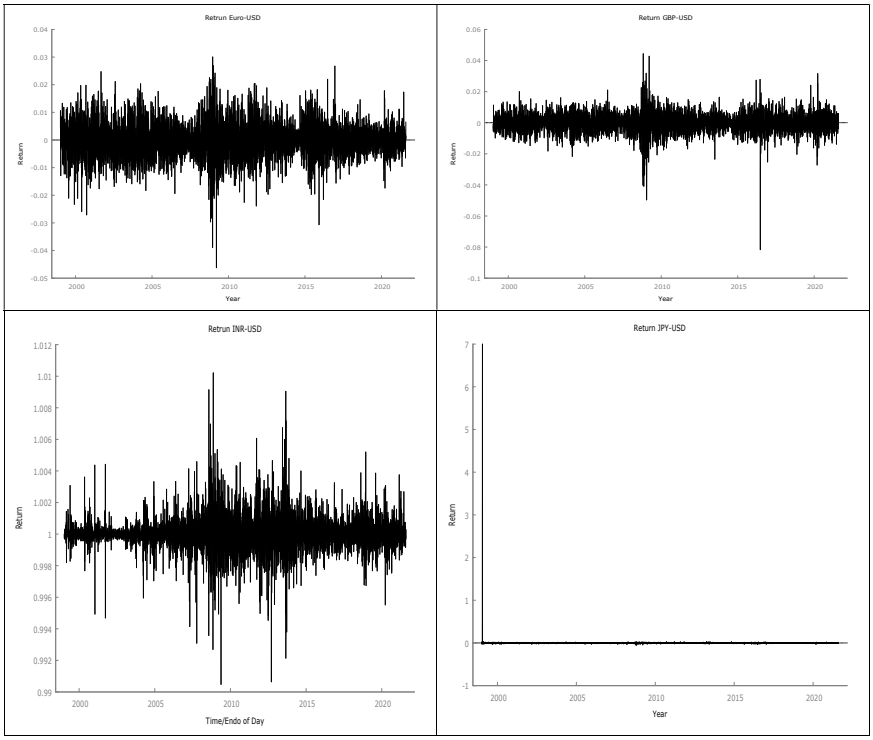


Figure 1: Time Series Plot, Daily Returns Selected Currency Pairs

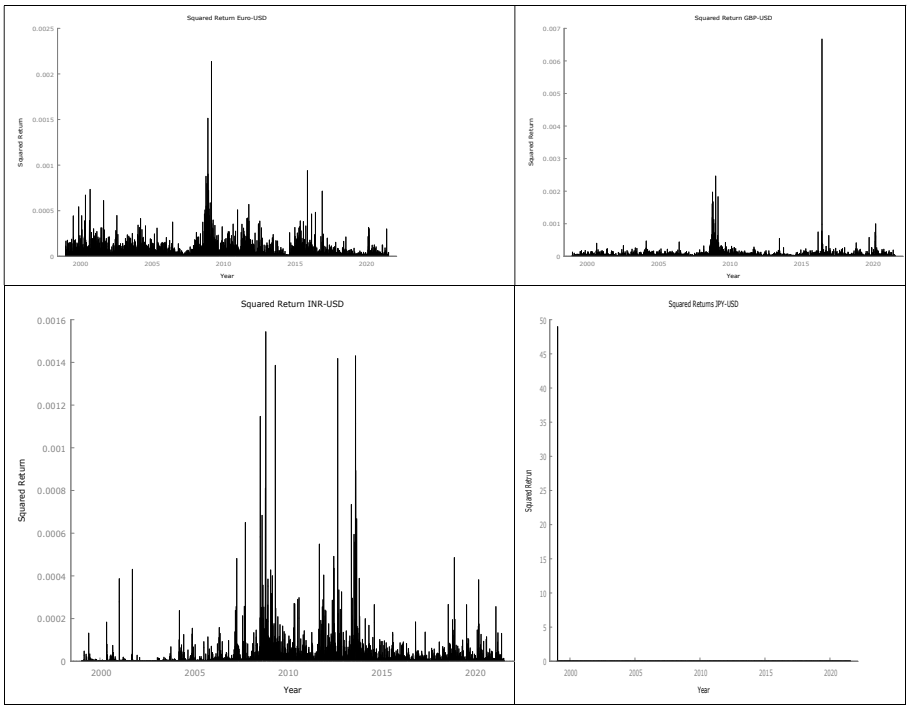


Figure 2: Square of Daily-Rate-of-Return Selected Currency Pairs

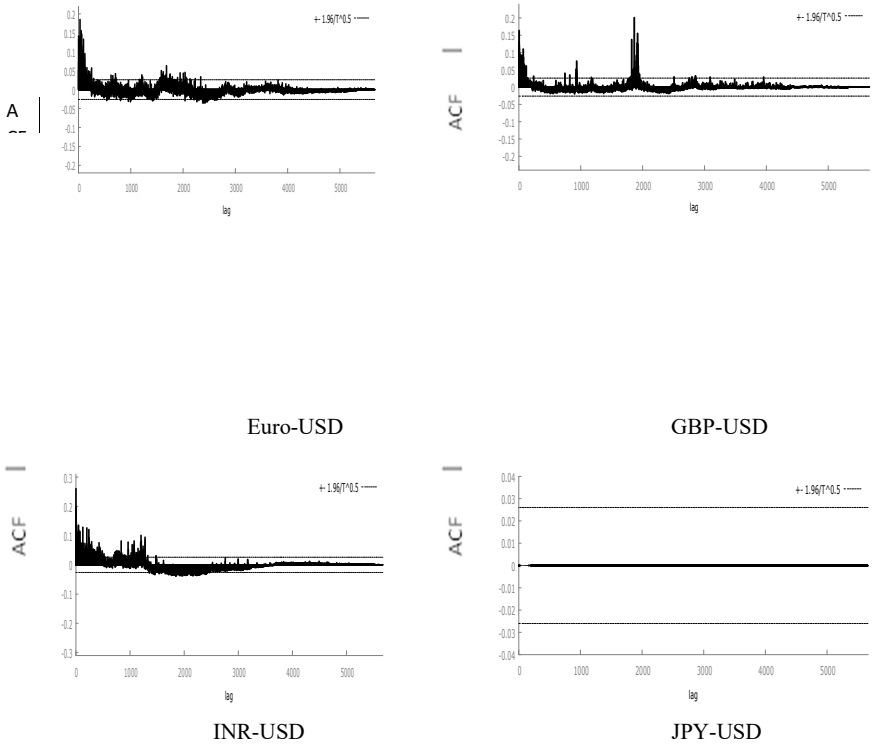


Figure 3: ACF, Squared Returns

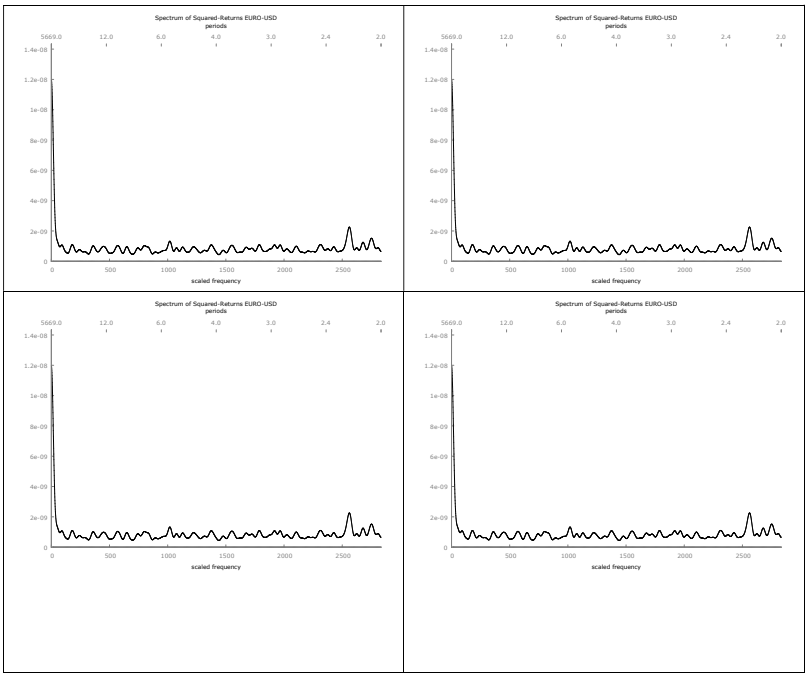


Figure 4: Periodogram, Squared Returns

View publication stats

Physica B 695 (2024) 416522

Contents lists available at ScienceDirect

Physica B: Condensed Matter

journal homepage: www.elsevier.com/locate/physb

Optical characteristics, Judd-Ofelt analysis of enhanced luminescence by flux in thermally stable, novel Eu^{3+}- doped $BaZr(PO_4)_2$ phosphor for indoor lighting applications

Shweta Yadav, A.S. Rao *, Deshraj Meena

Department of Applied Physics, Delhi Technological University, Bawana Road, New Delhi, 110042, India

ARTICLE INFO

Keywords:
Photoluminescence
White LEDs
XRD
Temperature dependent PL
Decay
Flux

ABSTRACT

A series of novel single phased Eu^{3+} doped $BaZr(PO_4)_2$ phosphor were synthesized employing simple solid-state reaction route and studied the influence of different chlorides as fluxes (NH_4Cl, KCl, LiCl and NaCl) on structural and photoluminescence properties. The structural characterization reveals pure phase formation of the as prepared phosphors with monoclinic crystal structure. The photoluminescence (PL) spectra recorded under 392 nm excitation depicts intense emission at 612 nm. Further, the concentration of Eu^{3+} ion was well tuned by carefully studying the PL spectra and the optimised amount of Eu^{3+} ion in $BaZr(PO_4)_2$ is found to be 3.0 mol%. The luminescence properties of phosphor with added aforementioned fluxes were also studied in detail. NH_4Cl was found to be the best among all the fluxes taken and the photoluminescence intensity of phosphor gets enhanced by 1.43 times of without flux. The Judd-Ofelt theory was employed for determination of spectral parameters from photoluminescent emission spectra. The temperature dependent photoluminescent studies establishes stability of phosphor emission at operating temperatures. The CIE coordinates were evaluated that confirms the pure red emission under n-UV excitation. All the studies indicate the potential of the said phosphor in phosphor converted white light emitting diodes as red emitter and indoor lighting.

1. Introduction

Historically, economic upsurge has been inextricably linked with elevated energy requirements thereupon raised up carbon emissions (greenhouse gas emissions), causing remarkable impact on environment [1]. So the energy efficient devices and clean energy sources play important role in attaining sustainable development goals [2]. It is widely known that a large portion of power consumption is directly contributed by lighting power demand. With emergence of solid state lighting, the traditional filament-based lamps and compact fluorescent lamps were proved incompatible for efficient lighting owing to their demerits of high energy consumption, low efficiency and short lifespan. White light emitting diodes (wLEDs) nevertheless stands tall on all fronts with its relevant merits of low power consumption, extended lifetime, environment friendliness, stable physical and chemical properties high luminous efficiency wide colour gamut, lower operating temperature and robustness [3–6]. Nowadays, two common routes are used for fabrication of wLEDs. The first method employ integrating Ce^{3+} activated $Y_3Al_5O_{12}$ (YAG:Ce^{3+}) yellow phosphor with a InGaN LED chip having blue emission. Though this approach gives high photoluminescence (PL) efficiency with blue light excitation, the obtained wLEDs have below par color-rendering indices (CRI) and elevated corelated color temperatures on account of missing emission in red region [7,8]. In the second approach improved CRI and tunable corelated color temperatures are achieved by devising Red, Green and Blue phosphor amalgamated on a LED chip having emission in near UV region [9]. The emission from this approach also possess weak emission in red region and suffers complex blending due to distinct characteristics of LEDs [10]. So efficient red phosphors need to be investigated for development of wLEDs by either of the approach. Commercially, for red phosphors, sulphides and nitrides-based materials are used. The synthesis of nitride-based phosphors is challenging and expensive due to high temperature and high pressure synthesis requirement [6,11]. The commercially used red emitting Y_2O_2S:Eu^{3+} phosphor has below par efficiency owing to low absorbance in near UV region. Furthermore, it suffers from poor chemical stability and decomposes into toxic by-products [12]. So, for development of highly efficient, environment friendly red phosphor to achieve high luminous efficiency of wLEDs, a

* Corresponding author.
E-mail address: drsallam@gmail.com (A.S. Rao).

https://doi.org/10.1016/j.physb.2024.416522
Received 10 July 2024; Received in revised form 30 August 2024; Accepted 7 September 2024
Available online 12 September 2024
0921-4526/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

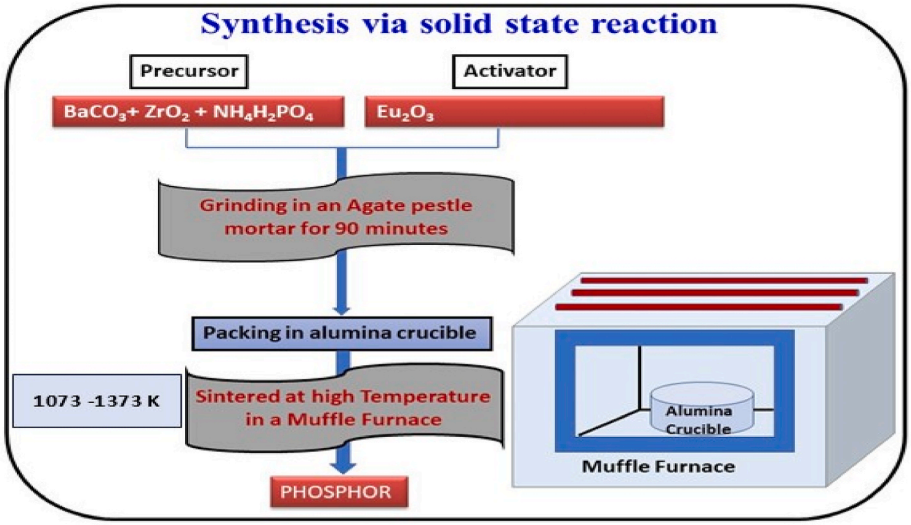


Fig. 1. Flowchart for solid-state reaction synthesis.

novel red emitting phosphor which is chemically and thermally stable is required. Among various activators, Eu^{3+} has significantly good luminescence with narrow emission band in red region, having absorption in near UV/blue region attributed to its 4f transition [3,4,6,8]. Moreover, in the past decades rare earth doped glasses and phosphors have drawn attention for their applications in solid state lasers, indoor lighting, and wLEDs [13,14].

Recently, many articles have been reported recognizing strong red emitted Eu^{3+} doped molybdates materials for their use in LEDs [5]. Eu^{3+} doped tungstate, zirconate, tantalate, phosphates etc have also been explored for potential material as efficient red phosphor [12,15–17]. Since the crystal field of host material can manipulate the emission and absorption from the activator ion so choice of host can affect the efficiency of phosphor [18]. In the last decade various works have been reported in which phosphates have emerged as a potential material for phosphors as they possess excellent thermal chemical and hydrolytic stability with low phonon energy [10,18–20]. Phosphate material offer favourable parameters like large optical band gap, high index of refraction, lower sintering temperatures and low synthesis cost to be considered as lucrative host material [10,21].

In the present study, we have investigated $BaZr(PO_4)_2$(BZP) as a novel host material for red emission using Eu^{3+} as activator. To the best of our knowledge, the as studied phosphate hasn't been reported for any PL studies. The effect of variation of Eu^{3+} concentration, and role of different chloride fluxes in improving luminescent properties has also been investigated systematically. Judd-Ofelt (JO) theory has been applied to understand various parameters and analysing them for indoor lighting applications.

2. Experimental &characterizations

A batch of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 1.0–5.0 mol%) powder samples were fabricated via simple solid reaction route. The schema chart depicting the process followed for optimisation of synthesis parameters to obtain pure phase is displayed in Fig. 1. The precursors Eu_2O_3(99.9 %), $BaCO_3$(99.99 %), ZrO_2(99.99 %) and $(NH_4)H_2PO_4$(99.99 %) taken were all highly pure and of analytical grade. First the undoped $BaZr(PO_4)_2$ was prepared by taking the stoichiometric amount of precursors mixed together in acetone in Agate mortar and grinded thoroughly for 90 min. The blend powder was sintered in a programmable high temperature furnace at variable temperatures from 1073 to 1373 K to obtain pure phase $BaZr(PO_4)_2$. The specimens were milled into fine powder for subsequent investigations. The optimum temperature for obtaining the pure phase was identified at 1373 K. Under the aforementioned temperature conditions, the Eu^{3+} doped $BaZr(PO_4)_2$ were synthesized by adding stoichiometric amount of Eu_2O_3 to the other precursors. The chloride flux assisted samples were prepared with different weight % of NH_4Cl (99.99 %), KCl (99.99 %), LiCl (99.99 %) and NaCl (99.99 %) added separately with the precursors. The nomenclature used in this paper for different flux assisted samples is given in Table 1.

The phase identification and structural studies were done by using Bruker D8-Advanced X- Ray diffraction which employs Cu-K_α radiation ($\lambda = 1.54$ Å). The surface morphology along with qualitative analysis of elemental composition using energy dispersive spectrometer (EDS) was studied using Scanning Electron Microscope (SEM) instrument Zeiss EVO50. The diffuse reflectance spectra were recorded in 200–800 nm range using Jasco UV–Vis-IR spectrometer. PerkinElmer spectrum 2 equipment was employed for recording the Fourier-transform infrared spectroscopy (FTIR) spectra. The Photoluminescence (PL) spectra and photoluminescence excitation (PLE) spectra were studied using JASCO FP-8300 Spectro-fluorophotometer. For the lifetime measurements decay traces were recorded on HORIBA scientific-Flourolog-3 instrument equipped with pulsed Xenon lamp.

The FLAME-S-XR1-ES model ocean optics spectrometer was employed for temperature dependent PL characteristics investigated in 300 K–420 K range with an interval of 30 K.

3. Results and discussion

3.1. XRD analysis

The powder X-Ray Diffraction (XRD) patterns for undoped BaZr

Table 1
Nomenclature for samples prepared with different weight percent of flux and Eu^{3+} concentration.

NAME	BZP1	BZP3	AC1	AC2	AC3	AC4	AC5	AC6	NC	LC	KC
Eu^{3+} conc (mol%)	1	3	1	1	1	1	1	3	1	1	1
Flux	–	–	NH_4Cl	NH_4Cl	NH_4Cl	NH_4Cl	NH_4Cl	NH_4Cl	NaCl	LiCl	KCl
Flux dosage (wt %)	–	–	2.5	5	10	15	20	5	5	5	5

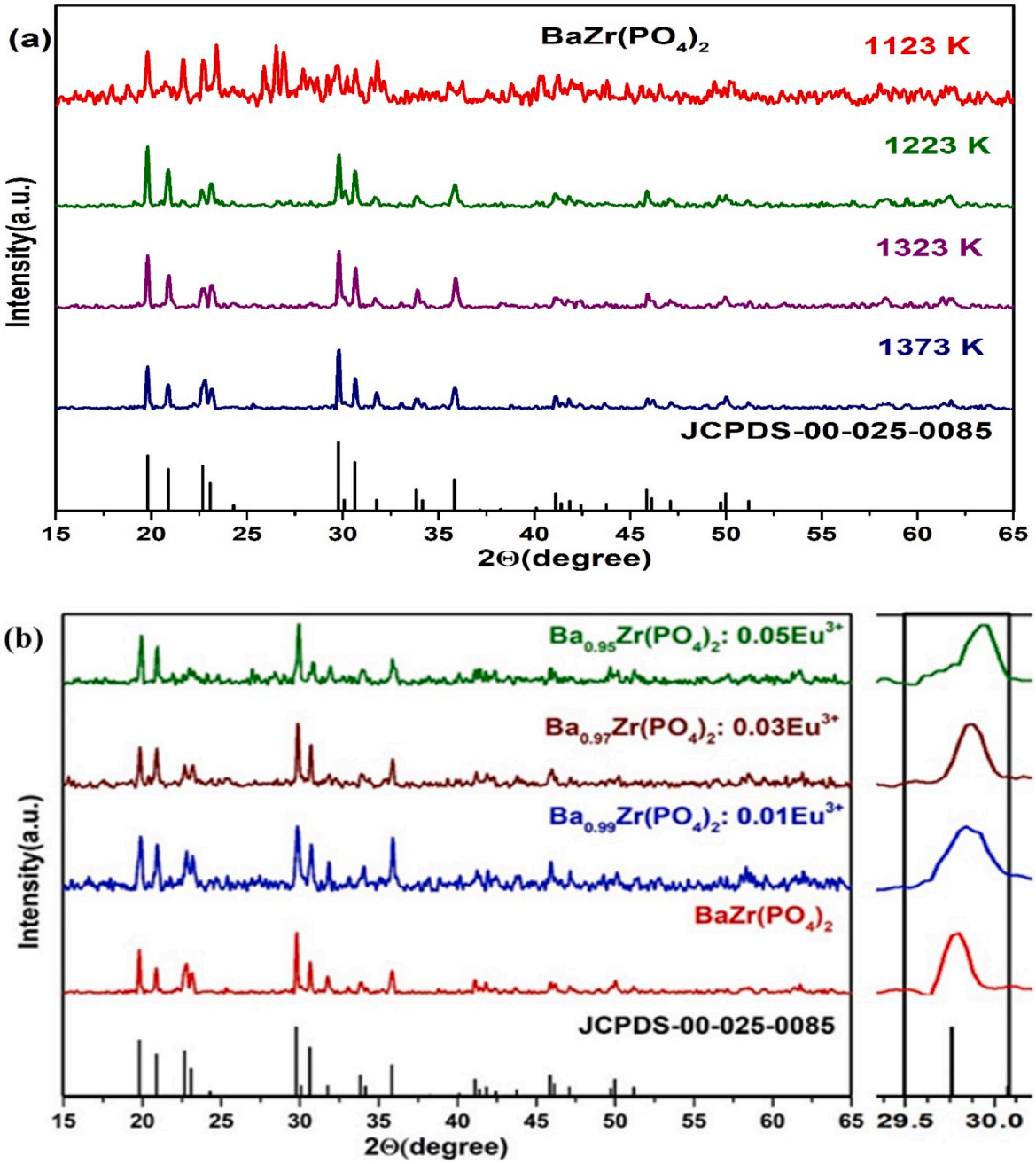


Fig. 2. (a): XRD patterns for $BaZr(PO_4)_2$ with variation of sintering temperature along with standard diffraction pattern. **(b):** XRD patterns for $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 0.0, 1.0, 3.0 and 5.0 mol%) along with standard diffraction pattern.

$(PO_4)_2$ sintered at different temperatures for obtaining pure phase are shown in Fig. 2(a). The peaks present due to impurities vanishes as the sintering temperature is raised from 1023 K and pure phase is obtained at a temperature of 1373 K, where the peaks are in agreement with standard data corresponding to JCPDS card no 00-025-0085, and with earlier reports, confirming the presence of pure phase [22]. The prepared compound has monoclinic structure with *C*2/*m* space group having lattice parameters of a = 8.5310 Å, b = 5.2930 Å, c = 7.8750 Å and cell volume of 354.81 $Å^3$ with 2 molecules per unit cell. The constituent units of structure include three polyhedrons viz. ZrO_6, BaO_{10} and PO_4. The PO_4 tetrahedra and BaO_{10} forms 2-D sheets in a plane parallel to (001) by sharing edges, which are stacked along c directions. The ZrO_6 octahedra is alternatively stacked with BaO_{10} along these sheets in c directions linked with its oxygen atoms O (1) and O (3) [23, 24].

The XRD patterns for Eu^{3+} doped $BaZr(PO_4)_2$ at optimised temperature for different dopant concentration is displayed in Fig. 2(b). The diffraction patterns with varied concentrations of dopant (Eu^{3+}) ion shows iso-structural phases, and no new peaks corresponding to any impurity phase emerge after doping with Eu^{3+} ions establishing integration of Eu^{3+} in crystal structure without altering the monoclinic phase. The enlarged XRD patterns in range 29.5^o to 30.0^o in Fig. 2(b) distinctly shows a slight peak shift to higher angles with augmented concentration of Eu^{3+}, which is attributed to substitution of smaller ionic radii Eu^{3+} in place of higher ionic radii Ba^{2+} [25]. Also considering the oxidation state and ionic radii difference Eu^{3+} will favourable occupy larger Ba^{2+} ions site rather than smaller Zr^{4+} ion site. The Scherrer formula given by $d = 0.9\lambda/\beta\cos\theta$ for crystallite size (d), was used to determine average crystallite size for the samples, where λ is wave-length of X-rays used (λ = 1.5406 Å), β is full width at half

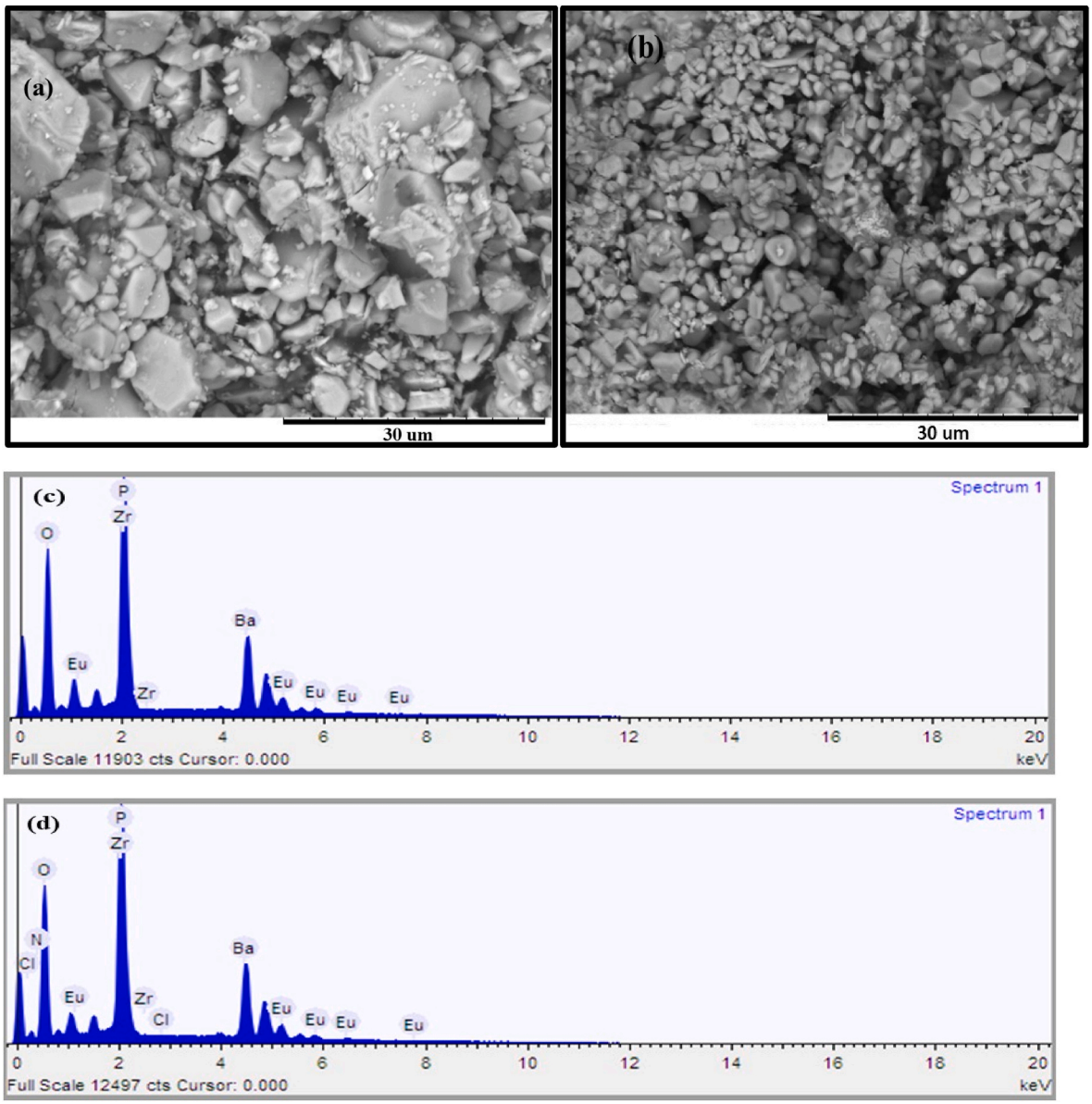


Fig. 3. (a)The SEM image of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} prepared 0 wt% of NH_4Cl **(b)** The SEM image of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} prepared 5 wt% of NH_4Cl **(c)** EDS profile of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} prepared 0 wt% of NH_4Cl (d) EDS of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} prepared 5 wt% of NH_4Cl.

maximum corresponding to diffraction angle θ. The calculated mean crystallite size for $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} for x = 0.0, 1.0, 3.0 and 5.0 mol % was obtained as 45.44, 38.79, 37.66 and 35.54 nm respectively. The shrinking crystallite size with increased concentration is attributed to smaller dopant size as compared to substituted ion leading to lattice contraction [26].

3.2. Surface morphology and elemental analysis

The surface morphology of the $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} synthesized with 0 and 5 wt percent of NH_4Cl (AC6) is displayed in Fig. 3(a) and (b) respectively. The image 3(a) reveal non uniform shape and size of the particle ranges from 3 to 15 μm suggesting that the high temperature sintering lead to agglomeration to bigger particle. In contrast the flux assisted AC6 sample as shown in Fig. 3(b) reveals smaller and nearly uniform shape of particles with size ranging in 1–5 μm range indicating role of flux in improving size distribution and the morphology of the phosphor. In fact, presence of flux at grain boundaries inhibit the steady crystal growth thus reducing the particle size, which also helps in enhancing the luminescence of the phosphor [27]. Also, the texture of the obtained powder phosphor also gets improved with flux. The morphological results suggests that the addition of flux has given favourable results in improving the morphology of the prepared phosphor. The particle size is found to be similar with the mostly available commercial phosphor [28].

Fig. 3(c) and (d) shows the energy dispersive spectrometer (EDS) graph for the $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} synthesized with 0 and 5 wt percent of NH_4Cl (AC6) depicting their elemental composition. The peaks in the graphs reveals that Ba, Zr, O, P and Eu are present in both the samples and no elemental impurity peaks were observed. Moreover, in both the samples the peak profiles are similar indicating that flux elements dissipated due to high temperature calcination as shown in Fig. 3(c) and (d). Thus, EDS finding also affirms that Eu^{3+} got incorporated in the host lattice effectively.

3.3. FTIR studies

The FTIR spectrum of $BaZr(PO_4)_2$ recorded for wave number

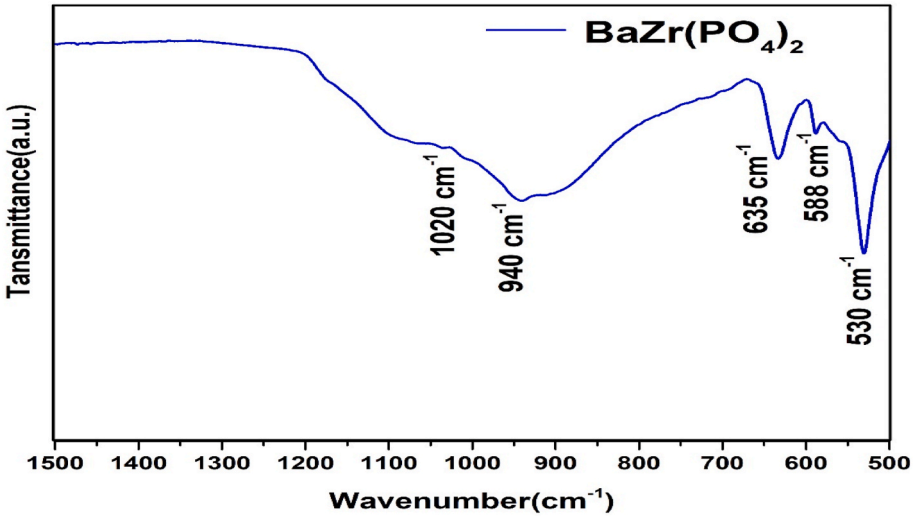


Fig. 4. FTIR spectrum of $BaZr(PO_4)_2$.

spanning 500 to 1500 cm^{-1} is depicted in Fig. 4. The two bands in 1120–940 cm^{-1} and 650–540 cm^{-1} region are characteristic IR absorption band of $(PO_4)^{3-}$ [29]. The recorded spectrum has bands present in both the ranges. The band around 940 cm^{-1} and 1020 cm^{-1} is attributed to symmetric and antisymmetric stretching vibrations respectively of PO_4 tetrahedra [30]. Similarly, the bands at 530 cm^{-1}, 580 cm^{-1}, and 635 cm^{-1} are due to asymmetric deformation mode of phosphate groups [31]. Thus, FTIR studies support the presence of phosphate groups in the phosphor.

3.4. Diffuse reflectance spectroscopy (DRS)

The diffuse reflectance spectroscopy graph of $BaZr(PO_4)_2$ along with the $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} is shown in Fig. 5(a). The contamination by impurity ions alters the optical properties of the semiconductors [32]. Both the samples shows an absorption band in 250–350 nm range peaking around the same point which corresponds to charge transition band of $(PO_4)^{3-}$ octahedra in the $BaZr(PO_4)_2$ host matrix. Past this range clearly apparent additional small drops are present in $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} spectra which are missing in undoped sample. These are located at 394 nm and 463 nm corresponding to $^7F_0$-$^5L_6$ and $^7F_0$-$^5D_2$ transitions of Eu^{3+} respectively well in agreement with excitation spectra. Moreover the reflectance spectrum for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} shifts to a lower value relative to $BaZr(PO_4)_2$ owing to $Eu^{3+}-O^{2-}$ charge transfer band [33].

The Tauc equation for determining the optical band gap E_g of a semiconductor using absorption coefficient (α) is given by:

$$\alpha h\nu = C\left(h\nu - E_g\right)^n \quad (1)$$

where h is Planck's constant, ν is frequency of photon, K is a constant, E_g is band gap, n = 1/2, 2, 1/3 and 3 for direct transition, indirect transition, inhibitory direct transition and inhibitory indirect transition respectively. Moreover, α can be replaced with Kubelka Munk function, F(R) by Kubelka Munk approximation:

$$F(R) = \frac{K}{S} = \frac{(1-R)^2}{2R} \quad (2)$$

where R, K and S are reflectance, absorption and scattering coefficients [34].For α is proportional to F(R), the Tauc relation becomes

$$F(R)h\nu = C\left(h\nu - E_g\right)^n \quad (3)$$

Fig. 5(b) shows Tauc plot between $(F(R)h\nu)^2$ vs $h\nu$ for $BaZr(PO_4)_2$ and $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+}. Band gap is then calculated from intercept of linear fit of linear portion of graph as shown. The band gap for $BaZr(PO_4)_2$ and $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} is extrapolated to 4.16 eV and 4.11 eV respectively. This red shift in optical band gap due to presence of dopant ion is attributed to generation of intermediate energy levels around valence and conduction band leading to narrowing of band gap.

3.5. Photoluminescence investigations

Fig. 6 illustrates the photoluminescent excitation spectra of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 1.0–5.0 mol%) monitored at emission corresponding to $^7F_0$ to $^5D_2$ transition of Eu^{3+} at 612 nm. Several intense peaks ascribed to 4f-4f transitions of Eu^{3+} can be clearly seen. These peaks centred around 360, 380, 392, 414 and 463 nm corresponds to $^7F_0$-$^5D_4$, $^7F_0$-$^5L_7$, $^7F_0$-$^5L_6$, $^7F_0$-$^5D_3$ and $^7F_0$-$^5D_2$ transitions respectively and are consistent with those obtained in diffuse reflectance spectra of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} [17,21,35]. Among these peaks the most intense corresponds to $^7F_0$-$^5L_6$ transitions at 392 nm which is four times intense than $^7F_0$-$^5L_7$ peak and 3.6 times intense than $^7F_0$-$^5D_2$ peak. Moreover, it is apparent that increasing Eu^{3+} ion concentration enhances the spectral intensity initially till x = 3.0 mol% where it is maximum, thereafter it decreases on subsequent addition of Eu^{3+} ion, without altering the emission spectra peak profile.

Fig. 7 (a) shows emission spectra of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} excited under 360, 380, 392 and 463 nm wavelengths. It is clear that the spectrum monitored at 392 nm wavelength is most intense, indicating that a UV-LED chip can be used as an effective excitation source for $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} phosphor. Fig. 7 (b) shows emission spectra of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 1.0–5.0 mol%) when samples were excited to $^5L_6$ levels by 392 nm wavelength. The emission profile comprises of narrow bands centred around 579, 591, 612, 652 and 700 nm ascribed to $^5D_0$-$^7F_{0\text{-}4}$ electronic transitions of trivalent Europium ion. The peak profile (position and shape) remains unaltered for different concentrations of Eu^{3+} due to shielding of $4f^6$ electrons of Eu^{3+} by outer 5s and 5p electrons [36].

The emission transition $^5D_0$-$^7F_0$ (597 nm), identified as electric dipole, as per Judd– Ofelt theory it is strictly not allowed (as ΔJ = 0). A low intensity peak corresponding to this may be attributed to J-mixing arising from perturbation in crystal field giving rise to mixing of wavefunctions of sublevels of different levels [37,38]. The transition $^5D_0$-$^7F_1$ (591 nm), lying in orange emission region, is a magnetic dipole transition and is permissible as per selection rules (ΔS = 0, ΔL = 0, ΔJ = 0, ±1, *with* 0 ⁄ 0 *forbidden*) [38]. The probability of this transition is not constrained by crystal environment around Eu^{3+} ion and doesn't vary significantly with host [39]. So, the peak corresponding to this transition

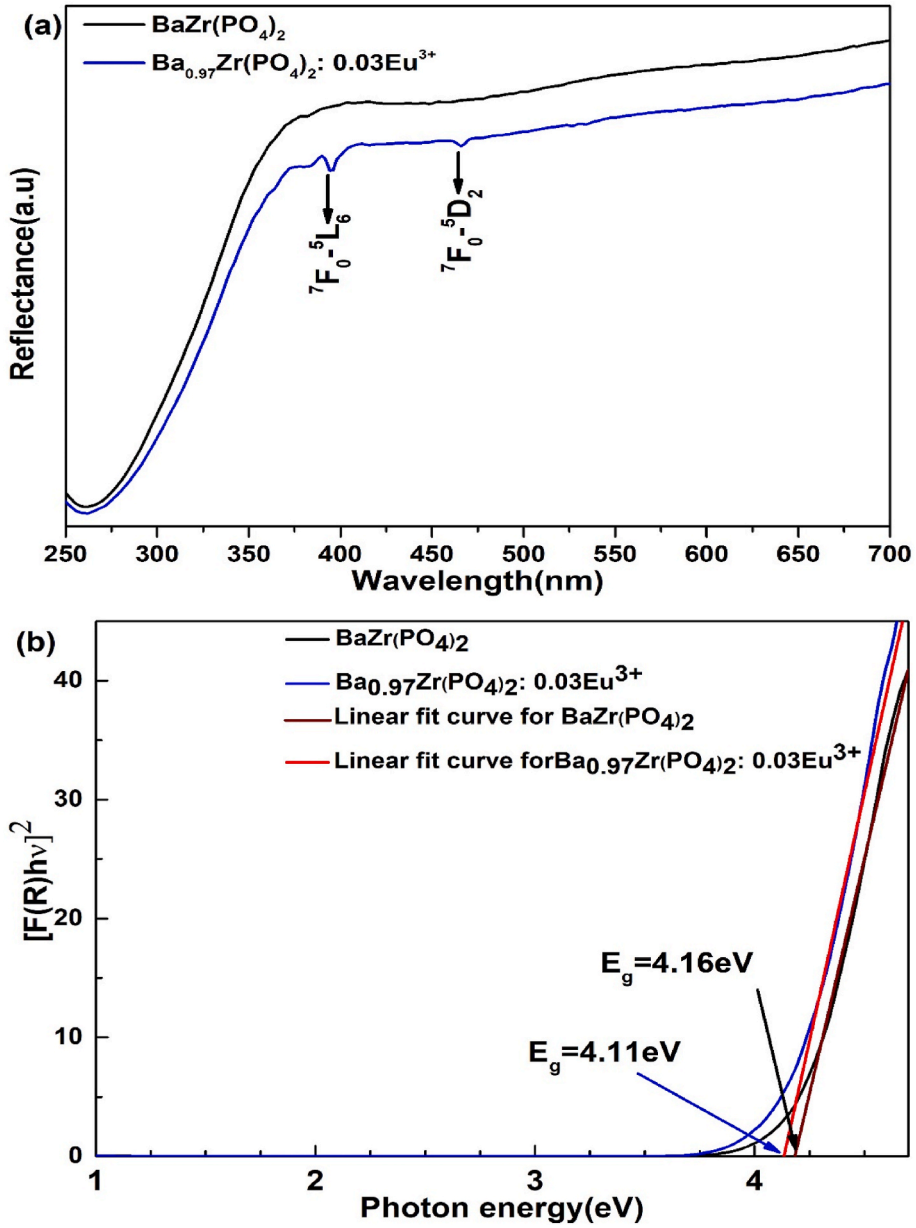


Fig. 5. (a): Diffuse Reflectance Spectra of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 0.0 and 3.0 mol%) **(b):** Tauc Plot for band gap determination for $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 0.0 and 3.0 mol%).

dominates when Eu^{3+} takes inversion symmetry sites. The $^5D_0$-$^7F_2$ (612 nm) transition corresponding to red emission is hypersensitive electric dipole transition allowed by the selection rules $|\Delta S| = 0$, $|\Delta L| \leq 2$, and $|\Delta J| \leq 2$, when Eu^{3+} occupy a low symmetry site. So, with the reduction in symmetry in local environment of Eu^{3+} the strength of emission of this transition increases. So, asymmetry ratio (A_{21}) which can be obtained from the ratio of electric dipole transition integral intensity to that of magnetic dipole transition (i.e. $A_{21} = I_{ED}/I_{MD}$), gives good judgement about symmetry around Eu^{3+} as well the red colour purity of phosphor. A high value of this ratio is a clear indication of low symmetry of Eu^{3+} site [40–43]. The emission spectra clearly show the dominating $^5D_0$-$^7F_2$ peak indicating presence of most of activator ions located at the asymmetric sites. The computed A_{21} values for different samples are tabulated in Table 2. A decrease in A_{21} value with increasing concentration indicates that Eu^{3+} ions present at two different symmetry sites [44]. Moreover, the stark splitting of $^5D_0$-$^7F_2$ level into two peaks centred at 587 nm and 593 nm indicates presence of some Eu^{3+} ions in sites having inversion symmetry [26].

The bar graph in Fig. 7(b) reveals the variation of intensity of the $^5D_0$-$^7F_2$ peak with Eu^{3+} ion concentration. The intensity is maximum for 3.0 mol% concentration of Eu^{3+} ions and falls off with variation on either side. As concentration of Eu^{3+} is increased the nearest activator ion distance decreases, and the probability of energy transfer among Eu^{3+} ions becomes more than probability of emission thereby leading to non-radiative loss of excitation energy among the Eu^{3+} ions [41]. This non-radiative loss can be in generally attributed to three different mechanisms viz. exchange interaction, multipole–multipole interaction and radiation reabsorption [3,8]. The critical distance between the nearest activator ions gives a good estimation of the energy loss mechanism involved, which can be evaluated using Blasse formulae mentioned elsewhere [40].

For $BaZr(PO_4)_2$ lattice, the value of R_c for Eu^{3+} in this system was found to be 22.44 Å, which being above 5 Å suggesting that multipole-multipole interaction mechanism accounts for non-radiative energy transfer. Further the nature of multipole interactions can be investigated by employing Dexter's relation [41]. A plot of log(I/x) vs log(x) is shown

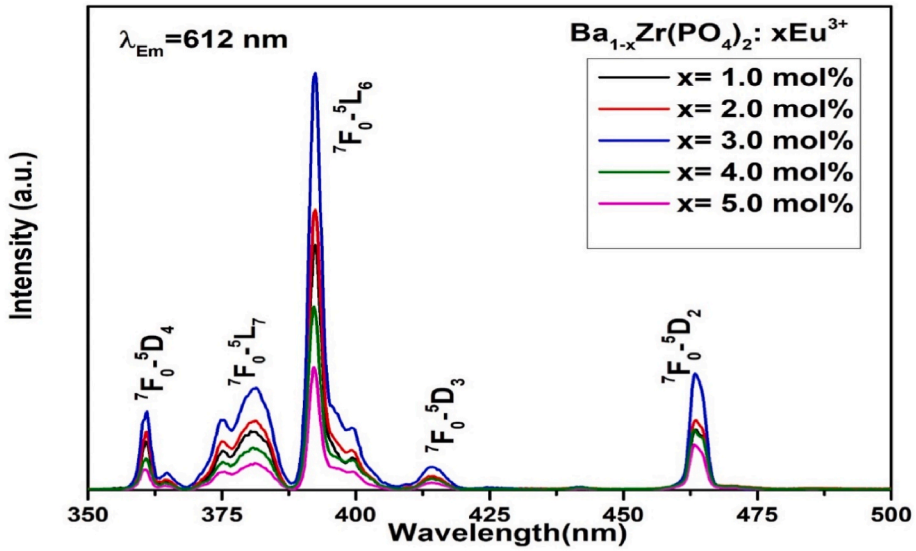


Fig. 6. The excitation spectra of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 1.0, 2.0, 3.0, 4.0 and 5.0 mol%) at 612 nm emission wavelength.

in Fig. 7(c) from where the value of Q is computed to be 10.13, thereby establishing quadrupole-quadrupole interactions contributing to non-radiative energy transfer.

3.5.1. *Influences of fluxes on emission spectra*

The role of fluxes in lowering the sintering temperature, augmenting the luminescent properties and improving the morphology of phosphors has been reported by many authors [45–49]. Fig. 8 shows emission spectra of $Ba_{0.99}Zr(PO_4)_2$:$0.01Eu^{3+}$ phosphor monitored under 392 nm excitation for BZP1, AC2, NC, LC and KC samples. It reveals that emission intensity increases only when NH_4Cl (AC2) is added as a flux while for rest of chlorides the emission intensity declines. The addition of NH_4Cl not only lowers the sintering temperature to 1223 K but also boost the luminescence of the phosphor by boosting the diffusion of Eu^{3+} ions in host lattice. The peak intensity for $^5D_0$-$^7F_2$ level of AC2 sample was found to be 1.31 times that of BZP1 sample and the spectrum profile was found to be similar for both AC2 and BZP1.

Further, the decline in luminous intensity of NaCl, KCl and LiCl assisted samples is attributed to crystal deformation and incorporation of impure phase in crystal structure which can be seen in Fig. 9 depicting the XRD profile of these phosphors. The higher amount of these fluxes further deteriorated the crystal structure so were not considered for further investigation. Also, the XRD pattern for NH_4Cl assisted sample clearly shows no such impurity peaks and any significant variation in XRD pattern is observed. Thus, among all the chlorides under investigation NH_4Cl emerges out as an obvious choice to be used as flux. With this, the amount of NH_4Cl was increased to identify the best possible amount of flux that can be used in the prepared phosphor. Further the correlation between luminescence and the quantity of flux was studied. Fig. 10 shows the variations in emission of $Ba_{0.99}Zr(PO_4)_2$:$0.01Eu^{3+}$ with variation in quantity (weight %) of NH_4Cl added as flux. It is clear from Fig. 10 that 5 wt % is the optimum quantity of flux as any variation on either side leads to fall in PL intensity.

Fig. 11 depicts the XRD profiles for different NH_4Cl assisted samples to look for any variation in the crystal structure with the addition of flux in the host. As it is seen in Fig. 11 that for AC1, AC3, AC4 and AC5 samples remarkable variation in peak intensity is clearly visible suggesting the phase distortion leading to decline in PL intensity for these samples [48–50]. Fig. 12 shows variation in intensity of PL spectra for optimised quantity of flux for 3.0 mol % Eu^{3+} doped $BaZr(PO_4)_2$. The obtained spectral profile is same for both the samples indicating that crystal phase is not distorted, which can be verified from the XRD profile of AC6 sample in Fig. 11. The mean crystallite size for AC6 sample using Debye Scherrer formula as done earlier was found to be around 35.87 nm which was 37.66 nm for BZP3 sample indicating that addition of NH_4Cl as flux leads to better crystallinity of the sample. It was found that intensity for $^5D_0$-$^7F_2$ transition got enhanced by 1.43 times in AC6 sample compared to BZP3 sample. The asymmetry ratio for BZP3 and AC6 samples was found to be 2.20 & 2.79 respectively indicating that flux assisted samples have more asymmetric nature of Eu^{3+} occupied sites. Moreover, for AC6 sample $^5D_0$-$^7F_2$, peak exhibits stark splitting with a shoulder peaked at 621 nm. The $^7F_2$ level is degenerate, so it can split in to a maximum of (2J + 1) levels due to crystal field splitting [51]. The flux assistance led to better incorporation of rare earth inside the host composition, thus altering the magnitude of local field acting on the rare earth ion resulting in stark splitting of the degenerate levels [52, 53]. Thus, the presence of spectral splitting in a dominant $^5D_0$-$^7F_2$ transition further affirms the presence of Eu^{3+} ions in asymmetric environment for flux assisted AC6 sample [54]. Hence, it can be concluded that NH_4Cl as a flux enhanced the luminescent performance of the phosphor.

3.5.2. *Temperature dependent PL*

The variation of the luminescent properties of a phosphor with temperature for assessing its thermal stability need to be investigated for considering commercial application of phosphor for wLEDs specially the high power one [3,9,55]. Fig. 13(a) depicts the temperature dependent photoluminescence of $Ba_{0.97}Zr(PO_4)_2$:$0.03Eu^{3+}$ for temperatures in range 300–420 K monitored at an excitation wavelength of 392 nm.

The thermal quenching behaviour of the phosphor is indicated by the observed decrease in luminescence as temperature rises. Moreover, the spectral profile remains unaltered at higher temperatures. Fig. 13(b) shows the normalised PL intensity at various temperatures and it is found that at 420 K the phosphor retained 75 % emission intensity. The reason for declining intensity can be related to the mechanism of temperature induced transition from excited state to the ground state and multiphonon relaxation [56,57]. Moreover, the dependence of luminescence on temperature is governed by the Arrhenius equation [55]:

$$I_T = I_0\left[1 + A\,exp\left(-\Delta E/KT\right)\right]^{-1} \quad (4)$$

where I_T is intensity of emission at temperature T, I_0 represents intensity of emission monitored at ambient temperature, A is arbitrary constant, ΔE is activation energy, K represents Boltzmann constant (8.629×10^{-5} eV/K). So, if a graph is plotted for $\ln((I_0/I_T)-1)$ against 1/KT, as

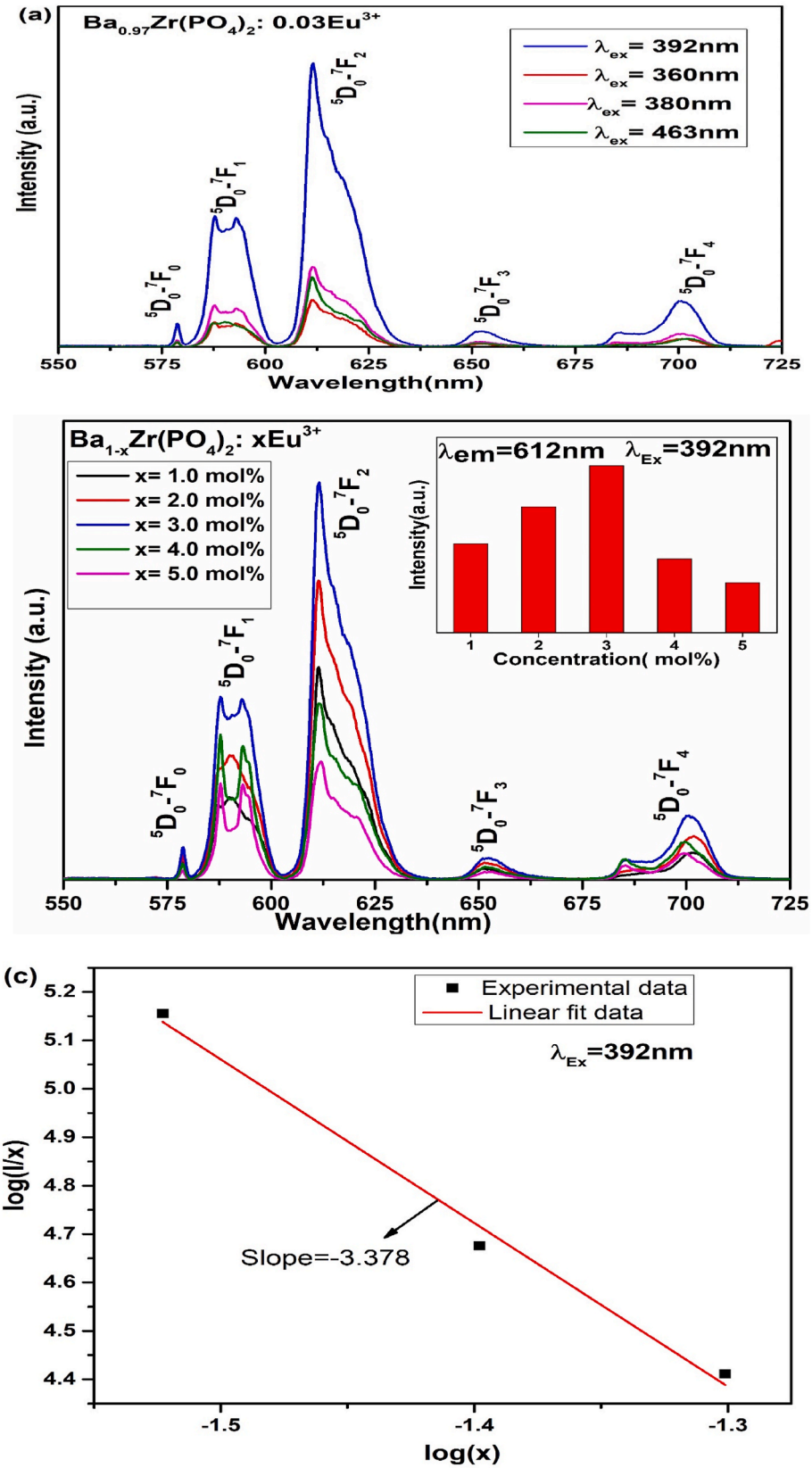


Fig. 7. (a): The emission spectra for $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} at (x = 3.0 mol%) monitored at different excitation wavelengths. (b): The emission spectra of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 1.0, 2.0, 3.0,4.0 and 5.0 mol%) under 392 nm excitation wavelength. (c): Plot of log(I/x) versus log(x) for $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} (x = 1.0, 2.0, 3.0, 4.0 and 5.0 mol%).

Table 2
Value of asymmetry ratio for various samples.

Sample	A21
$Ba_{0.99}Zr(PO_4)_2$:0.01Eu^{3+}	2.6
$Ba_{0.98}Zr(PO_4)_2$:0.02Eu^{3+}	2.4
$Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+}	2.2
$Ba_{0.96}Zr(PO_4)_2$:0.04Eu^{3+}	1.66
$Ba_{0.95}Zr(PO_4)_2$:0.05Eu^{3+}	1.63
AC6	2.79

displayed in Fig. 13(c), a straight line whose slope will give value of activation energy will be obtained. According to Fig. 13(c), the value of activation energy for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} phosphor is 0.333 eV, which is significantly better than the other reported Eu^{3+} activated phosphors like $Sr_{1.7}Zn_{0.3}CeO_4$:0.3$Eu^{3+}$and $Ca_2YZr_2Al_3O_{12}$: 0.08Eu^{3+} (E_a = 0.1788 eV and 0.16 eV) [9,33].

3.6. CIE chromaticity coordinates

The Commission Internationale de L'Eclairage (CIE) coordinate of phosphor is another measure to quantify the performance of phosphor. Therefore, using emission spectra monitored at 392 nm excitation, CIE coordinate for $BaZr(PO_4)_2$:xEu^{3+} (x = 1.0, 2.0, 3.0, 4.0, 5.0 mol%) were calculated and are shown in Table 3. Fig. 14 reveals that the CIE coordinates of all the samples belongs to pure red region of the plot thus indicating the pure red emission from the prepared sample.

The CIE chromaticity coordinates for flux assisted $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} were (0.650,0.354) which are close to that of commercial red phosphor Y_2O_3:Eu^{3+} (0.655, 0.345) [58]. Also the CIE coordinates of all the samples are proximate to the standard red color

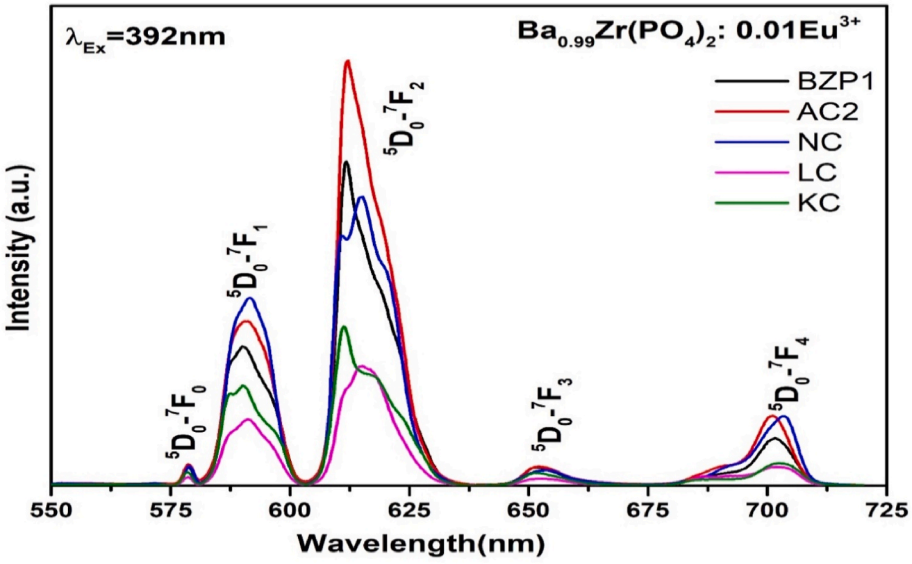


Fig. 8. Emission spectra of $Ba_{0.99}Zr(PO_4)_2$:0.01$Eu^{3+}$with different fluxes under 392 nm excitation wavelength.

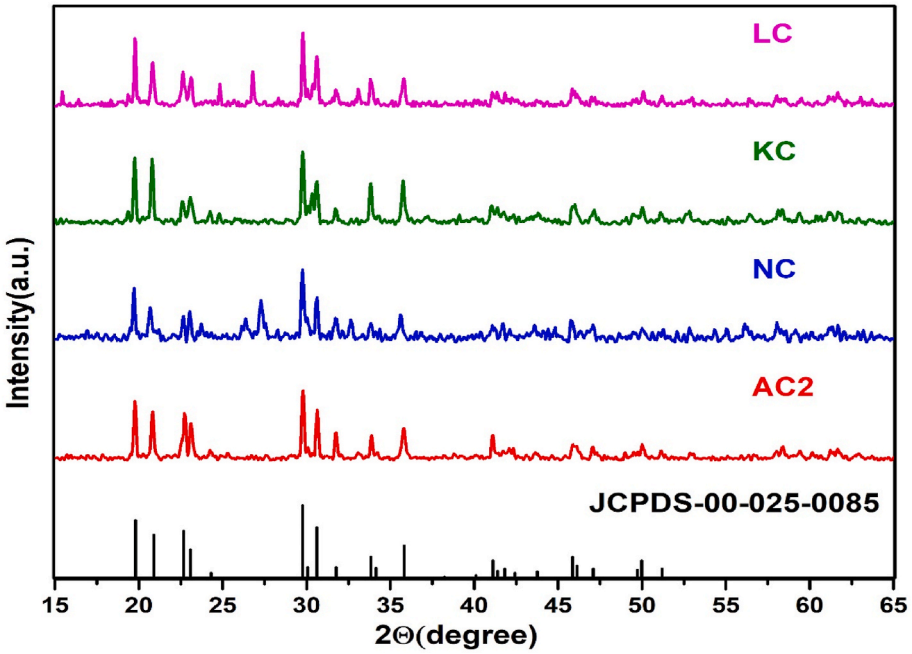


Fig. 9. X-ray diffraction pattern of various flux blended $Ba_{0.99}Zr(PO_4)_2$:0.01Eu^{3+}phosphors.

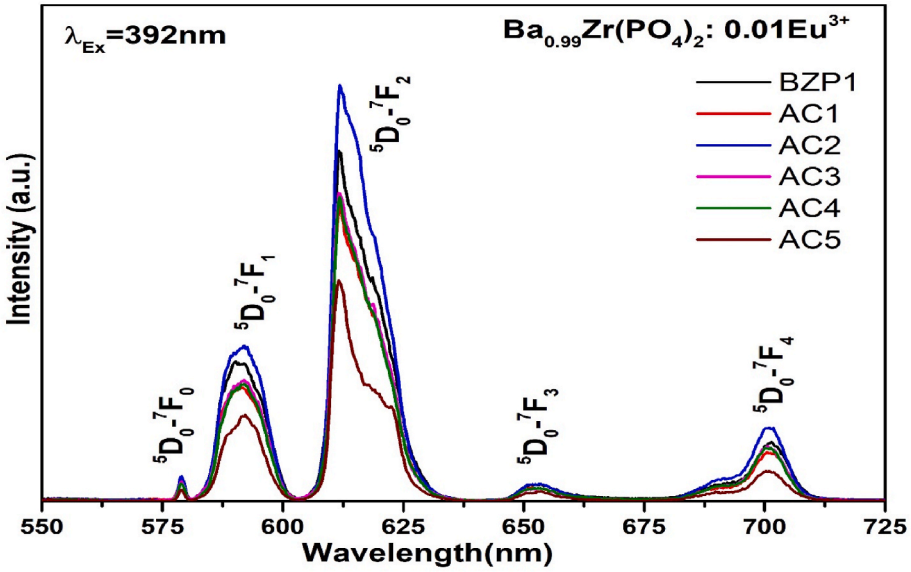


Fig. 10. Emission spectra of $Ba_{0.99}Zr(PO_4)_2$:0.01$Eu^{3+}$with different wt% of NH_4CL under 392 nm excitation wavelength.

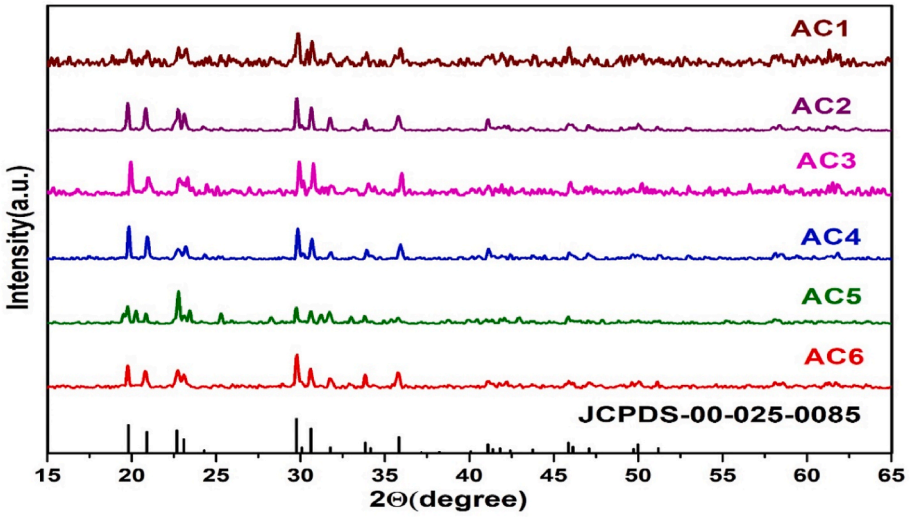


Fig. 11. X-ray diffraction patterns with different weight % of NH_4Cl assisted $BaZr(PO_4)_2$:xEu^{3+}phosphors.

coordinates value (0.67, 0.33) of National Television standard Committee (NTSC) system [8]. The color purity for red emission was computed employing the following equation [16]:

$$\text{Color purity} = \frac{\sqrt{(x-x_i)^2+(y-y_i)^2}}{\sqrt{(x_d-x_i)^2+(y_d-y_i)^2}} \times 100\% \quad (5)$$

where (x, y) are the color-coordinates of the sample, (x_i, y_i) corresponds to white point, and (x_d, y_d) are the coordinates corresponding to dominant wavelength point. Correlated color temperatures (CCT) were calculated using the McCamy's polynomial formula [26]. Both the CCT and color purity are tabulated in Table 3. The results suggests that the as prepared phosphor has a warm red emission and is a potential candidate for applications in wLEDs as a red component.

3.7. Decay curve

The photoluminescent decay traces was recorded for the $Ba_{(1-x)}Zr(PO_4)_2$:xEu (x = 1.0, 2.0, 3.0, 4.0, 5.0 mol%), AC2, and AC6 samples monitored for $^5D_0$-$^7F_2$ emission at $^7F_0$-$^5L_6$ excitation and are depicted in Fig. 15. The curves were fitted employing the biexponential fit equation which relates the luminescence intensity at time t, I_t with that at time t = 0 as:

$$I_t = I_0 + C_1 e^{\left(\frac{-t}{\tau_1}\right)} + C_2 e^{\left(\frac{-t}{\tau_2}\right)} \quad (6)$$

where C_1 and C_2 are constants, τ_1 and τ_2 are fast and slow lifetimes respectively. The average lifetime τ, for the phosphors was calculated from the fitted parameters using the relation:

$$\tau = \frac{C_1\tau_1^2 + C_2\tau_2^2}{C_1\tau_1 + C_2\tau_2} \quad (7)$$

The average lifetime values are tabulated in Table 3. The biexponential fit function with two different lifetimes suggests the presence of Eu^{3+} ions at more than one sites [59]. A decreasing trend for lifetime values with increasing concentration of Eu^{3+} ions is observed which is due to increase in non-radiative transitions among Eu^{3+} ions as the distance between them decreases [48].

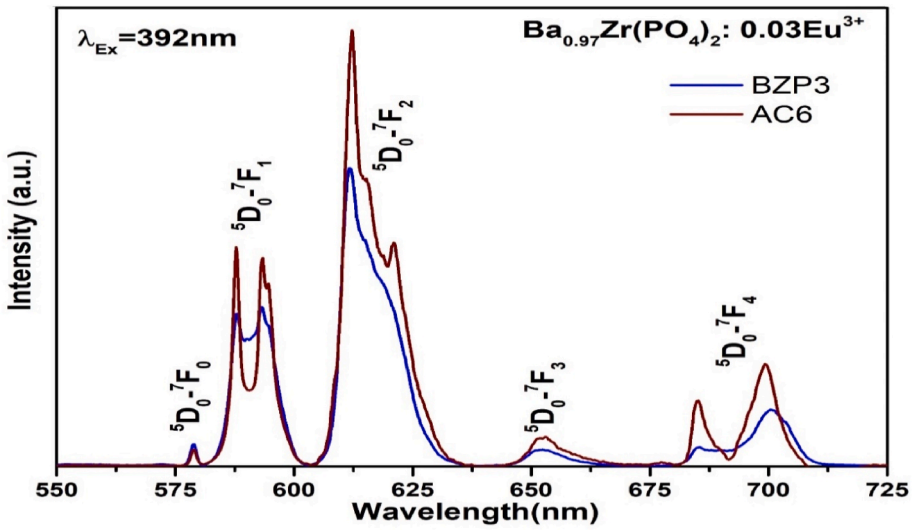


Fig. 12. Emission spectra of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+}without flux and with 5 wt% NH_4Cl under 392 nm excitation wavelength.

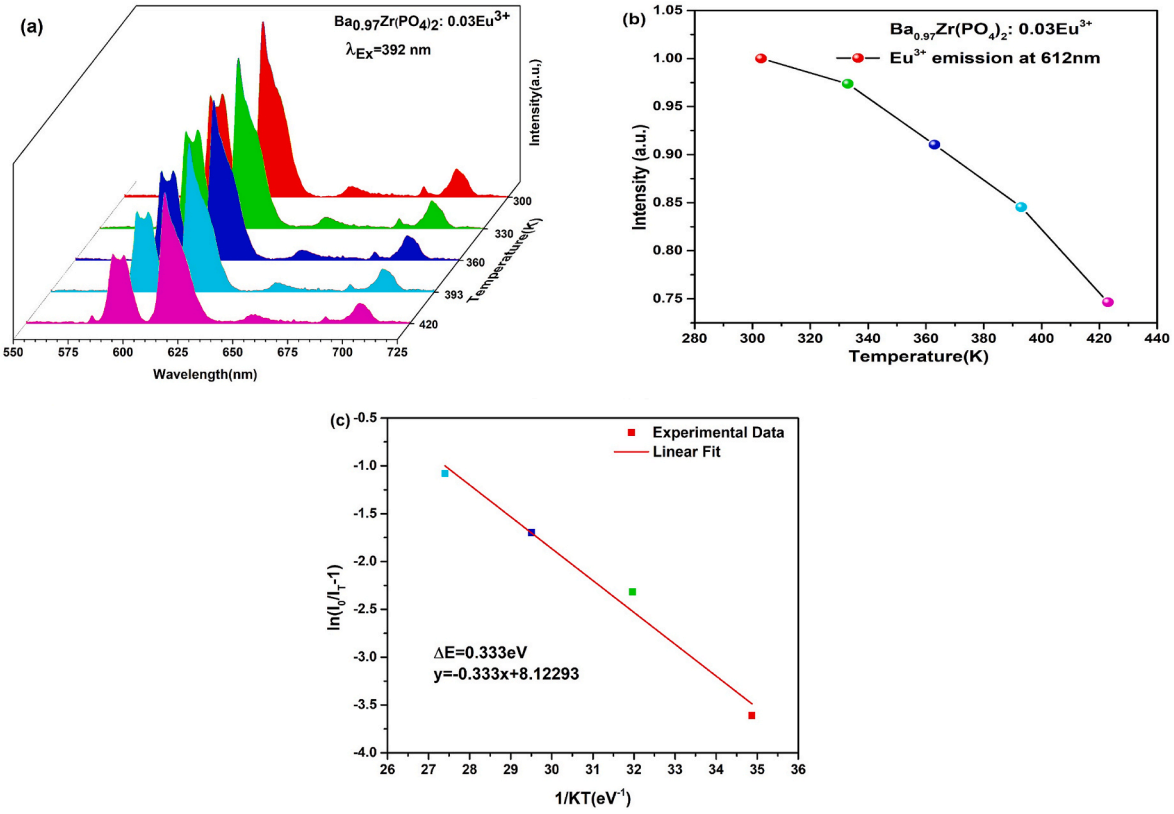


Fig. 13. (a): Temperature dependent PL spectra for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} phosphor at different temperatures between 300K and 420 K under 392 nm excitation. (b): Normalised PL intensity of $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} Phosphor at different temperatures. (c): linear fit plot of ln[(I_0/I)-1] versus 1/kT for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} phosphor.

3.8. Judd-Ofelt analysis of radiative properties

The Judd-Ofelt (J-O) theory provides significant insight about the spectroscopic behavior of trivalent rare earth ions in a particular coordination environment. The J-O Intensity parameters Ω_λ ($\lambda = 2$, 4, and 6) are evaluated to conclude radiative potential of RE ions in different hosts, and identifying bonding and local structure in vicinity of RE ions. Usually, these intensity parameters are derived from absorption spectrum but in case of Eu^{3+} activated host matrices we can evaluate Ω_2 and Ω_4 using emission spectrum due to special energy level structure of Eu^{3+} ions [60].

The transition $^5D_0$ to $^7F_1$ is magnetic dipole transition and is untrammelled from crystal field. As per J-O theory its spontaneous transmission probability (A_{01}) is related to magnetic dipole strength by following relation [60–62].

$$A_{01} = \frac{64\pi^4 \nu^3 n^3 S_{md}}{3h(2J+1)} \quad (8)$$

Table 3
CCT, CIE Co-ordinates, Color purity and decay time for of $Ba_{(1-x)}Zr(PO_4)_2$:xEu phosphors.

Sample	CCT(K)	CIE Co-ordinates (x,y)	Color purity(%)	Decay Time(ms)
$Ba_{0.99}Zr(PO_4)_2$:0.01Eu^{3+}	2292.485	(0.641, 0.358)	92.22	1.414
$Ba_{0.98}Zr(PO_4)_2$:0.02Eu^{3+}	2325.111	(0.642, 0.357)	92.56	1.381
$Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+}	2301.994	(0.642, 0.357)	92.37	1.361
$Ba_{0.96}Zr(PO_4)_2$:0.04Eu^{3+}	2177.306	(0.637, 0.362)	91.30	1.294
$Ba_{0.95}Zr(PO_4)_2$:0.05Eu^{3+}	2144.021	(0.636, 0.363)	90.96	1.280
AC2	2407.064	(0.645, 0.354)	93.20	1.365
AC6	2582.470	(0.650, 0.354)	94.35	1.191

where ν is energy in cm^{-1} corresponding to $^5D_0$ to $^7F_1$ transition, h is Planck's constant, (2J + 1) is degeneracy of initial state, s_{md} is magnetic dipole strength and n the refractive index for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} was estimated to be 1.85 using the formula [63]:

$$\frac{n^2-1}{n^2+1}=1-\sqrt{\frac{E_g}{20}} \tag{9}$$

where, E_g the optical band gap is obtained from DRS.

The transitions $^5D_0$ to $^7F_{J'}$ (J = 2,4 and 6) corresponds to electric dipole transitions and their corresponding spontaneous transmission probability(A_{0J}) is related to electric dipole strength by following relation

$$A_{0J}=\frac{64\pi^4\nu^3e^2}{3h(2J+1)}\frac{n(n^2+2)^2}{9}\sum_{\lambda=2,4,6}\Omega_\lambda\left|\langle 5_{D_0}||U^{(\lambda)}||7_{F_{J'}}\rangle\right|^2 \tag{10}$$

where ν is energy in cm^{-1} corresponding to $^5D_0$ to $^7F_{J'}$ transition, n is refractive index, J = 0 for $^5D_0$ state, e is electronic charge (4.8×10^{-10} esu) and $\left|\langle 5_{D_0}||U^{(\lambda)}||7_{F_{J'}}\rangle\right|^2$ is the squared reduced matrix elements, which are zero for all electric dipole transitions originating from $^5D_0$ level except for j' = 2, 4 and 6 with values 0.0032, 0.0023 and 0.0002 respectively. The ratio of radiative transition probability can be expressed in terms of ratio of integral intensity of corresponding peaks in emission spectra and hence Ω_λ can be obtained as per following equation [64]:

$$\Omega_\lambda=\frac{s_{md}{\nu_1}^2}{e^2{\nu_\lambda}^2}\frac{9n^3}{n(n^2+2)^2\left|\langle 5_{D_0}||U^{(\lambda)}||7_{F_{J'}}\rangle\right|^2}\frac{\int I_\lambda(\nu)d\nu}{\int I_1(\nu)d\nu} \tag{11}$$

The value of transition energy for calculation of J-O parameters has been obtained as an average over the selected transition region from emission spectra:

$$\nu_\lambda=\frac{\int \nu I(\nu)d\nu}{\int I(\nu)d\nu} \tag{12}$$

Thus obtained values of J-O parameters Ω_2 and Ω_4 for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} and flux assisted AC6 sample are listed in Table 4.The value of Ω_6 parameter cannot be estimated because the emission corresponding to $^5D_0$ to $^7F_6$ is located in Infrared region and could not be experimentally detected in our sample. However, the influence of Ω_6 on radiative property of Eu^{3+} can be neglected [64].

The Ω_2 parameter which is associated with hypersensitive electric dipole transition $^5D_0$ to $^7F_2$ is responsive to crystal field. The high value of Ω_2 is related to higher covalence of metal ligand bonds and indicates lower symmetry of Eu^{3+} sites which is consistent with the obtained higher asymmetry ratio. The parameters Ω_4 and Ω_6 on the other hand are attributed to macroscopic bulk properties such as rigidity and viscosity of the compounds [65].The increasing trend in value of Ω_2 from $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} to flux assisted AC6 sample indicates more covalency of Eu-O bond and low local symmetry of Eu^{3+} in AC6 sample. These findings are identical with the increasing trend in asymmetry ratio thus reaffirming more asymmetric nature of Eu^{3+} occupied sites in flux assisted AC6 sample.

The fluorescent branching ratio, which gives the ratio of emission of a level with respect to other levels, and is a measure of potential of emission of a level, for $^5D_0$ to $^7F_{J'}$ transitions can be obtained using the following relation [66]:

$$\beta=\frac{A_{J-J'}}{\sum A_{J-J'}} \tag{13}$$

The observed fluorescent lifetime (τ_{obs}) can be related to total transition rate (A_T), radiative transition rate (A_R) and non-radiative transition rate (A_{NR}) as:

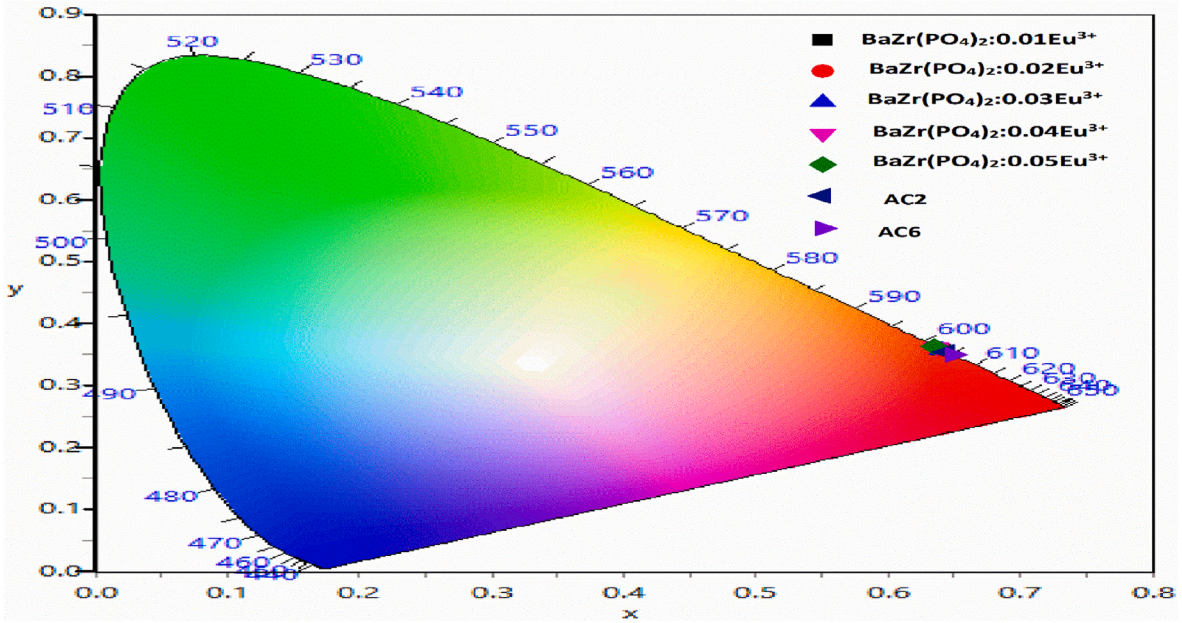


Fig. 14. CIE chromaticity coordinates diagram of $Ba_{(1-x)}Zr(PO_4)_2$:xEu^{3+} phosphors.

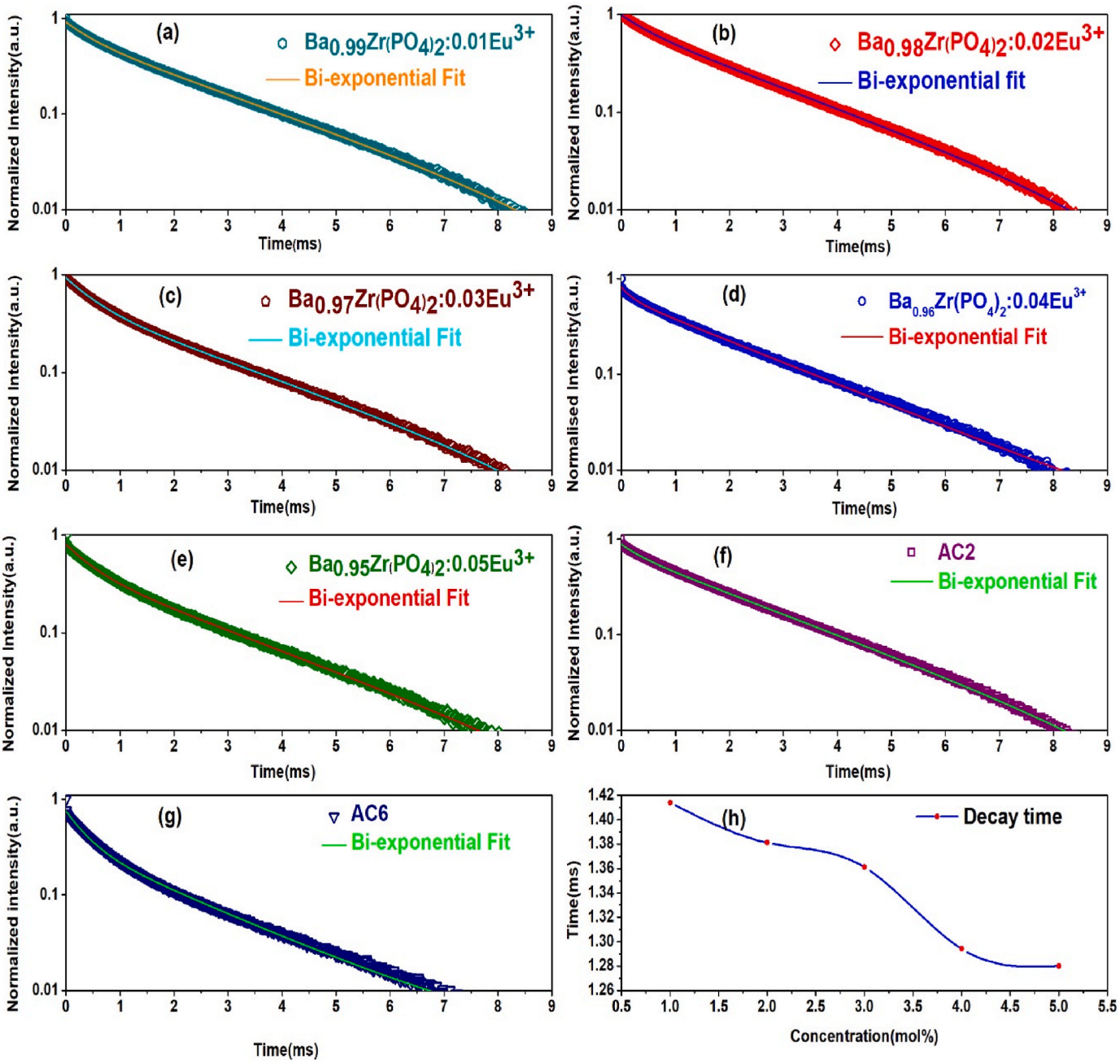


Fig. 15. (a–g): PL decay curve of all the phosphors (λ_{ex} = 494 nm and λ_{em} = 612 nm). **(h):** Variation of decay time with Eu^{3+} ions concentration.

Table 4
J-O parameters (Ω2, Ω4) of BZP3 and AC6 phosphors.

sample	Ω_2 (x 10^{-20}) cm^2	Ω_4 (x 10^{-20}) cm^2
$Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+}	3.36	1.33
AC6	4.31	1.96

$$\frac{1}{\tau_{obs}} = A_T = A_R + A_{NR} \quad (14)$$

The radiative transition rate can further be used to evaluate non-radiative transition rate. The quantum efficiency defined as fraction of photons emitted to those of absorbed by Eu^{3+} ion can thus be obtained by following equation:

$$\eta = A_R / (A_R + A_{NR}) \quad (15)$$

The evaluated radiative transition rates and branching ratios for $^5D_0$ to ^7F_J transitions and quantum efficiency for $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} and flux assisted AC6 sample are listed in Table 5. It can be observed that incorporation of flux marginally improves the branching ratio for $^5D_0$-$^7F_2$ transition. The branching ratio for $^5D_0$- $^7F_2$ transition in BZP3 and AC6 sample is coming out to be greater than 0.5 thus indicating that they can be a potential candidate for laser action for the said transition [66].

4. Conclusions

Single phased $Ba_{(1-x)}Zr(PO_4)_2$:xEu (x = 1.0–5.0 mol%) without and with added chloride fluxes having monoclinic structure belonging to *C2/m* space group has been synthesized via simple solid-state route. The morphological studies and crystal structure of the prepared phosphors was investigated using XRD and SEM. The band gap of the $BaZr(PO_4)_2$ and $Ba_{0.97}Zr(PO_4)_2$:0.03Eu^{3+} samples was computed to be 4.16 eV and

Table 5
Radiative transition rates, Branching ratio and quantum efficiency of BZP3 and AC6 phosphors.

	$^5D_0$- ^7F_J	Energy(cm^{-1})	Radiative Transition Rate($A_{0\text{-}J}$) (s^{-1})	β	A_R (s^{-1})	A_{NR} (s^{-1})	A_T (s^{-1})	η(%) (s^{-1})
BZP3	$^5D_0$- $^7F_1$	16905	92	0.28	332	402	734	45.23
	$^5D_0$- $^7F_2$	16224	201	0.60				
	$^5D_0$- $^7F_4$	14329	39	0.12				
AC6	$^5D_0$- $^7F_1$	16900	92	0.23	408	433	841	48.51
	$^5D_0$- $^7F_2$	16217	257	0.63				
	$^5D_0$- $^7F_4$	14344	59	0.14				

4.11 eV respectively. The photoluminescent studies revealed that UV LED chip can be used as a source of excitation for the prepared phosphor and have most dominant emission peak is in red region (612 nm). The critical activator concentration for Eu^{3+} was found to be 3.0 mol %, with quadrupole-quadrupole interaction contributing to radiation quenching. The incorporation of NH_4Cl as flux improved the luminescent performance of the phosphor without deforming the spectral profile as well as the crystal structure. The variation of photoluminescence with temperature revealed that the prepared phosphor exhibits superior thermally stable luminescent characteristics with 0.333 eV activation energy. The calculated CIE coordinates for the prepared samples lie proximate to NTSC standard for red colour. The J-O theory was employed for evaluation of J-O parameters, transition rates branching ratio and quantum efficiency. The branching ratio value for $^5D_0$- $^7F_2$ transition indicate the prepared phosphor as a promising prospect in laser and solid state lighting application. All the studies establish that 5 wt % NH_4Cl assisted 3 mol% Eu^{3+} activated $BaZr(PO_4)_2$ is a promising red emitter for wLEDs.

CRediT authorship contribution statement

Shweta Yadav: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **A.S. Rao:** Writing – review & editing, Validation, Supervision. **Deshraj Meena:** Validation, Supervision.

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.

Acknowledgements

Ms Shweta Yadav is thankful to Bhagini Nivedita College, University of Delhi for allowing her to carry out research work in Delhi Technological University.

References

[1] R. Vavrek, J. Chovancova, Decoupling of greenhouse gas emissions from economic growth in V4 Countries, Procedia Econ, Financ. Times 39 (2016) 526–533, https://doi.org/10.1016/s2212-5671(16)30295-7.
[2] F. Marechal, D. Favrat, E. Jochem, Energy in the perspective of the sustainable development: the 2000 W society challenge, in: Resour. Conserv. Recycl., Elsevier, 2005, pp. 245–262, https://doi.org/10.1016/j.resconrec.2005.01.008.
[3] J. Zhong, D. Chen, H. Xu, W. Zhao, J. Sun, Z. Ji, Red-emitting CaLa4(SiO4)3O:Eu3+phosphor with superior thermal stability and high quantum efficiency for warm w-LEDs, J. Alloys Compd. 695 (2017) 311–318, https://doi.org/10.1016/j.jallcom.2016.10.211.
[4] N. Shaishta, W.U. Khan, S.K.B. Mane, A. Hayat, D.D. Zhou, J. Khan, N. Mehmood, H.K. Inamdar, G. Manjunatha, Red-emitting CaSc2O4:Eu3+ phosphor for NUV-based warm white LEDs: structural elucidation and Hirshfeld surface analysis, Int. J. Energy Res. 44 (2020) 8328–8339, https://doi.org/10.1002/er.5376.
[5] W. Ran, H.M. Noh, S.H. Park, B.K. Moon, J.H. Jeong, J.H. Kim, J. Shi, Break the interacting bridge between Eu3+ ions in the 3D network structure of CdMoO4: Eu3+ Bright red emission phosphor, Sci. Rep. 8 (2018) 1–12, https://doi.org/10.1038/s41598-018-24374-3.
[6] Q. Zhang, X. Wang, X. Ding, Y. Wang, A potential red-emitting phosphor BaZrGe3O9:Eu3+ for WLED and FED applications: synthesis, structure, and luminescence properties, Inorg. Chem. 56 (2017) 6990–6998, https://doi.org/10.1021/acs.inorgchem.7b00591.
[7] S. Kaur, A.S. Rao, M. Jayasimhadri, Color tunability and energy transfer studies of Dy 3+/Eu 3+ co-doped calcium aluminozincate phosphor for lighting applications, Mater. Res. Bull. 116 (2019) 79–88, https://doi.org/10.1016/j.materresbull.2019.04.022.
[8] J. Zhao, H. Gao, H. Xu, Z. Zhao, H. Bu, X. Cao, L. He, Z. Yang, J. Sun, Structure and photoluminescence of Eu3+doped Sr2InTaO6red phosphor with high color purity, RSC Adv. 11 (2021) 8282–8289, https://doi.org/10.1039/d1ra00165e.
[9] X. Wang, Z. Zhao, Q. Wu, Y. Li, Y. Wang, A Garnet-based Ca2YZr2Al3O12:Eu3+ red-emitting phosphor for n-UV light emitting diodes and field emission displays: electronic structure and luminescence properties, Inorg. Chem. 55 (2016) 11072–11077, https://doi.org/10.1021/acs.inorgchem.6b01711.
[10] K. Munirathnam, G.R. Dillip, B. Ramesh, S.W. Joo, B.D. Prasad Raju, Synthesis, photoluminescence and thermoluminescence properties of LiNa3P2O7:Tb3+ green emitting phosphor, J. Phys. Chem. Solid. 86 (2015) 170–176, https://doi.org/10.1016/j.jpcs.2015.07.011.
[11] B. Wang, H. Lin, J. Xu, H. Chen, Y. Wang, CaMg2Al16O27:Mn4+-based red phosphor: a potential color converter for high-powered warm W-led, ACS Appl. Mater. Interfaces 6 (2014) 22905–22913, https://doi.org/10.1021/am507316b.
[12] A.K. Kunti, N. Patra, R.A. Harris, S.K. Sharma, D. Bhattacharyya, S.N. Jha, H. C. Swart, Structural properties and luminescence dynamics of CaZrO3:Eu3+ phosphors, Inorg. Chem. Front. 8 (2021) 821–836, https://doi.org/10.1039/d0qi01178a.
[13] R.A. Talewar, S. Mahamuda, K. Swapna, M. Venkateswarlu, A.S. Rao, Spectroscopic studies of Sm 3+ ions doped alkaline-earth chloro borate glasses for visible photonic applications, Mater. Res. Bull. 105 (2018) 45–54, https://doi.org/10.1016/j.materresbull.2018.04.033.
[14] Y. Tian, B. Chen, R. Hua, J. Sun, L. Cheng, H. Zhong, X. Li, J. Zhang, Y. Zheng, T. Yu, L. Huang, H. Yu, Optical transition, electron-phonon coupling and fluorescent quenching of La2(MoO4)3:Eu3+ phosphor, J. Appl. Phys. 109 (2011) 1–7, https://doi.org/10.1063/1.3551584.
[15] B.P. Singh Maheshwary, J. Singh, R.A. Singh, Luminescence properties of Eu3+-activated SrWO4 nanophosphors-concentration and annealing effect, RSC Adv. 4 (2014) 32605–32621, https://doi.org/10.1039/c4ra05903d.
[16] A.R. Kadam, R.L. Kohale, G.C. Mishra, S.J. Dhoble, Eu(iii)-Doped tri-calcium Ca3(1−X−Z)MZ(PO4)2AX:Xhost array: optical investigations of down-conversion red phosphor for boosting display intensity and high color purity, New J. Chem. 45 (2021) 7285–7307, https://doi.org/10.1039/d0nj05930g.
[17] F.H. Borges, F.J. Caixeta, R.R. Pereira, S. Ruella de Oliveira, R.A.S. Ferreira, R. R. Gonçalves, High Eu3+concentration quenching in Y3TaO7solid solution for orange-reddish emission in photonics, RSC Adv. 10 (2020) 16917–16927, https://doi.org/10.1039/d0ra01912g.
[18] B. Kumar Grandhe, V. Rao Bandi, K. Jang, H.S. Lee, D.S. Shin, S.S. Yi, J.H. Jeong, Effect of sintering atmosphere and lithium ion co-doping on photoluminescence properties of NaCaPO 4:Eu 2 phosphor, Ceram. Int. 38 (2012) 6273–6279, https://doi.org/10.1016/j.ceramint.2012.04.082.
[19] S. Choi, Y.J. Yun, H.K. Jung, Eu 2+ and Mn 2+ activated single phase white emitting phosphor Na(Sr,Ba)PO 4 for phosphor converted-LEDs, Mater. Lett. 75 (2012) 186–188, https://doi.org/10.1016/j.matlet.2012.02.019.
[20] R.Y. Yang, Y.M. Peng, H.L. Lai, Y.K. Su, S.J. Chang, The optimum sintering condition for KSrPO4:Eu3+ phosphors applied in WLEDs, Ceram. Int. 43 (2017) S682–S687, https://doi.org/10.1016/j.ceramint.2017.05.262.
[21] H. Pan, X. Li, J. Zhang, L. Guan, H. Su, F. Teng, Synthesis and luminescent properties of NaZnPO4:Eu3+ red phosphors for white LEDs, Mater. Lett. 155 (2015) 106–108, https://doi.org/10.1016/j.matlet.2015.04.084.
[22] K. Fukuda, A. Moriyama, T. Iwata, Crystal structure, phase transition and anisotropic thermal expansion of barium zirconium diorthophosphate, BaZr(PO4)2, J. Solid State Chem. 178 (2005) 2144–2151, https://doi.org/10.1016/j.jssc.2005.04.022.
[23] K. Popa, R.J.M. Konings, P. Boulet, D. Bouëxière, A.F. Popa, The high temperature behavior of barium zirconium diorthophosphate, Thermochim. Acta 436 (2005) 51–55, https://doi.org/10.1016/j.tca.2005.06.035.
[24] T. Geisler, K. Popa, R.J.M. Konings, A.F. Popa, A Raman spectroscopic study of the phase transition of BaZr(PO4)2: evidence for a trigonal structure of the high-

temperature polymorph, J. Solid State Chem. 179 (2006) 1490–1496, https://doi.org/10.1016/j.jssc.2006.01.065.
[25] R.D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta Crystallogr., Sect. A 32 (1976) 751–767, https://doi.org/10.1107/S0567739476001551.
[26] S. Kaur, M. Jayasimhadri, A.S. Rao, A novel red emitting Eu3+ doped calcium aluminozincate phosphor for applications in w-LEDs, J. Alloys Compd. 697 (2017) 367–373, https://doi.org/10.1016/j.jallcom.2016.12.150.
[27] Z. Xu, L. Fu, L. Liu, F. Du, Effects of single and composite fluxes on the morphology and luminescence intensity of Ce3+ doped Lu3Al5O12 phosphors, Mater. Chem. Phys. 248 (2020), https://doi.org/10.1016/j.matchemphys.2020.122918.
[28] R. Yu, Y. Guo, L. Wang, H.M. Noh, B.K. Moon, B.C. Choi, J.H. Jeong, Characterizations and optical properties of orange–red emitting Sm3+-doped Y6WO12 phosphors, J. Lumin. 155 (2014) 317–321, https://doi.org/10.1016/j.jlumin.2014.06.041.
[29] Y. Huang, X. Wang, H.S. Lee, E. Cho, K. Jang, Y. Tao, Synthesis, vacuum ultraviolet and ultraviolet spectroscopy of Ce 3+ ion doped olgite Na(Sr, Ba)PO4, J. Phys. D Appl. Phys. 40 (2007) 7821–7825, https://doi.org/10.1088/0022-3727/40/24/035.
[30] K. Vivekanandan, S. Selvasekarapandian, P. Kolandaivel, M.T. Sebastian, S. Suma, Raman and FT-IR spectroscopic characterisation of flux grown KTiOPO4 and KRbTiOPO4 non-linear optical crystals, Mater. Chem. Phys. 49 (1997) 204–210, https://doi.org/10.1016/S0254-0584(97)80165-4.
[31] K. Munirathnam, P.C. Nagajyothi, D. Prakashbabu, B. Deva Prasad Raju, J. Shim, X-ray photoelectron spectroscopy and optical analysis of pure white light emitting Dy3+ and Mn2+ codoped Na3Y(PO4)2 phosphors for solid-state lighting, Ceram. Int. 45 (2019) 686–694, https://doi.org/10.1016/j.ceramint.2018.09.230.
[32] Y. Li, Z. Zhao, Z. Song, R. Wan, J. Qiu, Z. Yang, Z. Yin, X. Liu, Q. Liu, Y. Zhou, Far-red-emitting BiOCl:Eu3+ phosphor with excellent broadband NUV-excitation for white-light-emitting diodes, J. Am. Ceram. Soc. 98 (2015) 2170–2176, https://doi.org/10.1111/jace.13589.
[33] H. Li, R. Zhao, Y. Jia, W. Sun, J. Fu, L. Jiang, S. Zhang, R. Pang, C. Li, Photoluminescence Properties, 2014, pp. 1–7.
[34] J.H. Nobbs, Kubelka—munk theory and the prediction of reflectance, Rev. Prog. Coloration Relat. Top. 15 (1985) 66–75, https://doi.org/10.1111/j.1478-4408.1985.tb03737.x.
[35] X. Zhang, J. Li, Y. Shu, K. Luo, H. Liu, P. Zhang, Y. Huang, G. Liu, Y. Fan, Photoluminescence properties of tunable full-color-emitting CaZrO3: ln (Ln = Tb3 +, Eu3+, Tb3+/Eu3+) phosphors for white LEDs, J. Lumin. 252 (2022) 119305, https://doi.org/10.1016/j.jlumin.2022.119305.
[36] D.A. Hakeem, K. Park, Enhanced photoluminescence properties of red emitting Ba1-xAl2Si2O8:xEu3+ phosphors by charge compensation, New J. Chem. 39 (2015) 7768–7772, https://doi.org/10.1039/c5nj00988j.
[37] O. Jamal Eddine, D. Boyer, M. El Bouchti, A. Boukhriss, O. Cherkaoui, R. Mahiou, H. Hannache, S. Gmouh, Effect of Eu3+ and Tb3+ ions concentration on the mechanical, structural, and photoluminescence properties of phosphate glass fibers, Opt. Mater. 122 (2021) 111664, https://doi.org/10.1016/j.optmat.2021.111664.
[38] K. Binnemans, Interpretation of europium(III) spectra, Coord. Chem. Rev. 295 (2015) 1–45, https://doi.org/10.1016/j.ccr.2015.02.015.
[39] A.V. Egorysheva, V.D. Volodin, A.A. Chistyakov, Y.A. Kuzishchin, V.M. Skorikov, T.D. Dudkina, Luminescence of europium-doped BaO-Bi 2O 3-B 2O 3 glasses, Inorg. Mater. 46 (2010) 1384–1390, https://doi.org/10.1134/S0020168510120204.
[40] Z. Liang, Z. Mu, Q. Wang, D. Zhu, F. Wu, The synthesis and luminescence properties of a novel red-emitting phosphor: Eu3+-doped Ca9La(PO4)7, Appl. Phys. Mater. Sci. Process 123 (2017) 1–7, https://doi.org/10.1007/s00339-017-1226-5.
[41] J. Lu, Z. Mu, D. Zhu, Q. Wang, F. Wu, Luminescence properties of Eu3+ doped La3Ga5GeO14 and effect of Bi3+ co-doping, J. Lumin. 196 (2018) 50–56, https://doi.org/10.1016/j.jlumin.2017.12.017.
[42] M. Liao, Z. Mu, S. Zhang, F. Wu, Z. Nie, Z. Zheng, X. Feng, Q. Zhang, J. Feng, D. Zhu, A red phosphor Mg3Y2Ge3O12: Bi3+, Eu3+ with high brightness and excellent thermal stability of luminescence for white light-emitting diodes, J. Lumin. 210 (2019) 202–209, https://doi.org/10.1016/j.jlumin.2019.02.038.
[43] X. Cai, Z. Mu, S. Zhang, D. Zhu, Q. Wang, Y. Yang, D. Luo, F. Wu, An investigation about the luminescence mechanism of SrGa2O4:Eu3+ showing no detectable energy transfer from the host to the dopant ions, J. Lumin. 200 (2018) 169–174, https://doi.org/10.1016/j.jlumin.2018.04.019.
[44] J. Zhang, G. Cai, W. Wang, L. Ma, X. Wang, Z. Jin, Tuning of emission by Eu3+ concentration in a Pyrophosphate: the effect of local symmetry, Inorg. Chem. 59 (2020) 2241–2247, https://doi.org/10.1021/acs.inorgchem.9b02949.
[45] D. Cui, Q. Xiang, Z. Song, Z. Xia, Q. Liu, The synthesis of narrow-band red-emitting SrLiAl 3 N 4 :Eu 2+ phosphor and improvement of its luminescence properties, J. Mater. Chem. C 4 (2016) 7332–7338, https://doi.org/10.1039/C6TC02093C.
[46] X. Li, Z. Yang, L. Guan, C. Liu, P. Li, Luminescent properties of Eu3+-doped La2Mo2O9 red phosphor by the flux method, J. Cryst. Growth 310 (2008) 3117–3120, https://doi.org/10.1016/j.jcrysgro.2008.03.006.
[47] Y. Feng, J. Huang, L. Liu, J. Liu, X. Yu, Enhancement of white-light-emission from single-phase Sr5(PO4)3F:Eu2+,Mn2+ phosphors for near-UV white LEDs, Dalton Trans. 44 (2015) 15006–15013, https://doi.org/10.1039/c5dt01794g.
[48] S. Kaur, A.S. Rao, M. Jayasimhadri, Enhanced red down-conversion luminescence and high color purity from flux assisted Eu3+ doped calcium aluminozincate phosphor, J. Lumin. 202 (2018) 461–468, https://doi.org/10.1016/j.jlumin.2018.06.015.
[49] A.K. Vishwakarma, M. Jayasimhadri, Significant enhancement in photoluminescent properties via flux assisted Eu3+ doped BaNb2O6 phosphor for white LEDs, J. Alloys Compd. 683 (2016) 379–386, https://doi.org/10.1016/j.jallcom.2016.05.052.
[50] L. Zhang, Z. Lu, P. Han, J. Lu, N. Xu, L. Wang, Q. Zhang, The evolution and role of NH 4 Cl flux used to synthesize Sr 2 SiO 4: Dy 3+ phosphor by solid-state reaction method, J. Am. Ceram. Soc. 95 (2012) 3871–3877, https://doi.org/10.1111/j.1551-2916.2012.05378.x.
[51] M.H.V. Werts, R.T.F. Jukes, J.W. Verhoeven, The emission spectrum and the radiative lifetime of Eu3+ in luminescent lanthanide complexes, Phys. Chem. Chem. Phys. 4 (2002) 1542–1548, https://doi.org/10.1039/b107770h.
[52] G.A. Gusev, S.M. Masloboeva, T.B. Popova, M.A. Yagovkina, M.V. Zamoryanskaya, Influence of gadolinium tantalum niobates doped with Eu3+ ions structural properties on luminescence spectrum features, J. Lumin. 252 (2022) 119281, https://doi.org/10.1016/j.jlumin.2022.119281.
[53] F. Rivera-López, I.R. Martín, I. Da Silva, C. González-Silgo, U.R. Rodríguez-Mendoza, V. Lavín, F. Lahoz, S.M. Díaz-González, M.L. Martínez-Sarrión, L. Mestres, J. Fernández-Urban, Analysis of the Eu3+ emission in a SrWO4 laser matrix under pressure, High Pres. Res. 26 (2006) 355–359, https://doi.org/10.1080/08957950601105085.
[54] R. Shukla, S.K. Gupta, H. Yadav, P. Ranjane, R.M. Kadam, S.N. Achary, A.K. Tyagi, A carnegieite type red emitting NaAlSiO4:Eu3+ phosphor: concentration dependent time resolved photoluminescence and Judd-Ofelt analysis, J. Lumin. 209 (2019) 283–290, https://doi.org/10.1016/j.jlumin.2019.02.002.
[55] Y. Wang, J. Ding, Y. Wang, Ca2-xY1+xZr2-xAl3+xO12:Ce3+: solid solution Design toward the green emission garnet structure phosphor for near-UV LEDs and their luminescence properties, J. Phys. Chem. C 121 (2017) 27018–27028, https://doi.org/10.1021/acs.jpcc.7b09783.
[56] Y. Zhao, C. Riemersma, F. Pietra, R. Koole, C. De Mello Donegá, A. Meijerink, High-temperature luminescence quenching of colloidal quantum dots, ACS Nano 6 (2012) 9058–9067, https://doi.org/10.1021/nn303217q.
[57] J. Sun, X. Zhang, Z. Xia, H. Du, Luminescent properties of LiBaPO 4:RE (RE Eu 2, Tb 3, Sm 3) phosphors for white light-emitting diodes, J. Appl. Phys. 111 (2012), https://doi.org/10.1063/1.3673331.
[58] Q. Liu, Y. Liu, F. Yang, B. Han, H. Feng, Q. Yu, A novel orange-red phosphor Ca3B2O 6:Sm3+, A+ (A = Li, Na, K) for white light emitting diodes, Funct. Mater. Lett. 7 (2014) 2–6, https://doi.org/10.1142/S1793604714500337.
[59] Z. Tang, Q. Zhang, Y. Cao, Y. Li, Y. Wang, Eu 2 + -doped Ultra-broadband VIS-NIR Emitting Phosphor, 2020, p. 388.
[60] B.J. Chen, E.Y.B. Pun, H. Lin, Photoluminescence and spectral parameters of Eu3+ in sodium-aluminum-tellurite ceramics, J. Alloys Compd. 479 (2009) 352–356, https://doi.org/10.1016/j.jallcom.2008.12.072.
[61] P. Du, X. Huang, J.S. Yu, Ethylene glycol assisted rapid preparation of NaEuF4 nanorods with splendid thermal stability for indoor illumination and optical displays, Dyes Pigments 153 (2018) 307–315, https://doi.org/10.1016/j.dyepig.2018.02.024.
[62] R.G.A. Kumar, S. Hata, K.G. Gopchandran, Diethylene glycol mediated synthesis of Gd2O3:Eu 3+ nanophosphor and its Judd-Ofelt analysis, Ceram. Int. 39 (2013) 9125–9136, https://doi.org/10.1016/j.ceramint.2013.05.010.
[63] A. Hooda, A. Khatkar, P. Boora, S. Singh, S. Devi, S.P. Khatkar, V.B. Taxak, Structural, optical and morphological features of combustion derived Ba3Y4O9: Dy3+ nanocrystalline phosphor with white light emission, Optik 228 (2021), https://doi.org/10.1016/j.ijleo.2020.166176.
[64] L. Dačanin, S.R. Lukić, D.M. Petrović, M. Nikolić, M.D. Dramićanin, Judd-Ofelt analysis of luminescence emission from Zn2SiO 4:Eu3 nanoparticles obtained by a polymer-assisted solgel method, Phys. B Condens. Matter 406 (2011) 2319–2322, https://doi.org/10.1016/j.physb.2011.03.068.
[65] B. Verma, R.N. Baghel, D.P. Bisen, N. Brahme, V. Jena, Microstructural, luminescence properties and Judd-Ofelt analysis of Eu3+ activated K2Zr(PO4)2 phosphor for lighting and display applications, Opt. Mater. 129 (2022) 112459, https://doi.org/10.1016/j.optmat.2022.112459.
[66] M.P. Hehlen, M.G. Brik, K.W. Krämer, 50th anniversary of the Judd-Ofelt theory: an experimentalist's view of the formalism and its application, J. Lumin. 136 (2013) 221–239, https://doi.org/10.1016/j.jlumin.2012.10.035.

SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Mechanics

Journal homepage:
https://semarakilmu.com.my/journals/index.php/appl_mech/index
ISSN: 2289-7895



Optimization of Lamps Configuration for a Large-scale Indoor Solar Simulator for PV Panels and Solar Collectors

Al Yaqdhan Al Umairi[1], Khalid Al Sultani[1], Abdullah Al Shuaili[1], Aisha Al Adawi[1], Afzal Husain[1,*], Mohd Zahid Ansari[2], Mohd Zunaid[3]

[1] Department of Mechanical and Industrial Engineering, Sultan Qaboos University, Muscat, PC-123, Oman
[2] MEMS and Microfluidics Lab, Mechanical Engineering Department, PDPM-Indian Institute of Information Technology Design and Manufacturing, Jabalpur, India
[3] Department of Mechanical Engineering, Delhi Technological University, Bawana Road, Delhi, India

ARTICLE INFO

Article history:
Received 3 June 2024
Received in revised form 21 July 2024
Accepted 29 July 2024
Available online 30 August 2024

Keywords:
Solar simulator; tungsten halogen lamp; solar spectrum; spectral uniformity; temporal uniformity

ABSTRACT

The current study presents investigations of a solar simulator design, development, and performance characterization for the application of photovoltaic (PV), thermal collector, and photovoltaic/thermal (PV/T) indoor experiments. The study highlights the research outcomes of the lamp configurations and numbers in the simulator to achieve an acceptable radiation distribution in the targeted area as per the ASTM standards. Tungsten Halogen lamps were used to mimic solar radiation due to their proximity to the solar spectrum of the visible and infrared band and relatively low cost. The results indicate that a solar simulator of Class-C for spatial uniformity and Class-A for temporal uniformity can be achieved with the application of Tungsten Halogen lamps if arranged properly and a range of radiation flux can be obtained to replicate the variation of solar flux during the daytime. It was observed that three to six lamps of 1kW arranged in specific configurations can provide solar radiation varying from 378 to 969 W/m^2 mimicking the daytime solar flux variation on an experimental area of 1200 × 520 mm^2.

1. Introduction

The clean energy transition is widely adopted at different word forums promoting wind, solar and other forms of renewable energy. Apart from several environmental, social, and sustainability factors, there are some adverse impacts, and it has some winners and some losers in the process. There are sparkling impacts on the coal and similar industries affecting the lives of millions [1]. However, the gradual adaptation of clean and sustainable energy resources is gaining great attention from the research community. Solar energy harvesting through solar thermal collectors and Photovoltaic (PV) panels is gaining momentum due to the reduced cost of PVs.

* *Corresponding author.*
E-mail address: afzal19@squ.edu.om

https://doi.org/10.37934/aram.123.1.195205

The testing of PV panels and thermal collectors in an indoor environment requires solar simulator to replicate solar light. The solar simulator is used to mimic realistic optical and thermal conditions of the sun. It is used for indoor testing of photovoltaic and thermal systems or ground testing of space applications. The main component of the solar simulator is the light source that is responsible for offering a spectral composition similar to the sunlight. Currently, several types of lamps are used in designing solar simulators, including tungsten halogen, arc (metal halide, and xenon), and light-emitting diode (LED) lamps. The choice of the light source is influenced by the solar simulator's intended characteristics and application [2]. The Tungsten halogen and arc lamps area mainly used to wider spectrum of solar radiation; whereas LED lamps are mainly used for visible lights.

A standard is known as ASTM E927-19 [3] was developed by the American Society for Testing and Materials (ASTM) to categorize solar simulators based on three factors: spectral match (R_{SM}), spatial non-uniformity (S_{NE}), and temporal instability (T_{IE}). The spectral match quantifies the similarity between the produced spectrum by the lamps and the solar spectrum. The amount at which the output power varies across the test plane is known as spatial non-uniformity, which is one of the major criteria of solar simulator design. The temporal instability measures the variation of the produced power within the tested time. As shown in Table 1, the criteria were developed to categorize solar simulators as Class A, Class B, Class C, or Class U (unclassified). The solar simulator is categorized as an AAA solar simulator if all parameters are in Class A [4].

Table 1
ASTM E927-19 Classification of Solar Simulators [3]

Classification	Spectral match	Spatial non-uniformity	Temporal instability
Class A	$0.75 \le R_{SM} \le 1.25$	$S_{NE} \le 2\%$	$T_{IE} \le 2\%$
Class B	$0.60 \le R_{SM} \le 1.40$	$S_{NE} \le 5\%$	$T_{IE} \le 5\%$
Class C	$0.40 \le R_{SM} \le 2.00$	$S_{NE} \le 10\%$	$T_{IE} \le 10\%$
Class U	$R_{SM} > 2.00$	$S_{NE} > 10\%$	$T_{IE} > 10\%$

Several researchers have investigated and proposed solar simulators based on lamp configurations and numbers. Meng *et al.,* [5] used 188 metal halide lamps with parabolic reflectors and 400 W power input to design a large-scale solar simulator. The amount of irradiance was controlled by switching off some lamps or by changing the distance between the lamps and the test plane. The highest irradiance measured was 1100 W/m^2. Regarding spectral match and spatial non-uniformity, the solar simulator's classifications were B and B, respectively. Xu *et al.,* [6] used 20, 150 W metal halide lamps to design a 130 cm × 80 cm solar simulator. The method of controlling the irradiance was changing the distance between each lamp and the distance between the lamps and the test area. It was found that the range of obtained irradiance was between 1008 W/m^2 and 1232 W/m^2 when the distance between the lamps and the test plane was 20 cm. In that situation, the calculated spatial non-uniformity and temporal instability were 9.56% (Class-C) and 1.2% (Class-B), respectively.

The tungsten halogen lamp is one of the most commonly used light sources in solar simulators. In this lamp, the electric current is passed through a tungsten filament which is surrounded by halogen gas, and the light is emitted due to the electric resistance of the filament. The emitted radiation from the halogen lamp is weaker in ultraviolet and stronger in infrared. However; Tungsten halogen lamps are widely used in solar simulators because of their low cost, and high intensity of light [7]. Grandi *et al.,* [8] proposed a solar simulator using a combination of halogen and LED lamps and achieved a Class B spectral match for a small experimental area of 12.5×12.5 cm^2. Several other authors have proposed solar simulators using different light sources in a single simulator [9–11]. However, these are mainly focused on small testing areas.

Hussain *et al.,* [12] fabricated a solar simulator to be used for indoor testing of solar collectors. The size of the target testing area was 0.636 m^2 and the simulator consists of 23 Philips 500 W halogen lamps. The achieved irradiance varied from 466 to 804 W/m^2. The intensity of radiation was controlled by turning off some of the lamps. It was stated that this strategy was adopted instead of utilizing a voltage regulator to maximize the life of the lights. The obtained spatial non-uniformity of each set of irradiances was satisfactory since it was less than 10%. Table 2 shows the characteristics of other solar simulators that were made of Tungsten halogen lamps.

Table 2
Characteristics of Tungsten Halogen lamp-based Solar Simulators

Author	Number of lambs	Lamp Power (W)	Test Plane Area (m^2)	Output Irradiance (W/m^2)
Moria *et al.,* [13]	8	500	0.4	500 - 1100
Salam *et al.,* [14]	36	50	0.55	1000
Yandri [15]	16	50	0.17	622
Agrawal *et al.,* [16]	28	500	2	600 - 800

Kim *et al.,* [2] designed a Class-C solar simulator using a combination of LED and halogen lamps to achieve a spectrum similar to the solar spectrum. Different colors (white, blue, cyan, purple) of LED lamps were used to emit the required ultraviolet and visible bands of the spectrum. The halogen lamps were used to provide light in the IR portion of the spectrum. Rosli *et al.,* [17] carried out an experimental investigation of an indoor solar simulator and evaluated spectral and spatial non-uniformity. The simulator used Tungsten Halogen lamps and the light spectrum closely matched with natural sunlight within specific wavelength ranges. For the 400-500nm range, the spectral match value was 0.40 (classified as Class C), and for the 500-600nm range, the value was 1.46 (classified as Class B) according to the international standard IEC 60904-9 [18]. The simulator demonstrated relatively uniform light distribution across the tested area with an average irradiance intensity of 981.98 W/m^2 and a spatial non-uniformity percentage of 8.42% over a tested area of 104 cm × 80 cm. The results ensure compliance with the minimum standard requirements set by IEC 60904-9.

Harun *et al.,* [19] conducted an experimental study on the flat plate absorber collector to assess its performance. The authors employed a solar simulator to imitate the solar spectrum, creating a controlled environment with a constant solar radiation intensity of 700 W/m^2. Horng *et al.,* [20] experimented to enhance the efficiency of solar panels utilizing PCM 36 as a cooling technology. They employed an indoor solar simulator (IEC 60904-9) having twelve 500 W Tungsten Halogen lamps to imitate the actual environment for the experiment.

In light of the above literature review, it can be concluded that a solar simulator based on halogen lamps can provide reasonable acceptability in terms of spectral match, spatial uniformity, and temporal stability to create large-scale solar experiments to mimic solar irradiance. The present study investigated and optimized several design parameters such as the number of lamps, arrangements of lamps, and distance between the simulator panel and the experimental plane.

2. Methodology

2.1 Design and Fabrication of Solar Simulator

The process of designing a solar simulator consists of the selection of the light source, the design and fabrication of the frame that holds the light source, and the thermos-electric control system, which is required for testing.

The light source is the heart of any solar simulator. Thus, the process of light source selection should be carried out carefully. The selection criteria of the light source in this study are color temperature, lifetime, cost, conversion efficiency, spectral match, and maintenance. The color temperature was chosen to be a selection criterion since the radiation of the lamps can be compared with the radiation of the sun based on color temperature. The color temperature of the sun was approximated to be 5777 K [7]. Several factors must be taken into account such as life cycle, cost, ease of operation, and heat produced apart from the standards discussed in Table 1 [21]. Table 3 presents a comparison among the different light sources for their maximum colour temperature, lifetime, conversion efficiency, and need for ballast and igniter.

Xenon lamps are the most suitable for colour temperature and closely match the solar spectrum; however, they suffer from high power consumption, short life cycle, high cost, and maintenance. Metal halide lamps come second to Xenon lamps for spectral match and colour temperature similar to tungsten Halogen lamps. The halogen lamps exhibit reasonable spectral match and lifetime but significantly lower cost, which makes it preferable for large-scale solar simulator applications where both infrared and visible parts of the spectrum are equally important.

Table 3
Properties of Different Light Sources and Selection Criteria

Selection Criteria	Xenon Lamps	Metal Halide Lamps	Halogen Lamps	LED
Maximum Colour Temperature	6000 K	5000 K	5000 K	3000 K
Lifetime	2000 hr.	3500 hr.	1000 hr.	20000 hr.
Conversion Efficiency	18.7 %	24.6 %	10.5 %	35 %
Need for Ballast and igniters	Yes	Yes	No	No

Further, maintenance was chosen to be a selection criterion since some types of lamps need additional components like ballast and igniters which require periodic maintenance. Based on that, criteria and considering the trade-off in the selection, the tungsten halogen lamps were selected with an input power of 1000 W.

The detailed drawing of the frame is shown in Figure 1 indicating the parts and their functions. The fabrication process of the frame was done in the engineering workshop at Sultan Qaboos University.



Fig. 1. Isometric View of the Solar Simulator Frame

3. Configuration Design

The aim of designing and testing the solar simulator is to select the best arrangement that achieves at most 10% special non-uniformity, 10% temporal instability, and 2% spectral match based on the ASTM standard that is specified in the ASTM Standards ASTM E927-19. The coming subsections describe the experiments that were conducted to achieve the required spatial non-uniformity and temporal instability.

3.1 Evaluation of Single Lamp Solar Simulator

This experiment aimed to evaluate and characterize the radiation distribution of a single lamp at different heights. In this experiment, the lamp was placed at four different heights from the test plane: 50, 75, 100, and 125 cm. At each height, the radiation sensor (Pyranometer) was moved in the test plane to measure the radiation along the axis of the lamp (X-axis) and perpendicular to the axis of the lamp (Y-axis). Also, 12 temperature sensors were distributed along the test plane to measure the temperature at different locations on the test plane. Figure 2 shows an illustration of the X and Y axis as well as the dimensions of the test plane and the floodlight. The results are presented in Figures 3 and 4, which show the radiation distribution of the single lamp at different heights along X and Y axes, respectively.

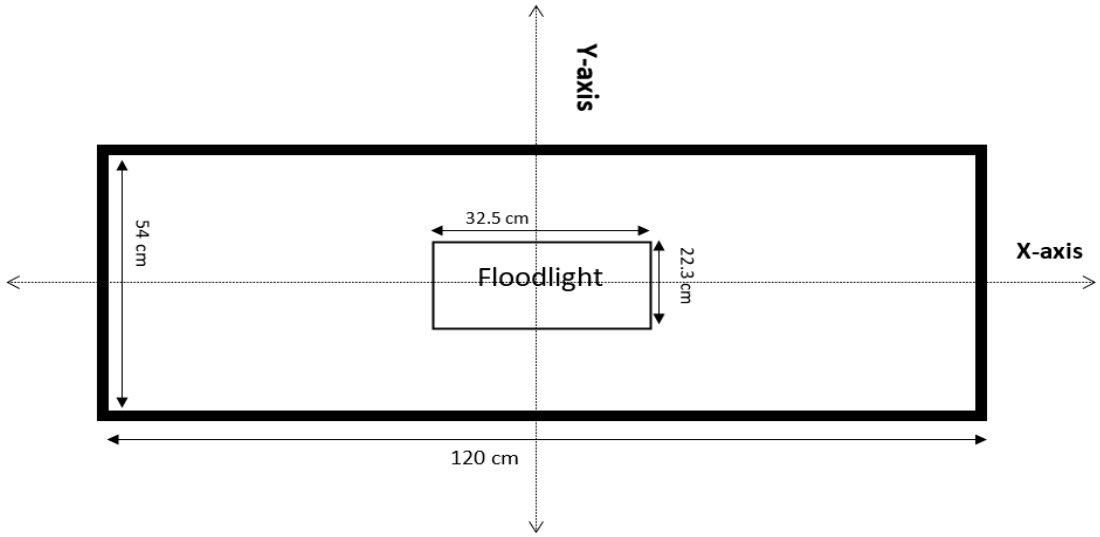


Fig. 2. Illustration of Single Lamp Radiation Test

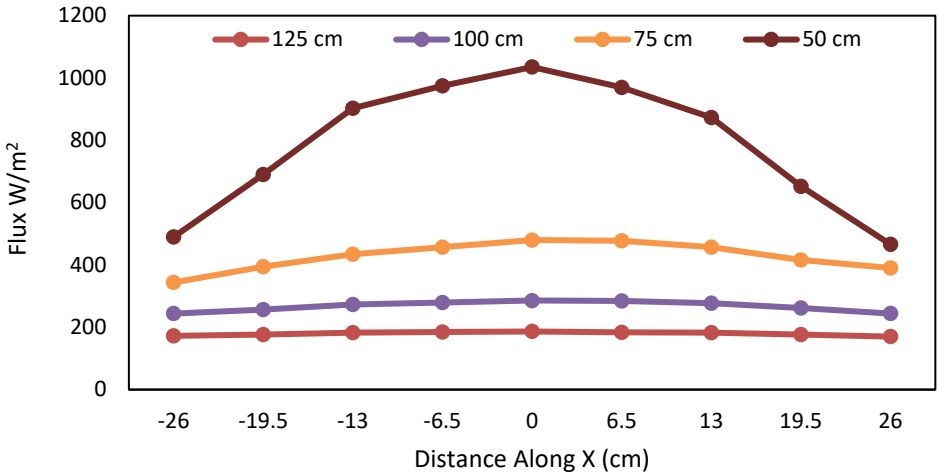


Fig. 3. Radiation Distribution of Single Lamp at Different Heights Along X-Axis

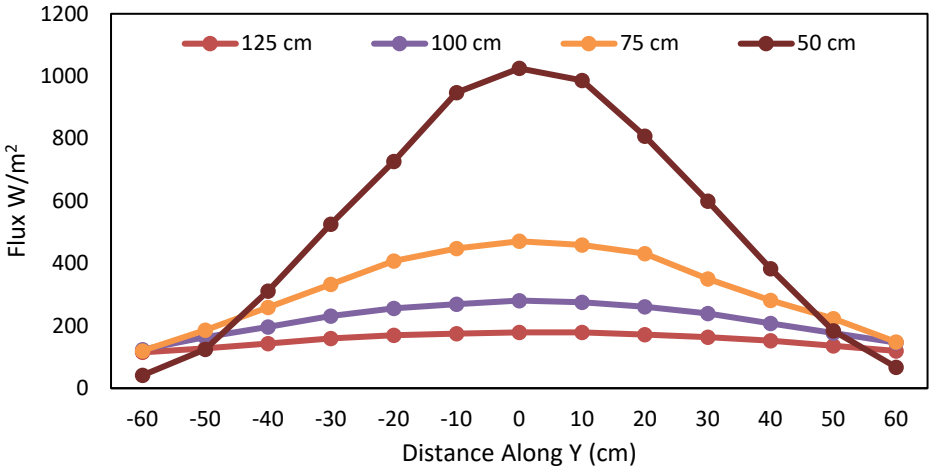


Fig. 4. Radiation Distribution of Single Lamp at Different Heights Along Y-Axis

From Figures 3 and 4, it is obvious that the radiation curve is almost flat at heights 100 and 125 cm along both X and Y axes in contrast with the radiation at heights 70 and 50 cm. However, at heights 100 and 125 cm the amount of radiation is small compared to the heights 50 and 75 cm. Moreover,

at heights 50 and 70 cm the temperature measurements show that there are some regions in which the temperature exceeds 60℃.

It can be concluded from the results that at heights 100 and 125 cm, the radiation can be increased by adding more lamps without significant change in the flatness of the radiation curve. Also, the radiations at heights 50 and 75 cm are not uniform. To achieve uniformity, more lamps should be added but that is not applicable due to the expected significant increase in the temperature along the test plane. However, this can be mitigated by using lower power lamps but more in numbers.

3.2 Evaluation of Three Lamps Solar Simulator

In this experiment, the radiation of three lamps was tested at heights 100 and 125 cm. One lamp was placed at the center of the test plane (X=Y=0) and the remaining two lamps were placed at a variable distance (K) from the lamp that is located at the center as shown in Figure 5. The spatial non-uniformity (S_{NE}) was calculated for each trial by measuring the radiation at different points along the X and Y axis. Then the maximum ($I_{max,S}$) and minimum ($I_{min,S}$) values of the measured irradiance were used in Eq. (1) to calculate the spatial non-uniformity (S_{NE}). Table 4 shows the spatial non-uniformity, and the average radiation in each trial is calculated as [21].

$$S_{NE} = \frac{I_{max,S} - I_{min,S}}{I_{max,S} + I_{min,S}} \times 100\% \quad (1)$$

The measurements for I were taken at a spatial distance of 5 cm over the entire experimental area.

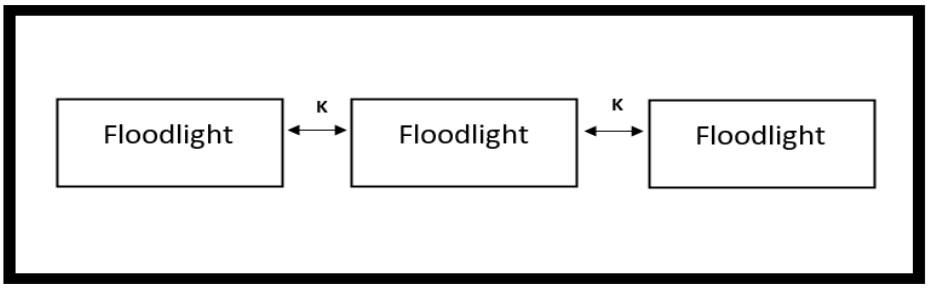


Fig. 5. The Arrangement of the Floodlights at Three Lamps Radiation Test

Table 4
Spatial Non-Uniformity and Average Radiation of Three Lamps at Different Arrangements

Trial number	Height (cm)	Description	Special Non-Uniformity	Average Radiation (W/m^2)
1	125	Three lamps, $K = 32cm$	20%	418.45
2		Three lamps, $K = 40cm$	18%	403.56
3		Three lamps, $K = 45cm$	14%	390.45
4		Three lamps, $K = 50cm$	9%	377.78
5	100	Three lamps, $K = 32cm$	26%	533.34
6		Three lamps, $K = 56cm$	17%	455.55
7		Three lamps, $K = 66\ cm$	10%	410.55

From Table 4, it can be concluded that the results are consistent. The spatial non-uniformity decreases when increasing the distance between the lamps because the radiation spread along the entire test plane rather than concentrating at the centre. Also, the average radiation decreases when

increasing the distance between the lamps. This is due to the dissipation of the radiation out of the test plane. The satisfied trials are numbers 4 and 7 in which the resulting spatial non-uniformity was not exceeding 10%, and the average radiations of these two trials are 377.78 and 410.55 W/m^2, respectively. However, the standard flux requirements for solar experiments are not met, which requires an increase in the number of lamps.

3.3 Evaluation of Four Lamps Solar Simulator

In This experiment, the radiation of four lamps was tested at heights 100 and 125 cm. The aim of using four lamps is to achieve higher average radiation than the maximum resulting average radiation of using three lamps (410.55 W/m^2). The approach to achieve the best arrangement was the same as the experiment of three lamps. The selected arrangement of the floodlights is shown in Figure 6, and it was used for both heights 125 cm and 100 cm. Table 5 shows the resulting spatial non-uniformity and average radiation after placing the lamps at heights 125 cm and 100 cm above the test plane. Based on the results, the required spatial uniformity was achieved in both heights but the uniformity at height 125 cm is higher than that of height 100 cm. Further, experiments were conducted with more lamps to achieve higher incidence flux.

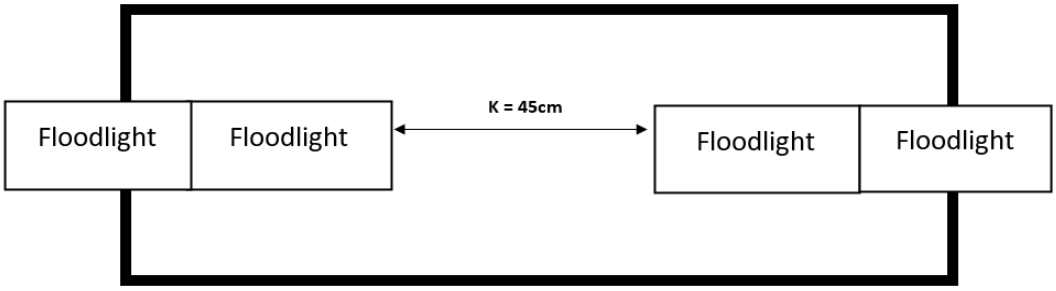


Fig. 6. The Arrangement of the Floodlights at Four Lamps Radiation Test

Table 5
Spatial Non-Uniformity and Average Radiation of Four Lamps at Different Heights

Height (cm)	Special Non-Uniformity	Average Radiation (W/m^2)
125	9%	472
100	10%	582.5

3.4 Evaluation of Five Lamps Solar Simulator

In this experiment, the radiation distribution of five lamps was tested at 125 cm height. The arrangement of the lamps is shown in Figure 7. Due to using a higher number of floodlights and to achieve more flexibility in changing the distance between them, the floodlight was rotated 90°. The best observed spatial non-uniformity and average radiation were 13% and 496 W/m^2, respectively at =10 cm. Thus, the experiment was not repeated at a height of 100 cm since the lower height results in lower uniformity as concluded from previous experiments.

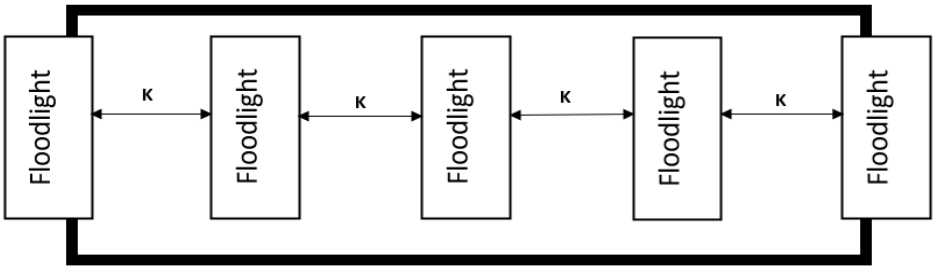


Fig. 7. The Arrangement of the Floodlights at Five Lamps Radiation Test

3.5 Evaluation of Six Lamps Solar Simulator

In this experiment, the radiation of six lamps was tested at heights 125, 100, and 75 cm. The arrangement of the lamps that are shown in Figure 8 was identical at each height. To select that arrangement, different values of *K* were tried until reaching the required results. Table 6 shows the resulting spatial non-uniformity and average radiation after placing the lamps at heights 125, 100, and 75 cm. Based on the results, the required spatial uniformity was achieved at all heights and the spatial non-uniformity is almost the same at all heights. The results are useful for designers testing new PV and PV/T modules intended to enhance electrical and thermal efficiency by conducting experiments in an indoor environment using solar simulator.

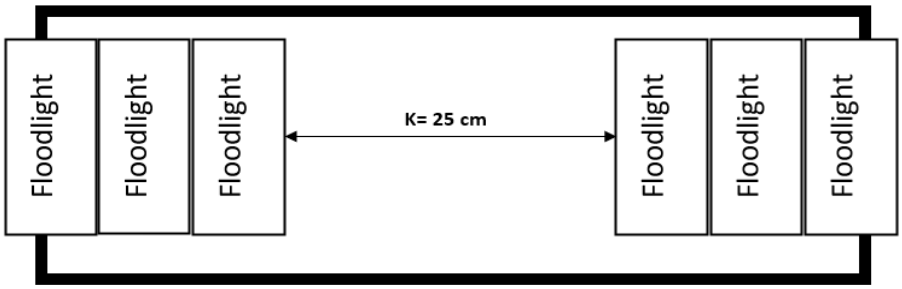


Fig. 8. The Arrangement of the Floodlights at Six Lamps Radiation Test

Table 6
Spatial Non-Uniformity and Average Radiation of Six Lamps at Different Heights

Height (cm)	Special Non-Uniformity	Average Radiation (W/m^2)
125	9%	571.2
100	9%	723.3
75	9%	969

3.6 Evaluation of Temporal Instability

The variation of emitted flux with time is critical and a measure of the temporal instability of irradiance (T_{IE}). The variation is measured over time and the degree of instability is defined the same way as that of spatial non-uniformity; however, using the maximum and minimum values over time. The temporal instability of irradiance can be expressed as:

$$T_{IE} = \frac{I_{max,T} - I_{min,T}}{I_{max,T} + I_{min,T}} \times 100\% \quad (2)$$

$I_{max,T}$ and $I_{min,T}$ are defined as maximum and minimum values of irradiation on the intended surface during experimentation.

Temporal instability is independent of the arrangement of the lamps. Thus, measuring the temporal instability in one arrangement is sufficient.

To measure the temporal instability of the lamps, the pyranometer was fixed on the test plane at the arrangement of four lamps and it is allowed to record the measurements of the radiation for 10 minutes with a sampling time of 1 second. After that the results were saved and the maximum and minimum values were used to calculate the temporal instability. The maximum and the minimum values were 620 and 604 W/m^2, respectively. The resulting temporal instability using those values is 1.3%, which is within the acceptable limit based on Table 1 for Class-A solar collectors. Further light configuration and numbers can be optimized and trade-offs can be obtained through an AI tool such as Deep Neural Network, which is a powerful tool to optimize complex engineering problems [22].

4. Conclusions

A solar simulator was designed based on tungsten halogen lamps. To validate the solar simulator, spatial non-uniformity, and temporal instability were measured based on ASTM E927-19 standards. The spatial non-uniformity of the designed solar simulator was tested at different heights and different arrangements. Table 7 summarizes the acceptable trials in which the special uniformity was satisfied. Also, the temporal instability of the light sources was tested, and it was found to be 1.3%. Based on the ASTM E927-19 classification of the solar simulators, the designed solar simulator belongs to Class-C and Class-A for spatial non-uniformity and temporal instability, respectively.

Table 7
Summary of the Acceptable Trials for Measuring Spatial Non-Uniformity

Height (cm)	Description	Spatial Non-Uniformity	Average Radiation (W/m^2)
125	Three lamps, $K = 50cm$	9%	377.78
	Four lamps, $K = 50cm$	9%	472
	Six lamps, $K = 25cm$	9%	571.2
100	Three lamps, $K = 66\ cm$	10%	410.55
	Four lamps, $K = 50cm$	10%	582.5
	Six lamps, $K = 25cm$	9%	723.3
75	Six lamps, $K = 25cm$	9%	969

Acknowledgments

The authors acknowledge support from SQU and MoHERI for funding this research through RC/RG-ENG/MIED/23/01 and RC/URG-ENG/MIED/23/01 grants.

References

[1] Carley, Sanya, and David M. Konisky. "The justice and equity implications of the clean energy transition." *Nature Energy* 5, no. 8 (2020): 569-577. https://doi.org/10.1038/s41560-020-0641-6

[2] Kim, Katherine A., Nathan Dostart, Julia Huynh, and Philip T. Krein. "Low-cost solar simulator design for multi-junction solar cells in space applications." In *2014 Power and Energy Conference at Illinois (PECI)*, pp. 1-6. IEEE, 2014. https://doi.org/10.1109/PECI.2014.6804544

[3] ASTM Standards (2019) E927 − 19 Standard Classification for Solar Simulators for Electrical Performance Testing of Photovoltaic Devices.

[4] Gunther, Matthew J. "Design and Validation of an Led-Based Solar Simulator for Solar Cell and Thermal Testing." Master's thesis, California Polytechnic State University, 2020.

[5] Meng, Q., Wang, Y., and Zhang, L. (2011) Irradiance characteristics and optimization design of a large-scale solar simulator. Solar Energy. 85 (9), 1758–1767. https://doi.org/10.1016/j.solener.2011.04.014

[6] Xu, Changwu, Kaicheng Huo, and Zhigang Ren. "Design and practical application of the solar radiation simulator." *Sensors & Transducers* 159, no. 11 (2013): 358.
[7] Tawfik, M., Xavier Tonnellier, and C. Sansom. "Light source selection for a solar simulator for thermal applications: A review." *Renewable and Sustainable Energy Reviews* 90 (2018): 802-813. https://doi.org/10.1016/j.rser.2018.03.059
[8] Grandi, Gabriele, Anastasiia Ienina, and Marinel Bardhi. "Effective low-cost hybrid LED-halogen solar simulator." *IEEE Transactions on Industry Applications* 50, no. 5 (2014): 3055-3064. https://doi.org/10.1109/TIA.2014.2330003
[9] Steiner, Myles A., John F. Geisz, Tom E. Moriarty, Ryan M. France, William E. McMahon, Jerry M. Olson, Sarah R. Kurtz, and Daniel J. Friedman. "Measuring IV curves and subcell photocurrents in the presence of luminescent coupling." *IEEE Journal of Photovoltaics* 3, no. 2 (2012): 879-887. https://doi.org/10.1109/JPHOTOV.2012.2228298
[10] Dong, Xue, Graham J. Nathan, Zhiwei Sun, Dahe Gu, and Peter J. Ashman. "Concentric multilayer model of the arc in high intensity discharge lamps for solar simulators with experimental validation." *Solar Energy* 122 (2015): 293-306. https://doi.org/10.1016/j.solener.2015.09.004
[11] Montgomery, Kyle H., David M. Wilt, Alex Howard, Bill Williams, and Bill Williams. "Characterization of a TS-Space quad-source solar simulator." In *2012 38th IEEE Photovoltaic Specialists Conference*, pp. 001517-001522. IEEE, 2012. https://doi.org/10.1109/PVSC.2012.6317884
[12] Hussain, Faridah, M. Y. H. Othman, B. Yatim, H. Ruslan, K. Sopian, Z. Anuar, and S. Khairuddin. "Fabrication and irradiance mapping of a low cost solar simulator for indoor testing of solar collector." (2011): 044502. https://doi.org/10.1115/1.4004548
[13] Moria, Hazim, Taib Iskandar Mohamad, and Fayez Aldawi. "Radiation distribution uniformization by optimized halogen lamps arrangement for a solar simulator." *Journal of Scientific and Engineering Research* 3, no. 6 (2016): 29-34.
[14] Salam, Z., M. Z. Ramli, and T. Zainal. "A low cost solar array simulator using halogen tungsten bulb with temperature control capability." In *2014 IEEE Conference on Energy Conversion (CENCON)*, pp. 170-174. IEEE, 2014. https://doi.org/10.1109/CENCON.2014.6967496
[15] Yandri, Erkata. "Uniformity characteristic and calibration of simple low cost compact halogen solar simulator for indoor experiments." *International Journal of Low-Carbon Technologies* 13, no. 3 (2018): 218-230. https://doi.org/10.1093/ijlct/cty018
[16] Agrawal, Sanjay, G. N. Tiwari, and H. D. Pandey. "Indoor experimental analysis of glazed hybrid photovoltaic thermal tiles air collector connected in series." *Energy and buildings* 53 (2012): 145-151. https://doi.org/10.1016/j.enbuild.2012.06.009
[17] Rosli, Mohd Afzanizam Mohd, Muhammad Aiman Danial Hamizan, Muhammad Zaid Nawam, Faridah Hussain, and Safarudin Gazali Herawan. "Measurement of Spectral Match and Spatial Non-uniformity for Indoor Solar Simulator." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 100, no. 1 (2022): 98-115. https://doi.org/10.37934/arfmts.100.1.98115
[18] Buchroithner, Armin, Bernhard Gerl, Richard Felsberger, and Hannes Wegleiter. "Design and operation of a versatile, low-cost, high-flux solar simulator for automated CPV cell and module testing." *Solar Energy* 228 (2021): 387-404. https://doi.org/10.1016/j.solener.2021.08.068
[19] Harun, Muhammad Amin, Zafri Azran Abdul Majid, Zairul Azrul Zakaria, Ahmad Faris Ismail, Sany Izan Ihsan, Kamaruzzaman Sopian, Amir Abdul Razak, and Ahmad Fadzil Sharol. "Performance Analysis of Flat Plate Base-Thermal Cell Absorber (FPBTCA): Low Thickness Design." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 96, no. 2 (2022): 122-133. https://doi.org/10.37934/arfmts.96.2.122133
[20] Horng, Chia Zhi, Mohd Afzanizam Mohd Rosli, Jayaprakash Ponnaiyan, Nurfarhana Salimen, Safarudin Ghazali Herawan, and Faridah Hussain. "The Performance Evaluation of Photovoltaic Integrated Organic Phase Change Material in a Single Container using Indoor Solar Simulator." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 109, no. 2 (2023): 168-183. https://doi.org/10.37934/arfmts.109.2.168183
[21] Esen, Vedat, Şafak Sağlam, and Bülent Oral. "Light sources of solar simulators for photovoltaic devices: A review." *Renewable and Sustainable Energy Reviews* 77 (2017): 1240-1250.\. https://doi.org/10.1016/j.rser.2017.03.062
[22] Xu, Yingliang, Fenglei Huang, Yan Liu, Junbo Yan, Fan Bai, Hao Yu, Baichuan Wang, and Jun Wang. "Effect of close-in successive explosions on the blast behaviors of reinforced concrete beams: an experimental study." In *Structures*, vol. 53, pp. 29-46. Elsevier, 2023. https://doi.org/10.1016/j.istruc.2023.04.043

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.2024.3463742

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.2024.0429000

Optimization of Wearable Biosensor Data for Stress Classification Using Machine Learning and Explainable AI

SHIKHA[1], DR. DIVYASHIKHA SETHIA[2], and Prof. S. Indu[3]
[1]Department of Computer Science Engineering, Delhi Technological University, India (e-mail: shikha.9126@gmail.com)
[2]Department of Software Engineering, Delhi Technological University, India (e-mail: divyashikha@dtu.ac.in)
[3]Department of Electronics and Communication, Delhi Technological University, India (e-mail: s.indu@dce.ac.in)

Corresponding author: Prof. S. Indu (s.indu@dce.ac.in).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT This work utilizes wearable devices for real-time stress detection and investigates the effectiveness of meditation audio in reducing stress levels after academic exposure. Physiological data, including Interbeat Interval (IBI)-derived Heart Rate Variability (HRV), Blood Volume Pulse (BVP), and electrodermal activity (EDA), are collected during the Montreal Imaging Stress Task (MIST). The stress classification methodology employs an integrated approach using Genetic Algorithm and Mutual Information to reduce feature set redundancy. It further uses Bayesian optimization to fine-tune machine learning hyperparameters. The results indicate that the combination of EDA, BVP, and HRV achieves the highest classification accuracy of 98.28% and 97.02% using the Gradient Boosting (GB) algorithm for 2-level and 3-level stress classification. In contrast, EDA and HRV alone achieve a comparable accuracy of 97.07% and 95.23% for 2-level and 3-level stress classification, respectively. Furthermore, the SHAP Explainable AI (XAI) analysis confirms that HRV and EDA are the most significant features for stress classification. The study also finds evidence that listening to meditation audio reduces stress levels. These findings highlight the potential of wearable technology combined with machine learning for real-time stress monitoring and management in academic environments.

INDEX TERMS Academic environment, Explainable AI, Feature Selection, Machine Learning, Mental stress, Wearable Device

I. INTRODUCTION

Students frequently face stressful situations in today's demanding academic environment, from class presentations to academic examinations. According to Prabu et al., [1], "academic stress is the psychological and emotional strain that arises from the demands and expectations associated with formal education and scholarly pursuits". This combination of rigorous academic workloads and competitive social dynamics can lead to significant mental stress, affecting their overall well-being and academic performance [2]. Studies have shown that unmanaged academic stress can lead to decreased motivation, anxiety, sleep problems, and even physical health issues [3] [4] [5]. Recognizing and addressing students' challenges in managing academic stress is crucial, as this has become a critical aspect of their educational journey.

However, the current focus in healthcare often lies in addressing and treating existing mental health issues and their associated symptoms. This approach is vital, but it is equally important to acknowledge the subtle indicators of stress and anxiety before they escalate. Recognizing these early signs empowers students to take action and manage their stress effectively, potentially preventing the need for more intensive interventions later.

To effectively identify these early signs of stress, researchers are increasingly emphasizing the relationship between variations in physiological signals and stress. When individuals undergo these physical and mental reactions to stress, their physiological signals also change [6] [7]. For example, Holtz et al. [8] introduced a sudden air horn sound and discovered that stress responses, such as an increase in Respiration Rate (RESP), Heart Rate (HR), and Electrodermal Activity (EDA) signals, occurred. Hence, it is vital to understand the connections between variations in physiological signals and stress to identify stress effectively. Concur-

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.2024.3463742

rently, there is a growing emphasis on using scientific and technological methods to monitor and analyze physiological signals, making it a significant research area in academia and industrial settings.

Wearable technology is promising for continuously and remotely monitoring mental health [9] [10]. These devices can gather physiological patient data and offer valuable contextual information. Additionally, the power of machine learning has enhanced data processing speed and the ability to derive meaningful insights from this physiological data. Recent advancements in machine learning have introduced ensemble techniques such as DeepAVP-TPPred [11], iAFPs-Mv-BiTCN [12], Deepstacked-AVPs [13], AIPs-SnTCN [14], and pAtbP-EnC [15], which have shown promising results in various applications. These models precisely predict peptide functionality by extracting features from sequential and evolutionary descriptors through ensemble learning techniques. These advanced predictors offer robust frameworks for analyzing physiological signals, providing more accurate and early detection of stress and anxiety indicators. One notable category of wearable devices is smartwatches. These devices function like miniaturized smartphones, boasting impressive computational capabilities. Moreover, they come equipped with various sensors capable of collecting physiological signals such as EDA, Photoplethysmography (PPG), Electrocardiography (ECG), and Skin Temperature (ST). This combination of technology and sensors makes them a valuable tool for monitoring mental health remotely and continuously.

Moreover, researchers have employed various stress-inducing tasks, including public speaking [16], questionnaires, programming contests [17], ice tasks, and games [18] while recording physiological signals. Some researchers have also delved into stress classification using logical thinking-based tasks such as pursuing mental arithmetic and the Montreal Imaging Stress Task (MIST) test [19].

The MIST, a well-validated and progressive stress-inducing test, has been extensively used for stress classification. It involves three distinct stages: resting, control, and experimental. The participants solve a series of mathematical tasks without time constraints during the control stage. However, in the experimental stage, participants are presented with challenging arithmetic problems that are time-constrained and aimed at inducing a higher stress level.

Various researchers [20] [21] explored EEG-based stress classification experiments using MIST. Minguillon et al. [22] proposed a portable system for real-time detection of stress levels. They utilized Electroencephalography (EEG), ECG, Electromyography (EMG), and Galvanic Skin Response (GSR) signals recorded during the MIST task. Using EEG alone, they achieved only 50% accuracy, but when combined with multimodal data, the accuracy increased to 86%. The results show the efficacy of multimodality, as it provides a comprehensive understanding of stress levels. Although EEG offers valuable insights into brain activity, it has some practical limitations, including noise susceptibility and discomfort, which hinder its everyday use. Wearable devices, especially smartwatches, emerge as a promising alternative for continuous, non-invasive monitoring of physiological signals like heart rate variability (HRV) and EDA. The unobtrusive nature of the signals makes them well-suited for real-time stress assessment in academic environments.

Table 1 shows the effectiveness of stress classification studies during tasks like MIST and incorporates wearable devices to record physiological signals such as EEG, ECG, EDA, and RESP. Notably, numerous studies explored EDA signals for stress classification.

Setz et al. [24] employed EDA and MIST as a stressor. The authors investigated six classifiers, including SVM (linear, Quad, RBF, Polynomial), LDA (linear discriminant analysis), and KNN (K-nearest Neighbours). They concluded that monitoring EDA allows discrimination between cognitive load and stress with an accuracy larger than 80% with leave-one-person-out cross-validation.

In another study, Zhu et al. [27] explored stress classification using EDA signal and MIST. They obtained the highest accuracy of 92.9% using the SVM classifier. Furthermore, Hsieh et al. [28] proposed a feature selection framework based on EDA signal. They conclude that the EDA signal contains the most significant information for stress detection and obtained the highest accuracy of 92.38%.

Despite the extensive use of EDA in stress classification studies, there needs to be more literature regarding the combined use of ECG and EDA in conjunction with MIST. Han et al. [25] explored the combination of ECG and RESP with MIST for stress classification and obtained the highest accuracy of 96% using RF. However, RESP may be less reliable as a stress biomarker due to challenges in accurately measuring breath rate in real-life environments and its susceptibility to confounding factors beyond stress, such as physical exertion, respiratory illnesses, changes in posture, or environmental conditions [35].

To the best of our knowledge, in the context of academic learning, no research has specifically investigated the combination of ECG and EDA during MIST. ECG provides insight into autonomic nervous system activity through HRV features, while GSR reflects sympathetic nervous system arousal through skin conductance changes [36]. Stress activates the sympathetic nervous system, leading to increased EDA and decreased HRV. Therefore, changes in these signals can help detect stress when an individual encounters a stressor. EDA is more sensitive when capturing acute stress [37] [38]. In contrast, HRV provides a more stable measure of chronic stress [39]. Combining data from EDA and IBI-derived HRV enables the development of more precise and resilient stress detection systems.

A. CONTRIBUTIONS

The study investigates the feasibility of IBI-derived HRV and EDA data in an academic setting by collecting student physiological data during MIST and analyzing the effectiveness of listening to meditation audio in alleviating stress. The study also employs an integrated feature selection, Genetic

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.2024.3463742

TABLE 1. Previous work on MIST and MAT stressors using Physiological Signals

Author	Signal	No. of Sub	Stressor Type	Phases	Experiment Duration	Classifiers	Result	Limitations
Minguillon et al. [22]	EEG, EMG, GSR, ECG	10	MIST	Relax, stress, Neutral	30 minutes	LDA	EEG Acc: 50%, Combined Sensors Acc (EEG, EMG, ECG, GSR): 86%	- only 4 channels are used. Did not work on Time Domain and Time-Frequency Domain features
Xia et al. [23]	EEG, ECG	22	MIST	Rest–Control-Experimental-Recovery	80 minutes	SVM	Acc: 79.54%, Sen: 81.00%, Spec: 78.00%	- Experiment duration is too long. Did not work on Time Domain and Time-Frequency Domain features
Setz et al. [24]	EDA	33	MIST	Rest, Control, Experiment, Recovery	50 minutes	SVM (Quad, RBF, Polynomial), LDA, Nearest class center	Accuracy: 82.8%	- Experiment Duration is too long. Utilized only EDA
Han et al. [25]	ECG, RSP	39	MIST	Rest - Control - Experiment - Recovery	18 minutes (excluding stress report filling time)	SVM, RF, LDA, Adaboost, KNN	3-level Acc:84% 2-level Acc:96%	- Can Utilize other physiological signals from wearable sensors.
Gjoreski et al. [26]	IBI, RSP, BVP, ST, EDA	22	MAT	Easy - Medium - Tough	30 minutes	SVM, RF, Boosting, Bagging, KNN, Naive Bayes, Decision Tree, Ensemble Selection	Recall:70% Precision:95%	- Considered only male participants.
Zhu et al. [27]	EDA	112	MAT	Easy - Medium - Tough	—	KNN, SVM, NB, LR, RF	Accuracy: 92.9%	- Other physiological signals results could be compared - Other evaluation metrics could be utilized
Hsieh et al. [28]	EDA	15	Public Speaking and MAT	Stress - Non stress	2 hr	XGboost	Accuracy: 92.38%	- Less sample size - Experimental duration is too long
Siirtola et al. [29]	ST, BVP, HR	15	Public Speaking and MAT	Stress - Non-stress	2 hr	LDA, QDA, RF	Accuracy: 87.4%	- Accuracy can be improved - Less sample size - Experimental duration is too long
Kumar et al. [16]	BVP, ECG, EDA, EMG	15	Public Speaking and MAT	Stress - Non-stress	2 hr	CNN hierarchal model	Accuracy: 87.77% F1 score: 83%	- Accuracy can be improved - Less sample size - Experimental duration is too long
Jaiswal et al. [30]	GSR	42	TSST, MAT,	Stress - Non stress	25 minutes	RF, SVM, KNN and AdaBoost.	Accuracy: 80%	- Other physiological biomarkers can be utilized.
Androutsou et al. [31]	PPG, GSR	32	SCWT, MAT	Rest, Control, Stress	45–60 min	SVM, KNN, DT, RF, XGB, RUSB, LGBM, ADB	LGBM Accuracy: 90.06% F1score: 90%	- reliance on self-reported stress levels, which may introduce subjective bias into the results
Campanella et al. [32]	PPG, EDA	29	TSST, MIST, Puzzle solving	Rest, Stress	30-35 min	RF, SVM, Log Regression	RF: 76.5%	- No hyperparameter tuning is adopted. - Imbalanced dataset.
Hemakom et al. [33]	ECG, EEG	66	MIST	Rest, Control, Stress	60 min	KNN, RBF-SVM, RF, Adaboost, MLP	87.60%	- Different models are trained for different stress levels which may result in misclassification in real-world scenarios.
Benchekroun et al. [34]	HRV	74 + 27	SCWT, MAT	Rest, Stress, Recovery	55 min	RF, LR	81%	- Self-report data label in second dataset could lead to biases. - Other physiological signals could be used.

Algorithm, and Mutual Information (GA+MI) to select the most important features from the data. Further, Bayesian optimization is used to tune machine learning classifiers and predict the stress level of each individual [40] [41]. The results obtained are validated using ten-fold cross-validation techniques. Lastly, the system's performance is analyzed for each individual to validate the impact of meditation audios on stress alleviation.

1) To conduct an innovative study to collect physiological signals such as EDA, IBI, and BVP from wearable devices across different stress levels and meditation phases.
2) To develop and optimize machine learning models for stress detection by implementing a hybrid feature selection approach combined with Bayesian optimization.
3) To employ the SHAP Explainable AI (XAI) technique to identify the most influential physiological features significantly impacting stress detection, promoting transparency in machine learning predictions.
4) To illustrate the impact of meditation on stress reduction through dynamic visualizations of physiological data collected from wearable devices among college students.

B. ROADMAP

The structure of this paper is as follows. Section 2 describes the experimental methodology employed in the study, which includes a detailed description of the data acquisition, preprocessing, and the proposed protocol design for developing a stress detection model. Section 3 presents the experimental results, and section 4 comprehensively discusses the results. Lastly, Section 5 explains the concluding remarks on this paper and potential future research directions.

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.2024.3463742

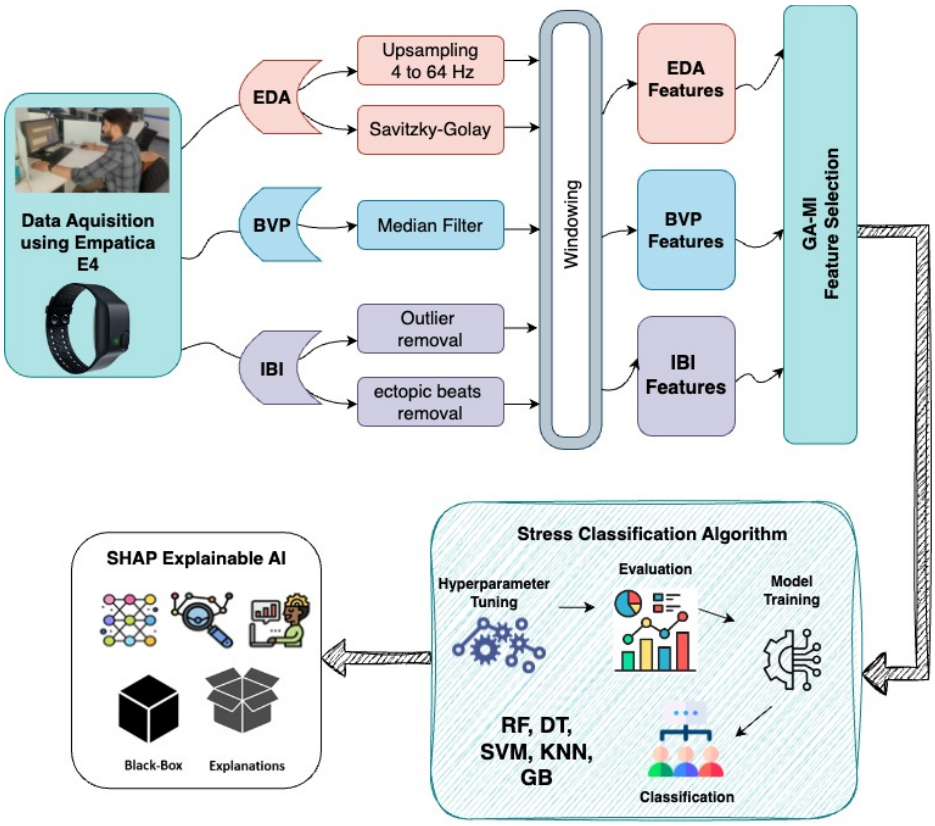


FIGURE 1. Proposed Classification Methodology

II. EXPERIMENTAL METHODOLOGY

This paper structures this section into five subsections shown in Figure 1. The methodology involves data acquisition, which outlines the procedure for collecting physiological signals, followed by data preprocessing to enhance signal quality. Subsequently, Feature Extraction and Feature Selection are performed. Further, the selected features undergo statistical analysis, and then these features are input into distinct classifiers, which are optimized using Bayesian optimization.

A. DATA AQUISITION:

This section involves the participant's details and the procedure during stress data collection.

1) Participants:

The study collects data from 36 engineering students aged between 17 and 25 (thirty-two males and four females) at Delhi Technological University, India. However, due to some participants missing IBI data caused by movements, only 30 student's information is considered for the study. During the data collection procedure, the participants are briefed about an overview of the study protocol and its objectives and asked to sign a consent form before starting the data collection.

The study adheres to strict ethical guidelines to ensure the well-being of participants. The DTU ethics committee, Delhi Technological University, Delhi, approves these guidelines. Before participation, all students are informed of the potential stress-inducing nature of the experiment, provide informed consent, and are asked to complete the Perceived Stress Scale (PSS) questionnaire to assess their self-reported stress levels [42]. Participants can also withdraw from the study at any time without penalty. Additionally, trained personnel monitored participants throughout the experiment and conducted debriefing sessions to address any concerns or discomfort experienced during the procedure. The test comprises five phases, as shown in Figure 2: Neutral phase, Rest phase, Controlled phase, Experimental phase, and Recovery phase.

1) **Neutral Phase:** The Neutral phase serves as the training phase, introducing participants to the MIST test and allowing them to familiarize themselves with the test's controls and answer format. Questions in this test have single-digit answers ranging from 0 to 9, and participants respond using a rotatory dial on the screen, controlled by the mouse's left and right buttons.
2) **Rest Phase:** During the rest phase, participants are instructed to sit comfortably in a relaxed position with minimal movements while a blank screen is displayed. This phase aims to establish a baseline physiological state for each participant.
3) **Controlled Phase:** In this phase, participants are presented with arithmetic questions of differing difficulty levels and no time constraints. They are encouraged to

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.2024.3463742

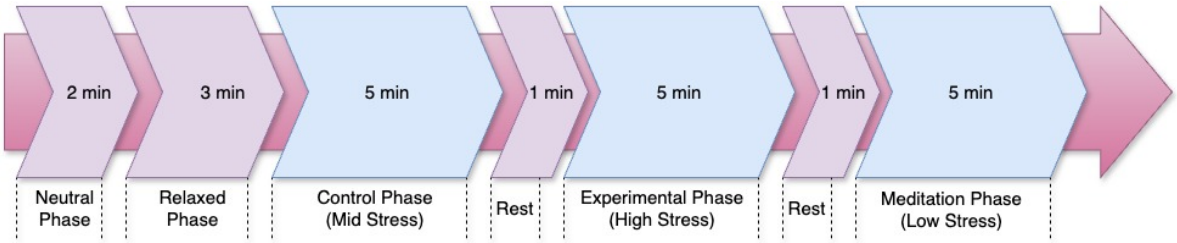


FIGURE 2. **Data Acquisition Procedure**

respond to as many questions as accurately as possible without receiving feedback. The study labels this phase as 'mild stress.'

4) **Experimental Phase:** The Experimental Phase presents adaptive arithmetic questions with an added element of social threat. Each question is timed, and participants are shown their current performance score alongside an artificially inflated average score of all participants. This artificial average score is intentionally set higher than the participant's achievable score, creating a sense of competition and heightened stress. Additionally, incorrect or untimely responses increase the time limit for the subsequent question, while three consecutive correct answers decrease the time limit for the next question. This phase effectively elicits greater stress in participants and is classified as 'high stress.'
5) **Recovery Phase:** The final phase assesses physiological changes as participants return to a relaxed state. Participants are engaged in a stress alleviation task involving listening to guided meditation audio with closed eyes while sitting comfortably. The study labels this phase as 'low stress'.

The current study includes a 2-minute training phase, followed by a 5-minute control section and a 5-minute experimental section. One-minute rest sections are interspersed between the control and experimental phases. The study concludes with a 5-minute recovery section for stress alleviation.

2) Empatica E4

This study continuously monitors physiological signals throughout the experiment using the Empatica E4 wristwatch. It consists of four sensors: a temperature sensor, an accelerometer, EDA sensors, and PPG sensors. Further, the PPG Sensor provides IBI and BVP signals. The collected data is transmitted via Bluetooth to a connected smartphone and monitored in real-time through the E4 Real-time app. Subsequently, this data is uploaded to Empatica's cloud platform, E4 Connect, for initial processing. The E4 device generates .csv files containing a variety of physiological measures, including skin temperature (ST), blood volume pressure (BVP), inter-beat-interval (IBI), EDA, heart rate (HR), and acceleration. Notably, the sampling rate for ST and EDA is 4 Hz, while the sampling rate for IBI, HR, and BVP is 64 Hz.

B. DATA PREPROCESSING

The primary objective of the current work is to explore physiological signals using the Empatica E4 and evaluate the accuracy of the proposed model using machine learning techniques. Even though the study collects the data inside the laboratory, its quality may be susceptible to various factors, such as sensor detachment, outliers, and missing information. Consequently, signal preprocessing is deemed necessary to improve the signal-to-noise ratio and safeguard valuable information from inadvertent filtering.

Labeling: Before the data preprocessing step, the data is clipped based on consecutive sessions. In three-level stress classification, the study labels the control phase as 'Mid Stress,' the experimental phase is labelled 'High Stress,' and the recovery phase as 'Low Stress.'

Signal Windowing: Signal windowing is a powerful technique to extract more detailed information from time-series data. This study implements a ten-second rolling window with a 50% overlap on the wristband data, maximizing the information extracted from the continuous data stream.

1. IBI (Interbeat Interval): The IBI.csv file consists of time intervals between successive heartbeats, and it is derived from the PPG Sensor. The data is preprocessed and analyzed using the *hrvanalysis* module in Python, which offers a comprehensive set of Heart Rate Variability (HRV) analysis tools [43]. These tools encompass preprocessing, R-peak detection, and feature extraction.

Data Cleaning: The IBI data is initially examined for missing values or outliers. These outliers are handled by employing interpolation techniques. Further, the IBI series may exhibit gaps due to discarded ectopic beats or undetected heartbeats, a common occurrence with wearable devices that can impact HRV analysis accuracy. These ectopic beats are addressed using the *"malik"* method [44].

2. EDA (Electrodermal Activity): As the five physiological signals are recorded at different sampling frequencies, it is essential to ensure that all the signals have the same sample size before analyzing [32]. The study employs upsampling, which increases the number of signal samples to address this sampling frequency discrepancy. Notably, among all the recorded features, the BVP and IBI are sampled at the highest frequency of 64 Hz. Consequently, this study upsamples the EDA signal from 4 to 64 Hz to standardize the sampling frequency for both signals. In addition, to eliminate potential

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.2024.3463742

TABLE 2. 58 Multi-domain features derived from EDA, IBI, and BVP signal

Feature Domain	Feature Description	Features	Feature ID
GSR-Statistical	Captures central tendencies and variability, providing important information about overall arousal levels and autonomic activity.	var_gsr, mean_gsr, kurtosis_gsr, skew_gsr, std_gsr	1-5
GSR-Tonic	Represents slowly varying baseline characteristics of the GSR signal, capturing continuous variations in skin conductance levels and reflecting an individual's overall arousal state and tonic sympathetic activity.	SCL_slope (tonic), var_scl, mean_scl, kurtosis_scl, skew_scl, slope_scl, std_scl	6-12
GSR-Phasic	Captures rapid and short-term fluctuations in GSR signal represent momentary responses to stimuli or events and reflect the phasic sympathetic activity.	scr_peaks (phasic), var_scr, mean_scr, kurtosis_scr, skew_scr, max_scr, std_scr	13-19
HRV-Time Domain	Statistical measures derived from the time intervals between consecutive heartbeats (RR intervals) in an HRV analysis.	median_nni, mean_nni, pnni_50, nni_50, nni_20, pnni_20, hrv_sdnn, hrv_rmssd, hrv_sdsd, range_nni, cvnni, cvsd, max_hr, mean_hr, std_hr, min_hr, tinn, triangular_index, cvi, sd1, csi, sd2, sampen, ratio_sd1_sd2, HR_max-HR_min	20-44
HRV-Frequency Domain	Captures the power distribution across different frequency bands, representing the relative contributions of sympathetic and parasympathetic activities in heart rate regulation.	total_power, lf, vlf, hf, lf_hf_ratio, hfnu, lfnu	45-51
BVP-Statistical Features	Numerical values that describe the BVP signal and its changes over time and frequency	Mean BVP, standard deviation, maximum, minimum, power VLF BVP, power LF BVP, power HF BVP	52-58

artefacts, the study applies a Savitzky-Golay [45] filter with a 40-point window and a sigma of 400 ms to smoothen the signal [46]. Subsequently, all the artefact-free segments underwent the feature extraction process.

3. BVP (Blood Volume Pressure):

The study derives the Blood Volume Pulse (BVP) signal from the PPG sensor by passing it through a high-pass filter. The specific cut-off frequency for this filter can vary, but it is typically set between 0.05 and 0.5 Hz. The data in this file represent BVP values calculated by a built-in algorithm, measured in nanowatts (nWatt) units. The BVP signals undergo filtering using a median filter, which involves scanning through the signal, point by point, and replacing each value with the median of neighbouring entries.

C. FEATURE EXTRACTION

The study retrieves numerical features from each data window in the feature extraction phase. Features are obtained by statistical functions and time and frequency analysis methods depending on the specific signal type.

The Empatica E4 device supplies IBI data, which is then employed to extract HRV features. The study gathers HRV features across various domains, including time, frequency, and non-linear, using the *hrvanalysis* Python module. In addition, statistical features are extracted from the BVP data using *numpy* module. The study employs the *PyEDA* package in the Python environment to extract phasic and tonic features from the processed EDA signal. Table 2 presents a total of 58 extracted features, encompassing HRV, EDA, and BVP, categorized by their respective domains. Additionally, Table 3 shows sample data example for 2-level stress classification.

D. FEATURE SELECTION

Feature selection is crucial in data analysis, particularly in clinical diagnosis systems [47]. Its primary goal is to identify and retain the most relevant features while discarding those that contribute minimally to the classification task. This process simplifies the data input, enhances model performance, and provides insights into the key factors influencing accurate stress classification. The primary advantage of the Feature Selection Algorithm (FSA) is reduced data complexity, improved model performance, enhanced interpretability, and reduced computational cost [48].

The success of feature selection relies on two critical factors: the search strategy and the selection criteria. Various methods are employed in different feature selection approaches to create subsets and drive the search processes. Extensive comparative studies have revealed that, especially in the case of large finite spaces, the search for the optimal solution becomes computationally demanding due to the exponential search space it entails. Consequently, most search strategies aim to identify sub-optimal or nearly optimal solutions. Recent research interest has gravitated towards global search algorithms, known as metaheuristic techniques.

The feature selection process's efficacy depends significantly on considering two key aspects: the search strategy and the criteria employed [49]. Various approaches to feature selection use distinct methods to generate subsets and advance the search processes. Comparative studies [50], [51], [52] have revealed that, particularly in large finite spaces,

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.2024.3463742

TABLE 3. Sample Data Example for 2-Level Stress Classification with Selected Features

GSR Statistical	GSR-Tonic		GSR Phasic	HRV Time domain			HRV Frequency domain	BVP Statistical	
mean_gsr	mean_scl	slope_scl	std_scr	RMSSD RR	Mean HR	Min HR	LF/HF	Mean BVP	Label
0.037	0.039	0.017	0.002	0.190	0.517	0.511	0.064	0.423	0
0.039	0.041	0.022	0.002	0.002	0.528	0.532	0.996	0.438	0
0.040	0.039	0.001	0.024	0.002	0.531	0.534	0.987	0.442	0
0.035	0.037	0.002	0.005	0.118	0.129	0.125	0.300	0.428	1
0.036	0.037	0.001	0.003	0.118	0.125	0.124	0.181	0.415	1
0.033	0.035	0.006	0.000	0.052	0.484	0.485	0.063	0.411	0
0.038	0.040	0.014	0.005	0.094	0.098	0.098	0.438	0.381	1
0.040	0.041	0.028	0.006	0.045	0.112	0.111	0.500	0.434	1
0.037	0.038	0.000	0.006	0.013	0.095	0.095	0.993	0.446	1
0.037	0.038	0.004	0.007	0.013	0.096	0.097	0.994	0.455	1
0.035	0.037	0.008	0.003	0.000	0.506	0.509	0.991	0.412	0

seeking the optimal solution is computationally demanding due to the resulting exponential search space. Consequently, most search strategies aim to identify sub-optimal or nearly optimal solutions [53]. Recent research has increasingly focused on global search algorithms, also known as metaheuristic algorithms. These strategies aim to find an optimal solution by exploring the entire space rather than limited feature spaces. Metaheuristic algorithms prove especially effective when dealing with uncertain and dynamic information. Hence, these advanced approaches efficiently manage the uncertainty and imprecision inherent in clinical data and medical images, particularly in biological predictive systems. Genetic Algorithm (GA) emerges as a natural choice for addressing feature selection challenges, and it is one of the most widely utilized optimization methods for navigating complex and non-linear search spaces [47].

1) Genetic Algorithm (GA)

GA represents a prominent branch of evolutionary computing and is a rapidly expanding field within Artificial Intelligence (AI). It is worth noting that the evolutionary computation community does not universally approve of a precise definition of "Genetic Algorithms" [54]. Nevertheless, GAs can be described as adaptive heuristic search algorithms, originally conceived by Holland in the 1960s, drawing inspiration from Darwin's theory of evolution, specifically, the concept of "survival of the fittest" [55]. Five fundamental factors can influence the execution of GAs: population creation, fitness function, selection schema, genetic operators, and stopping criteria. The algorithm commences with an initial population of binary strings known as chromosomes, potential candidates for solving optimization problems. In each iteration, these populations undergo evaluation based on their fitness, followed by applying crossover and mutation genetic operators to favour more robust solutions. Typically, candidate solutions are encoded within binary strings composed of 0s and 1s. In this representation, 0 denotes that a particular feature has not been selected, while 1 signifies the inclusion of the corresponding feature.

A study by Ramdania et al. [56] presents compelling evi-

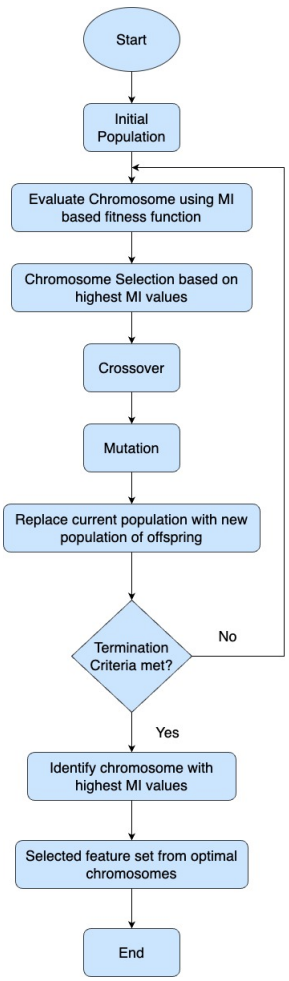


FIGURE 3. GA-MI Feature Selection Algorithm

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.2024.3463742

dence that Genetic Algorithms excelled over alternative methods in efficiently identifying nearly optimal features within vast datasets. Building on this, Rostami et al. [57] introduced a hybrid algorithm designed to enhance solutions nearest to those identified by the Genetic Algorithm. In another study, a comparative analysis of algorithms for feature selection in pattern recognition was carried out [58]. The findings concluded that Genetic Algorithms excel particularly in handling large-scale problems.

2) Mutual Information (MI)

In research, there is a need to carefully consider the criteria used to evaluate classification performance to perform effective feature selection. Feature selection methods rely on evaluation criteria and search strategies to identify the optimal feature set. MI has gained significant attention in this context because it excels at measuring the correlation between features and class labels, and it is less sensitive to noise or outliers than other methods [59]. MI-based feature selection algorithms aim to select the best subset of features by maximizing the joint MI between input features and the target output. However, estimating high-dimensional MI is complex and computationally demanding, limiting its practical use [60]. As a result, many methods opt for low-dimensional MI. Over time, numerous MI-based feature selection algorithms have been developed in this field.

3) Genetic Algorithm with Mutual Information (GA-MI)

The study elevates the MI-based feature selection algorithm by integrating a global search algorithm for stress classification. Figure 3 shows the core operation of the Genetic Algorithm incorporating Mutual Information in the context of the physiology-based stress dataset. Further, Table 4 summarizes the parameters used for implementing the Genetic Algorithm FSA. The detailed procedure is outlined as follows:

1. *Create initial population:* Define a population size p, representing the number of chromosomes in the initial population. Each chromosome c_i is a binary string of length n, where n is the total number of features in the physiology-based stress dataset. Each bit c_{ij} in the string corresponds to a specific feature f_j. For each bit c_{ij} in the chromosome c_i, randomly set its value to 0 or 1. This random initialization ensures that the initial population contains diverse potential solutions.

2. *Compute Mutual Information:* MI quantifies the statistical dependence between two random variables. In this context, the second step measures the relevance of a set of features, represented by a chromosome c_i, to the target class y. A higher MI value indicates a stronger association between the features and the class. Given a chromosome c_i, let X_{c_i} denote the set of features selected by c_i (i.e., features corresponding to bits set to 1 in c_i. The mutual information MI (c_i,y) is calculated as follows:

$$MI(c_i, y) = \sum_{x \epsilon c_i} \sum_{y \epsilon Y} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \tag{1}$$

where p(x,y) denotes the joint probability distribution of feature values $x \in X_{ci}$ and class labels y, p(x) represents the marginal probability distribution of feature values $x \in X_{ci}$, and p(y) represents the marginal probability distribution of class labels y. As a result, the highest MI value chromosomes are used in the next step.

3. *Fitness-Proportional Selection:* The selection probability $P(c_i)$ for each chromosome c_i is directly proportional to its mutual information $MI(c_i, y)$. It implies that chromosomes with higher MI values are more likely to be selected for reproduction, ensuring that the algorithm prioritizes features with greater relevance to the target class.

$$P(c_i) = \frac{MI(c_i, y)}{\sum_{j=1}^{p} MI(c_j, y)} \tag{2}$$

Where $MI(c_i, y)$ represents the mutual information of chromosome c_i, and the denominator represents the sum of mutual information values for all chromosomes in the population.

4. *Crossover and Mutation:* The crossover operator mimics genetic recombination by combining genetic information from two selected chromosomes (parents) to generate new offspring chromosomes. This exchange of information can potentially lead to the discovery of superior feature combinations. The mutation operator introduces random alterations to individual bits in the offspring chromosomes, potentially introducing new features or removing irrelevant ones. This mechanism helps the algorithm escape local optima and explore broader solutions.

5. *Iteration and Termination:* The algorithm iteratively executes steps 2 to 4, evolving the population over generations. The termination condition reaches the maximum number of generations G_{max}, typically determined experimentally based on the problem complexity and desired convergence.

6. *Optimal Solution:* The optimal feature c_{best} is identified as the chromosome with the highest mutual information $MI(c_{best}, y)$ in the final generation. It represents the features most relevant to the target class and is expected to yield the best classification performance.

E. ANALYSIS OF VARIANCE (ANOVA)

To further investigate the significance of the selected features, the study proceeded with an analysis using Analysis of Variance (ANOVA) test. ANOVA compares the differences in means among groups by examining the variance within and between groups, helping to determine if there are significant variations. This method assumes that observations are independent, normally distributed, and have equal variances among groups. It produces an F-statistic to assess significance of group differences. In this study, ANOVA is used to evaluate the distinctiveness of the selected feature subset by calculating degrees of freedom, p-values, and F-values for stress and non-stress classes. These statistical test parameters and the F-value are computed using following equations.

$$F = \frac{MS_B}{MS_w} \tag{3}$$

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.2024.3463742

TABLE 4. Parameters for Implementing Genetic Algorithm

Parameter	Description	Value
Initial Population	Number of chromosomes in the initial population	100
Evaluation Criteria	Metric used to assess chromosome fitness	Mutual Information (MI)
Reproduction Operators	Genetic operators for generating new offspring	Single-point crossover, mutation rate: 0.05
Crossover Rate	Probability of crossover occurring between chromosomes	0.8
Selection Criteria	Method for selecting chromosomes based on fitness	Fitness-proportional selection with elitism
Termination Criteria	Condition for stopping the GA iterations	100 generations or convergence

$$df_B = k - 1 \tag{4}$$

$$df_w = N - 1 \tag{5}$$

$$MS_B = \frac{Sumofsquarebetweensample}{df_B} \tag{6}$$

$$MS_w = \frac{Sumofsquarewithinsample}{df_w} \tag{7}$$

The p-value and the F-value have an inverse relationship – when one is low, the other tends to be high. Specifically, when the p-value is less than 0.05, it implies that each feature contributes individual and valuable information to the results.

F. CLASSIFICATION ALGORITHM

After feature selection, the data is divided into a ratio of 80% for training and 20% for testing. Further, the study employs five classification algorithms: Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K Nearest Neighbour (KNN), Xgboost (XGB), and Gradient Boosting (GB).

1) Bayesian Hyperparameter Optimization

Classification algorithms typically come with parameters that significantly impact their performance. These algorithms involve multiple hyperparameters that play a crucial role in determining the accuracy of the models. Therefore, it is vital to carefully fine-tune these hyperparameters, a process known as hyperparameter optimization. Traditionally, hyperparameter tuning has been viewed as more of an art than a precise science, relying on subjective judgments and iterative experimentation, which can be time-consuming [61]. While grid search represents a common hyperparameter optimization approach, it is computationally expensive and time-consuming, especially for problems with numerous hyperparameters [62] [63].

Bayesian optimization emerges as a systematic and efficient alternative, offering a more sophisticated solution to this challenge [64]. As evidenced by various studies [62] [65] [40], Bayesian optimization has demonstrated superior performance in various global optimization problems, particularly at hyperparameter tuning. It aligns with pursuing a more efficient and effective avenue for hyperparameter tuning in scenarios involving multiple hyperparameters. Therefore, the study leverages Bayesian hyperparameter optimization to fine-tune the hyperparameters of the classification models for stress classification. This approach allows the search for optimal hyperparameters to be systematic and can improve model performance.

TABLE 5. 2 Level stress: ANOVA test for EDA+IBI+BVP feature subset obtained using GA+MI feature selection algorithm

Features	dfb	dfw	F-Value	P-Value
Min HR	1	3764	2.77E+03	0.00E+00
mean_gsr	1	3764	1.53E+03	0.00e+00
Mean HR	1	3764	2.77E+03	0.00e+00
LF/HF	1	3764	7.45E+00	1.00E-14
RMSSD RR	1	3764	4.46E+01	3.05E-03
mean_scl	1	3764	2.77E+03	4.25E-04
slope_scl	1	3764	6.04E+01	6.36E-03
std_scr	1	3764	2.04E+01	6.51E-01
Mean BVP	1	3764	6.79E+00	9.69E-01

$*df_b$: degrees of freedom between classes, df_w: degrees of freedom within classes, p-value<0.05 indicate distinctive feature classes

TABLE 6. 3 Level stress: ANOVA test for EDA+IBI+BVP feature subset obtained using GA+MI feature selection algorithm

Features	dfb	dfw	F-Value	P-Value
Mean HR	2	5874	6.73E+04	0.00E+00
Min HR	2	5874	5.74E+03	0.00E+00
mean_gsr	2	5874	5.79E+03	0.00E+00
mean scl	2	5874	3.59E+03	3.10E-23
Power HF RR	2	5874	3.62E+03	4.42E-13
LF/HF	2	5874	3.59E+03	2.10E-04
RMSSD RR	2	5874	2.11E+03	1.25E-03
Mean BVP	2	5874	1.25E+03	3.90E-11
max_scr	2	5874	9.43E+02	3.08E-03
Slope scl	2	5874	1.25E+03	1.20E-05
STD RR/SDNN	2	5874	5.23E+01	2.87E-02

$*df_b$: degrees of freedom between classes, df_w: degrees of freedom within classes, p-value<0.05 indicate distinctive feature classes

III. EXPERIMENTATION AND RESULTS

A. PRELIMINARY ANALYSIS AND OPTIMIZATION

1. Statistical Feature Analysis: The study uses ANOVA to analyze the uniqueness of the optimal feature subset obtained during GA+MI feature selection. Tables 5 and 6 show the ANOVA results for binary and three-level stress classification, respectively.

Table 5 and Table 6 analysis reveal many features with statistically significant effects (p-value < 0.05) on differentiating between stress levels. Notably, features such as Min

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.2024.3463742

TABLE 7. List of hyperparameters used for stress classification

Classifier	Hyperparameter	Value Range	Selected Value
RF	n_estimators	[10, 40, 60, 80, 100, 150, 200]	200
	max_depth	[2, 4, 8, 10, 12, 14]	4
	min_samples_split	[3, 5, 10, 15]	3
	random_state		0
DT	max_depth	[2, 4, 8, 10, 12, 14]	8
	min_samples_split	[2, 5, 10, 15, 20, 25]	10
KNN	n_neighbors	[3, 4, 5, 6, 7, 9, 12, 14, 16]	6
	metric	["euclidean", "manhattan", "minkowski"]	"minkowski"
	weights	["uniform", "distance"]	"distance"
	C	[0.1, 1, 10]	1
SVM	kernel	["linear", "rbf"]	"rbf"
	gamma	[0.01, 0.001, 0.0001, scale, auto]	0.1
XGB	n_estimators	[50, 100, 200, 500, 700]	500
	learning_rate	[0.01, 0.001, 0.1, 0.2]	0.1
	max_depth	[2, 4, 8, 10, 15]	8
GB	n_estimators	[100, 200, 500, 700, 900, 1000]	200
	learning_rate	[0.01, 0.001, 0.1, 0.2]	0.1
	max_depth	[2, 4, 8, 10, 12, 14]	4

HR, mean_gsr, and Mean HR consistently emerged as significant across both 2-level and 3-level classifications. Their high F-values and low p-values further support their strong influence on the model's performance. Additionally, Table 6 shows other features with significant effects, including LF/HF, Power HF RR, and STD RR/SDNN.

2. Performance Indexes: The performance of machine learning classifiers in stress classification systems can be assessed using a statistical approach called the "confusion matrix. This method helps determine how well the classifiers are doing. This work measures the performance using key metrics, including accuracy, F1 score, and AUC commonly used in evaluating machine learning models. This study uses positive and negative samples from the confusion matrix to evaluate the stress detection system's results. This matrix helps to calculate accuracy and F1 score using equation 8 and 9.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \tag{8}$$

$$F1Score = \frac{2TP}{2TP + FP + FN} \tag{9}$$

Additionally, the study calculates the model's global performance by measuring the Area Under the Curve (AUC). A value of 0.5 indicates a model with random performance, while 1 signifies flawless performance.

3. Hyperparameter Tuning: To optimize the performance of the classifiers, Bayesian hyperparameter optimization techniques are utilized to fine-tune hyperparameters. Table 7 lists the optimal hyperparameters utilized in the study. However, a thorough experimentation process is conducted to ensure the reliability and consistency of the computational results. This involves determining which hyperparameters are best left with their default values and which require tuning, employing the Bayesian technique.

B. CLASSIFICATION RESULTS

Prior research has yet to investigate the integration of ECG and EDA in the academic environment within the context of the MIST. Therefore, the current study explores the potential of wearable sensors, including IBI, BVP, and EDA biomarkers, in facilitating a comprehensive evaluation of stress responses.

The stress classification framework of the study primarily centers on identifying distinctive features from a complex, high-dimensional dataset through the application of the GA+MI Feature Selection Algorithm. Subsequently, several machine learning classifiers, namely RF, DT, SVM, KNN, XGBoost, and GB, are employed to classify these features for stress classification. Further, the study validates the computational results' reliability and stability through a ten-fold cross-validation method to calculate the average performance metrics.

Table 8, 9, 10 and 11 present the experimental results of 2-level and 3-level stress classification models, with and without the incorporation of the feature selection method. These results provide valuable insights into the efficacy of these models in accurately identifying different stress levels.

In 2-level stress classification, comparing individual physiological signals presents compelling results. Table 8 indicates that EDA and BVP features exhibit comparable performance without feature selection. Remarkably, IBI-derived HRV features exhibit the highest performance, achieving an accuracy of 86.54%.

Moreover, Table 9 presents the results after employing the GA+MI FSA. The analysis reveals a 3.49% increment in accuracy for EDA and a 1.77% increment for BVP. Further,

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.2024.3463742

TABLE 8. Performance comparison of distinct physiological signals for 2-level stress classification with 58 original feature set

	RF			DT			SVM			KNN			XGB			GB		
Signals	**Acc**	F1	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	AUC	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**
EDA	76.4	77.69	84.4	76.32	75.61	78.3	55.7	55.69	60.2	60.15	59.57	69.67	78.93	77.91	87.3	**85.12**	85.64	94.63
IBI	83.84	83.62	89.42	80.37	79.11	83.65	78.72	78.88	86.52	80.5	80.33	83.98	82.92	81.4	86.83	**86.54**	86	93.02
BVP	83.91	83.92	88.54	78.61	78.75	78.37	54.7	55.03	64.29	78.83	78.73	86.66	82.88	83.97	88.29	**85.53**	85.91	91.4
EDA+IBI	93.35	92.54	97.03	90.22	90.47	95.03	83.96	83.5	87.1	83.5	81.45	87.4	92.94	92.98	96.7	**94.97**	95.35	97.86
IBI+BVP	89.76	89.67	95.02	86.22	87.14	92.64	79.71	80.78	84.98	83.88	83.75	89.75	90	90.1	95.69	**92.02**	92.56	96.94
EDA+BVP	88.3	88.13	93.76	87.37	87.35	91.54	62.15	62.03	67.7	80.91	82.55	87.32	88.21	88.16	94.72	**91.93**	91.91	97.38
EDA+IBI+BVP	93.37	92.43	97.83	90.54	90.87	94.15	84.54	83.21	89.26	84.59	85.63	90.41	92.23	91.67	97.36	**95.67**	95.73	98.59

TABLE 9. Performance comparison of distinct physiological signals for 2 level stress classification selected 9 features using GA-MI FSA

	RF			DT			SVM			KNN			XGB			GB		
Signals	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**
EDA	81.89	80.6	88.86	78.3	76.98	81.92	55.93	54.07	60.58	69.85	66.8	72.81	80.27	80.69	87.71	**88.61**	88.5	94.02
IBI	85.1	85.32	92.76	78.8	78.2	80.55	80.7	80.73	89.05	82.8	82.79	90.93	84.41	84.67	90.11	**90.86**	90.71	97.07
BVP	85.2	85.18	92.77	79	79.48	82.24	55.7	55.68	64.96	82.9	82.91	89.48	85.12	85.3	92.78	**87.3**	88.9	93.94
EDA+IBI	94.66	94.76	98.82	92.98	92.94	97.32	84.04	84.14	89.83	90.21	91.24	96.34	95.3	95.2	98.27	**97.07**	97.05	99.03
IBI+BVP	91.87	91.67	96.53	88.07	88.34	94.84	82.63	82.68	88.03	85.15	86.94	90.04	91.86	91.87	97.18	**94.83**	95.26	98.21
EDA+BVP	89.21	89.8	94.44	87.37	88.66	93.23	64.13	64.25	70.2	82.91	83.28	88.74	90.15	90.15	96.03	**93.42**	93.98	97.72
EDA+IBI+BVP	95.5	95.08	99.83	93.84	93.71	97.53	84.54	84	90.44	91.15	91.89	95.53	95.66	95.65	98.64	**98.28**	97.44	99.78

TABLE 10. Performance comparison of distinct physiological signals for 3-level stress classification with 58 original feature set

	RF			DT			SVM			KNN			XGB			GB		
Signals	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**
EDA	68.39	67.87	73.8	64.69	65.34	72.56	52.57	52.89	59.65	61.85	63.73	68.23	73.17	73.12	79.11	**75.01**	75.04	80.23
IBI	76.52	77.55	84.53	73.48	74.10	80.79	72.36	72.93	77.78	74.56	74.40	80.24	77.08	77.23	84.17	**79.74**	80.12	87.31
BVP	74.42	74.53	81.4	69.61	69.73	75.68	59.37	60.85	68.74	70.38	71.37	78.74	76.02	76.47	82.01	**79.70**	79.20	86.09
EDA+IBI	91.91	92.90	97.09	89.49	89.41	95.00	75.04	75.78	82.32	70.28	70.90	78.78	91.49	91.47	96.43	**93.87**	93.91	97.74
IBI+BVP	89.51	89.90	95.93	83.52	83.89	89.31	73.66	73.44	79.09	71.82	71.90	78.23	89.54	89.61	94.57	**91.92**	91.93	97.01
EDA+BVP	85.42	85.38	92.61	78.29	78.3	84.27	59.51	59.67	67.78	64.97	64.43	72.34	86.81	86.83	92.17	**89.43**	89.66	95.28
EDA+IBI+BVP	93.26	94.76	97.15	91.54	92.33	97.7	79.17	79.8	85.45	80.53	81.92	87.78	93.29	93.2	97.64	**95.16**	96.2	98.8

TABLE 11. Performance comparison of distinct physiological signals for 3-level stress classification with selected 11 features using GA-MI FSA

	RF			DT			SVM			KNN			XGB			GB		
Signals	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**
EDA	71.6	70.32	77.98	67.07	68.34	73.3	52.52	52.78	59.78	67.13	67.42	72.26	75.17	75.27	81.56	**79.23**	79.98	87.76
IBI	79.93	79.27	89.48	75.88	75.56	84.84	75.26	75.08	83.1	77.3	77.63	86.04	80.04	80.41	87.72	**82.93**	82.91	89.48
BVP	77.15	77.13	83.17	70.83	70.85	77.66	64.31	64.18	72.67	72.23	73.01	80.34	78.63	78.85	85.3	**82.45**	83.46	88.11
EDA+IBI	93.12	94.23	98.21	91.76	91.98	97.78	76.68	76.23	83.32	73.07	73.87	79.77	93.16	93.28	97.16	**95.23**	96.05	98.47
IBI+BVP	90.43	90.35	96.95	84.98	84.65	92.02	75.11	75.74	81.54	73.65	73.25	79.67	90.06	90.4	95.82	**92.78**	92.24	97.63
EDA+BVP	88.48	88.35	93.7	80.62	80.57	86.45	62.98	62.62	68.96	77.77	77.92	84.21	89.03	89.02	95.6	**91.45**	91.42	96.42
EDA+IBI+BVP	95.65	96.56	99.52	93.95	94.83	98.4	80.61	80.67	87.67	89.16	90.57	95.42	95.66	95.75	98.13	**97.02**	96.53	99.13

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.2024.3463742

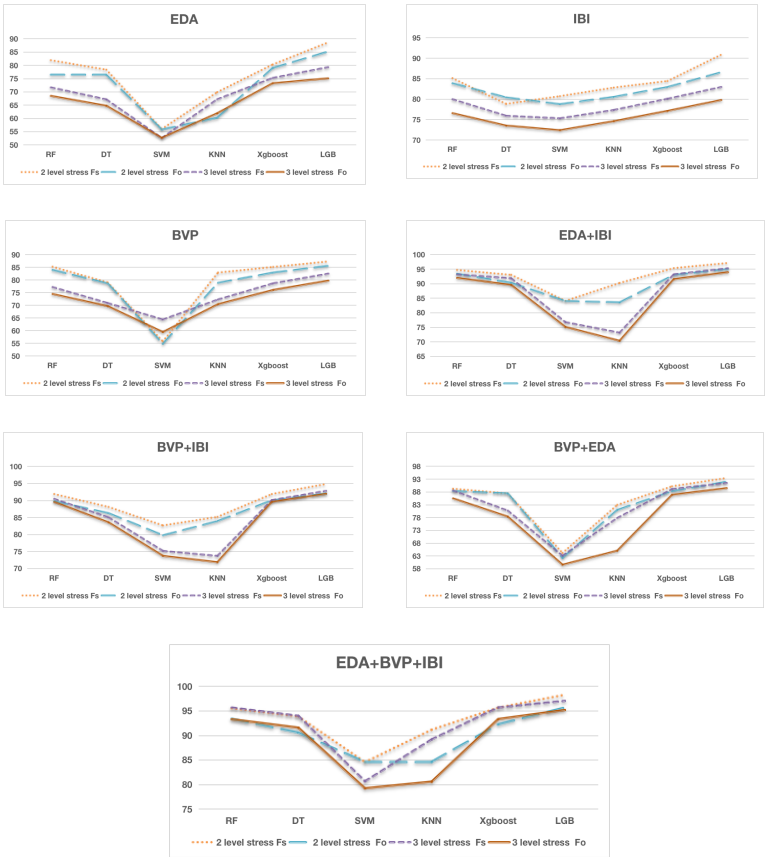


FIGURE 4. **Comparison of test accuracy of different classification models for 2-level and 3-level stress classification for the original feature set (F_o) and the selected feature set (F_s)**

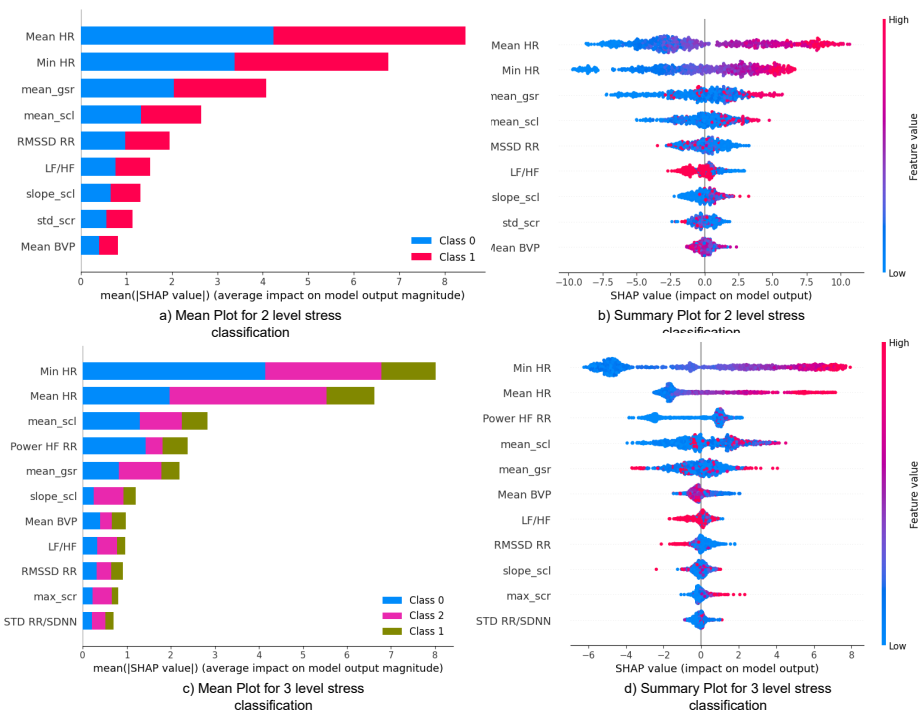


FIGURE 5. **Explaination of SHAP values**

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.2024.3463742

the accuracy and F1 score of IBI-derived HRV features show significant improvements after feature selection, with an increment of 4.32% and 4.71%, respectively.

The combination of EDA and HRV features provides an accuracy of 94.97% without feature selection and reaches the highest accuracy of 97.07% with the feature selection algorithm. In contrast, incorporating all three features (EDA+HRV+BVP) results in a modest 2.68% accuracy increase, highlighting the substantial effectiveness of EDA and IBI signals in stress classification. Additionally, the study observes that RF, DT, and KNN demonstrate comparable performance, while XGB and GB consistently outperform other models with or without the feature selection algorithm GA+MI.

Similarly, Table 10 and 11 present the results before and after employing the feature selection method, showcasing enhanced results in the three-level stress classification task. The three-level classification models introduce an additional layer of complexity by distinguishing low, medium, and high stress levels. While individual physiological signals such as EDA and IBI display moderate performance, the combination of EDA and IBI outperforms combinations that involve BVP. However, using the GB classifier, the inclusion of BVP consistently achieves the highest accuracy, f1 score, and AUC score of 97.02%, 96.53%, and 97.47%, respectively.

The findings indicate that incorporating feature selection, particularly the GA+MI FSA, significantly enhances classification accuracy. This highlights the impact of eliminating redundant features from the original feature set, which has been found to have a negative impact on the model's performance. A visual representation of the feature selection comparison is shown in Figure 4, contrasting the outcomes between the original feature set, F_o, and the selected feature set, F_s, for both 2-level and 3-level stress classification.

C. EXPLAINABLE AI

Figure 5a and 5c depict bar plots presenting SHAP (Shapley Additive Explanations) values, offering a visual representation of the one-to-one scenario for feature importance in 2-level and 3-level stress classification, respectively. Feature importance is determined by calculating the mean absolute value for each feature across all provided samples. Figure 5b and Figure 5d illustrate summary plots of SHAP values for the top 9 features in the case of 2-level stress classification and the top 11 features for 3-level stress classification, utilizing the entire Stress dataset.

The summary plots illustrated in Figure 5b and Figure 5d show the impact and contribution of a specific feature to the stress classification model. The horizontal axis represents the SHAP values and their connection to high or low-stress predictions. The vertical axis above the zero point implies minimal impact on prediction results. For example, a SHAP value of zero suggests that the feature has no significant influence on predictions, while values close to zero indicate less influential predictions. In contrast, high-quality predictions, characterized by SHAP values significantly distant from zero, indicate positive or negative correlations. The figures also display how each feature impacts model output for varying stress levels. Each row in these plots corresponds to an individual feature, ordered by their mean absolute SHAP values. These values can be either negative or positive, reflecting the significance of each feature. These dots are stacked to reveal density and are colour-coded to indicate the intensity, with red signifying high feature value and blue indicating low feature value.

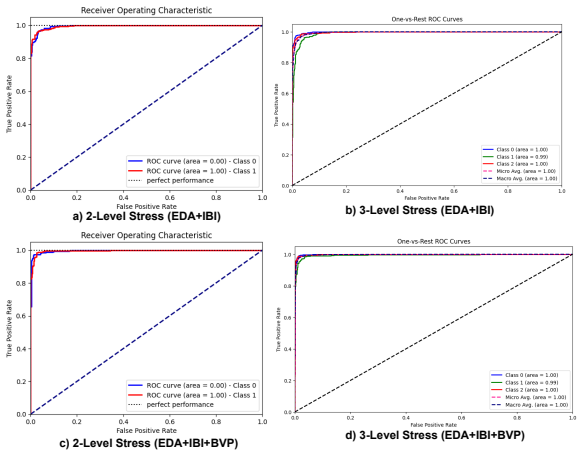


FIGURE 6. ROC of the best-evaluated results for integrated signals using Gradient Boosting

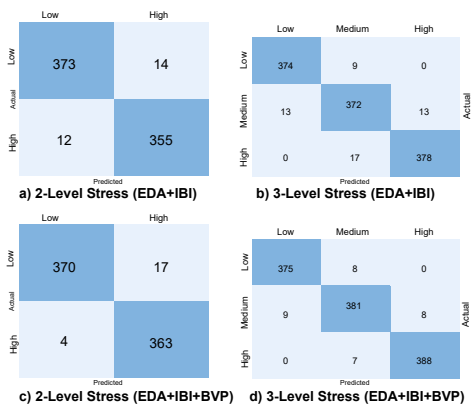


FIGURE 7. Confusion Matrix of the best-evaluated results for integrated signals using Gradient Boosting

Furthermore, Figure 5 shows that mean HR, Min HR, and mean_gsr are identified as highly important features for stress classification. This observation is consistent with the results of the ANOVA statistical analysis presented in Table 5 and 6. The low p-values and high F-score values in the ANOVA analysis support the conclusion regarding the significance of these features in predicting stress levels, providing a robust and reliable foundation for stress prediction.

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.2024.3463742

IV. DISCUSSION

The experimental results underscore the efficacy of EDA and IBI as robust biomarkers for stress classification when employing the Gradient Boosting (GB) algorithm. In the context of two-level stress classification, the combination of EDA and IBI achieved a commendable accuracy of 97.07%. The subsequent addition of BVP features to the model (EDA+IBI+BVP) led to a marginal increase in accuracy, reaching 98.28%. This observation is reinforced by the Receiver Operating Characteristic (ROC) curves and confusion matrices presented in Figure 6 and 7, respectively. The figures depict comparable true positive (TP) and true negative (TN) rates and high Area Under the Curve (AUC) values, affirming the model's discriminatory power. Expanding the stress categorization to three levels incorporating BVP features further refined the model's performance, achieving accuracies and F1 scores of 97.02% and 96.53%, respectively. The confusion matrices in Figure 7 demonstrate the model's proficiency in distinguishing between the three stress levels, corroborated by the ROC curves in Figure 6, which exhibit a high AUC for each stress category. Notably, the EDA+IBI combination exhibited a comparable accuracy of 95.23% and an F1 score of 96.06%.

These results highlight the effectiveness of EDA and IBI in distinguishing between stress and non-stress levels. The marginal difference in accuracy observed upon including BVP features implies that IBI is derived from BVP and may contribute redundant information, resulting in only marginal improvement. It reinforces the notion that combining EDA and IBI can comprehensively capture a diverse range of physiological responses to stress, thereby ensuring a robust and reliable performance in stress classification.

Further, the study found that gradient boosting achieves better classification than XGBoost. XGBoost, known for its effectiveness in handling high-dimensional data with numerous features, might not be fully optimized when applied to datasets with limited features, as observed in this study. With a relatively smaller feature set, XGBoost might need a greater breadth of information to perform optimally. Therefore, gradient boosting, which is less sensitive to the number of features, emerged as a more suitable choice for this study.

This paper consists of stress detection and stress alleviation stages. The overall system shows an accuracy of 95.8% for stress classification. During the stress recovery phase, individuals listened to meditation audio, and the results indicate that this stage effectively aids in stress reduction.

The experimental findings underscore the significance of IBI (Interbeat Interval) and EDA (Electrodermal Activity) as primary contributors to stress classification. Figure 8 illustrates the Fast Fourier Transform (FFT) spectrum, obtained through Kubios analysis, displaying various stress levels. The FFT scrutinizes the HRV (Heart Rate Variability) signal derived from IBI. HRV serves as a measure of the time intervals between successive heartbeats. These intervals are influenced by the Autonomic Nervous System (ANS), which governs numerous automatic bodily functions, including heart rate, blood pressure, and respiration.

The ANS is divided into two distinct systems: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). SNS is responsible for the "fight-or-flight" response, while PNS is responsible for the "rest-and-digest" response.

Figure 8 a) and 8 b) highlight a prominent peak in the low-frequency (LF) range associated with the activity of the sympathetic nervous system during moderate and high-stress situations. It also reveals a smaller peak in the high-frequency (HF) range, linked to parasympathetic nervous system activity. The figure indicates that even in the presence of stress, some parasympathetic activity continues, which may aid in alleviating the negative impacts of stress. Conversely, Figure 8 c) highlights a strong peak in the high-frequency (HF) range during low-stress situations, indicating increased activity in the parasympathetic nervous system.

Figure 9 shows the EDA graph to analyze EDA's phasic and tonic components. Several factors, including stress, arousal, and sweat, influence skin conductance. Phasic EDA refers to the quick changes in skin conductance that occur in response to specific stimuli or events. These changes are typically short-lived and are often associated with emotional arousal. Tonic EDA refers to the slower, more sustained changes in skin conductance that occur over time. Tonic EDA is thought to reflect the overall level of arousal and physiological activation.

The EDA graph is used to identify different stress states by analyzing EDA's phasic and tonic components. High stress shows a high concentration of phasic EDA responses and a high tonic EDA level. Moderate stress shows moderate phasic EDA responses and a moderate tonic EDA level. The figure shows the fall in phasic and tonic EDA levels, gradually decreasing EDA peaks for low-level stress.

A. COMPARISON WITH EXISTING WORK

The proposed work comprises stress detection and alleviation phases. The overall system achieves 98.28% and 97.02% accuracy for binary and 3-level stress classification. In the stress alleviation stage, participants listen to meditation audio. The study employs Explainable AI to identify influential features in stress detection.

Additionally, it investigates the impact of EDA and ECG on stress detection, analyzing the effects of stress and stress alleviation on the sympathetic and parasympathetic branches of ANS. The findings highlight the effectiveness of meditation audio in reducing stress within an academic setting.

Table 12 compares the previously proposed MIST studies with stress alleviation methods. It summarizes physiological signals, number of channels, subjects, stress alleviation technique, experiment duration, selected features, and results. Very few studies offer a solution for stress alleviation with MIST. Zhang et al. [20] classified stress levels with frequency domain features and employed the Fisher Ratio (FR) for feature reduction. The authors investigated distinct stress relief therapies such as watching and listening to alpha music,

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.2024.3463742

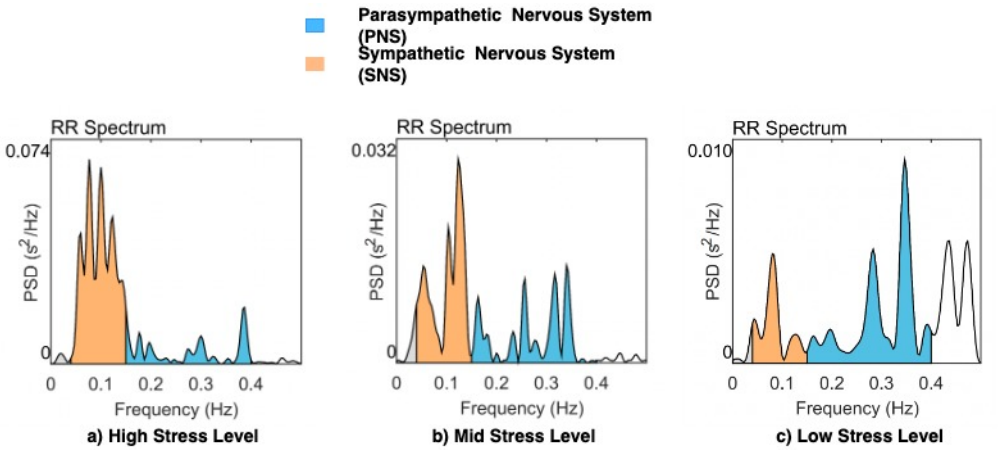


FIGURE 8. **Fast Fourier Transform Spectrum of distinct stress levels derived from IBI data and visualized using Kubios software: a) High-stress level b) Mid-stress level c) Low-stress level**

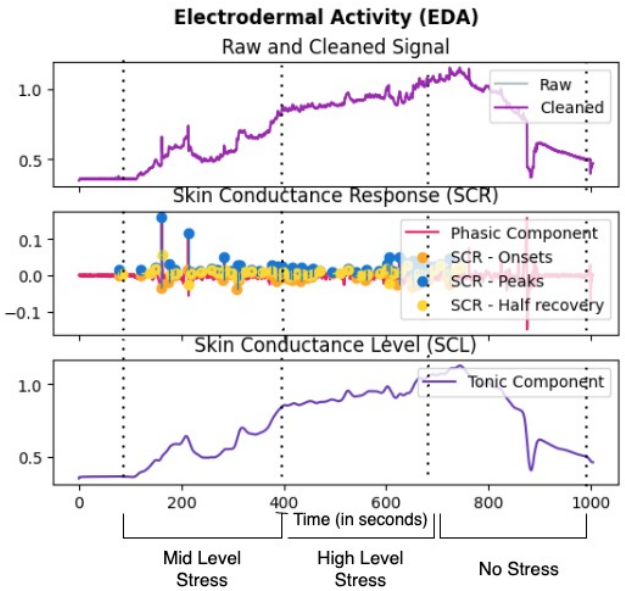


FIGURE 9. **EDA graph to analyze phasic and tonic components of EDA during stress classification**

videos of walking through a forest, bubble wrap popping, hugging a pillow, and squeezing a stress ball. The study found that hugging a pillow is the most effective method to mitigate stress, and the proposed stress alleviation technique works by comparing corresponding stress levels. Moreover, the study reported accuracy of 85.6 and 71.5 for 2-level stress and 3-level stress classification. Xia et al. [23] utilized EEG and ECG for stress detection and Fisher's discriminant ratio (FDR), minimum redundancy maximum relevance (mRmR), and information gain feature selection technique, achieving the highest accuracy of 79.5%. When the stress is induced via MIST, Sitting in a relaxed position and focusing on a circle on a computer screen is proposed for therapy. Moreover, the final selected features are not mentioned. In another study, Han et al. [25] utilized ECG and RESP and found sitting in a relaxed position to be stress-relieving therapy. However, sitting in an ideal relaxed position therapy can be influenced by environmental factors. External distractions or discomfort in the selected environment for therapy may hinder its efficacy. Sitting in a relaxed position is effective for the short term as some may find it difficult to induce stress or recall stressful situations genuinely.

Further, Setz et al. [24] proposed a general questionnaire and reading magazines as a stress relief method. However, reading a magazine for stress relief may vary in effectiveness due to individual preferences and the passive nature of the activity, which may only engage some equally. A quantitative evaluation of the effectiveness of the stress alleviation method needs to be presented in these studies.

The proposed work overcomes the limitations of the above studies. It employs an integrated feature selection algorithm, selecting only nine features and eleven features for 2-level- and 3-level stress classification. Further, it utilizes explainable AI to provide transparency in analysis. The detailed results show the significance of HRV and EDA features for stress detection systems. It also shows high classification accuracy using Bayesian optimization. Additionally, this study provides evidence that listening to meditation audio reduces stress levels, highlighting the effectiveness of such interventions in stress management.

V. CONCLUSION

Mental stress in academic environments adversely impacts students' well-being and academic performance. Through the exploration of wearable devices, particularly the utilization of IBI-derived HRV and electrodermal activity (EDA), the research demonstrates the potential for real-time stress detection and monitoring. The findings underscore the effectiveness of wearable technology in identifying stress among students, offering opportunities for timely interventions to alleviate its negative effects. Moreover, this paper has delved into the efficacy of stress reduction techniques, specifically the impact of listening to meditation audio. The results in-

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.2024.3463742

TABLE 12. Performance comparison of proposed MIST studies with stress alleviation approach in the literature

Author	Signals	Channels	Sub	Stress Alleviation Technique	Experiment Duration	Feature Selection	Features	Classifiers	Accuracy	F1 Score	Other Metrics
Zhang et al. [20]	EEG	4	25	Watching a video, listening to alpha music, squeezing a stress ball, bubble wrap popping, and hugging a pillow.	16 min approx	Fisher Ratio	4	LDA	2 level: 85.6%, 3 level: 71.5%	–	–
Xia et al. [23]	EEG, ECG	128	22	Sitting in a relaxed position and focusing on a circle appearing on a computer screen	80 min	FDR, mRMR, Information Gain	–	SVM	79.45%	–	Recall: 81%, Specificity: 78%
Setz et al. [24]	EDA	–	32	Questionnaire and reading magazines	50 min	Not mentioned	16	LDA, SVM, NCC	82.80%	–	–
Han et al. [25]	ECG, RSP	–	39	Sitting in a relaxed position	18 minutes (excluding stress report filling time)	OOB	36	SVM, KNN, LDA, Adaboost	84%	83%	Recall: 84, Precision: 83
Proposed Work	ECG, EDA	–	30	Listening to meditation audio	22 min	GA + MI	9	RF, DT, SVM, KNN, GB, XGB	2 levels: 98.28%, 3 levels: 97.02%	2 levels: 97.44%, 3 levels: 96.53%	AUC: 99.78, AUC: 99.13

dicate that such practices can be valuable tools for stress reduction, providing students with accessible means to manage their stress levels. By incorporating a hybrid feature selection approach with machine learning classifiers, fine-tuned through Bayesian optimization has illustrated the effectiveness of the proposed classification models. This hybrid feature selection method improved the classification accuracy by 2.61% for two-level stress classification and by 1.86% for three-level stress classification compared to existing models without hybrid feature selection. This suggests incorporating HRV and EDA signals in wearable devices can enhance stress prediction accuracy. In the future, researchers can focus on further refining the utilization of HRV and EDA signals, which can contribute to developing more advanced and accurate wearable stress monitoring systems, thereby improving the overall effectiveness of stress management interventions.

CONFLICT OF INTEREST

The authors confirm that there are no known financial conflicts of interest or personal relationships that could have potentially influenced the findings presented in this paper.

DATA AVAILABILITY

Data will be made available on reasonable request.

REFERENCES

[1] P. S. Prabu, "A study on academic stress among higher secondary students," *International journal of humanities and social science invention*, vol. 4, no. 10, pp. 63–68, 2015.

[2] S. D. Ghatol, *Academic Stress among School Students*. Allied Publishers, 2019.

[3] A. Waqas, S. Khan, W. Sharif, U. Khalid, and A. Ali, "Association of academic stress with sleeping difficulties in medical students of a Pakistani medical school: a cross sectional survey," *PeerJ*, vol. 3, p. e840, 2015.

[4] V. M. Bhujade, "Depression, anxiety and academic stress among college students: A brief review.," *Indian Journal of Health & Wellbeing*, vol. 8, no. 7, 2017.

[5] S. Maji, A. Chaturmohta, D. Deevela, S. Sinha, S. Tarsolia, and A. Barsaiya, "Mental health consequences of academic stress, amotivation, and coaching experience: A study of india's top engineering undergraduates," *Psychology in the Schools*.

[6] S. Byun, A. Y. Kim, E. H. Jang, S. Kim, K. W. Choi, H. Y. Yu, and H. J. Jeon, "Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol," *Computers in biology and medicine*, vol. 112, p. 103381, 2019.

[7] M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh, A. Shalbaf, D. L. García, J. M. Gorriz, and U. R. Acharya, "Emotion recognition in EEG signals using deep learning methods: A review," *Computers in Biology and Medicine*, p. 107450, 2023.

[8] L. Holtz, M. Martinez, K. Paton, K. Rosich, and E. Schnittka, "Effects of physiological stress response on short-term memory recall," 2017.

[9] C. Anders and B. Arnrich, "Wearable electroencephalography and multimodal mental state classification: A systematic literature review," *Computers in Biology and Medicine*, p. 106088, 2022.

[10] A. Arsalan and M. Majid, "Human stress classification during public speaking using physiological signals," *Computers in biology and medicine*, vol. 133, p. 104377, 2021.

[11] M. Ullah, S. Akbar, A. Raza, and Q. Zou, "DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm," *Bioinformatics*, vol. 40, no. 5, p. btae305, 2024.

[12] S. Akbar, Q. Zou, A. Raza, and F. K. Alarfaj, "iAFPs-Mv-BiTCN: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks," *Artificial Intelligence in Medicine*, vol. 151, p. 102860, 2024.

[13] S. Akbar, A. Raza, and Q. Zou, "Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model," *BMC bioinformatics*, vol. 25, no. 1, p. 102, 2024.

[14] A. Raza, J. Uddin, A. Almuhaimeed, S. Akbar, Q. Zou, and A. Ahmad, "AIPs-SnTCN: Predicting anti-inflammatory peptides using fasttext and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks," *Journal of chemical information and modeling*, vol. 63, no. 21, pp. 6537–6554, 2023.

[15] S. Akbar, A. Raza, T. Al Shloul, A. Ahmad, A. Saeed, Y. Y. Ghadi, O. Mamyrbayev, and E. T. Eldin, "pAtbP-EnC: identifying anti-tubercular peptides using multi-feature representation and genetic algorithm based deep ensemble model," *IEEE Access*, 2023.

[16] A. Kumar, K. Sharma, and A. Sharma, "Hierarchical deep neural network for mental stress state detection using IoT based biomarkers," *Pattern Recognition Letters*, vol. 145, pp. 81–87, 2021.

[17] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, 2019.

[18] B. Egilmez, E. Poyraz, W. Zhou, G. Memik, P. Dinda, and N. Alshurafa, "UStress: Understanding college student subjective stress using wrist-based passive sensing," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 673–678, IEEE, 2017.

[19] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien, and J. C. Pruessner, "The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," *Journal of Psychiatry and Neuroscience*, vol. 30, no. 5, pp. 319–325, 2005.

[20] Y. Zhang, Q. Wang, Z. Y. Chin, and K. K. Ang, "Investigating different stress-relief methods using electroencephalogram (EEG)," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2999–3002, IEEE, 2020.

[21] F. Al-Shargie, "Prefrontal cortex functional connectivity based on simultaneous record of electrical and hemodynamic responses associated with mental stress," *arXiv preprint arXiv:2103.04636*, 2021.

[22] J. Minguillon, E. Perez, M. A. Lopez-Gordo, F. Pelayo, and M. J. Sanchez-Carrion, "Portable system for real-time detection of stress level," *Sensors*, vol. 18, no. 8, p. 2504, 2018.

[23] L. Xia, A. S. Malik, and A. R. Subhani, "A physiological signal-based method for early mental-stress detection," *Biomedical Signal Processing and Control*, vol. 46, pp. 18–32, 2018.

[24] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.

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.2024.3463742

[25] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang, and Z. Zhao, "Detecting work-related stress with a wearable device," *Computers in Industry*, vol. 90, pp. 42–49, 2017.

[26] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of biomedical informatics*, vol. 73, pp. 159–170, 2017.

[27] L. Zhu, P. Spachos, P. C. Ng, Y. Yu, Y. Wang, K. Plataniotis, and D. Hatzinakos, "Stress Detection Through Wrist-Based Electrodermal Activity Monitoring and Machine Learning," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[28] C.-P. Hsieh, Y.-T. Chen, W.-K. Beh, and A.-Y. A. Wu, "Feature selection framework for xgboost based on electrodermal activity in stress detection," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 330–335, IEEE, 2019.

[29] P. Siirtola, "Continuous stress detection using the sensors of commercial smartwatch," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pp. 1198–1201, 2019.

[30] D. Jaiswal, D. Chatterjee, R. Gavas, R. K. Ramakrishnan, and A. Pal, "Person and Stressor Independent Generic Model for Stress Detection Using GSR," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 7195–7198, IEEE, 2021.

[31] T. Androutsou, S. Angelopoulos, E. Hristoforou, G. K. Matsopoulos, and D. D. Koutsouris, "Automated Multimodal Stress Detection in Computer Office Workspace," *Electronics*, vol. 12, no. 11, p. 2528, 2023.

[32] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, "A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques," *Sensors*, vol. 23, no. 7, p. 3565, 2023.

[33] A. Hemakom, D. Atiwiwat, and P. Israsena, "ECG and EEG based machine learning models for the classification of mental workload and stress levels for women in different menstrual phases, men, and mixed sexes," *Biomedical Signal Processing and Control*, vol. 95, p. 106379, 2024.

[34] M. Benchekroun, P. E. Velmovitsky, D. Istrate, V. Zalc, P. P. Morita, and D. Lenne, "Cross dataset analysis for generalizability of HRV-based stress detection models," *Sensors*, vol. 23, no. 4, p. 1807, 2023.

[35] M. Quiñones, A. Van den Broeck, and H. De Witte, "Do job resources affect work engagement via psychological empowerment? A mediation analysis," *Revista de Psicología del Trabajo y de las Organizaciones*, vol. 29, no. 3, pp. 127–134, 2013.

[36] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Cañón, T. Aljama-Corrales, S. Charleston-Villalobos, and K. H. Chon, "Power spectral density analysis of electrodermal activity for sympathetic function assessment," *Annals of biomedical engineering*, vol. 44, pp. 3124–3135, 2016.

[37] A. Greco, G. Valenza, J. Lazaro, J. M. Garzon-Rey, J. Aguilo, C. de la Cámara, R. Bailón, and E. P. Scilingo, "Acute stress state classification based on electrodermal activity modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 788–799, 2021.

[38] Y. Liu and S. Du, "Psychological stress level detection based on electrodermal activity," *Behavioural brain research*, vol. 341, pp. 50–53, 2018.

[39] C. Schubert, M. Lambertz, R. Nelesen, W. Bardwell, J.-B. Choi, and J. Dimsdale, "Effects of stress on heart rate complexity—a comparison between short-term and chronic stress," *Biological psychology*, vol. 80, no. 3, pp. 325–332, 2009.

[40] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020," in *NeurIPS 2020 Competition and Demonstration Track*, pp. 3–26, PMLR, 2021.

[41] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, 2019.

[42] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of health and social behavior*, pp. 385–396, 1983.

[43] R. Champseix, L. Ribiere, and C. Le Couedic, "A Python Package for Heart Rate Variability Analysis and Signal Preprocessing.," *Journal of Open Research Software*, vol. 9, no. 1, 2021.

[44] S. Föll, M. Maritsch, F. Spinola, V. Mishra, F. Barata, T. Kowatsch, E. Fleisch, and F. Wortmann, "FLIRT: A feature generation toolkit for wearable data," *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106461, 2021.

[45] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[46] G. Cosoli, A. Poli, L. Scalise, and S. Spinsante, "Measurement of multimodal physiological signals for stimulation detection by wearable devices," *Measurement*, vol. 184, p. 109966, 2021.

[47] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in biology and medicine*, vol. 112, p. 103375, 2019.

[48] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020.

[49] M. M. Kabir, M. Shahjahan, and K. Murase, "A new local search based hybrid genetic algorithm for feature selection," *Neurocomputing*, vol. 74, no. 17, pp. 2914–2928, 2011.

[50] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1424–1437, 2004.

[51] L. I. Kuncheva and L. C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection," *Pattern recognition letters*, vol. 20, no. 11-13, pp. 1149–1156, 1999.

[52] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 44–49, 1998.

[53] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE transactions on evolutionary computation*, vol. 4, no. 2, pp. 164–171, 2000.

[54] M. Mitchell, "An introduction to genetic algorithms mit press," *Cambridge, Massachusetts. London, England*, vol. 1996, 1996.

[55] J. J. Grefenstette, "Genetic algorithms and machine learning," in *Proceedings of the sixth annual conference on Computational learning theory*, pp. 3–4, 1993.

[56] D. Ramdania, M. Irfan, F. Alfarisi, and D. Nuraiman, "Comparison of genetic algorithms and Particle Swarm Optimization (PSO) algorithms in course scheduling," in *Journal of Physics: Conference Series*, vol. 1402, p. 022079, IOP Publishing, 2019.

[57] M. Rostami, K. Berahmand, and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection," *Journal of Big Data*, vol. 8, no. 1, pp. 1–27, 2021.

[58] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern recognition*, vol. 33, no. 1, pp. 25–41, 2000.

[59] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[60] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, pp. 175–186, 2014.

[61] P. Koch, O. Golovidov, S. Gardner, B. Wujek, J. Griffin, and Y. Xu, "Autotune: A derivative-free optimization framework for hyperparameter tuning," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 443–452, 2018.

[62] M. R. Hossain and D. Timmer, "Machine learning model optimization with hyper parameter tuning approach," *Glob. J. Comput. Sci. Technol. D Neural Artif. Intell*, vol. 21, no. 2, 2021.

[63] S. Andradóttir, "A review of random search methods," *Handbook of Simulation Optimization*, pp. 277–292, 2014.

[64] V. Nguyen, "Bayesian optimization for accelerating hyper-parameter tuning," in *2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE)*, pp. 302–305, IEEE, 2019.

[65] F. Fatih, Z. En-Naimani, and K. Haddouch, "Comparative Study of Bayesian Optimization Process for the Best Machine Learning Hyperparameters," in *International Conference On Big Data and Internet of Things*, pp. 239–249, Springer, 2022.

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.2024.3463742



SHIKHA received B.Tech degree in Information Technology from the Guru Gobind Singh Indraprastha University (GGSIPU) Delhi in 2014 and an M.Tech degree in Computer Science and Engineering from Indira Gandhi Delhi Technical University (IGDTUW) Delhi in 2016. She is pursuing a doctorate of philosophy from Delhi Technological University(DTU), Delhi. Her research interests are Human-computer Interaction in healthcare and Machine Learning.



DR. DIVYASHIKHA SETHIA received her B.Tech degree in Computer Science and Engineering from the Maharaja Sayajirao University of Baroda, India, in 1997, and her M.Tech degree in the same field from the Indian Institute of Technology Delhi, India, in 2006, with a focus on distributed systems and mobile computing. She has over five years of industry experience in the software development sector. Since 2010, she has been an Assistant Professor in the Department of Computer Science and Engineering at Delhi Technological University (DTU) Delhi. Her research interests encompass telemedicine, security in distributed systems, and the application of brain-computer interfaces in healthcare. She is the author of more than 13 journal publications and 50 conference papers.



PROF. S. INDU (Senior Member, IEEE) received the B.Tech. degree from the TKM College of Engineering, University of Kollam, India, in 1987, the M.Tech degree from the College of Engineering Trivandrum, University of Kerala, Thiruvananthapuram, India, in 1990, and a Ph.D. degree from the University of Delhi, India, in 2012. She joined the Electronics and Communication Engineering Department at Delhi College of Engineering, Delhi, in 1999. She works as Dean (student Welfare) at DTU and a Professor of the ECE Department at Delhi Technological University (DTU), Delhi. Her research interests include computer vision, sensor networks, and image processing. Prof. Indu received the Commendable Research Award from DTU in 2018 and 2019. She has over 150 publications in books, book chapters, and peer-reviewed journals.

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

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Optimizing Carrier Concentration for Enhanced Thermoelectric Performance in AgSbS 2 Monolayer

Stephen Shaiju Mathew
Delhi Technological University
Sangeeta .
Delhi Technological University
Rajesh Kumar
Delhi Technological University
Mukhtiyar Singh
Delhi Technological University
Manish K. Kashyap
manishdft@gmail.com

Jawaharlal Nehru University

Research Article

Keywords: Thermoelectrics, DFT, Electronic Transport, Thermal Conductivity

Posted Date: September 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4892551/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

This work investigates the structural, electronic, elastic, and transport properties of the $AgSbS_2$ monolayer, using density functional theory in conjunction with semiclassical Boltzmann transport theory. In this study, we proposed a strategy to enhance the figure of merit (*ZT*) by optimizing the carrier concentrations. The monolayer of $AgSbS_2$ is found to be both mechanically and thermodynamically stable. The calculated electronic band structure shows a semiconducting nature of $AgSbS_2$ with an indirect band gap of 1.31 eV using the Heyd-Scuderia-Ernzerhof (HSE06) exchange-correlation functional. The investigated monolayer is found to be anisotropic, hence we analyzed its thermoelectric properties at various carrier concentrations along *a*– and *b*–directions at 300 K. It attained a high value of Seebeck coefficient of 360 μVK^{-1} and 370 μVK^{-1} in the *a*– and *b*–directions at room temperature, respectively. The low thermal conductivity and high power factor result in an appreciable *ZT* value. The maximum *ZT* of $AgSbS_2$ monolayer at an optimized carrier concentration of 3×10^{19} cm^{-3} is found to be 0.5 in the *b*-direction for the n-type monolayer, whereas, for the p-type, it is predicted as 0.3 at 300 K. The present work supports the potential use of $AgSbS_2$ in room-temperature energy harvesting applications.

1. Introduction

In a challenging era of excessive fossil fuels consumption and their consequential environmental impacts, achieving a net zero target requires minimum reliance on conventional energy sources. To achieve this, developing clean and secure energy alternatives is imperative. New technologies can play a pivotal role in energy generation and storage. These encompass diverse areas such as photovoltaics, wind and water turbines, and energy-efficient technologies like low-loss electronics, piezoelectrics, and thermoelectrics (TEs). TE materials, for instance, can efficiently convert waste heat into usable electrical energy, powering TE generators. TE devices show unique advantages as compared to other energy conversion devices, including stability, long service life, and noiselessness [1–3]. The dimensionless figure of merit (ZT) represents the efficiency of a TE material and can be calculated from the formula $ZT = S^2\sigma T/\kappa$, where T is the absolute temperature, *S* denotes the Seebeck coefficient, σ represents electrical conductivity, and κ encompasses the combined effects of electronic (κ_e) and lattice (κ_l) contributions to thermal conductivity [4, 5]. Achieving a high ZT necessitates a significant power factor ($S^2\sigma$) and reduced κ value. A significant challenge lies in optimizing $S^2\sigma$ and κ_e, as they are interdependent through the carrier concentration [6, 7]. Various strategies have been devised to overcome this challenge to decouple these parameters. These include increasing the power factor (PF) via carrier concentration adjustments, tuning band structures, and reducing lattice thermal conductivity through dimensional reduction or introducing defects [8–12].

Two-dimensional (2D) semiconductors are highly effective for TE applications and possess exceptional electronic, thermal, and transport properties, along with quantum confinement effects. 2D materials like graphene, Mxenes (e.g., Ti_4N_3), and allotropes have shown excellent electronic properties, making them incredibly promising for TE use. Apart from these, transition metal mono– and di–chalcogenides have also shown remarkable potential in 2D TE materials [13]. Notably, the Mg_3Sb_2 monolayer has shown a favorable *ZT* of 2.5 at 900 K [14]. Further, enhanced *ZT* values ≥ 2.5 have been achieved in materials such as PbTe and SnSe [15]. It has been shown that Ag-E1-E2 (E = S/Se/Te) compounds exhibit optimal TE performance compared to their binary counterparts, due to the effective scattering of short-wavelength phonons by atomic defects, which

are generated by alloying [16]. Silver-based chalcogenides, like Ag_2E (where E = S, Se, Te) and their other derivatives and hybrids have attracted significant research interests due to their high TE performance near room temperature. These materials are expected to drive substantial advancements in energy harvesting for self-powered wearable intelligent medical devices [17, 18]. Generally, these materials exhibit compelling attributes due to their high electron mobility, lower electron effective mass, and very narrow band gap, contributing to a high σ [19, 20]. Consequently, they have *ZT* values ranging from 0.6 to 1.1 across different temperature ranges. Particularly, Ag_2S has a low σ value of the order 10^{-1} Sm^{-1}, a low *S*, and an appreciably high κ_e value; these parameters resulted in a constrained TE performance [21, 22]. Therefore, efforts have been made to maximise the carrier concentration of the Ag_2S [23]. In the Te-Sb phase, the excess Sb readily combines with Te to yield Sb_2Te_3, which is recognized as a top-performing p-type TE material. Consequently, an abundance of Sb is anticipated to generate Sb_2Te_3 impurities, fostering high-concentration boundary interfaces within the material. Consequently, this scenario is poised to lower κ_l by intensifying phonon scattering at these interfaces, which is beneficial for TE applications [24]. Lower κ_l minimizes heat transfer through the material, allowing for better TE efficiency. Sb is not only expected to dope Te but also produce dense boundary interfaces for improving TE performance [25]. It is found that a ~0.5% Sb-doping increases the hole concentration from $\sim 10^{18}$ cm^{-3}. The I-V-VI_2 semiconductors, where I = Cu or Ag, V = Sb or Bi, VI = chalcogens (S, Se or Te), are found to be promising TE materials due to their intrinsically low κ_l. ZT values of $Ag_{1+x}Sb_{1-x}Se_2$ (x = 0.00, 0.01, 0.02, 0.03) samples due to the combined effect of enhanced PF and lowered κ_l, a *ZT* value of ~ 1.02 was achieved for $Ag_{1.02}Sb_{0.98}Se_2$ at 673 K [26].

In this work, we have studied the rectangular monolayer of $AgSbS_2$, which is derived from its bulk orthorhombic phase. The rectangular monolayer is an iso-structure of the layered group $CuMN_2$ (M = Bi, Sb; N = S, Se) that has shown enhanced *ZT* values, ranging from 0.70–1.50 for n-type and 0.90–1.17 for p-type $CuMN_2$ at 300 K, via carrier concentration optimization [27]. These *ZT* values are higher or somewhat similar to many 2D TE materials, such as PdSe (0.9) [28] and InSe (0.5) [29]. Thus, the present study aims to investigate the transport properties of novel rectangular $AgSbS_2$ monolayer by carrier concentration optimization. To the best of our knowledge, the $AgSbS_2$ rectangular monolayer has not yet been explored for TE properties. We believe that this work will pave the way for future experimental and theoretical investigations to design this monolayer for efficient TE performance.

2. Computational methodology

The structural, elastic, and electronic properties of the $AgSbS_2$ monolayer were assessed using density functional theory (DFT) within the Vienna *ab-initio* Simulation Package (VASP), employing the projector augmented wave method (PAW) [30, 31]. A vacuum of 10 Å was added in the *c*-direction for further calculations. The Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation (GGA) was utilized for structural optimization where the Kohn-Sham equations were solved self consistently till a convergence criterion of 10^{-6} eV was met [32]. *k*-point mesh of size 11×11×1 with a Γ-centred sampling scheme was used in the simulations. A plane wave cutoff of 450 eV, validated through energy optimization compass self-consistent field calculations using both PBE and the Heyd-Scuseria-Ernzerhof hybrid functional (HSE06) [33]. The hybrid functional helps to compute the electronic band gap more accurately.

The transport properties were calculated using the *ab-initio* scattering and transport method, implemented in the AMSET code [34], which uses the electronic Boltzmann transport equation (BTE) to calculate scattering rates and mobilities using the Born approximation. This work considered the acoustic deformation potential (ADP), ionized impurity scattering (IMP), and polar optical phonon scattering (POP), which are the most common scattering types found in semiconductor-like materials [34]. The ADP is responsible for the electron-phonon interactions. IMP represents the charge carriers scattering through the ionization of lattice, and POP includes the exchange of electrons and polar optical phonons. Matthiessen's rule can find out the resultant carrier relaxation time, the inverse of scattering rates by the following equation:

$$\frac{1}{\tau} = \frac{1}{\tau_{ADP}} + \frac{1}{\tau_{IMP}} + \frac{1}{\tau_{POP}} \quad (1)$$

where $\tau_{ADP},\ \tau_{IMP},\ \text{and}\ \tau_{POP}$ are the relaxation times from the ADP, POP, and IMP scattering. In ADP and IMP scatterings, electrons undergo elastic scattering, where they neither gain nor lose energy. On the other hand, POP scatterings show inelastic scattering behavior characterized by phonon emission or absorption-induced energy changes. The Fermi Golden rule is applied to compute both elastic and inelastic scattering rates. This rule allows us to calculate the transition rates from an initial quantum state $|nk\rangle$ to a final quantum state$mk+q$.

$$\tilde{\tau}^{-1}_{nk\to mk+q} = \frac{2\pi}{\hbar}|g_{nm}(k,q)|^2\delta\left(\varepsilon_{nk} - \varepsilon_{mk+q}\right)$$

2

where *ħ* stands for reduced Planck's constant, *δ* represents a Dirac delta function, and *n* relates to the Bose-Einstein distribution. The quantity ε_{nk} pertains to the energy of the state $|nk\rangle$, while $g_{nm}\left(k,q\right)$ signifies the coupling matrix element, governing the specific scattering process under consideration. To attain the scattering rates, it is necessary to have deformation, static and high-frequency constants along with elastic constants and dense and uniform band structures. In our calculations, these were all obtained using DFT and density functional perturbation theory (DFPT) [35]. The average relaxation time was calculated from all the scattering rates and is incorporated to find the transport parameters *σ*, and κ_e by solving BTE using an interpolation factor of 10. We obtained the elastic constants using the energy-strain relationship using VASP and checked the mechanical stability of the investigated monolayer.

3. Results and Discussion

3.1. Structural properties

This study is focused on $AgSbS_2$ in its orthorhombic phase (space group, *Pnma* #62). Within this structure, Ag^{1+} ions are coordinated with four S^{2-} ions, forming AgS_4 tetrahedra that share corners with eight identical SSb_5 square pyramids and also share corners with other four identical AgS_4 tetrahedra, along with one SbS_5 square pyramid. The structural optimization of the bulk $AgSbS_2$ was performed and then the resultant optimized structure was further used to obtain its monolayer. $AgSbS_2$ monolayer exhibits a rectangular phase where Ag is bonded with three Sb atoms and three S atoms as depicted in Fig. 1, characterized by two-fold rotations and orthogonal mirror planes. The optimized lattice parameters for the monolayer of this compound in its equilibrium state were obtained as listed in Table 1. So far no previous studies of the investigated structure have been done,

therefore, we have compared its structural parameters to its iso-structure $CuSbS_2$ [27]. In addition, the crystal structure shows anisotropic properties, making it applicable for tuneable structural, elastic and transport properties that might help to improve TE performance.

Table 1
Obtained structural parameters (*a* and *b* in Å and V in $Å^3$) and bond lengths (in Å) of $AgSbS_2$ in comparison with its iso-structure $CuSbS_2$

Material	Structural parameters			Bond lengths	
	a	*b*	*V*	Ag/Cu-S_1	Sb-S_1
$AgSbS_2$	6.367	3.927	390.403	2.327	2.435
$CuSbS_2$ [27]	6.150	3.810	338.226	2.308	2.484

3.2. Stability analysis

It is necessary to assess the stability of studied monolayer before doing further calculations as it is not experimentally realized yet. The monolayer was confirmed to meet the thermodynamic stability criteria by calculating the cohesive energy (E_{coh}) through the following equation [36]:

$$E_{coh} = \left[E_{AgSbS_2} - \left(E_{Ag} - E_{Sb} - 2E_S\right)\right] / N$$

1

The ground state energy (E_{AgSbS_2}) of $AgSbS_2$ can be expressed as a function of the total number of atoms in its unit cell (N) and the individual total energy per atom for Ag, Sb, and S in their respective bulk states (E_{Ag} and E_S). The calculated $E_{coh}$for $AgSbS_2$ monolayer was determined to be -6.868 eV/atom, hence confirming it as thermodynamically stable.

The elastic constants were evaluated by the energy strain relationship to ascertain the mechanical stability. Since, the monolayer forms a rectangular structure, it possesses four independent elastic constants, C_{11}, C_{12}, C_{22} and C_{66} where C_{11} = 11.057, C_{12} = 1.801, C_{22} = 15.252 and C_{66} = 2.533, all in units of Nm^{-1}. The Born stability criteria for this structure is given as [37]:

$C_{11} > 0, C_{66} > 0$ and $C_{11} \times C_{22} > C_{12} \times C_{12}$ (2)

All the stability criteria for the rectangular structure were met, suggesting that the rectangular form remains mechanically stable in ambient conditions. The Young's modulus (*Y*), bulk modulus (*B*), shear modulus (*G*), and Poisson's ratio (*v*) were found to infer various mechanical properties such as brittleness, ductility, stiffness, and strength of $AgSbS_2$ monolayer. The values of the elastic constants are presented in Table 2. The values of *B* and G were approximated using Voigt-Reuss hill approximation [38]. *B* dominates the mechanical stability as it has a higher value than the *G*. The Pugh ratio (*B/G)* is higher than 1.75, which shows the ductile nature of $AgSbS_2$ [39]. The anisotropy in $AgSbS_2$ results in a *v* of approximately 0.118 along the *a*-direction the *v* and 0.484 along the *b*-direction. *v* serves as an indicator of the bonding characteristics of the material; when *v* falls within the range of

0.25 to 0.5, it suggests the prevalence of central forces in the material's bonding [40–42]. In the case of the $AgSbS_2$ monolayer, the presence of central forces found to be present along the *b*-direction and a presence of noncentral forces along the *a*-direction which signifies a covalent nature in *a*- direction. These elastic constants were used to determine the Debye temperature (θ_D) using following expression:

$$\theta_D = (h/k_B)\left[\left(\frac{3n}{4\pi}\right)(N_A\rho/M)\right]^{\frac{1}{3}}.\nu_m$$

(3)

Where ρ is the density, N_a is the Avogadro number, k_B is the Boltzmann constant, *n* is the number of atoms in the monolayer, *M* is the molecular weight of the compound, and is the average sound velocity. The longitudinal (ν_l) and transversal (ν_t) sound both contribute to the average sound velocity (ν_m) by using elastic constants through the following equations [43]:

$$\nu_m = \left[\frac{1}{3}\left(\frac{2}{v_t^3} + \frac{1}{\nu_l^3}\right)\right]^{\frac{-1}{3}}$$

4

$$\nu_l = \sqrt{\frac{G}{\rho}}$$

5

$$\nu_t = \sqrt{\frac{3B + 4G}{\rho}}$$

6

By solving these equations, we got the values ofν_t, ν_m, ν_l, and θ_D along both the directions as presented in Table 2. The maximum values of *Y*, *v* and θ_D of $CuSbS_2$ were 66.53 Nm^{-1}, 0.14 and 90.10 K respectively [27]. The maximum value *Y* for $AgSbS_2$ was found to be much lower than $CuSbS_2$, hence resulting in a comparatively higher value Debye temperature.

Table 2

Obtained Young's Modulus *(Y)*, Bulk Modulus (*B)*, Shear Modulus *(G)* in Nm^{-1}, Poisson's ratio (*v*), the transverse (ν_t) and longitudinal (ν_l) sound velocity, mean sound velocity (v_m) in Kms^{-1} and Debye temperature (θ_D) in K of $AgSbS_2$.

Direction	*Y*	*B*	*G*	*v*	v_t	v_l	v_m	$\varvec{\theta_D}$
a-direction	7.518	7.478	2.553	0.118	1.93	0.54	0.767	114
b-direction	14.958	7.478	5.530	0.484	2.93	0.796	0.892	132.29

3.3. Electronic structure

The band structure and the density of states (DOS) are essential parameters that give us information about the electronic and transport properties. The band structure was calculated using both PBE and HSE06 functionals as depicted in Fig. 2. Even though the band gap values calculated by both methods differ, the behavior of the electronic band structure remains the same. We identified $AgSbS_2$ as an indirect band gap semiconductor as the top of its valence band is at Γ and the bottom of the conduction band (CB) is at *X*. The band gap values of 0.58 eV using PBE and 1.31 eV using HSE06 were estimated which turned out be comparable to the other ternary and binary dichalcogenide monolayers like $CuSbS_2$ (1.10 eV), $CuBiS_2$ (0.83 eV) and $SnSe_2$ (0.715 eV) that are well known for their TE performance [27, 44]. For VB maxima, the deformation potential (E_l) has a value of 3.53 eV along the *a*-direction and 5.25 eV along the *b*-direction, whereas for the CBM, it is 0.89 eV along the *a*-direction and 4.59 eV along the *b*-direction. The carrier mobility increases when E_l is large, conversely, it may decrease if the deformation potential attains lower values [45].

The analysis of the density of states (DOS) using HSE06 functional, as shown in Fig. 3 reveals that the lower states in the VB (from − 5 eV to -2 eV) are primarily occupied by the Ag atoms. Further examination of the partial density of states (PDOS), in Fig. 3, shows that Ag-*d* states dominate the deeper region of the VB whereas in the top of the VB, the major contributors are the S-*p* states with a minimal contribution from the Sb-*p* states. The CB is predominantly consisting of a S-*p* orbitals (from 0.5 eV to 2.5 eV) whereas the lower regions have almost equivalent contributions by S-*p* and Sb-*p* states (from 2.5 eV to 4 eV) along with a negligible contribution from Sb-*s* states.

3.4. Transport Properties

The transport properties were investigated with the help of the *ab-initio* scattering and transport (AMSET) method. We found the carrier relaxation time and overall mobilities by considering different scattering phenomena. We studied these with varying carrier concentrations at 300 K for both n- and p-type $AgSbS_2$ monolayers. The three main scattering phenomena are ADP, POP, and IMP; hence, the average scattering time can be found by Matthiessen's rule, which is already taken into consideration while calculating the transport properties in AMSET. The calculated scattering rates ($1/\tau$) are shown for both p- and n-type in Fig. 4.

It can be observed from Fig. 4 that both ADP and POP scattering rates exhibit no dependency on the chosen carrier concentrations. Conversely, IMP scattering rates progressively rise with increasing carrier concentration, ultimately driving the overall trend in the net scattering rate for both n- and p-type at 300 K. Altogether, the

electrons contribute to higher scattering rates than the hole for the investigated $AgSbS_2$ monolayer. To further understand the transport properties, we examined how the mobility contributions from ADP, POP, and IMP scatterings vary with carrier concentrations at room temperature using the relation $\mu = e\tau/m^*$, where m^* is the effective mass, and *e* is the electronic charge. The *μ* of a charge carrier is a direct result of scattering rates and dispersion of bands, [46] as depicted in Fig. 5 at 300 K. In the case of n-type material, we observed that as carrier concentrations were increased, the *μ* shows a noticeable increase in the contribution from POP scattering. However, for IMP and ADP, *μ* exhibited a decreasing trend and eventually stabilized, reaching a nearly constant value, resulting in the overall trend driven by the *μ* contributions of POP. On the other hand, for the p-type at *μ* contributions for all ADP, POP, and IMP attained a saturated state with each of them almost having constant values, driving the entire trend to be almost independent of carrier concentrations.

Next, we found *S*, *σ*, κ_e, and $S^2\sigma$ with varying carrier concentrations for both the crystallographic directions *a*- and *b*- of the $AgSbS_2$ monolayer at room temperature. Our findings align with the anticipated trend: the *S* decreases as carrier concentrations increase, which is consistent with the Mott equation [47, 48]. In the case of n-type material, we observed that the *S* is higher in the *b*-direction as compared to the *a*-direction at 300 K (Fig. 6a). However, for p-type material, we have a very minimal difference in the value variation of *S* along both the directions with carrier concentration. The highest value for the *S* for p-type is 361.62 μVK^{-1} in the *a*-direction and 370.72 μVK^{-1} in the *b*-direction, whereas for n-type was 327.47 μVK^{-1} in the *a*-direction and 352.82 μVK^{-1} for 300 K. These values are comparable to various state-of-the-art TE monolayers like WS_2 (328.15 μVK^{-1}) [49] and $PtTe_2$ (400.15 μVK^{-1}) at 300 K [50]. The value of *σ* in the n-type material depicts a notable increase with the increase in the carrier concentration along the *b*-direction as compared to the *a*-direction, attaining an optimal value of 274.65 × $10^4\ \Omega^{-1}\mathrm{m}^{-1}$ in the *b*-direction at 300 K due to dispersive bands in the conduction region and high-value of *μ* for the charge carriers. On the other hand, in the *a*-direction, the values are significantly low as compared to the *b*-direction, hence making it independent of direction (Fig. 6b). Conversely, the p-type material exhibits a similar trend, it also displays a substantial increase in the *b*-direction as compared to the *a*-direction, attaining value up to 136.95 × $10^4\ \Omega^{-1}\mathrm{m}^{-1}$ but unlike n-type, the gap between the values along both directions is significantly less, hence making it direction dependent. The value of κ_e is observed to increase significantly with the increase in carrier concentrations in the *b*-direction for the n-type, as shown in Fig. 6c. This same trend is supported by the p-type (Fig. 6c). The divergent patterns observed in *σ* and *S* yield high $S^2\sigma$ values for n-type materials in the *b*-direction at 300 K, attaining a high value of 10.66 *m*$Wm^{-1}K^{-2}$ and 11.13 *m*$Wm^{-1}K^{-2}$, which is significantly higher than in the *a*-direction. For p-type material, a similar trend is observed in both directions but the values are lower compared to the n-type shown in Fig. 6d. Due to the weak bonding characteristics and high atomic mass of $AgSbS_2$, we have an appreciably low value of Debye temperature; therefore the κ_l cannot be found out using the Slack's model as it holds for higher values of Debye temperature. Hence we calculated it using the Cahill model, given by the following expression [51].

$$\kappa_l = k_B p^{2/3} \left(v_l + 2v_t\right) \Big/ 2.48$$

7

Here, k_B stands for the Boltzmann constant, *p* is the total number of atoms per unit volume, v_l and v_t are longitudinal and transverse sound velocities, respectively at room temperature. We found the value of κ_l as 0.205 $Wm^{-1}K^{-1}$ along the *a*-direction and 0.306 $Wm^{-1}K^{-1}$ along the *b*-direction.

3.5. Thermoelectric figure of merit (*ZT*)

The *ZT* values are found in the *a*- and *b*-directions at 300 K as a function of carrier concentrations for both n- and p-type $AgSbS_2$ monolayer. As the carrier concentration increases, the *ZT* value first increases, reaches a peak value, and then decreases from its peak value to attain a lower value in both directions at 300 K, as shown in Fig. 7. *ZT* values are obtained for both n- and p- type at a carrier concentration of 5×10^{18} – 5×10^{20} cm^{-3}, having the values going up to ~ 0.3 and ~ 0.2 in both *a*- and *b*- directions, respectively, for p-type at 300 K. Further, for n-type, these values go as high as 0.5 in the *b*-direction at 300 K for a doping concentration of 3×10^{19} cm^{-3}. The n-type material expresses higher *ZT* value than p-type materials in the *b*-direction. On closer examination, we observed particularly elevated *ZT* values in the *b*-direction. These ZT values are undoubtedly comparative to various TEs like $RbZn_4P_3$ (0.33), MoS_2 (0.6), and Mg_3Sb_2 (0.6) at room temperature [52–54]. There is a further potential for increasing the thermoelectric efficiency of the material through strain engineering application and at elevated temperatures, which we intend to study in our further works.

4. Conclusion

In this study, we have used first-principles calculations in conjunction with semiclassical Boltzmann transport theory to realize the high *ZT* of $AgSbS_2$ monolayer at 300 K along *a*- and *b*- crystallographic directions. We have analyzed the effect of carrier concentration on its TE performance. $AgSbS_2$ is thermodynamically and mechanically stable, confirmed by its cohesive energy and elastic constants. The elastic properties have confirmed its covalent bonding in the *a*-direction; it has also been found that $AgSbS_2$ is anisotropic and ductile. The electronic band structure has shown that $AgSbS_2$ is an indirect semiconductor, having an indirect band gap of around 1.31 eV using hybrid functional. The calculated unit cell parameters and various bond lengths are comparable with a previous similar study on the $CuSbS_2$ monolayer. The n-type $AgSbS_2$ monolayer at an optimized carrier concentration of 5×10^{18} cm^{-3} exhibited peak ZT values of 0.5 in the *b*-directions at 300 K These promising *ZT* values make $AgSbS_2$ an optimal thermoelectric material for the future. Our findings on the electronic and transport properties of the $AgSbS_2$ monolayer may open new horizons for the design and synthesis of more efficient silver-based chalcogenide thermoelectric materials.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding Declaration

There was no Funding available for this manuscript.

Author's contribution

S.S.M. and S. performed Investigation and wrote the main manuscript text.

R.K. prepared figures 1-7.

M.S. and M.K.K. reviewed and edited the manuscript text.

Acknowledgment

Authors are thankful to RAFM-2024 for providing this platform for submission. The authors would like to thank Delhi Technological University, Delhi for providing a computational facility. Some part of the computational work has been carried out by using the resources of the National PARAM Supercomputing Facility (NPSF) at the Center for Development of Advanced Computing (C-DAC) Pune, which is an Autonomous Society of Ministry of Electronics and Information Technology (MeitY), Govt. of India. MKK is grateful to Ultra International India Pvt. Ltd., Ghaziabad, India for providing the lab furniture for Renewable Energy Laboratory at SPS, JNU.

References

1. T. Mori, S. Priya, Materials for energy harvesting: At the forefront of a new wave, MRS Bull. 43 (2018) 176–180. https://doi.org/10.1557/mrs.2018.32.
2. F. Zareef, M. Rashid, A.A.H. Ahmadini, T. Alshahrani, N.A. Kattan, A. Laref, Optoelectronic and thermoelectrical and mechanical properties of $CdLu_2X_4$ (X = S, Se) using first-principles calculations for energy harvesting applications, Mater. Sci. Semicond. Process. 127 (2021) 105695. https://doi.org/10.1016/j.mssp.2021.105695.
3. M. Roknuzzaman, K. Ostrikov, H. Wang, A. Du, T. Tesfamichael, Towards lead-free perovskite photovoltaics and optoelectronics by ab-initio simulations, Sci. Rep. 7 (2017) 14025. https://doi.org/10.1038/s41598-017-13172-y.
4. G. Tan, M. Ohta, M.G. Kanatzidis, Thermoelectric power generation: from new materials to devices, Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 377 (2019) 20180450. https://doi.org/10.1098/rsta.2018.0450.
5. L.E. Bell, Cooling, Heating, Generating Power, and Recovering Waste Heat with Thermoelectric Systems, Science 321 (2008) 1457–1461. https://doi.org/10.1126/science.1158899.
6. Y. Pei, A.D. LaLonde, N.A. Heinz, X. Shi, S. Iwanaga, H. Wang, L. Chen, G.J. Snyder, Stabilizing the Optimal Carrier Concentration for High Thermoelectric Efficiency, Adv. Mater. 23 (2011) 5674–5678. https://doi.org/10.1002/adma.201103153.
7. Sangeeta, M. Singh, Augmented thermoelectric performance of LiCaX (X = As, Sb) Half Heusler compounds via carrier concentration optimization, J. Phys. Chem. Solids 174 (2023) 111182. https://doi.org/10.1016/j.jpcs.2022.111182.
8. R. Kumar, R. Kumar, A. Vij, M. Singh, A first-principle study of electronic, thermoelectric, and optical properties of sulfur doped c-HfO_2, Phys. Scr. 97 (2022) 075813. https://doi.org/10.1088/1402-4896/ac7678.
9. A. Mehdizadeh Dehkordi, M. Zebarjadi, J. He, T.M. Tritt, Thermoelectric power factor: Enhancement mechanisms and strategies for higher performance thermoelectric materials, Mater. Sci. Eng. R Rep. 97 (2015) 1–22. https://doi.org/10.1016/j.mser.2015.08.001.
10. J. Yang, L. Xi, W. Qiu, L. Wu, X. Shi, L. Chen, J. Yang, W. Zhang, C. Uher, D.J. Singh, On the tuning of electrical and thermal transport in thermoelectrics: an integrated theory–experiment perspective, Npj Comput. Mater. 2 (2016) 15015. https://doi.org/10.1038/npjcompumats.2015.15.
11. R. Kumar, R. Kumar, M. Singh, D. Meena, A. Vij, Carrier concentration mediated enhancement in thermoelectric performance of various polymorphs of hafnium oxide: a plausible material for high

temperature thermoelectric energy harvesting application, J. Phys. Appl. Phys. 55 (2022) 495302. https://doi.org/10.1088/1361-6463/ac9986.

12. Sangeeta, M. Singh, Computational study of the thermoelectric properties and lattice dynamics of Li_2MN_2 (M = Zr or Hf), Mater. Res. Bull. 172 (2024) 112650. https://doi.org/10.1016/j.materresbull.2023.112650.
13. D. Li, Y. Gong, Y. Chen, J. Lin, Q. Khan, Y. Zhang, Y. Li, H. Zhang, H. Xie, Recent Progress of Two-Dimensional Thermoelectric Materials, Nano-Micro Lett. 12 (2020) 36. https://doi.org/10.1007/s40820-020-0374-x.
14. S. Huang, Z. Wang, R. Xiong, H. Yu, J. Shi, Significant enhancement in thermoelectric performance of Mg_3Sb_2 from bulk to two-dimensional mono layer, Nano Energy 62 (2019) 212–219. https://doi.org/10.1016/j.nanoen.2019.05.028.
15. Y. Xiao, L.-D. Zhao, Charge and phonon transport in PbTe-based thermoelectric materials, Npj Quantum Mater. 3 (2018) 55. https://doi.org/10.1038/s41535-018-0127-y.
16. H. Wu, X. Shi, J. Duan, Q. Liu, Z.-G. Chen, Advances in Ag_2Se-based thermoelectrics from materials to applications, Energy Environ. Sci. 16 (2023) 1870–1906. https://doi.org/10.1039/D3EE00378G.
17. K. Hoang, S.D. Mahanti, J.R. Salvador, M.G. Kanatzidis, Atomic Ordering and Gap Formation in Ag-Sb-Based Ternary Chalcogenides, Phys. Rev. Lett. 99 (2007) 156403. https://doi.org/10.1103/PhysRevLett.99.156403.
18. R. Detemple, D. Wamwangi, M. Wuttig, G. Bihlmayer, Identification of Te alloys with suitable phase change characteristics, Appl. Phys. Lett. 83 (2003) 2572–2574. https://doi.org/10.1063/1.1608482.
19. S.Y. Tee, D. Ponsford, C.L. Lay, X. Wang, X. Wang, D.C.J. Neo, T. Wu, W. Thitsartarn, J.C.C. Yeo, G. Guan, T. Lee, M. Han, Thermoelectric Silver-Based Chalcogenides, Adv. Sci. 9 (2022) 2204624. https://doi.org/10.1002/advs.202204624.
20. K. Lv, C. Shi, Y. Yang, Q. Wang, X. Sun, W. Chen, The pyrolysis preparation of the compact and full-coverage $AgSbS_2$ thin films and the photovoltaic performance of the corresponding solar cells, J. Mater. Sci. Technol. 98 (2022) 268–271. https://doi.org/10.1016/j.jmst.2021.04.070.
21. M. Jin, J. Liang, P. Qiu, H. Huang, Z. Yue, L. Zhou, R. Li, L. Chen, X. Shi, Investigation on Low-Temperature Thermoelectric Properties of Ag_2Se Polycrystal Fabricated by Using Zone-Melting Method, J. Phys. Chem. Lett. 12 (2021) 8246–8255. https://doi.org/10.1021/acs.jpclett.1c02139.
22. L. Mehdaoui, R. Miloua, M. Khadraoui, M.O. Bensaid, D. Abdelkader, F. Chiker, A. Bouzidi, Theoretical analysis of electronic, optical, photovoltaic and thermoelectric properties of $AgBiS_2$, Phys. B Condens. Matter 564 (2019) 114–124. https://doi.org/10.1016/j.physb.2019.04.006.
23. M. Zhu, X.-L. Shi, H. Wu, Q. Liu, Z.-G. Chen, Advances in Ag_2S-based thermoelectrics for wearable electronics: Progress and perspective, Chem. Eng. J. 475 (2023) 146194. https://doi.org/10.1016/j.cej.2023.146194.
24. S. Lin, W. Li, X. Zhang, J. Li, Z. Chen, Y. Pei, Sb induces both doping and precipitation for improving the thermoelectric performance of elemental Te, Inorg. Chem. Front. 4 (2017) 1066–1072. https://doi.org/10.1039/C7QI00138J.
25. T. Chen, K. Zhang, H. Wang, W. Su, F. Mehmood, T. Wang, J. Zhai, X. Wang, T. Huo, C. Wang, The high thermoelectric performance of slightly Sb doped PbTe alloys, J. Mater. Chem. C 8 (2020) 1679–1685. https://doi.org/10.1039/C9TC06124J.
26. X.-C. Liu, Y.-M. Wang, M.-L. Qi, M.-Y. Pan, Enhanced thermoelectric properties in Ag-rich AgSbSe2, J. Solid State Chem. 288 (2020) 121454. https://doi.org/10.1016/j.jssc.2020.121454.

27. W. Fang, Y. Chen, K. Kuang, M. Li, Excellent Thermoelectric Performance of 2D $CuMN_2$ (M = Sb, Bi; N = S, Se) at Room Temperature, Materials 15 (2022) 6700. https://doi.org/10.3390/ma15196700.
28. Z. Huang, Y. Zhang, H. Wu, S.J. Pennycook, L.-D. Zhao, Enhancing Thermoelectric Performance of p-Type PbSe through Suppressing Electronic Thermal Transports, ACS Appl. Energy Mater. 2 (2019) 8236–8243. https://doi.org/10.1021/acsaem.9b01708.
29. S.W. Lee, T. Kim, H.-S. Kim, O. Park, D.H. Kim, S. Kim, Enhanced thermoelectric properties of InSe through simultaneous increase in electrical conductivity and Seebeck coefficient by Cl doping, J. Mater. Res. Technol. 19 (2022) 2077–2083. https://doi.org/10.1016/j.jmrt.2022.05.180.
30. G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50. https://doi.org/10.1016/0927-0256(96)00008-0.
31. P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953–17979. https://doi.org/10.1103/PhysRevB.50.17953.
32. J.P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77 (1996) 3865–3868. https://doi.org/10.1103/PhysRevLett.77.3865.
33. J. Heyd, G.E. Scuseria, Efficient hybrid density functional calculations in solids: Assessment of the Heyd–Scuseria–Ernzerhof screened Coulomb hybrid functional, J. Chem. Phys. 121 (2004) 1187–1192. https://doi.org/10.1063/1.1760074.
34. A.M. Ganose, J. Park, A. Faghaninia, R. Woods-Robinson, K.A. Persson, A. Jain, Efficient calculation of carrier scattering rates from first principles, Nat. Commun. 12 (2021) 2222. https://doi.org/10.1038/s41467-021-22440-5.
35. K. Nakamura, S. Higuchi, T. Ohnuma, Density Functional Perturbation Theory to Predict Piezoelectric Properties, in: İ. Bakırtaş (Ed.), Perturbation Methods Appl. Sci. Eng., InTech, 2018. https://doi.org/10.5772/intechopen.76827.
36. M. Mushtaq, M.A. Sattar, S.A. Dar, Phonon phase stability, structural, mechanical, electronic, and thermoelectric properties of two new semiconducting quaternary Heusler alloys CoCuZrZ (Z = Ge and Sn), Int. J. Energy Res. 44 (2020) 5936–5946. https://doi.org/10.1002/er.5373.
37. M. Born, K. Huang, M. Lax, Dynamical Theory of Crystal Lattices, Am. J. Phys. 23 (1955) 474–474. https://doi.org/10.1119/1.1934059.
38. R. Hill, The Elastic Behaviour of a Crystalline Aggregate, Proc. Phys. Soc. Sect. A 65 (1952) 349–354. https://doi.org/10.1088/0370-1298/65/5/307.
39. F. Parvin, M.A. Hossain, I. Ahmed, K. Akter, A.K.M.A. Islam, First-principles calculations to investigate mechanical, optoelectronic and thermoelectric properties of half-Heusler p-type semiconductor BaAgP, Results Phys. 23 (2021) 104068. https://doi.org/10.1016/j.rinp.2021.104068.
40. H. Fu, D. Li, F. Peng, T. Gao, X. Cheng, Ab initio calculations of elastic constants and thermodynamic properties of NiAl under high pressures, Comput. Mater. Sci. 44 (2008) 774–778. https://doi.org/10.1016/j.commatsci.2008.05.026.
41. J.J. Zhang, Rock physical and mechanical properties, in: Appl. Pet. Geomech., Elsevier, 2019: pp. 29–83. https://doi.org/10.1016/B978-0-12-814814-3.00002-2.

42. Y.M. Poplavko, Mechanical properties of solids, in: Electron. Mater., Elsevier, 2019: pp. 71–93. https://doi.org/10.1016/B978-0-12-815780-0.00002-5.
43. O.L. Anderson, A simplified method for calculating the debye temperature from elastic constants, J. Phys. Chem. Solids 24 (1963) 909–917. https://doi.org/10.1016/0022-3697(63)90067-2.
44. N.D. Hien, N.Q. Cuong, L.M. Bui, P.C. Dinh, C.V. Nguyen, H.V. Phuc, N.V. Hieu, H.R. Jappor, L.T.T. Phuong, B.D. Hoi, L.C. Nhan, N.N. Hieu, First principles study of single-layer SnSe2 under biaxial strain and electric field: Modulation of electronic properties, Phys. E Low-Dimens. Syst. Nanostructures 111 (2019) 201–205. https://doi.org/10.1016/j.physe.2019.03.025.
45. P. Yan, G. Gao, G. Ding, D. Qin, Bilayer MSe_2 (M = Zr, Hf) as promising two-dimensional thermoelectric materials: a first-principles study, RSC Adv. 9 (2019) 12394–12403. https://doi.org/10.1039/C9RA00586B.
46. A.R. Natarajan, L. Ponvijayakanthan, M.K. Gupta, R. Mittal, D.J. Singh, V. Kanchana, High thermoelectric performance of layered LaAgXO (X = Se, Te) from electrical and thermal transport calculations, Phys. Rev. Mater. 7 (2023) 025405. https://doi.org/10.1103/PhysRevMaterials.7.025405.
47. J.P. Heremans, V. Jovovic, E.S. Toberer, A. Saramat, K. Kurosaki, A. Charoenphakdee, S. Yamanaka, G.J. Snyder, Enhancement of Thermoelectric Efficiency in PbTe by Distortion of the Electronic Density of States, Science 321 (2008) 554–557. https://doi.org/10.1126/science.1159725.
48. M. Cutler, N.F. Mott, Observation of Anderson Localization in an Electron Gas, Phys. Rev. 181 (1969) 1336–1340. https://doi.org/10.1103/PhysRev.181.1336.
49. A.N. Gandi, U. Schwingenschlögl, WS_2 As an Excellent High-Temperature Thermoelectric Material, Chem. Mater. 26 (2014) 6628–6637. https://doi.org/10.1021/cm503487n.
50. Anisha, R. Kumar, S. Srivastava, K. Tankeshwar, Thermoelectric properties of PtX_2 (X = Se, Te) monolayers, Indian J. Phys. (2023). https://doi.org/10.1007/s12648-023-02727-7.
51. D.G. Cahill, S.K. Watson, R.O. Pohl, Lower limit to the thermal conductivity of disordered crystals, Phys. Rev. B 46 (1992) 6131–6140. https://doi.org/10.1103/PhysRevB.46.6131.
52. Sangeeta, R. Kumar, M. Singh, Realizing high thermoelectric performance in p-type $RbZn_4P_3$ Zintl compound: a first-principles investigation, J. Mater. Sci. 57 (2022) 10691–10701. https://doi.org/10.1007/s10853-022-06953-y.
53. G. Tan, F. Shi, S. Hao, L.-D. Zhao, H. Chi, X. Zhang, C. Uher, C. Wolverton, V.P. Dravid, M.G. Kanatzidis, Non-equilibrium processing leads to record high thermoelectric figure of merit in PbTe–SrTe, Nat. Commun. 7 (2016) 12167. https://doi.org/10.1038/ncomms12167.
54. S.-D. Guo, A.-X. Zhang, Potential 2D thermoelectric materials ATeI (A = Sb and Bi) monolayers from a first-principles study, Nanotechnology 28 (2017) 445702. https://doi.org/10.1088/1361-6528/aa8741.

Figures

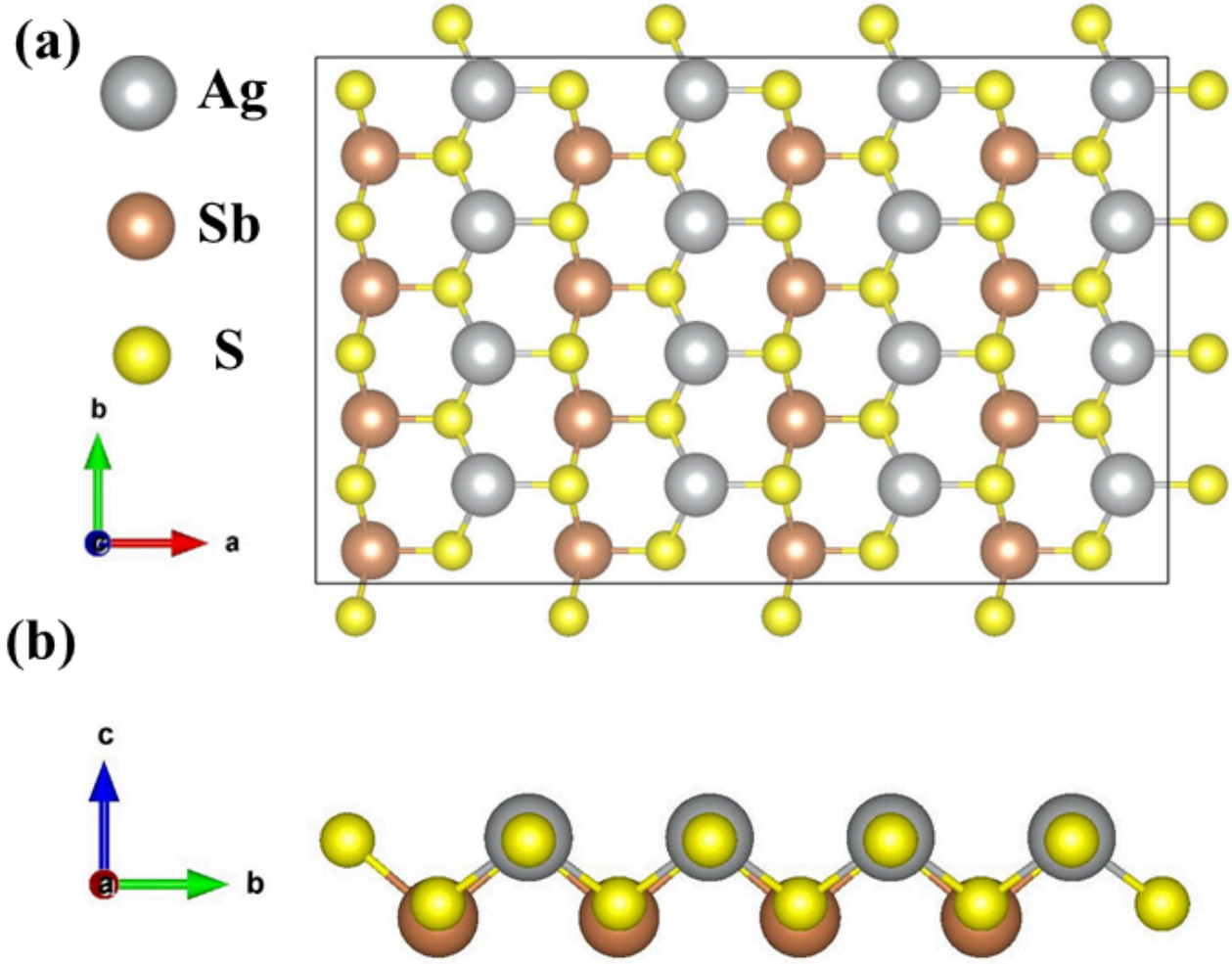


Figure 1

(a) Top and (b) Side view of $AgSbS_2$ monolayer

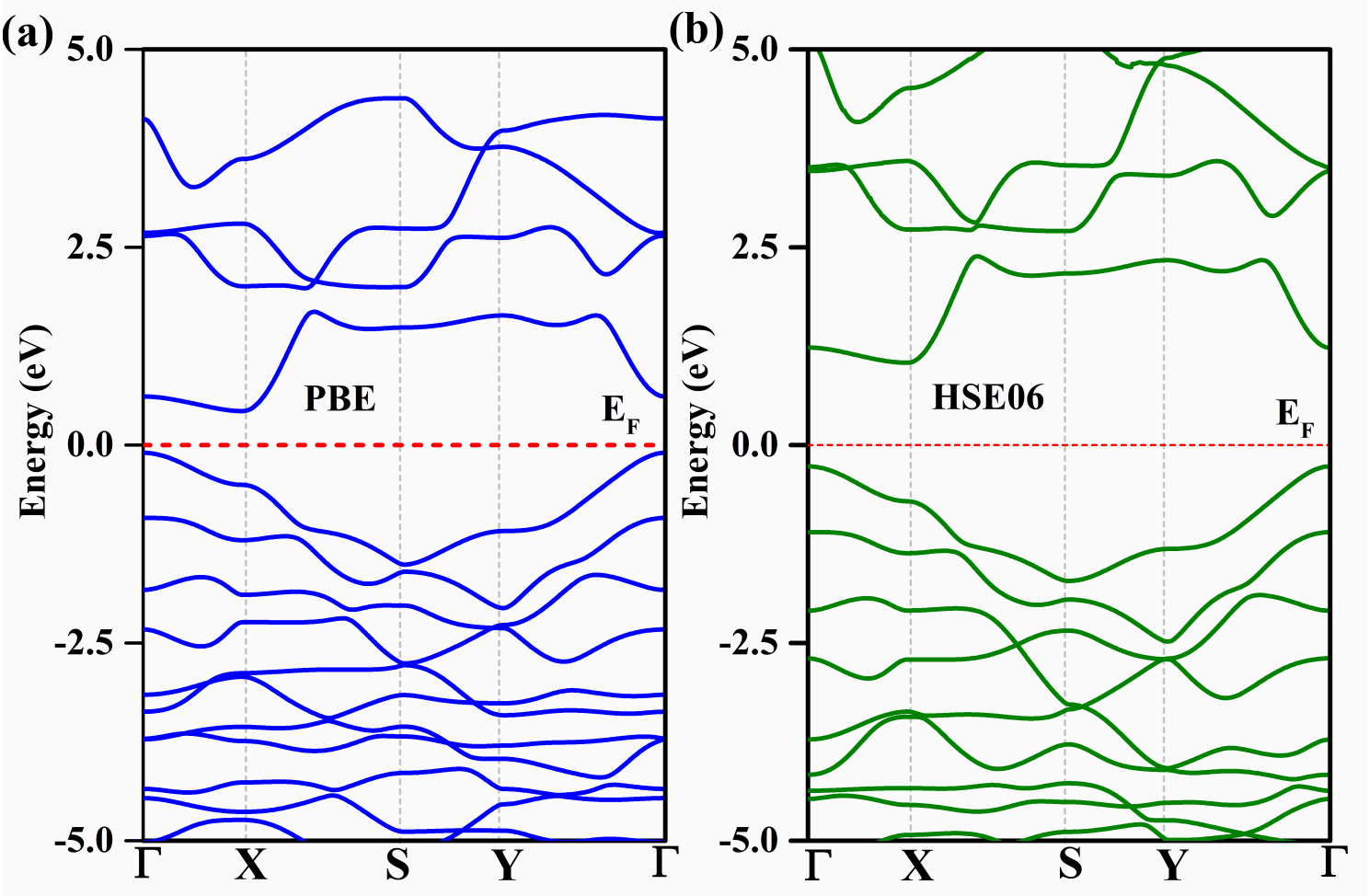


Figure 2

Electronic band structure of $AgSbS_2$ using (a) PBE and (b) HSE06

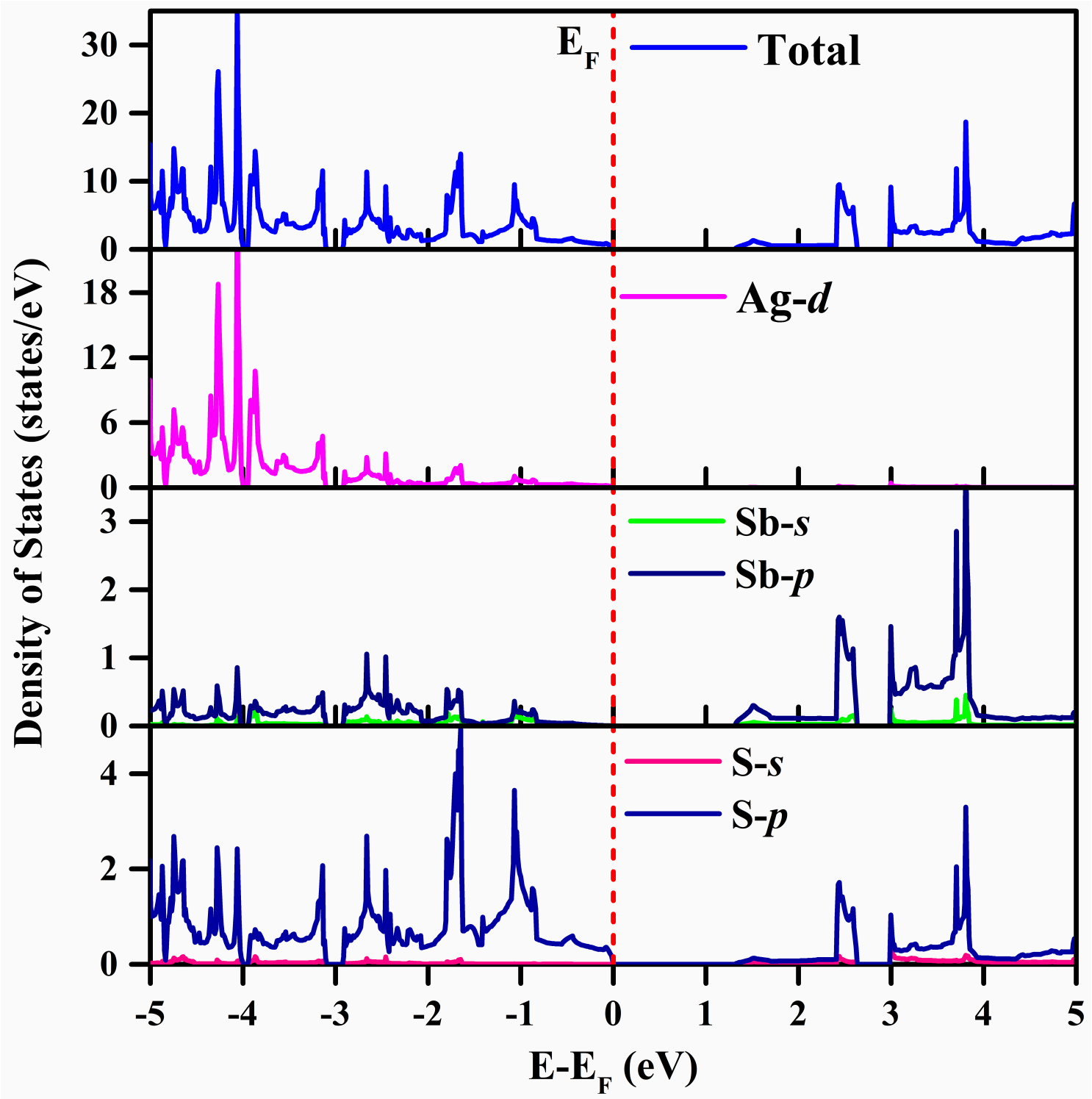


Figure 3

The total and partial density of States of $AgSbS_2$ using HSE06 functional. The Fermi level (E_F) is sets to 0 eV

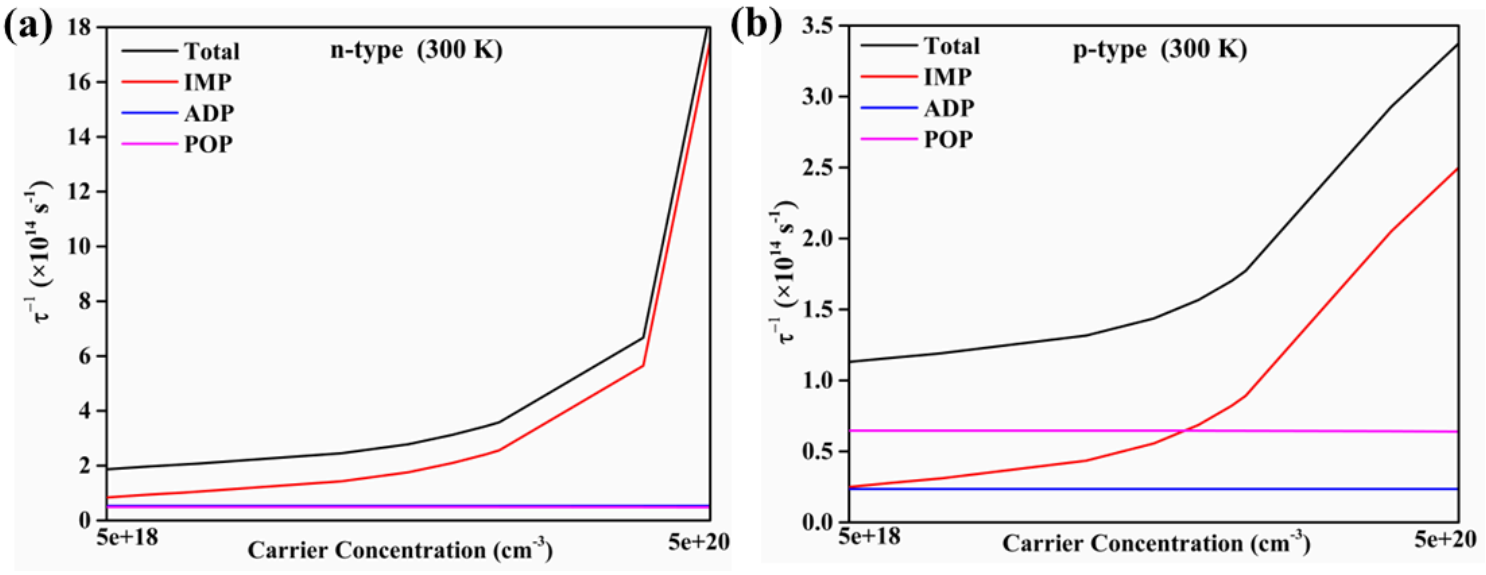


Figure 4

Scattering rates due to different scatterings by varying carrier concentration of $AgSbS_2$ for **(a)** p-type **(b)** n-type at 300 K

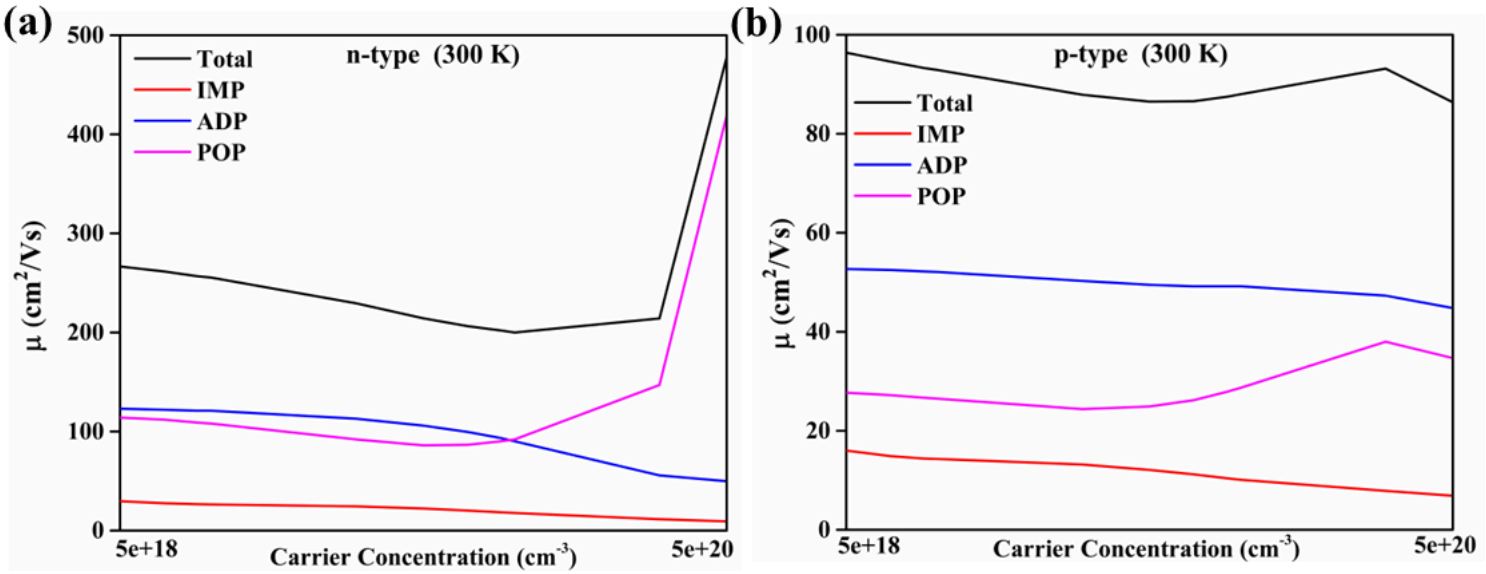


Figure 5

The carrier mobility due to different scatterings by varying carrier concentration for **(a)** p-type and **(b)** n-type $AgSbS_2$ monolayer at 300 K

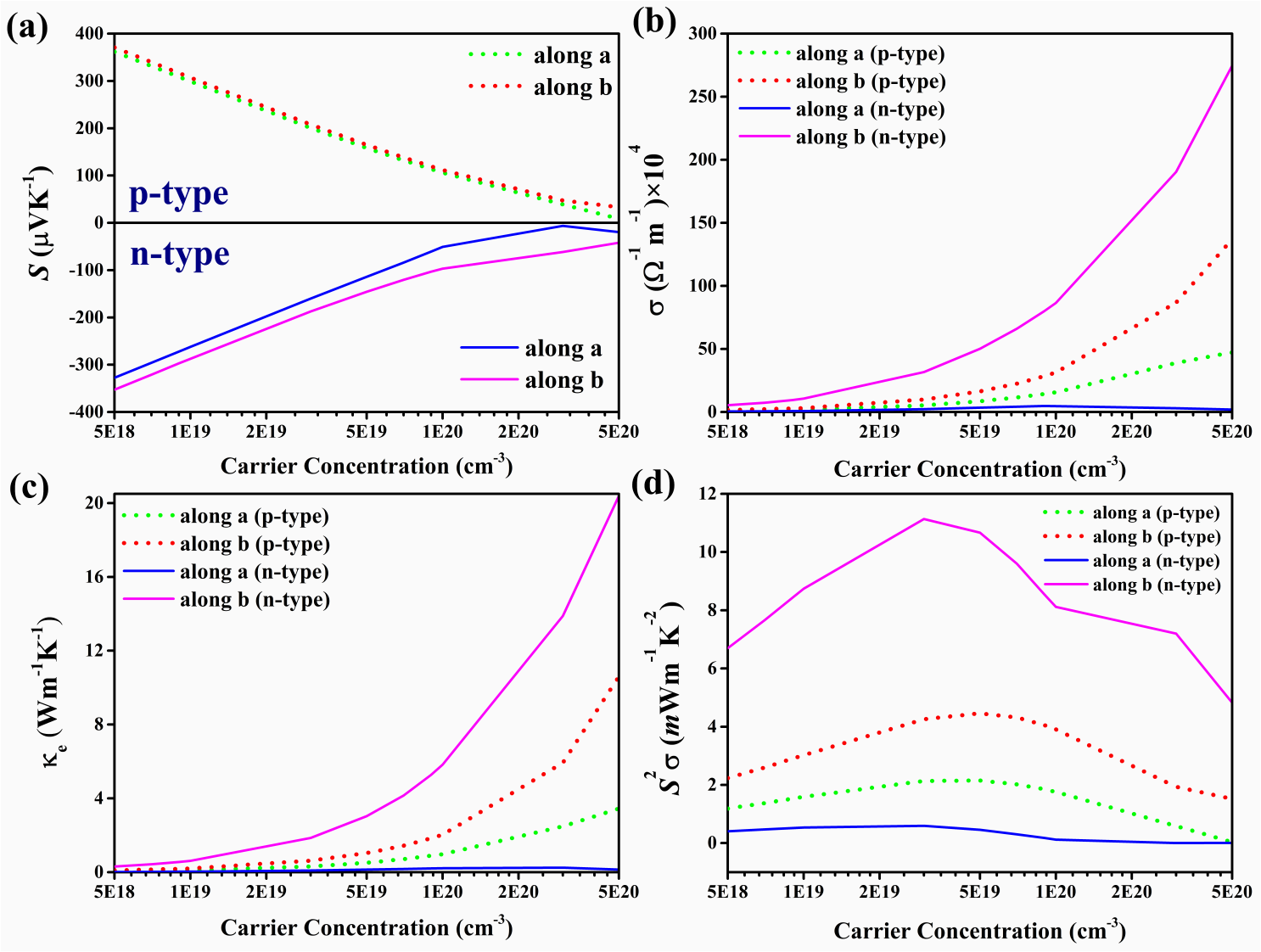


Figure 6

(a) Seeback coefficient, **(b)** electrical conductivity, **(c)** electronic contribution to thermal conductivity and **(d)** power factor for both p- and n-type $AgSbS_2$ monolayer along both *a*- and *b*-directions at 300 K

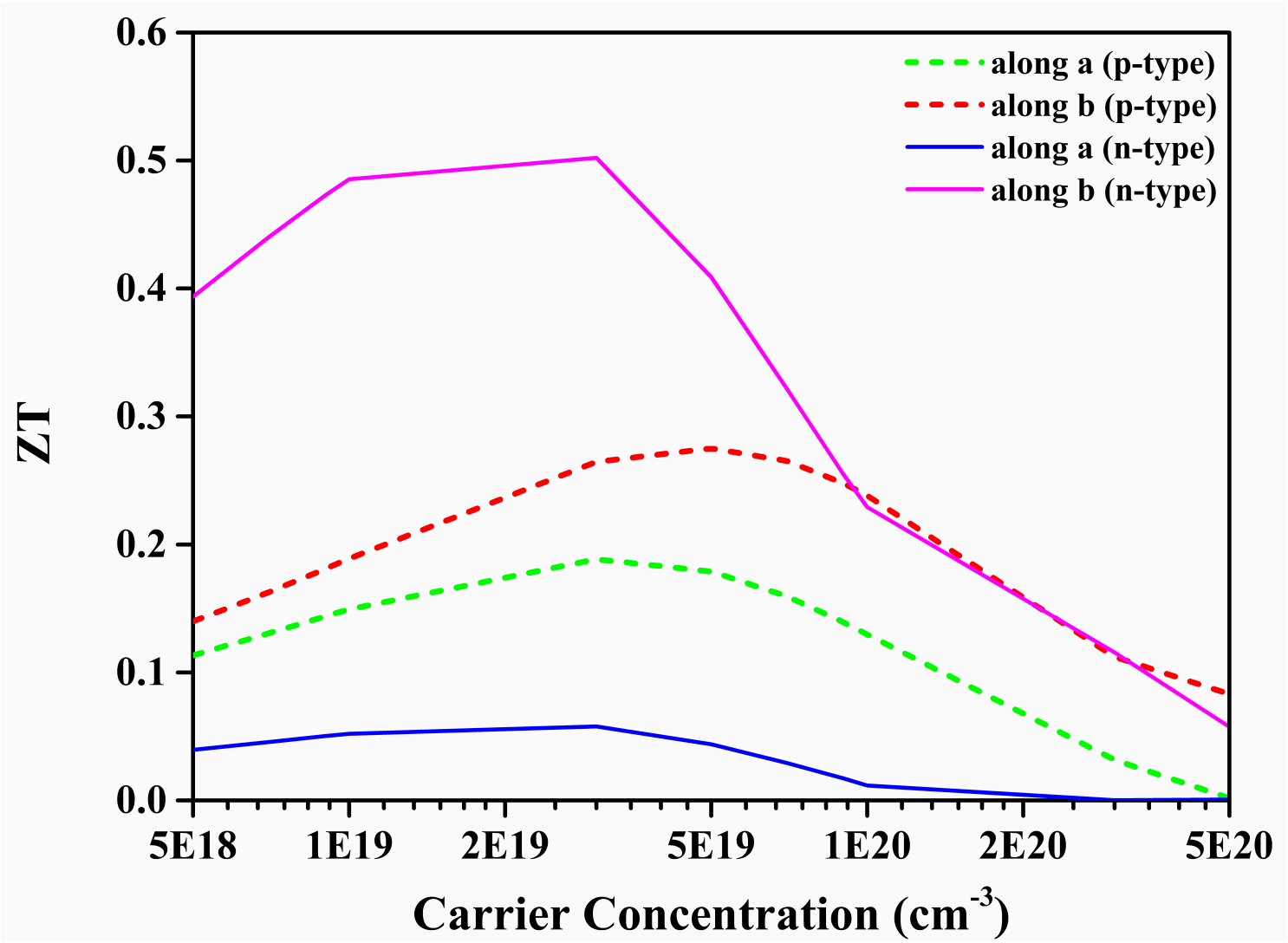


Figure 7

The ZT at 300 K in both *a*- and *b*-directions with the variation of carrier concentration for n- and p-type $AgSbS_2$ monolayer

Rapid #: -23205805

CROSS REF ID: **6420162**

LENDER: **IPL (Purdue University) :: Main Library**

BORROWER: **HUL (Harvard University) :: Widener Library**

TYPE:	Book Chapter
BOOK TITLE:	Advances in Digital Health and Medical Bioengineering
USER BOOK TITLE:	Advances in Digital Health and Medical Bioengineering
CHAPTER TITLE:	Predicting Breast Cancer Survival Rate Based on Genetic Data: A Machine Learning Approach
BOOK AUTHOR:	Yadav, Saanya
EDITION:	
VOLUME:	
PUBLISHER:	Springer Nature Switzerland
YEAR:	2024
PAGES:	393-399
ISBN:	9783031625015
LCCN:	
OCLC #:	
PATRON NOTES:	109

Processed by RapidX: 9/27/2024 9:50:33 AM

This material may be protected by copyright law (Title 17 U.S. Code)

Predicting Breast Cancer Survival Rate Based on Genetic Data: A Machine Learning Approach

Saanya Yadav and Yasha Hasija(✉)

Department of Biotechnology, Delhi Technological University, New Delhi, India
yashahasija@dtu.ac.in

Abstract. Breast cancer is the predominant form of cancer affecting women worldwide, and there is a relatively high success rate (70–80%) in treating individuals diagnosed with early-stage, non-metastatic breast cancer. Unfortunately, advanced breast cancer that has spread to distant organs continues to pose a significant challenge and is associated with a lower survival rate, despite the existing treatment modalities. However, since its molecular characteristics were thoroughly identified, There has been a profound transformation in the perception of breast cancer and now encompasses immunohistochemical markers, genomic markers, and immunomarkers, such as TIL and PD-L1. These markers include ER, PR, HER2, and proliferation marker protein Ki-67. This cancer can now be classified into variants using markers thanks to improvements in health care technology, which also make it possible to better detect and characterize the disease. These markers can then be used to predict treatment outcomes, risk of distant recurrence, and prognosis using an assortment of molecular and clinical variables. In this paper, we have tried building a machine learning model that can predict the rate of survival of a patient based on important genetic and clinical markers information. We analyzed the patients' records and have discovered interesting patterns amongst the patients. Upon further cleaning of the model and working with only clinical data, the best accuracy came with a Logistic Regression mode having 82.5% accuracy. These results can provide a deeper insight into the severity of the disease and can indicate the kind of treatment or therapy that is required or should be recommended to the patient, as the clinical data of patients is constantly measured. Not only such a model can be incorporated in the current treatment of patients, but can also be used to constantly track patient's progress. However, it is necessary to conduct more validation studies and clinical trials to evaluate the suggested model's resilience and generalizability in a larger patient population.

Keywords: Breast Cancer · ER · HER · Mutation · Machine Learning

1 Introduction

Breast cancer is a formidable illness that impacts countless women on a global scale. It is a type of malignancy that develops within the breast tissue, predominantly arising in either the milk ducts or lobules. The incidence of breast cancer has been on the rise in recent decades, making it a significant public health concern [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H.-N. Costin et al. (Eds.): EHB 2023, IFMBE Proceedings 109, pp. 393–399, 2024.
https://doi.org/10.1007/978-3-031-62502-2_46

The specific etymology of breast cancer remains elusive, but various risk factors have been characterized. Genetic predisposition plays a crucial role, with mutations in specific genes like BRCA1 and BRCA2 increasing the likelihood of getting breast cancer. Other hormonal factors, such as early menses, late menopause, and invasive hormone addition therapy, also contribute to the risk [2]. Other risk factors include obesity, excessive alcohol consumption, less exercise, and exposure to ionizing radiation.

Several diagnostic methods are employed to find breast cancer. Mammography is the most common screening tool, involving X-ray imaging of the breast tissue to identify abnormalities. Ultrasound imaging and magnetic resonance imaging (MRI) may be used as additional imaging techniques for further evaluation [3]. Biopsy, which involves the elimination of a small tissue sample for probation under a microscope, provides a definitive diagnosis of breast cancer and helps determine the specific type and stage of the disease.

Breast cancer has a profound impact on both individuals and society as a whole. For individuals, the diagnosis of breast cancer often brings fear, uncertainty, and emotional distress. Treatment regimens can be physically and emotionally demanding, affecting quality of life. Furthermore, breast cancer survivors may experience long-term side effects and face the constant fear of recurrence [4]. On a societal level, breast cancer awareness campaigns, research initiatives, and support networks have emerged to raise awareness, promote early detection, and provide support to those affected by the disease.

The dataset that we have taken here, METABRICS have some very important features relevant for Breast Cancer like HER 2 receptor status, ER status, Pam-50 values, grade for neoplasm histologic, given hormone therapy, integrative clusters, examination of lymph nodes etc. It also tells about mRNA z score, which is a statistical metric used to evaluate the expression levels of mRNA (messenger RNA) in a given sample compared to a reference population.

In this study, we propose a machine learning model for predicting the survival rate of breast cancer patients based on their genetic data. The model leverages a comprehensive set of genetic features, including gene expression levels, type of breast cancer surgery already performed, single nucleotide polymorphisms (SNPs), copy number variations (CNVs), size and stage of tumor, and hormone receptor status. By integrating these genetic features, our model aims to capture the complex and heterogeneous nature of breast cancer. Such a model can have many advantages-a newly diagnosed patient's chance of survival will clearly indicate the severity of the disease, analysis of a patient's data with ongoing treatment will indicate the patient's progress, among others. Thus, the objective of this paper is to develop a model which can predict a breast cancer patient's likelihood of survival on the basis of clinical records.

2 Dataset and Methods

2.1 Dataset

The METABRICS dataset was chosen which had 1904 records, which is available on Kaggle and is openly accessible. It is a comprehensive collection of molecular and clinical data from a massive assembly of breast cancer patients. Its integration of diverse data types and inclusion of different breast cancer subtypes make it a valuable resource

for studying the phylogenomics of breast cancer, identifying prognostic markers, and guiding personalized treatment approaches.

This dataset contained clinical information along with genetic information. There were m-RNA levels z-score for three hundred thirty one genes and mutations for one hundred seventy five genes.

2.2 Methodology

The data source was the METABRIC dataset, which is publicly available, which contains information about 700 attributes of upto 2000 breast cancer patients. While carrying out the initial analysis, the clinical and genomic data was segregated out and analyzed separately.

The first step involved looking for missing values in the dataset. The genomic data did not have any missing values whereas about 1100 data points were missing from the clinical data. These attributes were imputed with the respective modal values of the columns.

Next, the distribution of various columns was studied to understand the spread of values in different columns (refer Figs. 1 and 2). These graphs were plotted mostly with the help of python libraries (matplotlib and seaborn) and aided with the help of data visualization tools (Tableau).

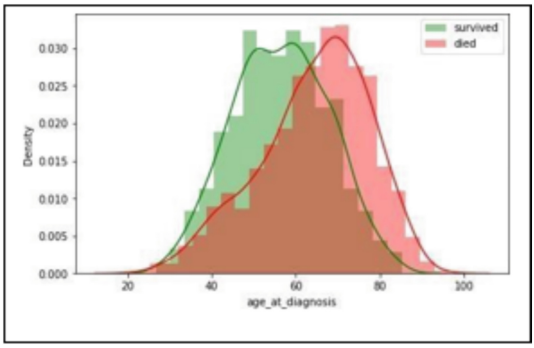


Fig. 1. Delayed diagnosis decreases chance of survival

After the analysis, a correlation matrix was formulated to gain better insights at the attributes which had the most relation with the outcome i.e. survival. Similar correlation matrices were formulated for mutations and genetic data. However, correlation values were found to be even lower than those found for clinical attributes, ranging between + 0.2 and −0.2.

Post-analysis of the different 3 types of data found in the dataset, the values were translated to numerical data type so as to mold the data to be read to be fed into machine learning models. This was done via encoding: both nominal and ordinal.

The data was then scaled down using the MinMaxScaler, which helped standardize the features. This was done to reduce the range in order to reduce the tendency of machine learning algorithms to assign incorrect weights to values.

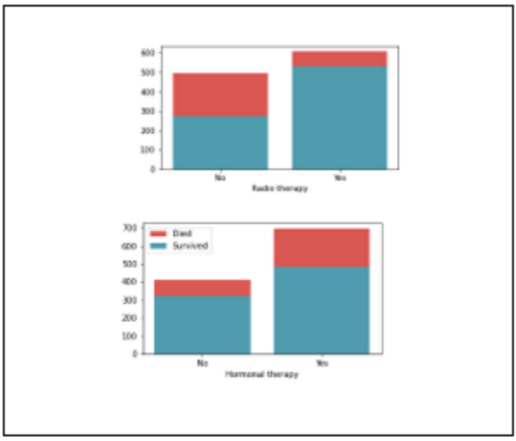


Fig. 2. Demonstration of patients who took therapy and survived

Before feeding the data to a machine learning algorithm, PCA (principal component analysis) was performed in order to reduce complexity in the dataset. With an aim to predict the survival rates, the model can make predictions either on the basis of instance-based methods or classification algorithms. As a result, we considered KNN, Logistic Regression and decision tree algorithms along with ensemble methods such as Random Forest and Extra Trees. The data was fed to these five algorithms. The results are shown in Fig. 3:

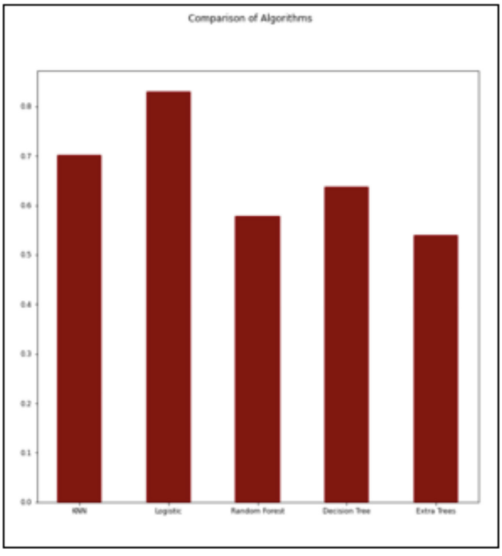


Fig. 3. Accuracy shown by various ML models

3 Results

In the data exploration phase of the project, while visualizing various clinical attributes, the findings matched with one's intuition of the subject. The machine learning model with best results across 5-fold cross validation was logistic regression, with accuracy up to 82.5%.

Some of the findings are shared below with their respective graphs (Figs. 4 and 5):

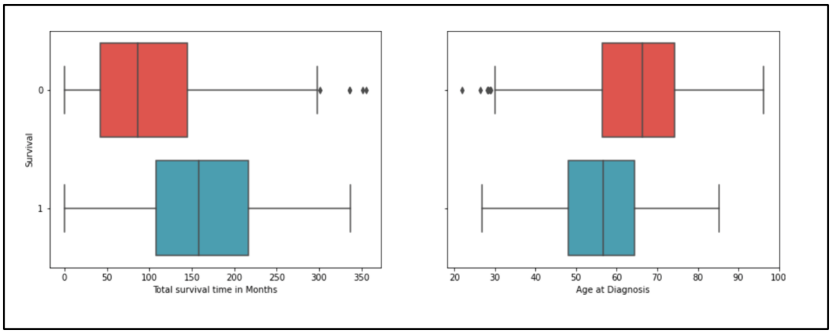


Fig. 4. Distribution of Survival time in months and age with target attribute

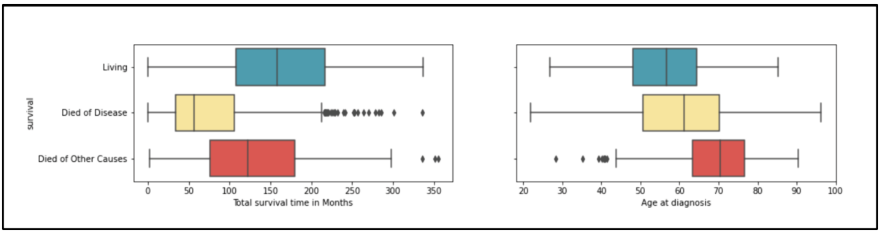


Fig. 5. Distribution of survival and recurrence with target attribute

- Median survival time in months of patients who died from breast cancer is low compared to the other two classes and it's distribution is right-skewed with a lot of outliers.
- Patients who died from other causes than cancer tend to be older than the other two classes. The distribution of it is left-skewed with some younger outliers.
- When the tumor is at lower stages, the chances of survival are greater compared to when the patient reaches the fourth stage.
- As the tumor stage increases, the tumor size increases (Fig. 6).
- When total survival time increases, the probability of survival increases (Fig. 7).
- The probability of dying from reasons other than cancer decreases slightly.

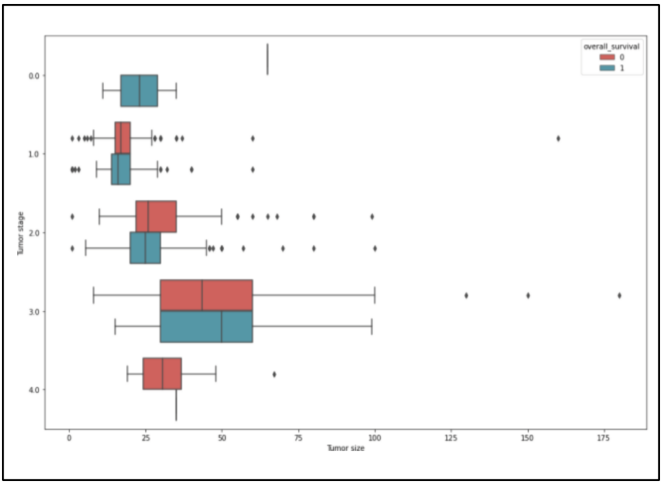


Fig. 6. Tumor stage vs tumor size and overall survival

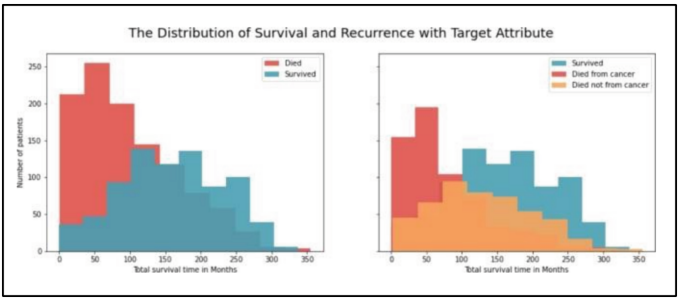


Fig. 7. The distribution of survival and recurrence with Target attribute

4 Conclusions

This study used machine learning techniques to analyze the clinical and genetic aspects of breast cancer survival. In comparison to previous algorithms, model evaluation using logistic regression produced results with higher accuracy. Variable selection procedures in healthcare research, particularly when utilizing machine learning techniques, may produce diverse outcomes depending on the dataset, geography, and patient lifestyle. In this way, our study has identified key factors impacting breast cancer patients' survival rates as well as the model performance, which can be applied in clinical practice. Building prediction apps for survival analysis of various diseases in the medical field can be done using visualizations of the data.

Such predictive models can be used to predict survival rate of cancer patients diagnosed along with the confidence values associated with the final prediction, which can be used to deliver better attention and care.

We would also like to acknowledge the support of Cillian Buckley in aiding our access to the data used in this paper and information about the original study.

References

1. Mahvi, D.A., Liu, R., Grinstaff, M.W., Colson, Y.L., Raut, C.P.: Local cancer recurrence: the realities, challenges, and opportunities for new therapies. CA Cancer J. Clin. **68**(6), 488–505 (2018). https://doi.org/10.3322/caac.21498
2. Clark, B.Z., Onisko, A., Assylbekova, B., Li, X., Bhargava, R., Dabbs, D.J.: Breast cancer global tumor biomarkers: a quality assurance study of intratumoral heterogeneity. Mod. Pathol. **32**(3), 354–366 (2019). https://doi.org/10.1038/s41379-018-0153-0
3. Peintinger, F., Reitsamer, R., Smidt, M.L., Kühn, T., Liedtke, C.: Lymph nodes in breast cancer - What can we learn from translational research. Breast Care **13**(5), 342–347 (2018). https://doi.org/10.1159/000492435
4. Kitamura, M., et al.: Progression potential of ductal carcinoma in situ assessed by genomic copy number profiling. Pathobiology **86**(2–3), 92–101 (2019). https://doi.org/10.1159/000492833

Uncorrected Proof

Journal of
Water & Climate Change

Check for updates

IWA PUBLISHING

© 2024 The Authors

Journal of Water and Climate Change Vol 00 No 0, 1 doi: 10.2166/wcc.2024.078

Prediction of aeration performance of different types of piano key weirs using different machine learning models

Niyati Narwal[a], Nipun Bansal[a], Deepak Singh [b,*] and Munendra Kumar[c]

[a] Department of Computer Science & Engineering, Delhi Technological University (DTU), New Delhi, India
[b] Department of Civil Engineering, NIT Delhi, New Delhi, India
[c] Department of Civil Engineering, Delhi Technological University (DTU), New Delhi, India
*Corresponding author. E-mail: kumaindeepaksingh@gmail.com

DS, 0000-0001-9397-3220

ABSTRACT

Aeration is a cost-effective and efficient method for increasing the available oxygen or dissolved oxygen content in water bodies, which is crucial for the existence of the aquatic life. However, conventional techniques for estimating aeration in different hydraulic structures are time-consuming and incorrect ways to approximate aeration. Therefore, new, computationally more efficient, and more accurate methods are required. In this article, three machine learning models are presented: (1) ELM (extreme learning machine) model, (2) online sequential extreme learning machine model, and (3) I-ELM (incremental extreme learning machine) model. These models assess the air conditioning capacity of the three variants of PKWs, denoted as A, B, and C, about Cd, Cs, and Cu, which are the three most important parameters for aeration efficiency at different temperatures. The model performance is evaluated and compared based on mean squared error, root-mean-square error, correlation coefficient, mean absolute error, and Nash–Sutcliffe efficiency. This research concludes that I-ELM is the best-performing model for complete available data that are time invariant.

Key words: aeration efficiency, dissolved oxygen content, ELM, I-ELM, OS-ELM, PKWs

HIGHLIGHTS

- Aeration is crucial for maintaining water quality, particularly in rivers and streams.
- The utilization of multiple machine learning models enables to predict aeration efficiency over various types of piano key weirs.
- This study offers a valuable tool for optimizing the design and advancement of piano key weirs.

1. INTRODUCTION

Aeration is the process that alludes to air incorporation into things such as water and soil. In the process of water treatment, aeration enriches oxygen levels in the water and purifies water by removing impurities. Importantly, cavitation does not occur through aeration. Aeration supports hydraulic structure design and sewerage to ensure their good performance and environmental sustainability surely. Dissolved oxygen (DO) required for different species of aquatic organisms depends on species, habitat type, and specific ecological considerations. Crabs and oysters that live at the bottom need a minimum of 4 mg/L DO, while shallow-water fish require up to 4–15 mg/L DO for survival. The necessity of DO in aquatic organisms relies upon factors like temperature of water, salinity, nutrient concentration, as well as the presence of pollutants (Baylar & Bagatur 2000; Emiroglu & Baylar 2003). In general, aeration is the process that enhances the DO level in water bodies and is crucial for the existence of the aquatic life. It prevents the formation of anaerobic conditions (i.e., those that produce harmful gases like methane or hydrogen sulphide). It saves our environment and also plays a crucial role in climate change.

In hydraulic engineering, aeration serves to improve the flow characteristics of water by reducing its weight so that it can become more buoyant (Baylar *et al.* 2011). This prevents vapour pockets as well as reduces pressure fluctuation promoting efficiency, stability, and safety for hydraulic structures (Singh & Kumar 2022b). Hydraulic structures are built to decrease cavitation and increase oxygen content in flowing water (Rathinakumar *et al.* 2014) during extreme weather events such as heavy rainfall and droughts. Weirs and other water management structures can be used to manage water flow, reduce

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/).

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

the risk of flooding, and maintain water levels during droughts. Properly managed weirs can also help mitigate the impacts of temperature increases on water quality by promoting aeration and maintaining oxygen levels. The last 10 years have seen growing interest in the study of aeration performance of various free flow structures such as triangular, rectangular, labyrinth, and PKWs for efficiency improvement and stability reasons. These studies aim to understand the flow dynamics of different types of weirs and how they affect the aeration and mixing of air in the water while the flow occurs over such types of weirs.

Although the water flows through the hydraulic structure for a short duration, the DO level increases significantly due to turbulence flow that facilitates the mixing of air with flowing water (Emiroglu & Baylar 2003; Guenther *et al.* 2013). However, these effects are transient, and the DO concentration returns to the normal level as the water flows in plane areas that are far away from the hydraulic structure (Wormleaton & Tsang 2000). This phenomenon mitigates greenhouse gas emissions and prevents the Earth's surface temperature from rising.

The efficient use of a weir for aeration in rivers or streams was initially shown by Gameson (1957). Weirs are known to stir the water and cause turbulence, which raises the concentration of DO. The work by Baylar & Bagatur (2000) was a landmark in the fields of hydrology and for improving the water quality. The effectiveness cost of the weir aeration depends on many factors such as the weir's size and placement, water flow rate, desired level of the aeration, and the cost of equipment and maintenance (Baylar *et al.* 2001).

This study has several advantages as a continuation of earlier works on the impact of hydraulic structures on water quality. Experimental studies (Apted & Novak 1973; Ervine & Elsawy 1975; Gulliver & Rindels 1993; Baylar & Bagatur 2000; Baylar *et al.* 2001) examined the aeration capabilities of various types of weirs and found that the labyrinth weir exhibits a higher entrainment rate compared with the linear weir (Avery & Novak 1978; Nakasone 1987; Wormleaton & Soufiani 1998; Emiroglu & Baylar 2003). Many attempts have been made to estimate the ratio of air/water flow and specifically aeration performance of various hydraulic structures like weirs in detail (Bagatur & Sekerdag 2003; Baylar & Ozkan 2006; Baylar *et al.* 2010, 2011). From a range of PKW experiments, it was discovered that the water with low heads flowing down through these weirs gets naturally aerated. In fluid dynamics, 'nappe' represents a sheet of water passing over a structure such as a weir or a dam. The characteristics of the nappe while flowing is another important aspect to consider the design of weirs based on vibrations, the flow of nature that could be laminar or turbulent, fluctuations, or surging of the flow (Baylar *et al.* 2011; Jaiswal & Goel 2019; Jaiswal & Goel 2020).

Crookston & Tullis (2012) performed research on the labyrinth weir by examining four different ways that the nappe was ventilated: clinging, aerated, partially aerated, and drowned. Their examination presented the understanding of the processes involved in the creation of nappes. Moreover, the research emphasizes the substantial impact of fluctuating negative pressures beneath the overspill nappe, underscoring their capacity to trigger seismic tremors. In addition, researchers examined the consequences of intentionally increasing the oxygen content in the flow of the nappe (Leite Ribeiro *et al.* 2007). Usually, the size of the nappe depends on how high it falls and how much it spills over the top. When dealing with bigger runs or narrow outlet keys, there can be conflicts between water from lateral ridges, but these conflicts would not significantly affect the efficiency of the output if flow restrictions are kept constant (Machiels 2012).

In addition, Jaiswal & Goel (2020) applied various soft computing methods to forecast air movement around triangular weirs. They noticed that the Gaussian process and M5P techniques accurately predict the amount of oxygen added. Moreover, other approaches include adaptive neuro-fuzzy inference systems (ANFIS) and linear regression indicating PKW oxygenation (Komal *et al.* 2017).

In the specific field of hydraulic engineering, the scope of using soft computing techniques for the prediction of aeration efficiency for different hydraulic structures has been increased recently. The reason for their choice is that they are intelligent, fast, and precognitive. These methods find favour on account of their intelligence, speed computation, and predictive ability. If implemented, these models would reduce scale effects and simplify the processes involved in model building. In water resources management and hydraulic engineering, different tools of soft computing techniques have been used (Luxmi *et al.* 2022; Srinivas & Tiwari 2022). Machine learning and soft computing methodologies have recently received attention from more researchers regarding estimating aeration potentials for different weirs (Verma *et al.* 2022).

Bayler *et al.* (2011) investigated using gene expression programming (GEP) to estimate the oxygen transfer efficiency in stepped cascade aeration. Singh *et al.* (2021) attempted to analyse and assess labyrinth weir aeration efficiency by adopting diverse methods like the M5P model and support vector machine. Kumar *et al.* (2020) implemented a plunge jet analysis using artificial neural networks. Some of the other tools used for this purpose include generalized regression neural network, ANFIS, and multivariate adaptive regression splines and many others, especially for estimating oxygen transfer rates into

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

water and lakes. It is with proper recognition of these facts that studies by Baylar *et al.* (2009a, 2009b) have shown that it is feasible to apply a support vector machine for predicting the efficiency of aeration by plunging overfall jets from weirs. Other forms of machine learning techniques such as K-nearest neighbours were also tried in the process of aeration performance estimation using for predicting aeration performance using Parshall flumes also included random forest regression (Sangeeta *et al.* 2021). To derive these models, Aradhana *et al.* (2021) sought to identify the machine learning-based methodologies, which were adaptable for subsequent analysis of aeration capability of labyrinth weir. Bansal *et al.* (2023) performed energy calculations on PKW with the help of extreme gradient boosting (XGBoost) and GEP. A recent literature review indicates that many machine leaning techniques have been applied to predict the performance of different hydraulic systems in terms of aeration. By promoting aeration, weirs help to improve water quality by encouraging the breakdown of organic materials through aerobic processes. This can reduce the levels of pollutants and prevent the growth of harmful algae, which thrive in low-oxygen environments.

In this continuation, this article used three different machine learning techniques to predict the aeration performance of various kinds of PKWs, i.e., type A, B, and C (refer to Figure 1). In this work, the models are built by utilizing the data collected by Singh & Kumar (2022a). The various parameters C_d, C_s, and C_u are used in the training procedure of extreme learning machine (ELM), online sequential extreme learning machine (OS-ELM), and I-ELM. The aeration efficiency is considered a target parameter. Ultimately, the authors suggested comparing ELM-, OS-ELM-, and I-ELM-based strategies to conventional techniques.

2. MATERIALS AND METHODS

2.1. Laboratory setup and dataset used

This study makes use of data published by Singh & Kumar (2022a) on the aeration performance of various types of PKWs. Singh & Kumar (2022a) performed a series of laboratory experiments using a horizontal flume of a rectangular cross-section of width 0.516 m, height 0.6 m, and length 10 m. The discharge flow rate was determined using an electromagnetic flowmeter. Also, its accuracy tolerance should be within the range of $\pm 0.2\%$. The aeration efficiency and flow regime over the PKWs were investigated. To ensure proper aeration throughout the operation, the level of the water pool downstream is kept above the bubble's deepest penetration. The DO & temperature were measured upstream (u/s) and downstream (d/s) of the PKW using a Thermo-Scientific Orion Star A223 DO Portable Meter (Figure 2) that is calibrated. The water that flows over the weir is clean. The storage tank is filled with clean water before starting each experiment. The details of the experiment setup for the different PKWs are presented in Table 1.

2.2. Methodology

The quantity of air that gets dissolved in the water determines aeration efficiency. As water flows through hydraulic structures, the oxygen concentration rate changes with time in the air–water phase and may be described as follows (Gameson 1957; Gulliver *et al.* 1990):

$$\frac{dm}{dt} = V\frac{dC}{dt} = K_L A(C_s - C) \tag{1}$$

where dm/dt is the gas molecule mass transfer rate, K_L is the coefficient of liquid-mass transfer, C is the degree of oxygen concentration, A is the area of the surface that is available for aeration, V is the volume of water with which gas molecules

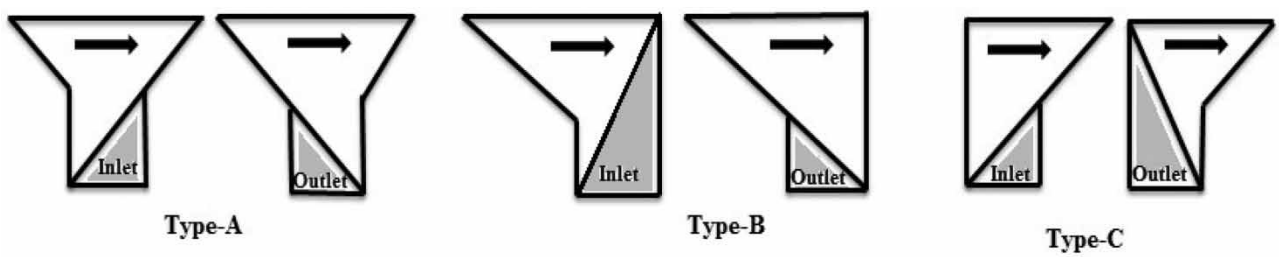


Figure 1 | PKW (types A, B, and C) (derived from Lempérière *et al.* 2011).

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

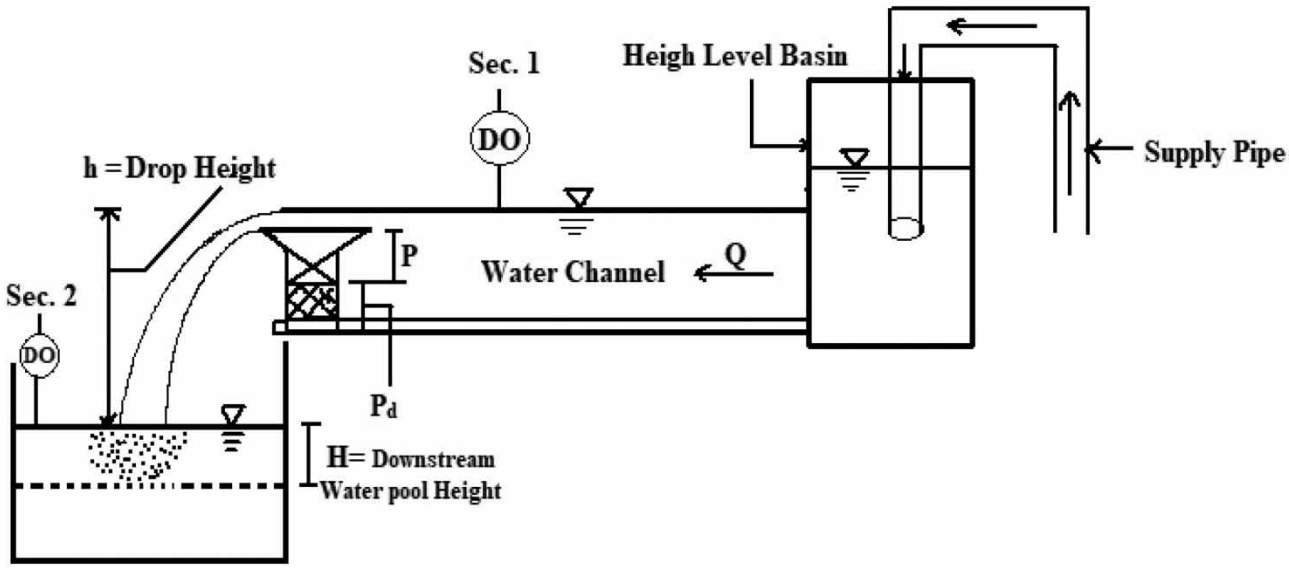


Figure 2 | Laboratory schematic PKW aeration apparatus (adopted from Singh & Kumar 2022a).

Table 1 | Range and different specifications of the experimental setup taken to collect data for different PKWs

Model No.	Range of Q (m^3/s)	$\frac{W_i}{W_o}$	W_i (m)	W_o (m)	P (m)	$\frac{L}{W}$	B (m)	B_i (m)	B_o (m)	Range of drop height h (m)	Range of aeration efficiency (E_{20})	N (no. of cycles)
PKW-A	0.003–0.0155	1.28	0.088	0.069	0.20	6	0.427	0.142	0.142	0.20–0.40	0.185–0.983	3
PKW-B	0.003–0.0155	1.28	0.088	0.069	0.20	6	0.427	0	0.284	0.20–0.40	0.157–0.631	3
PKW-C	0.003–0.0155	1.28	0.088	0.069	0.20	6	0.427	0.284	0	0.20–0.40	0.151–0.771	3

transfer, C_S is the concentration at saturation point in equilibrium or ambient conditions, and t is the time. The aeration efficiency is determined by integrating Equation (1) and is given as follows:

$$E = \frac{C_d - C_u}{C_s - C_u} = 1 - \frac{1}{r} \tag{2}$$

where C_d, C_u, and C_s are the concentrations of DO downstream of the hydraulic framework, DO upstream of the hydraulic framework, and the DO at a saturated level, respectively, and r is the oxygen deficit ratio.

The aeration efficiency $E = 0$ and 1.0 indicate that zero transfer and the transfer up to the saturation point had appeared at the structure, respectively. Researchers employ a temperature adjustment factor to assess aeration efficiency because temperature significantly impacts aeration efficiency. The aeration efficiency is adjusted at 20 °C by creating the mass exchange comparability relationship and denoted as E_{20} (Gulliver *et al.* 1990).

$$E_{20} = 1 - (1 - E)^{1/f} \tag{3}$$

where the exponent f is expressed as follows and depends on instantaneous temperature.

$$f = 1 + 0.02103(T - 20) + 8.261 \times 10^{-5}(T - 20)^2. \tag{4}$$

where T denotes the measured temperature in degrees Celsius. The machine learning models are fed with C_d, C_s, and C_u at different temperatures for all three PKWs to calculate their aeration efficiencies. The different machine learning models applied are explained in the following subsection.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

2.3. ELM and its application

ELM is a kind of feedforward neural network initially suggested by Huang *et al.* (2004, 2006). It is designed for simplicity and fast training. It has been demonstrated that the typical ELM with additive or RBF activation function (Huang *et al.* 2006, 2007, 2008) can approximate any function universally. These models have been applied to many problems, including classification, regression, and feature extractions (Rong *et al.* 2008; Huang *et al.* 2010; Wang *et al.* 2011; Lim *et al.* 2013). The ELM model's concept is to randomly assign hidden biases and input weights, which are solved analytically for the output weights, instead of using iterative optimization methods like backpropagation, gradient descent, etc. The model is applied to predict the aeration efficiency E for a given value of C_d and C_u at different heights of PKWs and C_s at temperature 33°C.

The structure of the ELM is depicted in Figure 3, where ω signifies the input weights assigned between the input and the hidden layers. The threshold for the hidden layer node is denoted by b. Initially, both the input weights and biases have been selected randomly, and β symbolizes the output weight that bridges the output layer to the hidden layer, which is determined analytically. I, H, and P denote the number of nodes in the input, hidden, and output layers, respectively. Y is the prediction of the model.

The given dataset $D = \{(x_i, t_i)|x_i \in R^I, t_i \in R^P\}^N$ in the ELM model with H nodes of the hidden layer and with activation function can be designed as follows:

$$\sum_{i=1}^{H} \beta_i G_i(x_j) = \sum_{i=1}^{H} \beta_i G(a_i.x_j + b_i) = t_j,\ j = 1, 2, \ldots N \tag{5}$$

where $a_i = [a_{i1}, a_{i2}, \ldots, a_{iI}]^T$ is the weight vector that bridges the input node and the ith hidden node, and b_i signifies the threshold of the ith hidden node. The weight vector $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{iP}]^T$ connects the ith hidden node to the output nodes. $a_i.x_j$ signifies the inner product. The activation functions include 'Sigmoid' and 'Sine'.

Equation (5) can be written as follows:

$$L\beta = T \tag{6}$$

where $L = \begin{bmatrix} G(\omega_1.x_1 + b_1) \ldots\ldots\ldots\ldots G(\omega_H.x_1 + b_H) \\ \vdots \quad \ldots\ldots\ldots\ldots \quad \vdots \\ G(\omega_1.x_N + b_1) \ldots\ldots\ldots\ldots G(\omega_H.x_N + b_H) \end{bmatrix}_{N\times H}$ and $\beta = \begin{Bmatrix} \beta_1^T \\ \vdots \\ \beta_H^T \end{Bmatrix}_{H\times P}$, $T = \begin{Bmatrix} t_1^T \\ \vdots \\ t_N^T \end{Bmatrix}_{N\times P}$.

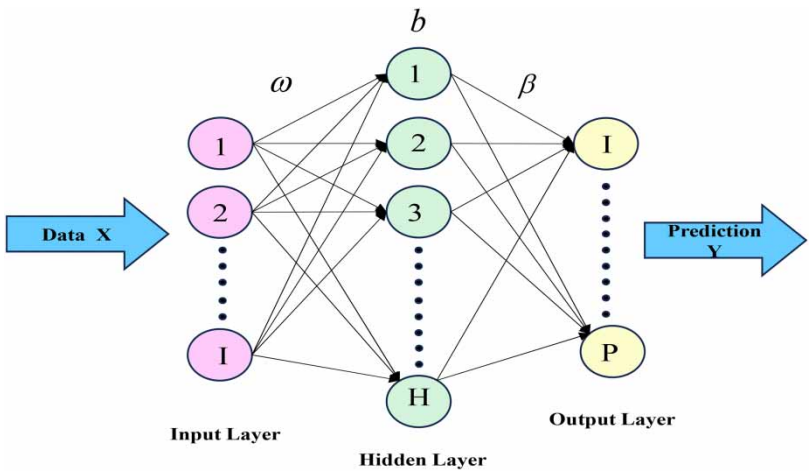


Figure 3 | ELM model structure with a single hidden layer.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

The optimization object of the ELM model is given as follows:

$$\min_{||\beta||}(\min\sum_{i=1}^{N}||t_i - o_i||^2) \tag{7}$$

where o_i is the prediction made by the ELM model for the ith sample with the sample having an accurate label t_i. The variable N stands for the total number of samples in the dataset. The optimization challenge at hand can be effectively addressed using the Moore–Penrose generalized inverse method, which is expressed mathematically as follows:

$$\beta = L^{-1}T \tag{8}$$

2.4. OS-ELM and its application

Liang *et al.* (2006) suggested an online version of ELM. OS-ELM is a modification of ELM algorithms. Single hidden layer feedforward neural networks, such as online sequential ELM model and ELM model, are well known for their ability to understand complex nonlinear functions. The critical difference between the two is that the OS-ELM is designed for online learning, while the ELM is intended for batch learning. All training data are fed once in ELM, while in OS-ELM, large training data are fed in modules. OS-ELM is used in case of memory constraints. OS-ELM is relatively fast in learning as ELM, while it has an advantage over ELM in learning from the newly arrived data. The hidden layer weights of OS-ELM are created at random and maintained as constant, whereas the adjustment of output weights takes place incrementally with the arrival of new data. This enables the model to acquire knowledge rapidly and adjust to shifting trends in the data. OS-ELM can use existing data to update the model, without having to train the whole dataset again, which can be faster and cheaper. OS-ELM can be used for various tasks such as recognizing images and sounds, analysing biological data, and controlling systems. The specifications of algorithms are outlined below. The OS-ELM algorithm's mathematical equations can be classified into two steps: the initialization step and the online learning step.

Initialization step:

- A set of training data is fed to start the training process, involving N_O unique and arbitrary samples such that $N_O > H$.
- Input parameters are assigned randomly, as in the ELM model, and initial the output matrix is computed:
  L_O from the hidden layer
- Calculate the initial output weight matrix from Equation (8):

$$\beta_O = L_O^{-1}T_O$$

- Set $P_O = (L_O^T L_O)^{-1}$ and $k = 0$, where k represents the count of data chunks.

Online learning step:

- For N_{k+1} distinct arbitrary samples, the given $(k+1)$th segment of new training data is considered.
- Determine the hidden layer output matrix L_{k+1} similarly as described in Equation (6).
- Determine the output matrix β_{k+1} by applying the following mathematical relation:

$$\beta_{k+1} = \beta_k + P_{k+1}L_{k+1}^T(T_{k+1} - L_{k+1}\beta_k)$$

where $P_{k+1} = P_k - P_k L_{k+1}^T (I + L_{k+1}P_k L_{k+1}^T)^{-1} L_{k+1}P_k$.

The algorithm used in OS-ELM is as follows:

Input:

H: hidden nodes count;
$K+1$: count of data chunks;
The training data chunk is denoted by $\{(X_k, T_k)|k = 0, \ldots, K\}$:

$$(X_k, T_k) = \{(x_j, t_j)|x_j \in R^I, t_j \in R^P),\ j = 0, 1, \ldots N_k.$$

$$N_k = |(X_k, T_k)|$$

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

Outputs:

1. //Initialization;
2. for ($i = 1$ to H);
3. set input weight ω_i randomly;
4. set hidden layer bias b_i randomly;
5. compute L_0 (the hidden layer output matrix);
6. determine $\beta_o = L_O^T T_O$ (the output weight matrix);
7. $P_O = (L_O^T T_O)^{-1}$; $k = 0$;
8. //Learning process;
9. while k is less than $K(k \leq K)$;
10. determine the output matrix of the hidden layer L_{k+1}; and

$$P_{k+1} = P_k - P_k L_{k+1}^T (I + L_{k+1} P_k L_{k+1}^T)^{-1} L_{k+1} P_k$$

11. calculate the output weight matrix β_{k+1};
12. $\beta_{k+1} = \beta_k + P_{k+1} L_{k+1}^T (T_{k+1} - L_{k+1} \beta_k)$;

$$k = K + 1$$

2.5. I-ELM and its application

The incremental ELM is an extension of the ELM algorithm that is introduced (Huang & Chen 2008). It is designed to handle incremental learning. A limited number of hidden nodes are initially present in the model. The hidden layer's input weights and biases are chosen at random. The model is trained on the first batch of the data samples using the ELM training method. The new hidden nodes are added, and adjustments in the biases and the input weights of the already current hidden nodes have been updated incrementally. It uses simple linear least-square methods to calculate the output weight, considering the new data samples and previously learned data.

In addition to being able to handle the nonlinearity of the data and high dimensionality in problems with large scale, it can also quickly and efficiently adapt to new data samples. However, some drawbacks include its dependence on initial conditions and the need for fine-tuning for optimal performance. The structure of the incremental ELM model is given in Figure 4. It consists I (inputs), H (hidden nodes), and P (outputs).ω_i represents the weight vector of $(1 \times I)$ order for the currently hidden layer neuron, whose elements are evenly spaced random numbers between $[-1, 1]$.

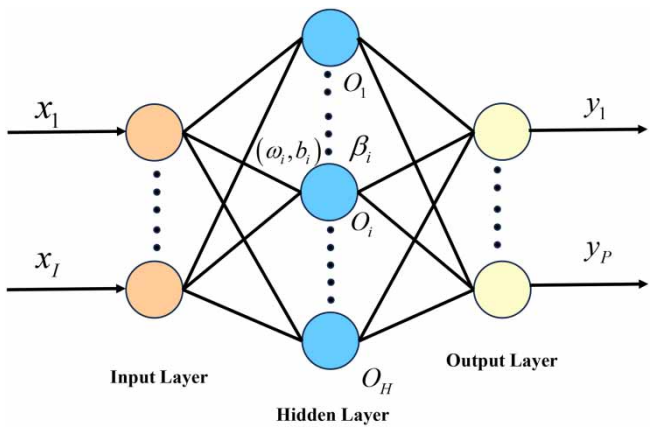


Figure 4 | The structure of incremental extreme learning machine model with I inputs, H hidden nodes, and Poutputs.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

b_i is the ith's node bias and its value is taken randomly from the evenly divided set [−1, 1] when the hidden layer's activation function is additive and calculated as given in Equation (9):

$$g(x) = \frac{1}{1 + e^{-x}} \tag{9}$$

where x represents an input vector. β denotes the output weight vector of the order $(1 \times P)$. The dataset to train the model $\{(X, Y)\}$ is considered, where X represents I inputs in each Ndataset and Y represents P output of each N dataset. I-ELM algorithm's training steps are outlined here as follows:

Step 1: Initially, let $I = 0$ and H are the maximum hidden nodes. The predicted accuracy is predetermined to be $\in$. The output Y is set as the initial values for the residuals E, which represents the variance between the network's actual and desired outputs.

Step 2: Steps taken to train the model: While $I < H$ and $||E|| > \in$.

1. $H = H + 1$, where H is the count of hidden nodes.
2. The newly expanded hidden layer neuron O_i's input weight ω_i and bias b_i are generated at random.
3. Determine the output of the activation function $g(x')$ for O_i.

$$x' = \omega_H X + b_H \tag{10}$$

4. Determine the output vector $\bar{H}$:

$$\bar{H} = g(x') \tag{11}$$

5. Determine the output weight for O_H as given in Equation (12):

$$\bar{\beta} = \frac{E.\bar{H}^T}{\bar{H}.\bar{H}^T} \tag{12}$$

6. Once the new hidden node has been increased, the residual error is determined as follows:

$$E' = E - \bar{\beta}\bar{H} \tag{13}$$

The output weight, as determined in Equation (12), decreases the error of the network quickly. Repeat the previous steps until the residuals are less than the expected error ϵ. Else, the training process should be restarted when $H > I$ and the error is more than the anticipated error $\in$. Restarting the process of training is advised since it is typically because of the random calculation of the input weights (i.e.,ω_H) and biases (i.e., b_H). The user can check to see if the trained network satisfies the requirements using the provided testing set (X', Y').

2.6. Model performance metrics

The performances of the proposed models are determined using five different statistical measures: (1) mean squared error (MSE), (2) root-mean-square error (RMSE), (3) mean absolute error (MAE), (4) correlation coefficient (CC), and (5) Nash–Sutcliffe efficiency (NSE). The MSE determines the average square of the error, which is the deviation between the actual values and predicted ones. A better model will have a value of MSE close to 0. MSE is calculated as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(E_{\text{Actual}} - E_{\text{Predicted}})^2 \tag{14}$$

where N is the count of data points in the given dataset, and E_{Actual} and $E_{\text{Predicted}}$ are actual and predicted aeration efficiency, respectively.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

RMSE is utilized to calculate the error between actual and predicted values. The value falls within the range of 0 and 1. A value nearer to 0 results in a better and well-fit model. RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(E_{\text{Actual}} - E_{\text{Predicted}}\right)^2} \tag{15}$$

MAE calculates the mean of the absolute error, and its lower value indicates a better prediction of a given machine model.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|E_{\text{Actual}} - E_{\text{Predicted}}\right| \tag{16}$$

The CC indicates the relation between two variables' relative movement. It ranges from -1 to $+1$, whereas a correlation of 0 means no relationship between two variables. The CC is calculated as follows:

$$\text{CC} = \frac{N\sum_{i=1}^{N} E_{\text{Actual}}E_{\text{Predicted}} - \left(\sum_{i=1}^{N} E_{\text{Actual}}\right)\left(\sum_{i=1}^{N} E_{\text{Predicted}}\right)}{\sqrt{\left[N\left(\sum_{i=1}^{N} E_{\text{Actual}}{}^2\right) - \left(\sum_{i=1}^{N} E_{\text{Actual}}\right)^2\right]\left[N\left(\sum_{i=1}^{N} E_{\text{Predicted}}{}^2\right) - \left(\sum_{i=1}^{N} E_{\text{Predicted}}\right)^2\right]}} \tag{17}$$

The NSE is a statistical measure utilized to measure the effectiveness of environmental and hydrological systems, representing the ratio of the model's error to potential error. The value of NSE is from $-\infty$ to 1. A value closer to 1 represents a perfect match of the model to the actual data. The NSE is calculated as follows:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(E_{\text{Actual}} - E_{\text{Predicted}}\right)^2}{\sum_{i=1}^{N}\left(E_{\text{Actual}} - \text{Average of }(E_{\text{Actual}})\right)^2} \tag{18}$$

The aforementioned five metrics compare the degree of accuracy of the applied machine learning techniques. The section on simulation and results presents a comparison of the outcomes. The aforementioned three described models are trained and simulated for the given dataset, and results are presented in the following section.

3. SIMULATION AND RESULTS

The fundamental concept of fluid discharge over a Piano Key Weir involves three primary elements: flow across the inlet key, lateral flow, and outlet key along the side crest (Machiels 2012; Khassaf *et al.* 2015). These components work together, leading to a complex three-dimensional flow pattern illustrated in Figure 5. Water tends to flow more readily from the side crest toward the outlet key as the head increases, reducing hydraulic efficacy until two discharging streams merge. Consequently, the PKW exhibits behaviour akin to a linear weir.

Observation of water movement along the edge of the crest displays two discernible patterns. The initial pattern, closer to the upstream region, shows no signs of aeration. In contrast, the second pattern, located far from the sidewall crest, exhibits clear aeration. This aeration zone becomes more pronounced as the water discharge progresses downstream. As depicted in Figure 5, the flow over the PKW structure has a high degree of ventilation and a three-dimensional nature. Notably, areas of splashing and spraying are evident both inside the outlet keys and in the vicinity of the base of the structure (Singh & Kumar 2022b).

Concerning the trajectory of H_t(head), there is a gradual expansion in the area where spraying and sprinkling occur. The planar jet starts far from the crest. In contrast, the region of air circulation undergoes a substantial increase as H_t rises. This expansion is partly attributed to the rise in local velocity, which results in elevated and turbulent mixing and advection levels.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

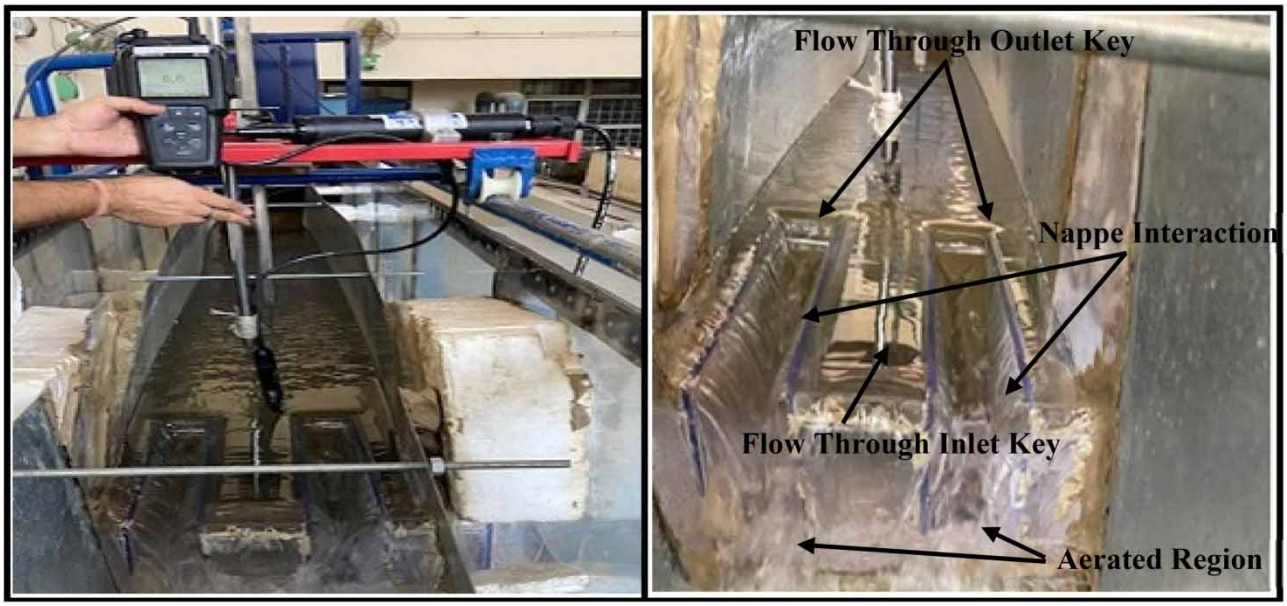


Figure 5 | Flow pattern over PKW.

This particular flow pattern enhances the aeration efficiency, as the intricate flow behaviour leads to an increase in air bubbles over a prolonged duration and distance (Eslinger & Crookston 2020).

As per existing research, hydraulic structures' effectiveness in aeration is impacted by factors that include the temperature of the flowing water, water quality, depth of the tailwater, height of the drop, and the discharge of water. It was seen that the weir's aeration performance is highly affected by the drop height as well as the discharge rate across it (Singh & Kumar 2022a).

The experimental data collected by Singh & Kumar (2022a) for type A, B, and C are used to train ELM, OS-ELM, and I-ELM models. The ELM and its variant are used as powerful tools for regression tasks. Here, regression establishes the relationship between the output and the corresponding input dataset. Aeration efficiency (E) depends on many factors, as explained in Section 1; the objective of applying different machine models is to examine how accurately the machine models establish a relation between the output E and inputs C_u. The dataset contains three features and one output. The features affect the aeration efficiency predominantly as given in Equation (2) and are also compatible with ELM and its variants. There are two groupings within the entire dataset: (1) training dataset and (2) testing dataset. Eighty percent of the overall dataset is the training dataset, which is chosen at random, with the remaining 20% going toward testing.

3.1. Graphical visualization

The results demonstrate the predicted outcome of different models and their correlation with varying input parameters with several graphs. The findings based on the various parametric studies over type A PKW and its correlation with input parameters are presented in the following subsection.

3.1.1. Simulation results for type A PKW

The input dataset of 3,207 different sets of three input attributes C_s is taken to simulate different models. Hence, the size of the input dataset is (3207×3). The total 2,565 number of different sets, which is 80% of the total available dataset, is employed randomly for the training of the model. The remaining 20% of different sets are used to test the model. The aeration efficiency, E, which is the output of machine models, is a vector of size (3207×1). Figure 6(a) compares the predicted aeration efficiency with the actual aeration efficiency when the ELM model is used. Figure 6(b) and 6(c) presents the results when OS-ELM and I-ELM are employed to forecast the aeration efficiency of type A PKW.

Figure 6(d) represents the results of all three applied machine models in a single graph overlaying each other to compare the results at a glance. The value of deviation of the predicted aeration efficiency from the actual aeration efficiency is

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

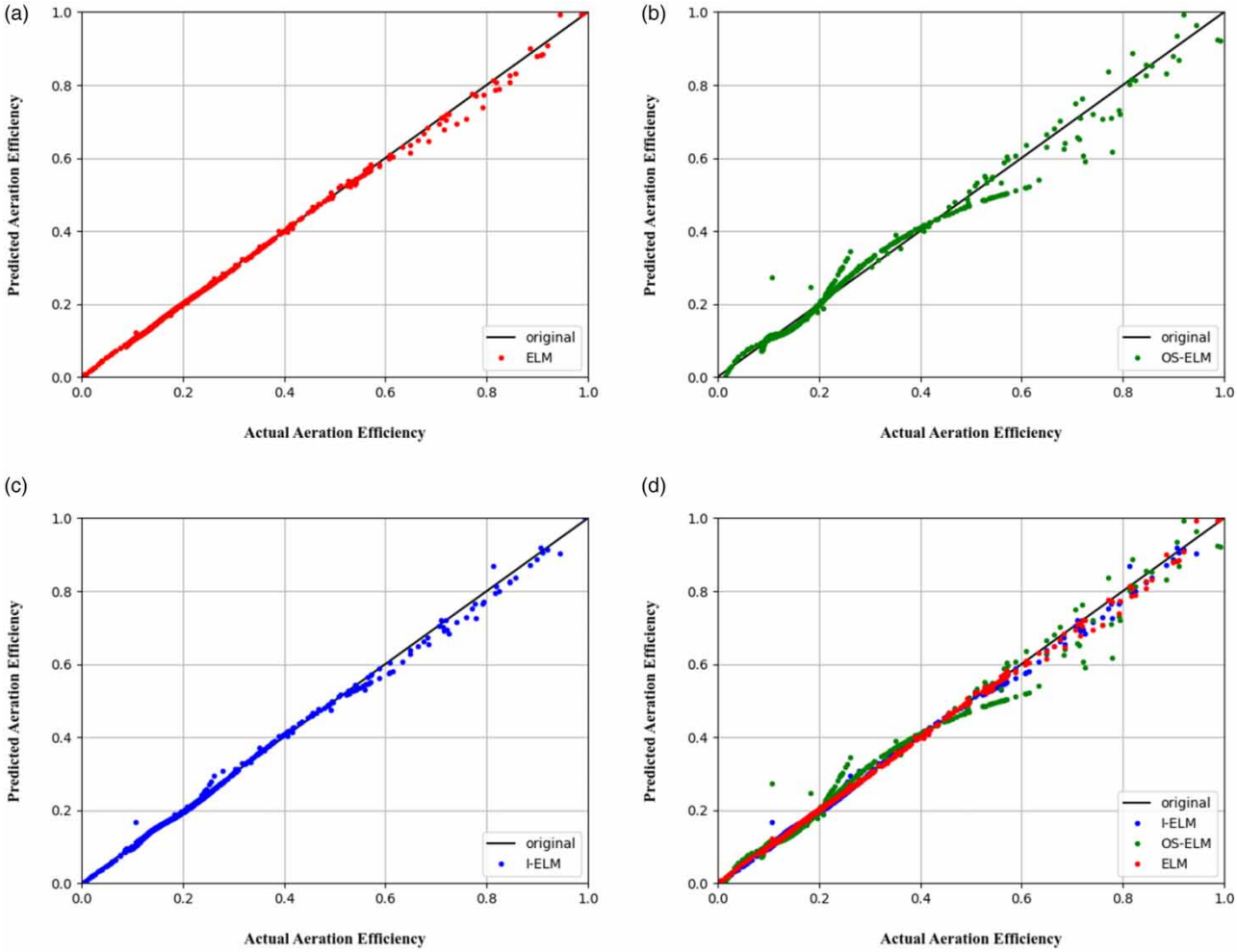


Figure 6 | The predicted aeration efficiency versus actual aeration efficiency of type A PKW (a) when ELM model is applied, (b) with OS-ELM model, and (c) with I-ELM model. (d) Results all the three models overlaid in a single plot.

calculated by MSE. The minimum MSE means the corresponding model is more efficient. Figure 7 represents absolute error, i.e. the variation between the actual and predicted values for all testing dataset.

3.1.2. Simulation results for type B PKW

The simulation results for PKW-B are presented similarly to those for type A PKW mentioned earlier. All three machine models are then used to evaluate the aeration efficiency of PKW-B, and the results are presented in this section. Figure 8(a)–8(c) shows the predicted aeration efficiency versus actual aeration efficiency as predicted by ELM, OS-ELM, and I-ELM, respectively. Figure 8(d) plots the results of all three models overlaying on each other but represented with different colour schemes.

The deflection of the predicted value from the actual value for the same testing dataset corresponding to the PKW-B is plotted similarly and indicated by the absolute error, as given in Figure 9.

3.1.3. Simulation results for PKW-C

This section represents the results when all three machine learning algorithms are applied to predict the aeration efficiency of type C similar to PKW (types A and B), and simulations are presented in Figures 10 and 11.

After analysing the overall pattern of data points on the graphs plotted between actual and predicted aeration efficiency, it is found that the data points are clustered diagonally in the case of ELM and I-ELM, suggesting that the predictions nearly

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

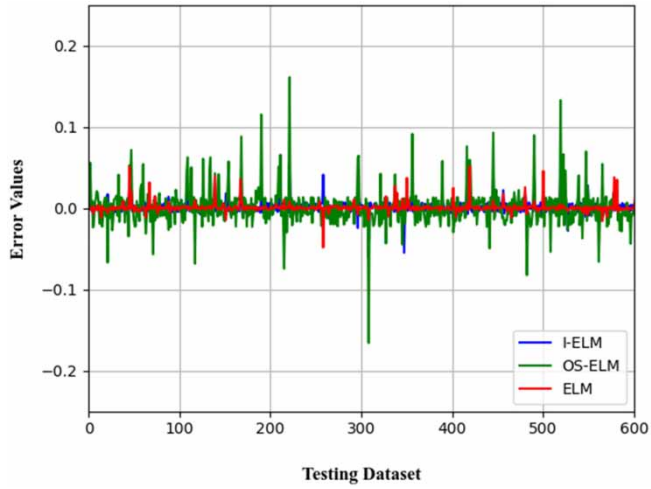


Figure 7 | The absolute error of each of 642 testing datasets predicted by ELM, OS-ELM, and I-ELM in case of type A PKW.

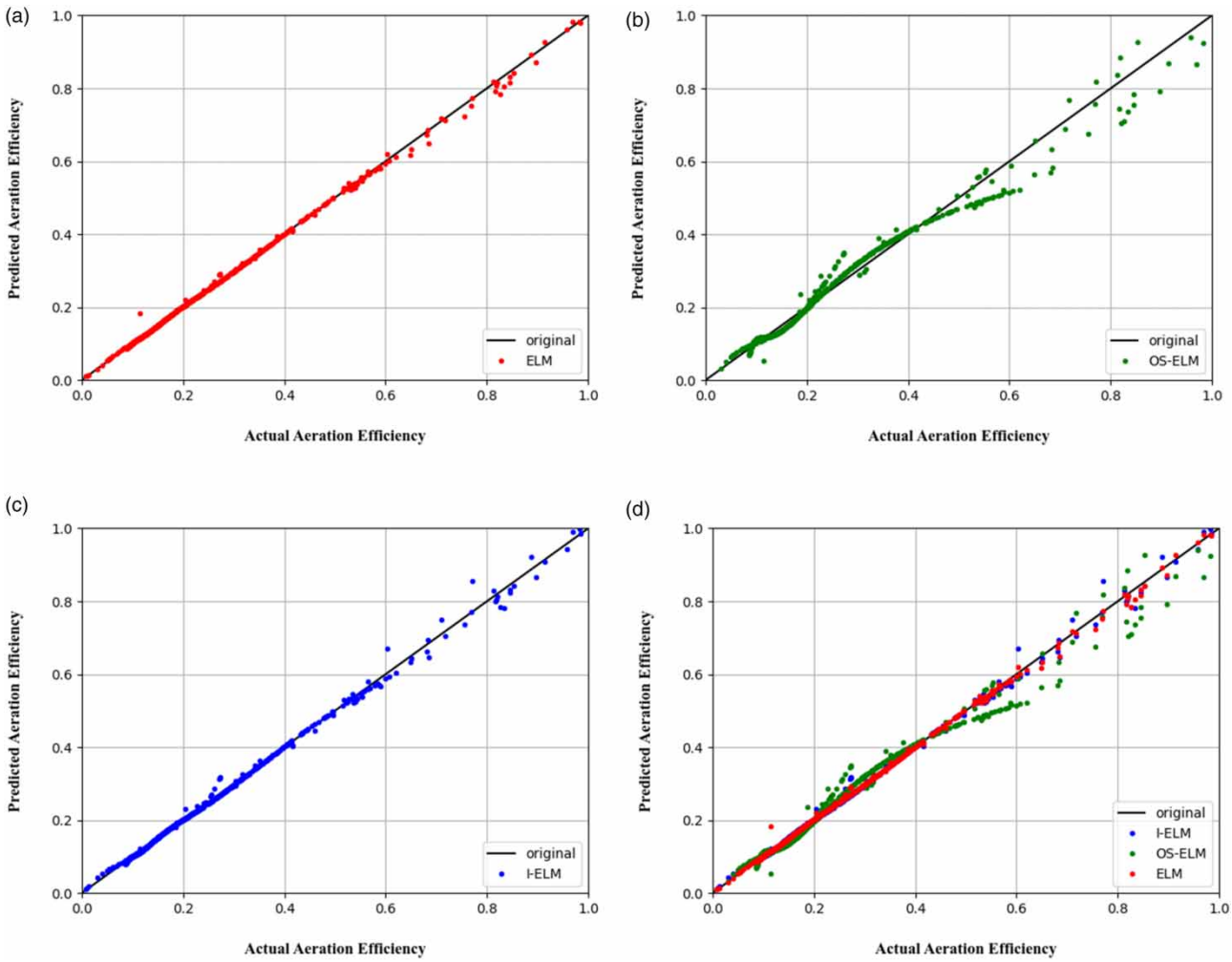


Figure 8 | The predicted aeration efficiency versus actual aeration efficiency of PKW-B with the application of (a) ELM, (b) OS-ELM, and (c) I-ELM model. (d) All results are presented on the same plot for comparison.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

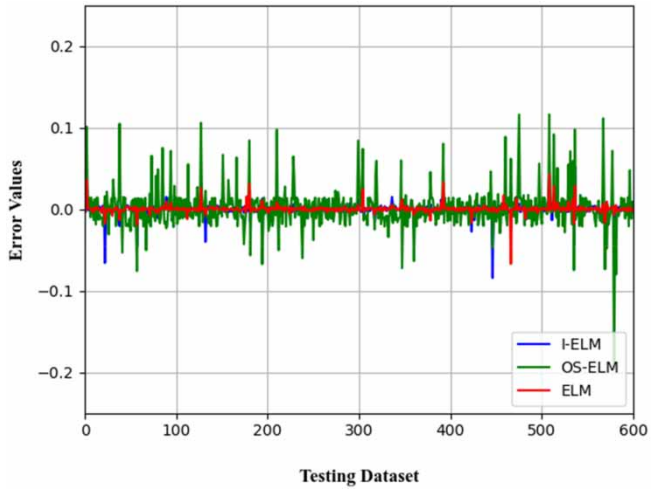


Figure 9 | The absolute error of each of 642 testing datasets predicted by ELM, OS-ELM, and I-ELM in case of PKW-B.

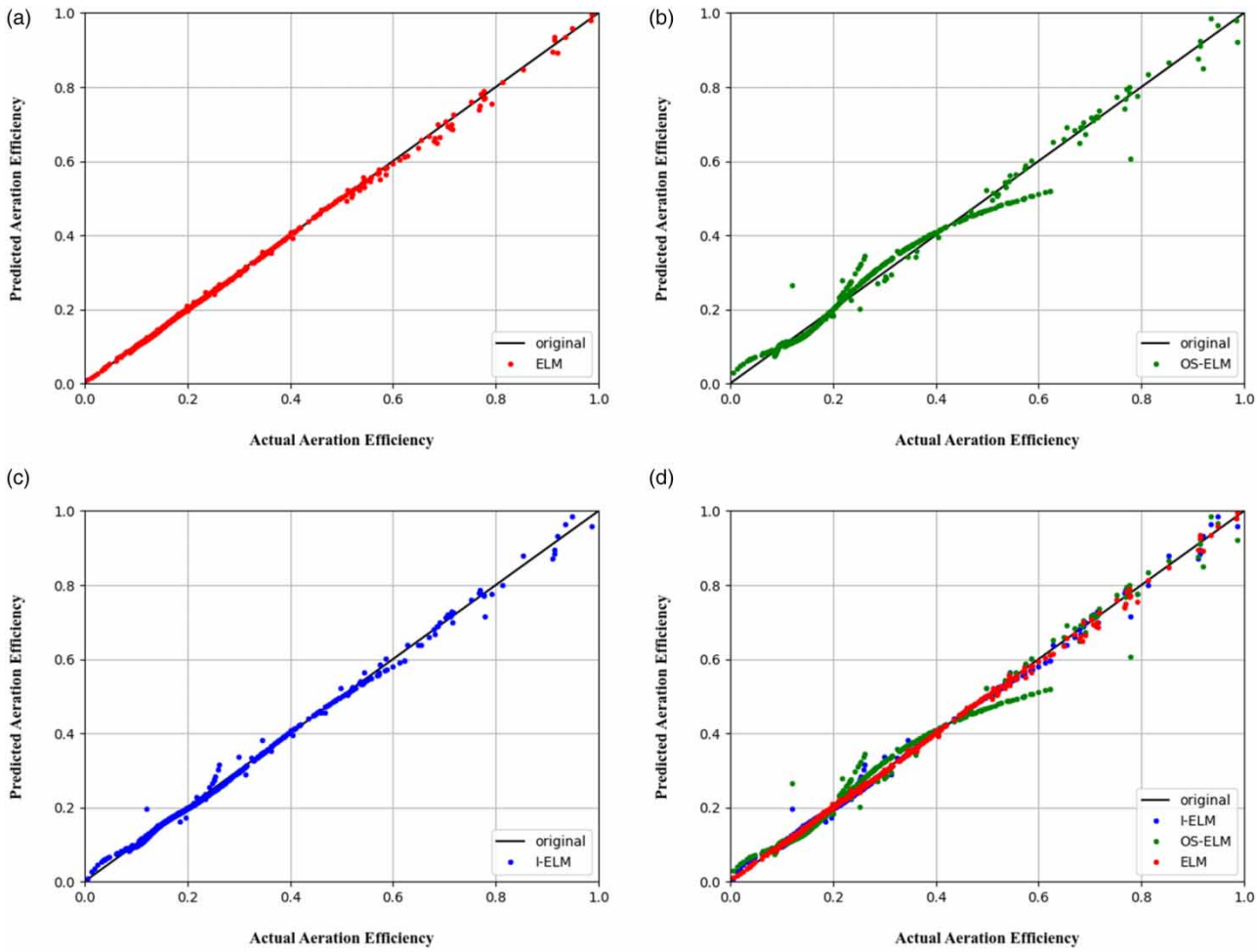


Figure 10 | The predicted aeration efficiency versus actual aeration efficiency when (a) ELM model, (b) OS-ELM model, and (c) I-ELM model are used for PKW-C. (d) All results are presents in a single plot for better visual comparison.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

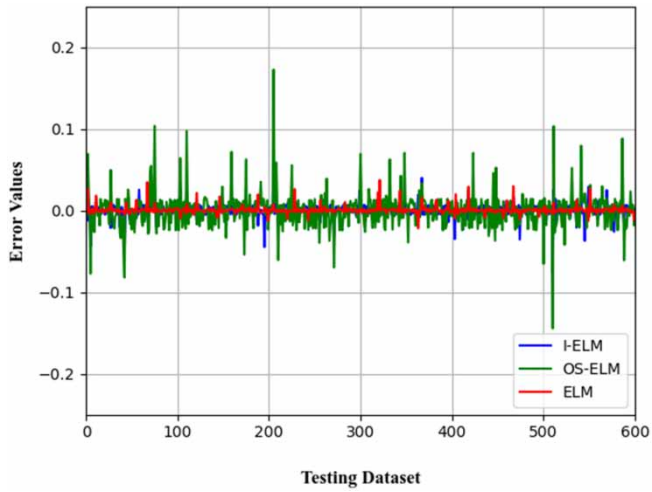


Figure 11 | The absolute error for testing datasets predicted by ELM, OS-ELM, and I-ELM in case of PKW-C.

Table 2 | MSE and RMSE of predicted aeration efficiency made by ELM, OS-ELM, and I-ELM for type A PKW, PKW-B, and PKW-C

Model	ELM			OS-ELM			I-ELM		
PKW type	A	B	C	A	B	C	A	B	C
MSE	0.0000204	0.000029	0.000024	0.00054	0.00064	0.00049	0.0000201	0.000023	0.000019
RMSE	0.00451	0.00540	0.00493	0.0232	0.0252	0.0221	0.00448	0.00485	0.00426
MAE	0.005176	0.00487	0.004084	0.01522	0.013949	0.012374	0.002446	0.002097	0.002404
CC	0.998982	0.998855	0.999318	0.992188	0.992143	0.992197	0.999668	0.99975	0.999643
NSE	0.997842	0.997584	0.998556	0.984389	0.984231	0.984106	0.99932	0.999465	0.999169

reflect the actual values. Hence, ELM and I-ELM predict the results more precisely and accurately. The results deviate from the accurate values, and therefore, the absolute error is maximum in the case of OS-ELM as the input data are not streaming time series data.

3.2. Comparison of prediction of ELM, OS-ELM, and I-ELM

In Section 3.1, three machine learning techniques – ELM, OS-ELM, and I-ELM – are provided for comparison in terms of prediction accuracy. Then their performances are evaluated using five metrics described in Section 2.6, which are MSE, RMSE, MAE, CC, and NSE. The accuracy of the prediction is inversely proportional to MSE, RMSE, and MAE, while CC and NSE are directly proportional to the accuracy of the models in prediction aeration efficiency.

MAE, MSE, and RMSE values are the minimum for I-ELM and the maximum for OS-ELM. The value of CC and NSE is the maximum and close to unity for I-ELM compared with the other two models, as shown in Table 2. Hence, the prediction made by I-ELM is nearer to the actual value of aeration efficiency in all three cases of PKWs. The results given by OS-ELM deviate the most from the actual value, hence the least efficient model for the given dataset and conditions. The accuracy of the ELM lies between the remaining two. The complete dataset is available to feed at once in the models and does not change rapidly with time. Hence, ELM is more efficient than OS-ELM, which is more suitable when the data arrive sequentially. It is observed that for the available complete dataset, ELM and I-ELM perform better than OS-ELM. I-ELM can process data in batch and incremental modes that can handle vast data. It provides better results for the conditions in this research work. The model can be placed as OS-ELM, ELM, and I-ELM to increase their performance, as concluded in Table 2.

4. CONCLUSIONS

Accurately assessing the aeration performance of the weir structures is critical for hydraulic engineers and designers because aeration is the most effective way to enhance oxygen content in flowing water and prevent the anaerobic conditions in the

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

water bodies, which are critical for the survival of aquatic life. To this end, the dataset obtained experimentally to calculate the aeration efficiency of different types of PKWs is used to train different extreme machine learning models and their variants like OS-ELM and I-ELM. Hyperparameters of all the applied models are correctly tuned. The major portion of the dataset is allocated for training the model, while the residual portion is utilized for testing the learning of different models. The prediction efficiency of three models, ELM, OS-ELM, and I-ELM, is compared while calculating the aeration efficiency of different hydraulic structures of type A, B, and C PKWs. It is found that for a complete available dataset that does not change rapidly with time, I-ELM performs the best, whereas, for these conditions, the prediction made in OS-ELM deviates the most from the actual value. The presented work can be extended to establish a well-defined relation between the input and output variables where the input dataset is unorganized. The performance of the models can be further evaluated under different data types, conditions, and problems from different domains.

ACKNOWLEDGEMENTS

The authors express their gratitude to the faculty, staff, and technical team of Delhi Technological University's Department of Civil Engineering for their help and assistance in conducting this research.

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

Apted, R. W. & Novak, P. 1973 Some studies of oxygen uptake at weirs. In: *Proceedings of the XV Congress, IAHR Paper B23*, Vol. 23. Istanbul, Turkey: International Association for Hydraulic Research, pp. 177–186.

Aradhana, A., Singh, B. & Sihag, P. 2021 Predictive models for estimation of labyrinth weir aeration efficiency. *Journal of Achievements in Material and Manufacturing Engineering* **105** (1), 18–32.

Avery, S. & Novak, P. 1978 Oxygen transfer at hydraulic structures. *Journal of Hydraulic Engineering, ASCE* **104** (HY11), 1521–1540.

Bagatur, T. & Sekerdag, N. 2003 Air-entrainment characteristics in a plunging water jet system using rectangular nozzles with rounded ends. *Water SA* **29** (1), 35–38.

Bansal, N., Singh, D. & Kumar, M. 2023 Computation of energy across the type-C piano key weir using gene expression and extreme gradient boosting (XGBoost). *Energy Reports* **9** (4), 310–321.

Baylar, A. & Bagatur, T. 2000 Study of aeration efficiency at weirs. *Turkish Journal of Engineering and Environmental Sciences* **24** (4), 255–264.

Baylar, A. & Bagatur, T. 2006 Experimental studies on air entrainment and oxygen content downstream of sharp-crested weirs. *Water Environmental Journal* **20** (4), 210–216.

Baylar, A. & Ozkan, F. 2006 Applications of Venturi principle to water aeration systems. *Environmental Fluid Mechanics* **6** (4), 341–357.

Baylar, A., Bagatur, T. & Tuna, A. 2001 Aeration performance of triangular-notch weirs. *Water Environmental Journal* **15** (3), 203–206.

Baylar, A., Hanbay, D. & Ozpolat, E. 2008 An expert system for predicting aeration performance of weirs by using ANFIS. *Expert Systems with Applications* **35** (3), 1214–1222.

Baylar, A., Unsal, M. & Ozkan, F. 2009a Hydraulic structure in water aeration process. *Water, Air, & Soil Pollution* **210**, 87–100.

Baylar, A., Hanbay, D. & Batan, M. 2009b Application of least square support vector machine in the prediction of aeration performance of plunging overall jets from weirs. *Expert Systems with Applications* **36** (4), 8368–8374.

Baylar, A., Unsal, M. & Ozkan, F. 2010 Hydraulic structures in water aeration processes. *Water, Air, & Soil Pollution* **210** (1–4), 87–100.

Baylar, A., Unsal, M. & Ozkan, F. 2011 GEP modeling of oxygen transfer efficiency prediction in aeration cascades. KSCE. *Journal of Civil Engineering* **15** (5), 799–804.

Crookston, B. M. & Tullis, B. P. 2012 Hydraulic design and analysis of Labyrinth weirs: Part-II: Nappe Aeration, Instability and Vibration. *Journal of Irrigation & Drainage Engineering, ASCE* **138** (8), 757–765.

Emiroglu, M. E. & Baylar, A. 2003 The effect of broad crested weir shape on air entrainment. *Journal of Hydraulic Research* **41** (6), 649–655.

Ervine, D. A. & Elsawy, E. M. 1975 Effect of a falling nappe on river aeration. In *16th Congress of the International Associations for Hydraulic Research Sao Paulo, Paper No. C45*, Brazil, pp. 390–397.

Eslinger, K. R. & Crookston, B. M. 2020 *Energy Dissipation of Type A Piano Key Weirs*. Basel, Switzerland: MDPI.

Gameson, A. L. H. 1957 Weirs and aeration of rivers. *Journal of the Institution of Water Engineers* **11** (5), 477–490.

Guenther, P., Felder, S. & Chanson, H. 2013 Flow aeration, cavity processes and energy dissipation on flat and pooled stepped spillways for embankments. *Environmental Fluid Mechanics* **13** (5), 503–525.

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

Gulliver, J. S. & Rindels, A. J. 1993 Measurement of air-water oxygen transfer at hydraulic structures. *Journal of Hydraulic Engineering ASCE* **119** (3), 327–349.

Gulliver, J. S., Thene, J. R. & Rindels, A. J. 1990 Indexing gas transfer in self-aerated flows. *Journal of Environmental Engineering, ASCE* **116** (3), 503–523.

Huang, G. B. & Chen, L. 2008 Enhanced random search based incremental extreme learning machine. *Neurocomputing* **71**, 3060–3068.

Huang, G. B., Zhu, Q. Y. & Siew, C. K. 2004 Extreme learning machine: A new learning scheme of feedforward neural networks, Vol. 2 (25–29). In: *Proceedings of International Joint Conference on Neural Networks (IJCNN2004)*, pp. 985–990.

Huang, G. B., Zhu, Q. Y. & Siew, C. K. 2006 Extreme learning machine: Theory and applications. *Neurocomputing* **70** (1–3), 489–501.

Huang, G. B., Ding, X. & Zhou, H. 2010 Optimization method based extreme learning machine for classification. *Neurocomputing* **74**, 155–163.

Jaiswal, A., Goel, A., 2019 Aeration through weirs: A critical review. In: *Sustainable Engineering, Lecture Notes in Civil Engineering*, Vol. 30 (Agnihotri, A., Reddy, K. & Bansal, A., eds.). Singapore: Springer.

Jaiswal, A., & Goel, A., 2020 Evaluation of aeration efficiency of triangular weirs by using Gaussian process and M5P approaches. In: *Advances Engineering Optimization Through Intelligent Techniques. Advances in Intelligent Systems and Computing*, Vol. 949 (Venkata, R. & Taler, J., eds.). Singapore: Springer. https://doi.org/10.1007/978-981-13-8196-6_66.

Khassaf, S. I., Aziz, L. J. & Elkatib, Z. A. 2015 Hydraulic behavior of piano key weir type B under free flow conditions. *International Journal of Scientific & Technology Research* **4** (8), 158–163.

Komal, K. M., Tiwari, N. K. & Ranjan, S. 2017 Aeration performance evaluation of piano key weir using linear regression and adaptive neuro-fuzzy inference system. In: *ICET: EITM-2017, NIT Hamirpur*, December 16–18, 2017, India.

Kumar, M., Tiwari, N. K. & Ranjan, S. 2020 Soft computing based predictive modelling of oxygen transfer performance of plunging hollow jets. *ISH Journal of Hydraulic Engineering* **28** (sup1), 223–233.

Leite Ribeiro, M., Boillat, J. L., Schleiss, A., Laugier, F. & Albalat, C. 2007 Rehabilitation of St-Marc Dam – Experimental Optimization of a Piano Key Weir. In: *Proceedings of the 32nd Congress of IAHR*.

Lempérière, F., Vigny, J. P., Ouamane, A., 2011 General comments on labyrinth and piano key weirs: The past and present. In: *Proceedings of the International Conference on Labyrinth and Piano Key Weirs (PKW 2011)* (Erpicum, S., Laugier, F., Boillat, J.-L., Pirotton, M., Reverchon, B. & Schleiss, A., eds.). Leiden, Netherlands: CRC Press, Taylor & Francis Group.

Liang, N. Y., Huang, G. B., Saratchandran, P. & Sundararajan, N. 2006 A fast and accurate on-line sequential learning algorithm for feedforward networks. *IEEE Trans Neural Netw* **17** (6), 1411–1423.

Lim, J. S., Lee, S. & Pang, H. S. 2013 Low complexity adaptive forgetting factor for online sequential extreme learning machine (OS-ELM) for application to nonstationary system estimations. *Neural Computing and Applications* **22** (3–4), 569–576.

Luxmi, K. M., Tiwari, N. K. & Ranjan, S. 2022 Application of soft computing approaches to predict gabion weir oxygen aeration efficiency. *ISH Journal of Hydraulic Engineering* 1–15. https://doi.org/10.1080/09715010.2022.2050311.

Machiels, O. 2012 *Experimental Study of the Hydraulic Behaviour of Piano Key Weirs*. *PhD thesis*, Belgium: HECE Research Unit, University of Liège.

Nakasone, H. 1987 Study of aeration at weirs and cascades. *Journal of Environmental Engineering, ASCE* **113** (1), 64–81.

Rathinakumar, V., Dhinakaran, G. & Suribabu, C. R. 2014 Assessment of aeration capacity of stepped cascade for selected geometry. *International Journal of Chemical Technology Research* **6** (1), 254–262.

Rong, H. J., Ong, Y. S., Tan, A. H. & Zhu, Z. 2008 A fast pruned-extreme learning machine for classification problem. *Neurocomputing* **72**, 359–366.

Sangeeta, Haji Seyed Asadollah, S. B., Sharafati, A., Sihag, P., Al-Ansari, N. & Chau, K. W. 2021 Machine learning model development for predicting aeration efficiency through Parshall flumes. *Engineering Applications of Computational Fluid Mechanics* **15**, 889–901.

Singh, D. & Kumar, M. 2022a Study on aeration performance of different types of piano key weir. *Water Supply* **22** (5), 4810–4821.

Singh, D. & Kumar, M. 2022b Energy dissipation of flow over the type-B piano key weir. *Flow Measurement and Instrumentation* **83**, 102–109.

Singh, A., Singh, B. & Sihag, P. 2021 Experimental investigation and modeling of aeration efficiency at labyrinth weirs. *Journal of Soft Computing in Civil Engineering* **5** (3), 15–31.

Srinivas, R. & Tiwari, N. K. 2022 Oxygen aeration efficiency of gabion spillway by soft computing models. *Water Quality Research Journal* **57** (3), 215–232.

Verma, A., Ranjan, S., Ghanekar, U. & Tiwari, N. K. 2022 Soft computing techniques for predicting aeration efficiency of gabion stepped weir. In: *Proceedings of the International Conference on Industrial and Manufacturing Systems (CIMS2020)*. Cham: Springer, pp. 117–122.

Wang, L., Huang, Y. P., Luo, X. Y., Wang, Z. & Luo, S. W. 2011 Image deblurring with filters learned by extreme learning machine. *Neurocomputing* **74**, 2464–2474.

Wormleaton, P. R. & Soufiani, E. 1998 Aeration performance of triangular platform labyrinth weirs. *Journal of Environmental Engineering* **124** (8), 709–719.

Wormleaton, P. & Tsang, C. 2000 Aeration performance of rectangular planform labyrinth weirs. *Journal of Environmental Engineering* **126** (5), 456–465.

First received 17 January 2024; accepted in revised form 17 June 2024. Available online 9 September 2024

Downloaded from http://iwaponline.com/jwcc/article-pdf/doi/10.2166/wcc.2024.078/1477054/jwc2024078.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Realizing In-Memory Computing using Reliable Differential 8T SRAM for Improved Latency

AYUSH DAHIYA, Department of Electronics and Communication, Delhi Technological University, Delhi, India
POORNIMA MITTAL, Department of Electronics and Communication, Delhi Technological University, New Delhi, India
RAJESH ROHILLA, Department of Electronics and Communication, Delhi Technological University, Delhi, India

Traditional von Neumann computing architectures suffer from high energy and lower speed as compared to the requirements of modern applications like those required in neural network accelerators. A modified differential eight transistor (8^+T) static random access memory (SRAM) based in-memory computing (IMC) structure was presented for realizing bit-wise boolean logic operations. The 8^+T SRAM-IMC is designed at 32 nm technology node with throughput of 2.1849, 2.4815, 2.5795, 2.6240, 2.6495, 2.6619, 2.6690, 2.6732 and 2.6749 giga outputs per second (GOPS) for 0.5-1.3 V supply voltage range respectively. The differential 8T cell used to implement logic operations supports NAND and NOR operations with minimal overhead while also performing the standard storage operation with added stability over the conventional 6T and 8T SRAM cells. The SRAM-IMC offers reliable boolean logic operations by using asymmetric sensing strategy for all process corners, TT, SS, SF, FS and FF for an operating temperature range of 220 K to 400 K. Monte carlo simulation considering global threshold voltage deviation of 50 mV were performed for various operating condition in order to study the impact of variations on design parameters such as latency.

CCS Concepts: • **Hardware** → **Very large scale integration design**; **Process variations**.

Additional Key Words and Phrases: SRAM Cell, In-memory Computing, von Neumann Bottleneck, Latency, Throughput

1 Introduction

Technology scaling has paved the way for computing cost to be decreased to such an extent that we are surrounded by intelligent connected devices [Bai et al. 2022] but this constant struggle to make computing even faster, cost-efficient and lower power can not continue perpetually. This limitation of complementary metal oxide semiconductor (CMOS) scaling is countered by moving towards beyond-CMOS devices [Wong and Salahuddin 2015; Zeng et al. 2022] but the sheer disruption associated with introducing new abstraction levels at various stages of the design hierarchy as well as novel manufacturing processes will have tremendous "costs". The training and tools for designers or industry-wide adoption to compete with the status quo, i.e., CMOS-based technologies, are a few such hindrances attached to the beyond-CMOS shift. Thus, for the foreseeable future, we assume CMOS-like technologies to continue to dominate computing for the near future. The conventional von Neumann architecture shown in Fig. 1a comprises of a seperate processing unit and memory unit whereby the

Authors' Contact Information: Ayush Dahiya, Department of Electronics and Communication, Delhi Technological University, Delhi, Delhi, India; e-mail: ayushdahiya010@gmail.com; Poornima Mittal, Department of Electronics and Communication, Delhi Technological University, New Delhi, Delhi, India; e-mail: poornimamittal@dtu.ac.in; Rajesh Rohilla, Department of Electronics and Communication, Delhi Technological University, Delhi, Delhi, India; e-mail: rajesh@dce.ac.in.

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 1557-7309/2024/9-ART
https://doi.org/10.1145/3696666

Just Accepted

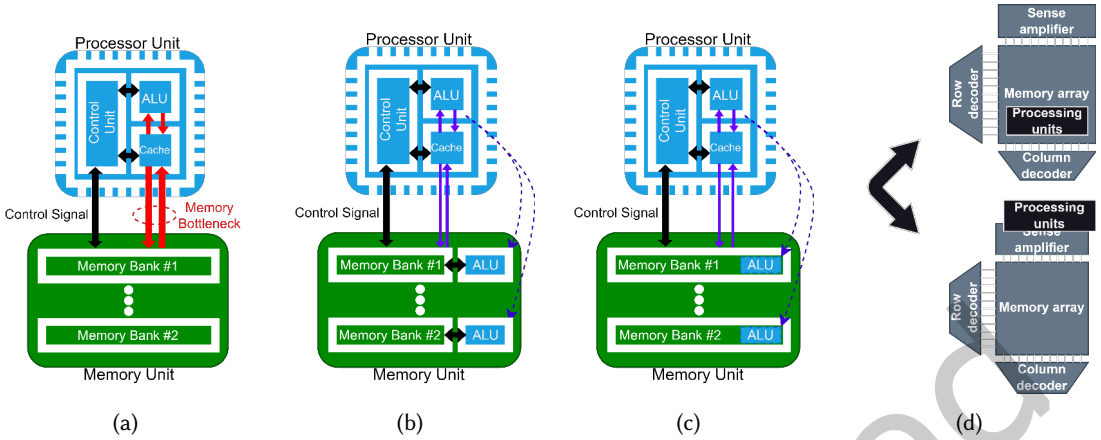


Fig. 1. Memory bottleneck dilemma and its solutions. (a) von Neumann Architecture. (b) Near-memory computation. (c) In-memory computing architecture. (d) two configurations for implementation of In-memory computing.

instructions are fetched for each clock cycle and the computation is performed by the Arithmetic and logic unit (ALU). This constant transfer of large amounts of data between memory and processor causes a bottleneck for the performance of the system. To overcome the energy costs of the memory system as well as improve performance a specialized hardware approach called Compute-in-memory (CIM) which employs both near-memory computing (NMC) and in-memory computing (IMC) structures is being explored in the scientific domain. The NMC and IMC structures are shown in Fig. 1b and 1c respectively. A number of memory technologies have been explored in CIM such as SRAM [Agrawal et al. 2018; Chen et al. 2022a; Lin et al. 2023a; Rawat and Mittal 2023], RRAM [He et al. 2020; Wan et al. 2022], DRAM [Yu et al. 2021; Zhang et al. 2023], spin-transfer torque (STT) magnetic random access memory (MRAM)[Wu et al. 2023, 2024]. Although these emerging technologies promise highly dense low power memory chips, the reality is that their viability in real-world on-chip memories is still under scientific scrutiny. There are two domains for IMC operation, Analog IMC [Han et al. 2022; Jiang et al. 2020; Xuan et al. 2023] and Digital IMC [Lin et al. 2023b; Yin et al. 2020]. The analog IMC offer lower precision owing to low-cost Analog-to-digital converters (ADC) and thus are only suitable for low accuracy applications like CNN inference. Digital IMC have higher computational accuracy but they suffer for process, voltage and temperature (PVT) as they calculate logic operations by turning on multiple word lines and discharging the bitlines. This limits the benefits of using IMC designs in the first place. The IMC architectures can either be SRAM cells with peripherals added beyond the sense amplifier to obtain the arithmetic operation outputs or the SRAM cells can be modified using logic cells within the memory array to obtain a IMC macro. These two different approaches to IMC are shown in Fig. 1d.

The remainder of this article is organized as follows. Section 2 introduces the related work of SRAM IMC using standard 6T, 8T cells (differential and single ended) and illustrates the working principle and process of the circuit with examples. Section 3 analyzes how to optimize the 8^{+}T cell for IMC to perform logical operations and check the reliabiltiy of the cell in memory and compute modes. Section 4 presents the comparative results of all the discussed IMC structures. Finally, Section 5 provides the conclusion.

2 In-Memory Computing using Standard SRAM Bit cells

The conventional SRAM bit cell lends itself very well to IMC architectures due to the fact that two concurrently activated wordlines intrinsically perform NOR operation for the corresponding stored data values. This is clearly

Just Accepted

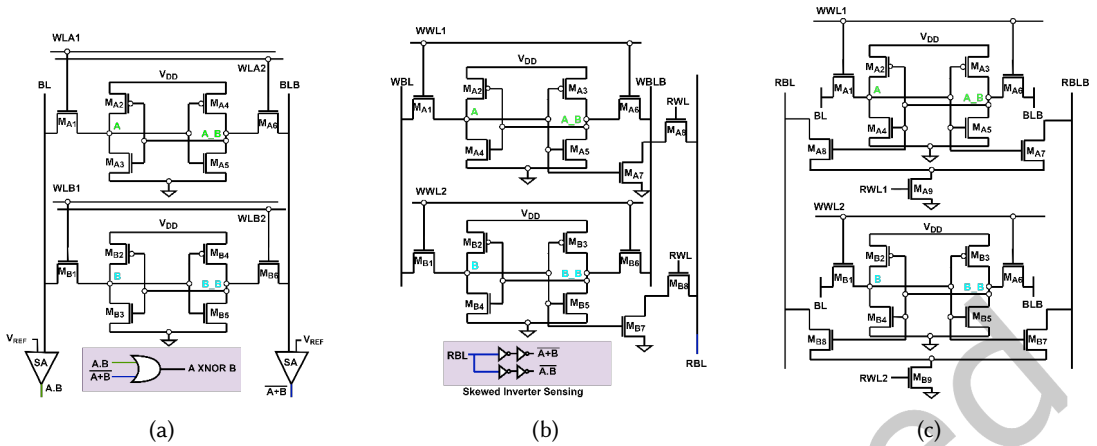


Fig. 2. Circuit schematics for implementation of single bit boolean logic for two operands A,B along with the sensing strategy employed. (a) 6T SRAM-IMC (b) 8TC SRAM-IMC (c) 8TD SRAM-IMC

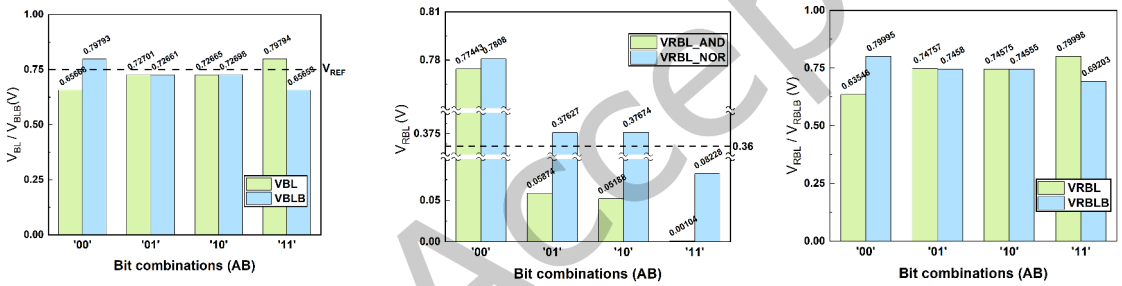
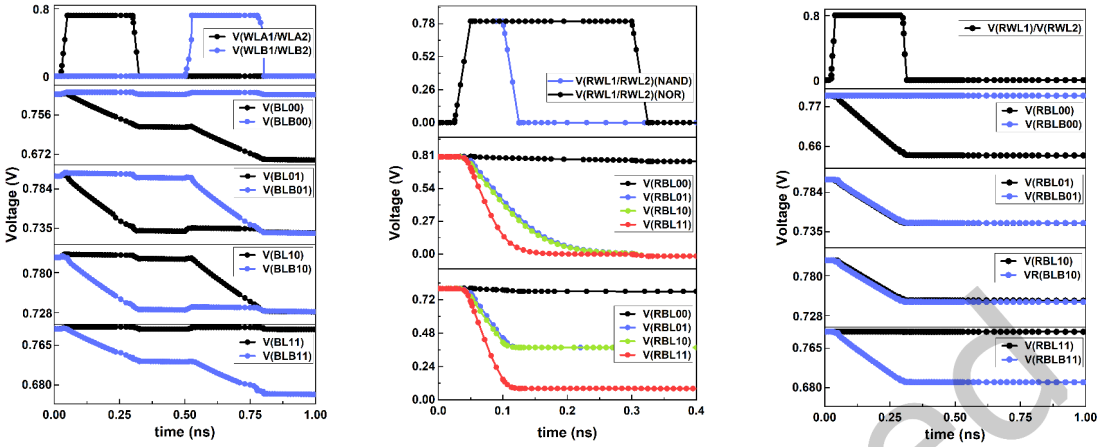


Fig. 3. SRAM bit lines (BL/BLB) voltage levels to be sensed by SAs for IMC bitwise logical operations for (a) 6T SRAM cell (b) 8TC SRAM cell (c) 8TD SRAM cell.

evident for standard 6T SRAM cell illustrated in Fig. 2*a* with the sense amplifiers being used to perform the compute operation shown as their digital gate equivalent. The content addressable memory circuitry (CAM) implemented using standard 6T push rule cells [Jeloka et al. 2016] with the wordline split into two different control signals, WLA1 and WLA2 as shown in Fig. 2*a*, performs XNOR logic operation to compare the search string with the stored values in the corresponding bit cells. This enables parallel search by applying the search string on the wordlines and comparing the column-wise stored values. Other iterations of 6T bit cell being used in IMC have also used sequential wordline activation instead of concurrently turning on multiple WLs which might cause read/ compute disturbance due to the common read/write path of conventional 6T cells. The cells in the same column will flip if the bitline discharges to a relatively low value, i.e., lower than the switching threshold of the single ended inverter based sense amplifiers and false read might occur. The 6T SRAM with under-driven multiple wordline control during the compute operation has also been used which helps avoid the read disturbance encountered when the stored datum 'A' and 'B' are different (AB = 01/10). However, the usage of low WL signals yields inherently slower speeds for NMOS-access transistor based 6T. A solution to is to

Just Accepted



(a) (b) (c)

Fig. 4. Transient response of SRAM bitcells in compute-mode for all possible bit combinations of two single bit operands (A,B) '00', '01', '10 and '11' at V_{DD} = 0.8 V for (a) 6T SRAM cell (b) 8TC SRAM cell (c) 8TD SRAM cell.

use a dual port 8T(6+2T) SRAM cell [Chen et al. 2022b] with a isolated compute port comprised of two PMOS access transistors together with split WL which avoid the reliability degradation and increased latency suffered by 6T-based IMC architectures. The Fig. 4*a* illustrated the transient simulation for the two 6T cells accessed using sequentially activated WL signals and the corresponding discharging of the BL/BLB line based on the stored data combinations. The AND and NOR operations are implemented by using single ended SAs with V_{REF} taken such we get a high logic output for V_{BL} in case where AB = 11 for AND implementation and V_{BLB} where AB = 00 for NOR logic as shown in Fig. 3*a*.

SRAM cell structures have evolved from the conventional 6T SRAM cell to single ended cell structures as well as differential ended modified architectures with increased transistor count. These are designed around an isolated read and write path which allows for modulation of the Static Noise Margin (SNM) in both the hold and read modes separately. The isolated read and write ports are also useful when multiple WL activations occur as is the case in Computation modes for standard SRAM cell structures. The read and write disturbance is mitigated using the isolated ports. The two conventional 8T (8TC) cells as shown in Fig. 2*b* store the corresponding operands for the two bit logic operations supported in-memory. The 8TC transient response for the logic operations supported is depicted in Fig. 4*b* where the Read bitline (RBL and RBLB) voltages are shown for all the possible bit combinations, i.e., 00,01,10,11. The charge and discharge of the RBL/RBLB is used as the output for implementing the logic operations. Skewed inverter sensing is used for performing the logic operation with the conventional cross-coupled inverter pair sharing the same bitlines acting as a wired NORed structure similar to the 6T SRAM-IMC. The resulting voltages and the outputs of the skewed inverter sensing is shown in Fig. 3*b*. The differential 8T cell (8TD) [Kulkarni et al. 2011] structure can be used in IMC for implementing logic operations like NAND, NOR and XOR. The cell's differential structure is designed using two extra NMOS transistors together with a footswitch common to the whole row for WWL control signal connected in the pulldown path for the cross-coupled inverter pair. The circuit schematic for 8TD SRAM-IMC is shown in Fig. 2*c*. The boolean logic functions are implemented using an asymmetric differential sense amplifier [Agrawal et al. 2018] which senses the differential Read bitlines (RBL/RBLB). The transistors are skewed by sizing the access transistors in the senese amplifier. The complement read bitline (RBLB) access transistor sized greater than RBL access transistor yield OR (NOR) logic combinations

Just Accepted

at the SAout (SAoutb) while a wider RBL access transistor results in AND (NAND) logic combination at the SAout (SAoutb). The RBL and RBLB voltages for different combinations of operands 'AB' are shown in Fig. 3*c*. The transient response of the 8TD SRAM-IMC structure is depicted in Fig. 4*c* for input control signal WL/WWL at V_{DD} = 0.8 V.

3 Implementation of In-Memory Computations using 8$^+$T SRAM cell

The recently proposed differential SRAM 8$^+$T cell [Pal and Islam 2016] and its adjacent structures lend themselves very well to In-memory computational needs due to their isolated read/write dual port architecture. A modified version of the cell is designed using the conventional cross-coupled inverter pair with two extra control buffer PMOS transistors being used together with a tail NMOS as shown in Fig. 5*a*. The read and write operations are illustrated using the equivalent circuit for the 8$^+$T cell in Fig. 5*b*, 5*c* respectively. One of the control buffer transistors (Q=1 connected to its gate terminal) forms a conducting path through the tail transistor during the access mode of the differential SRAM cell. While for the write operation to be performed the bitlines need a conduction path to the internal storage nodes the WLB = 0, and tail is turned ON.

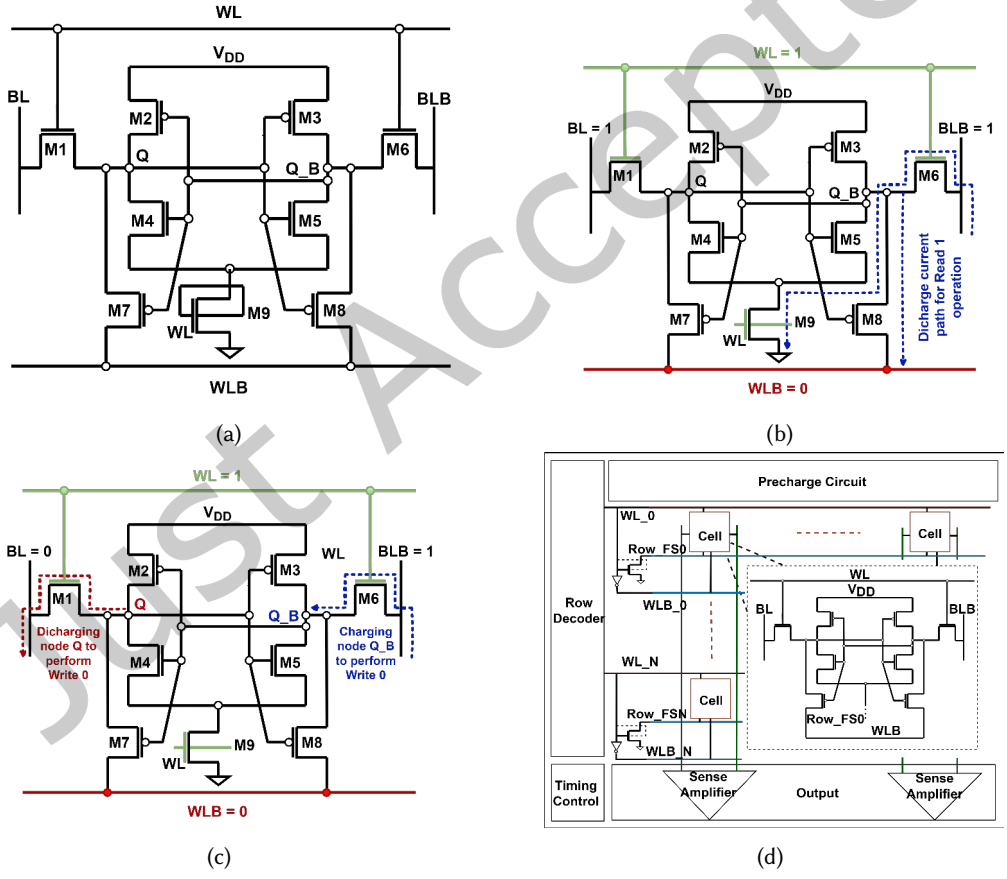


Fig. 5. (a) Circuit schematic of 8$^+$T SRAM cell. The equivalent circuits for (b) read and (c) write operation of the SRAM cell. (d) The SRAM architecture for the 8$^+$T Cell showing M9 transistor common to each row of the memory array.

Just Accepted

The storage nodes Q/QB are then charged/discharged by the voltages present at the bitlines (BL/BLB). The cell exhibits lower leakage as compared to the conventional 6T SRAM cell. This is due to the stack effect which dictates that when two transistors that are connected in series; with one of them being conducting (ON) while the other is turned OFF, this stack causes the intermediate node voltage to increase to some positive value. This in turn causes a reduction in sub-threshold leakage component. The high WLB in the data retention (hold) mode also helps reduce the static power consumption. The SRAM array architecture for the 8$^+$T cell reuses the WL tail transistor for all the cell in a single rows as depicted in Fig. 5*d* which results in reduced area overhead as compared to conventional cell and array architectures.

3.1 Robustness of the 8$^+$TP Cell under Threshold Voltage Variations in Memory Mode

The reliability of SRAM cells is a major concern for sub-45nm CMOS due to the presence of variations introduced by photo-lithographic fabrication steps wherein the precision is low due to the small-feature sizes. These are characterised into two types, global (systematic) variations and local (random) variations. The global variations stem from die-to-die variations which are usually uniform and can thus be accounted for and the design process tuned accordingly. The local variations are however limited to within the die.

These result in changes in threshold voltage (V_{TH}) and effective gate length (L_{min}). A strong dependence of V_{TH} on temperature changes also results in V_{TH} shifts causing increased variability. The impact of ΔV_{TH} on the 8$^+$T SRAM cell's static noise margin is shown in Fig. 6. The temperature is varied from 220 to 400 K in steps of 12 K for all five process corners, named as NMOS-PMOS, typical-typical, slow-slow, slow-fast, fast-slow, and fast-fast (abbreviated as TT, SS, SF, FS, FF respectively). The cell stability increases with an increase in temperature for all process corners except SS corner. 6k Monte Carlo simulations for all the process corners were performed and the resulting distribution of the SNM was plotted. These results show a conformation to 6σ deviation under V_{TH}

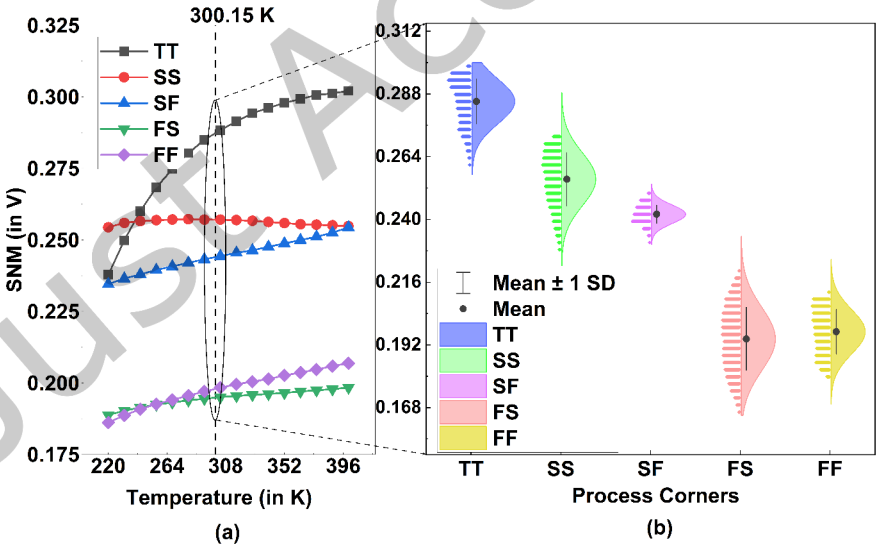


Fig. 6. (a) Impact of temperature variations on the 8$^+$T SRAM Cell's static noise margin (b) distribution of SNM for different process corners at T = 300.15 K. Each distribution is obtained for 6k Monte Carlo simulation runs for ΔV_{TH} = 50 mV and V_{DD} = 0.8 V.

Just Accepted

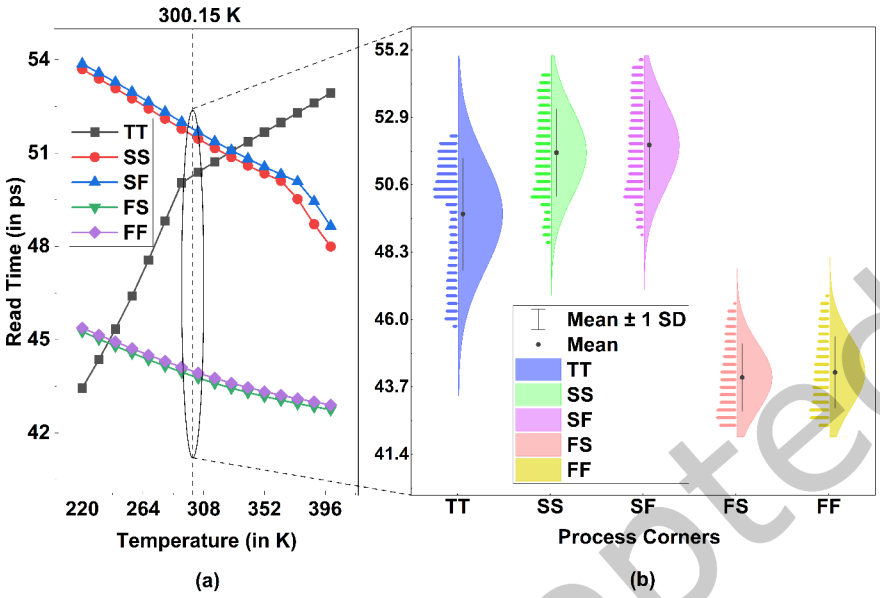


Fig. 7. (a) Impact of temperature variations on the 8⁺T SRAM Cell's read time (b) distribution read time for different process corners at T = 300.15 K. Each distribution is obtained for 6k Monte Carlo simulation runs for ΔV_{TH} = 50 mV and V_{DD} = 0.8 V.

variations. The Fig. 6*b* shows normal distribution curves for the SNM values obtained from the Monte-Carlo simulation runs. The SNM has 3.0054%, 3.9780%, 1.4752%, 6.2175%, 4.3189% coefficient of variation (CV) for TT, SS, SF, FS, FF corners respectively. A lower CV equates to a dataset with lower variability and higher stability. The read time for SRAM cell is defined as the time taken from 50% point of the wordline activation to the instant where the differential outputs at the bitlines reaches half of the supply voltage. The variation of read time for 8⁺T SRAM cell is depicted in Fig. 7a along with the distributions for the read time at various process corners in Fig. 7b. The read time highlights the read performance of the SRAM cell against process variations at V_{DD} = 0.8 V. The read time CV for the TT, SS, SF, FS, FF corners are 3.8475%, 2.8878%, 2.9131%, 2.5962%, 2.7335% respectively.

3.2 Robustness of the 8⁺TP Cell under Threshold Voltage Variations in Compute Mode

In order to implement single bit boolean logic functions like NAND/NOR, the 8⁺T SRAM cell schematic shown in Fig. 8a is used. The two operands stored in the bit cells are A and B respectively. There are two possibilities for the control wordline that lead to NAND/NOR logic operations similar to the conventional 6T/8TD SRAM cells discussed in Section 2. When both the wordlines of the two cells storing the operands are activated at the same instant of time, i.e. concurrent wordline activation, the transient response shown in Fig. 8*b* for the voltages at the bit line (BL) and complementary bit line (BLB) which is analogous to 6T SRAM-IMC. This is illustrated for all possible combinations of A,B, namely '00,01,10,11'. Correspondingly, for sequentially activated wordlines of cell A followed by cell B, which helps mitigate the problem of read disturbance where the stored values might change when two or more wordlines in the same column are turned on concurrently, the resulting transient simulation is depicted in Fig. 8c.

An asymmetric sense amplifier (SA) [Agrawal et al. 2018] is used with the two NMOS bitline access transistors

Just Accepted

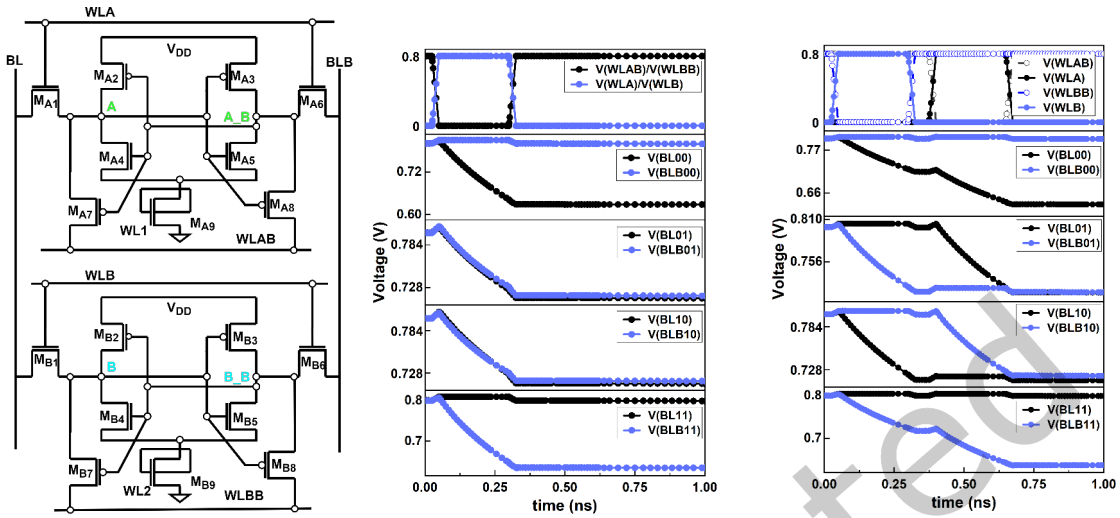


Fig. 8. (a) Circuit schematics for implementation of single bit boolean logic for two operands A,B using 8$^+$T SRAM cell. Transient response for compute mode of 8$^+$T SRAM cell for all possible input operand combinations (AB), '00,01,10,11' using (b) concurrent word-line and (c) sequential word-line activations at V_{DD} = 0.8 V.

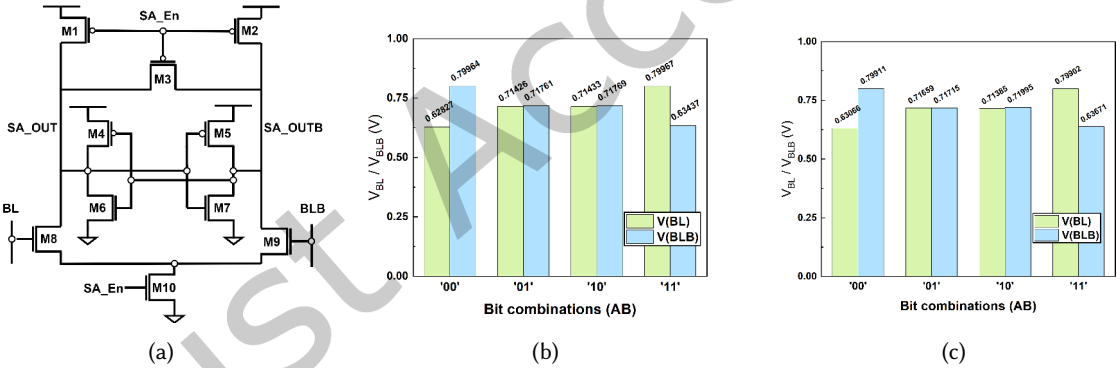
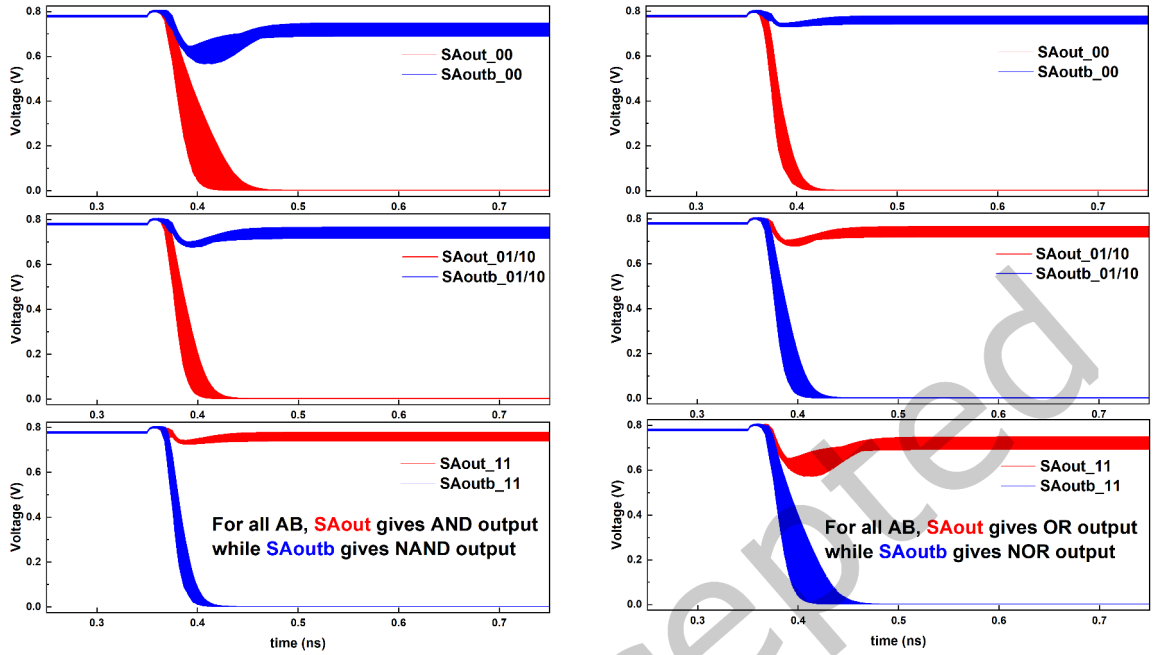


Fig. 9. SRAM bit lines (BL/BLB) voltage levels to be sensed by (a) asymmetric SAs [Agrawal et al. 2018] for IMC bitwise logical operations for 8$^+$T SRAM cell using (b) concurrent wordline and (c) sequential wordline activation.

(M8 and M9) sized non-uniformly based on the logic operation (NAND/NOR) to be realised. The voltages at BL/BLB for enabling NAND/NOR at V_{DD} = 0.8 V are shown in Fig. 9b and 9c respectively. The asymmetric sense amplifier (SA) operates as follows with M8 being wider than M9 and the stored operands AB. For AB = '00', BL discharges to GND while BLB remains at V_{DD}, causing the cross-coupled inverter pair to stabilize with SAout = 0. When AB = '01' or '10', both bitlines are connected to similar storage values, leading to comparable discharge rates. However, since M8 is wider than M9, SAout discharges faster, resulting in SAout = 0. For AB = '11', BL remains precharged at V_{DD}, while BLB is pulled down to GND, producing SAout = 1. This output combination clearly aligns with AND logic when sensing SAout and NAND logic when considering SAoutb. Similarly, another

Just Accepted



(a) (b)

Fig. 10. Transient response of the asymmetric sense amplifier for 6k Monte carlo simulations at TT corner for T = 300.15 K and V_{DD} = 0.8 V to implement (a) NAND and (b) NOR outputs for all possible input cases of input operands (AB), '00,01,10,11', at for ΔV_{TH} = 50 mV.

Just Accepted

SA in parallel with M9 wider than M8 generates the OR/NOR logic. The transient response obtained from the simulations demonstrates that the sensing mechanism produces accurate results, with no erroneous values detected throughout the operation. This ensures that the design yields zero computational errors, confirming its resilience against threshold variations in performing logic operations. This configuration effectively supports all possible input combinations for operands A and B, providing correct logical outputs without any errors. To further validate the robustness of the design, Monte carlo simulations for 6k trials were performed for showcasing the reliability of the NAND/NOR logic operations using the 8$^+$T SRAM cell at T = 300.15 K for TT corner with V_{DD} = 0.8 V and ΔV_{TH} = 50 mV to account for process variations. The results, presented in Fig. 10*a* and 10*b*, illustrate the transient output voltage for the asymmetric sense amplifier when performing NAND and NOR operations, respectively. These figures highlight the stability and consistency of the outputs across multiple simulation runs, further reinforcing the reliability of the 8$^+$T SRAM-IMC architecture.

The latency of the logic operations derived from the Monte carlo simulation runs are then plotted as histograms for both NAND and NOR logic operations. Fitted Gaussian distributions curves are then superimposed over the histograms. This is illustrated in Fig. 11 for both NAND and NOR logic, covering all possible input cases for operands AB. The NOR logic has a CV of 0.7645%. 1.1570% and 2.4873% for case 00, 01/10 and 11 respectively. Similarly, NAND logic has CV of 2.3857%, 1.1098% and 0.7217%. The worst-case average latency for NAND and NOR logic are 403.372 ps and 402.209 ps for case '00' and case '11' respectively.

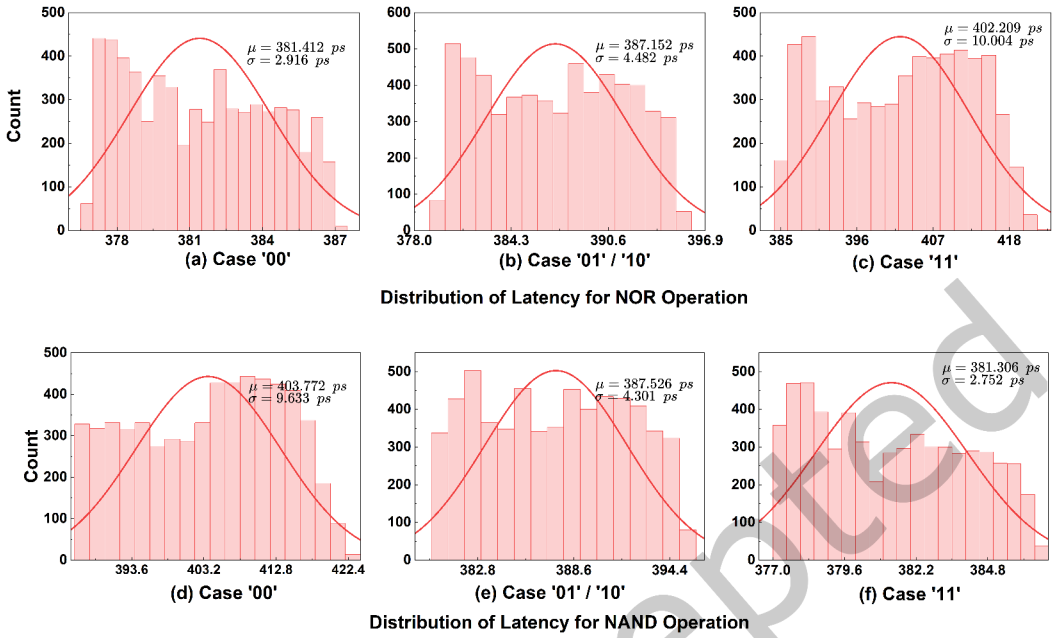


Fig. 11. Distribution of latency for NOR and NAND logic operation for all possible combinations of input operands (AB), '00,01,10,11' for 6k Monte Carlo simulation runs at T = 300.15 K and V_{DD} = 0.8 V with ΔV_{TH} = 50 mV.

4 Performance Comparison of In-memory SRAM cell architectures

This section presents a comparative analysis of the various conventional SRAM-IMC architectures discussed earlier, all designed using the Predictive Technology Model [Arizona State University and Group 2008] at the 32 nm technology node. All the various SRAM cells were designed using consistent transistor sizing in order to perform an objective analysis for supply voltages ranging from near-threshold, i.e, 0.5 V to 1.3 V. The static noise margin, latency and average energy consumption per bit for various SRAM-IMC have been calculated based on the simulation results. The identical word line (WL) and sense amplifier enable (SA_En) timing signals are utilized while performing all simulations. A π-3 model is used for the bitline capacitance to simulate large SRAM array structures. The bitline capacitance of 50 fF is assumed for all three SRAM-IMC being compared.

There are a number of SRAM-IMC implementations for enabling single bit boolean logic without much additional overhead as compared to conventional SRAM array architectures. The choice of cell selections and the sensing employed to realize the single bit NAND/NOR operations can be made depending on two critical design metrics, latency and average energy per bit of computation. The stability of all the cells is compared against global threshold variations (ΔV_{TH} = 50 mV) at V_{DD} = 0.8 V in Fig. 12a. It can be clearly seen that all the cells exhibit very comparable stability due to the inherent similarities in their cell structures. The 8^{+}T cell depicts best reliability against variations and is the most stable among the four cell designs discussed. Moreover, the reliability of the IMC design against process variations is also an essential aspect when selecting the bit cell. The impact of process variation on the latency of the cells discussed are shown in Fig. 12*b* and 12*c*. The 6T, 8TC, 8TD and 8^{+}T cells show a CV of 0.7077%, 2.4224%, 0.7384% and 0.7580% respectively.

The conventional 6T SRAM used in IMC yields a very simple design which allows computation in-memory but

Just Accepted

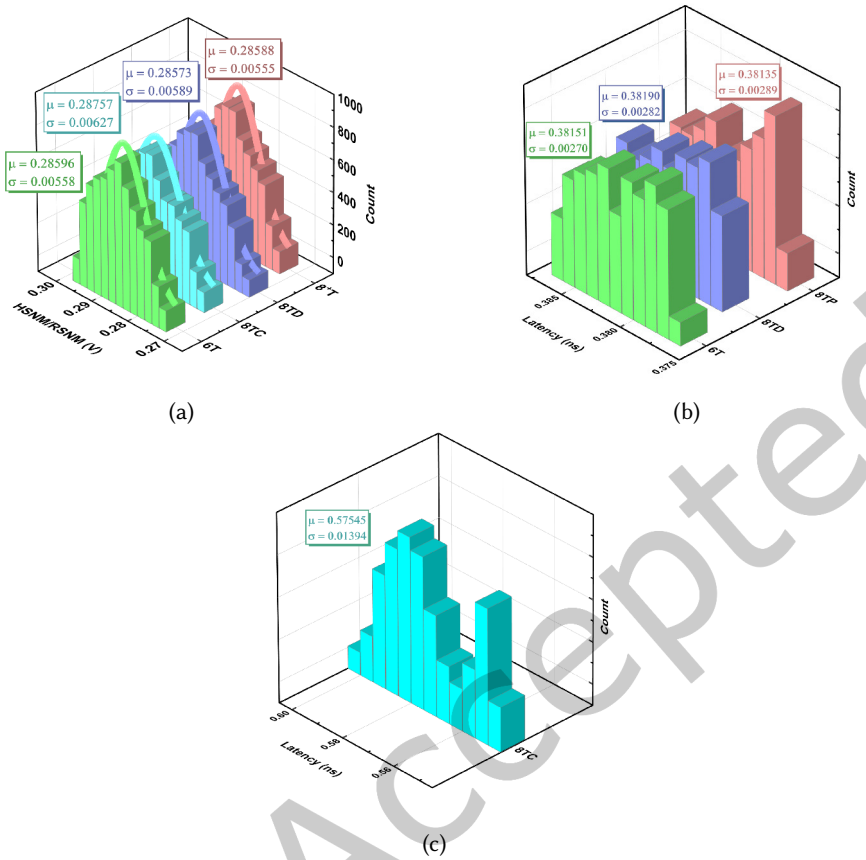


Fig. 12. Distribution of (a) SNM for 6T, 8TC, 8TD, 8$^+$T (b) latency for 6T, 8TD, 8$^+$T and (c) latency for 8TC SRAM cell derived from 6k Monte Carlo simulation for ΔV_{TH} = 50 mV at V_{DD} = 0.8 V, TT corner and T = 300.15 K.

concurrent wordline activation for multiple cells often leads to read disturbance; which can inadvertently lead to change in the stored bits of the selected cells. This can be managed by making use of sequential wordline activation which improves robustness by reducing read disturbance at the expense of speed. The 8TC and 8TD cells have the benefit of read decoupled paths in their respective architectures whereby multiple cells of the same column can be selected without any read disturbance. The read operation of the single ended 8TC allows for lower power consumption (i.e. improved energy/bit) at increased latency. The 8TD cell has better reliability due to the use of differential sensing for a lower latency implementation. The 8$^+$T cell also offers good reliability towards global variations including threshold voltage mismatch and temperature changes as these variations tend to have a symmetrical effect on both the sides of the differential sense amplifier. Fig. 13a, 13b and 13c show the trend of latency, average energy per bit and throughput respectively for 6T, 8TC, 8TD and 8$^+$T SRAM-IMC structures for typical process corner at room temperature. The single ended inverter based sensing used in 8TC cell shows the slowest performance across all the supply voltages while consuming least energy per bit compared to all the other cells. The three differential cells considered; 6T, 8TD and 8$^+$T all give very similar latency and average energy per bit with 8$^+$T cell being slightly faster compared to others. The 8$^+$T SRAM-IMC structure

Just Accepted

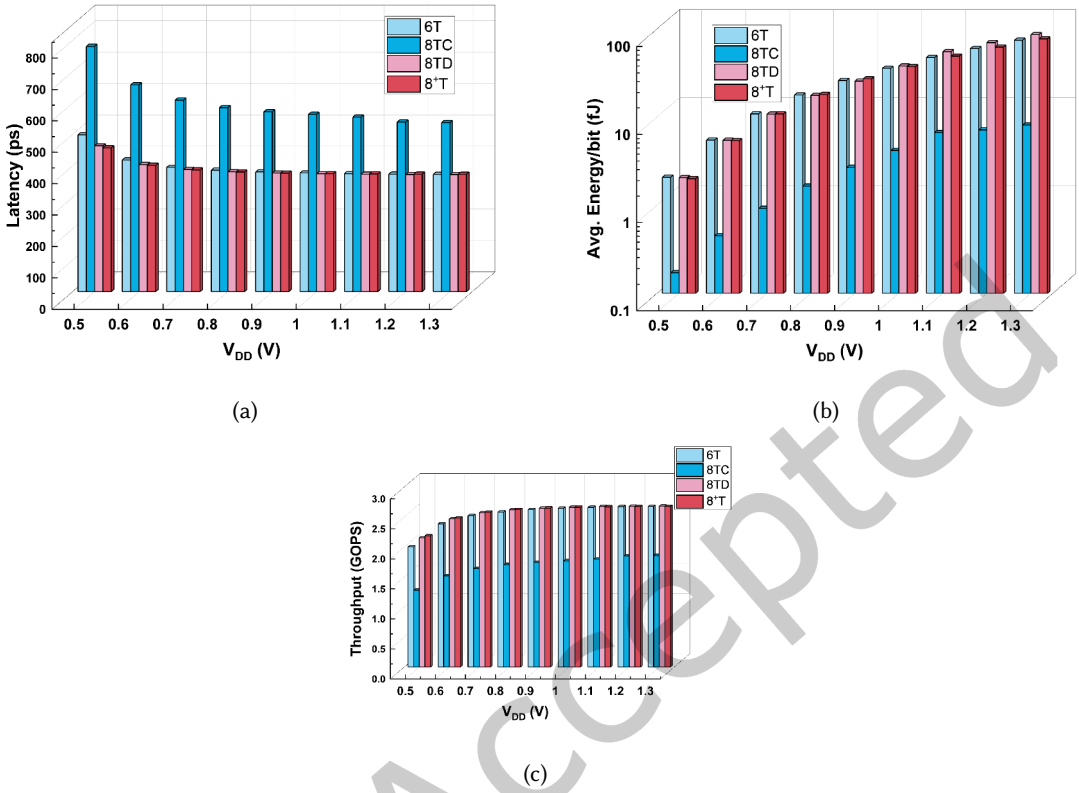


Fig. 13. Performance comparison of (a) Latency, (b) Average energy/bit and (c) Throughput for all the four cells; 6T, 8TC, 8TD, 8$^+$T SRAM cells for implementation of logic functions (NAND and NOR) as a function of supply voltage at TT corner and T = 300.15 K.

Just Accepted

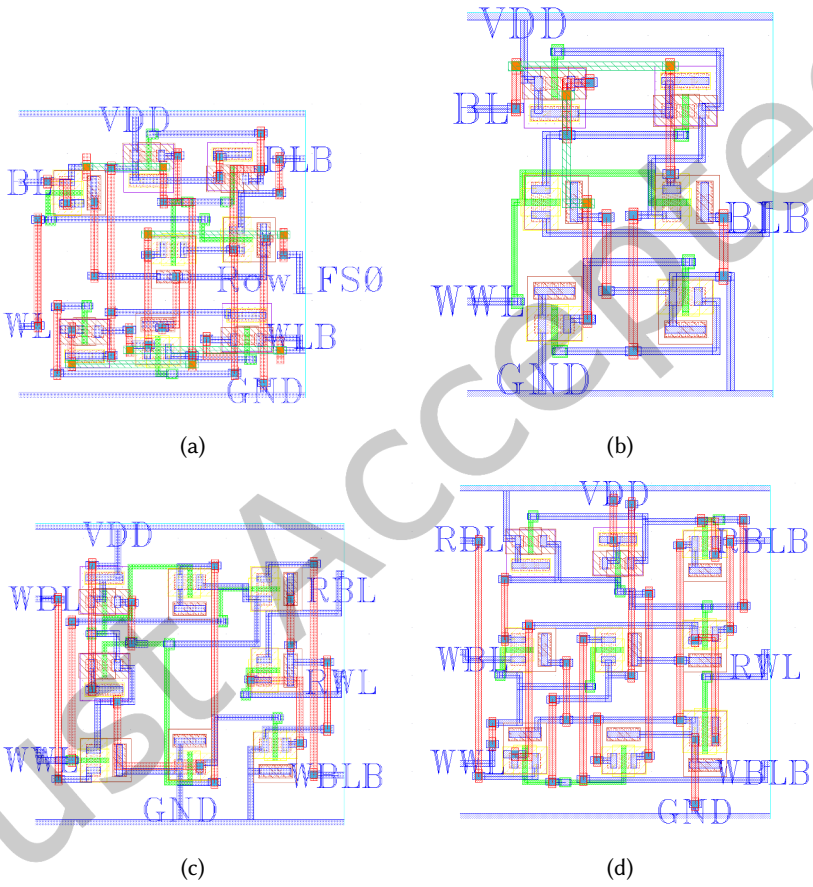
offers best latency and throughput over other conventional SRAM cell based IMC structures while consuming nearly same energy per bit computed with very minimal overhead over the conventional 6T SRAM cell area. These parameters are highlighted in Table 1. The layouts of the cells are also shown in Fig. 14 where all the cells are designed using design rules for 32 nm technology and show no DRC (design rule check) errors and have undergone layout versus schematic check. The layout of the conventional SRAM cells are designed using automated place and route procedure for a minimum separation of 1 while a fully custom layout is drawn for the 8$^+$T SRAM-IMC to reduce area overhead.

5 Conclusion

The recent state-of-the art dictates high throughput machines which focus on data-intensive applications such as cryptography, machine learning, smart vision devices etc. are the essential. These necessitate the use of alternatives to the conventional von Neumann architectures that have been the mainstay in computing. In-memory computation (IMC) has been brought to the foray as a promising candidate which aims to reduce the constant memory access in conventional architectures together with it's associated energy consumption by

Table 1. Performance Summary of Conventional SRAM-IMC designs with 8⁺T SRAM at V_{DD} = 0.8 V

SRAM-IMC	Latency (ns)	Avg. Energy/bit (fJ)	Throughput (GOPS)	Normalized Area
6T	386.982	17.805	2.584	1
8TC	586.166	1.647	1.706	1.264
8TD	381.675	17.628	2.620	1.394
8⁺T	**381.097**	**18.0895**	**2.624**	**1.06**



JSC Accepted

Fig. 14. Layout of all the four cells SRAM Cells; (a) 8⁺T SRAM (b) 6T (c) 8TC and (d) 8TD considered for implementation of logic functions.

reducing instead allowing computations to take place within the memory array with only slight changes to the cell structures or the peripheral circuitry employed such as the sense amplifier. This work presents a modified SRAM-IMC strategy using conventional 8⁺T bit cell to realize a computationally complete logic operations (NAND and NOR). The reliability of the logic operations is established using Monte Carlo simulations against

global variations in operational conditions such as process, voltage and temperature using Predictive Transistor Models (PTM) at 32 nm technology node. The 8$^+$T SRAM-IMC's logic performance is compared to other IMC structures built around conventional 6T and 8T SRAM cells. The 8$^+$T SRAM-IMC offers 41.42%, 38.82% and 36.43% improvement in latency over conventional 8T SRAM-IMC at a supply voltage of 0.5, 0.6 and 0.7 V respectively.

References

Amogh Agrawal, Akhilesh Jaiswal, Chankyu Lee, and Kaushik Roy. 2018. X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories. *IEEE Transactions on Circuits and Systems I: Regular Papers* 65, 12 (2018), 4219–4232. https://doi.org/10.1109/TCSI.2018.2848999

Nanoscale Integration Arizona State University and Modeling Group. 2008. Predictive Technology Models. http://ptm.asu.edu/latest.html

Yunkai Bai, Andrew Stern, Jungmin Park, Mark Tehranipoor, and Domenic Forte. 2022. RASCv2: Enabling Remote Access to Side-Channels for Mission Critical and IoT Systems. *ACM Trans. Des. Autom. Electron. Syst.* 27, 6, Article 65 (jun 2022), 25 pages. https://doi.org/10.1145/3524123

Jian Chen, Wenfeng Zhao, Yuqi Wang, Yuhao Shu, Weixiong Jiang, and Yajun Ha. 2022a. A Reliable 8T SRAM for High-Speed Searching and Logic-in-Memory Operations. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 30, 6 (2022), 769–780. https://doi.org/10.1109/TVLSI.2022.3164756

Jian Chen, Wenfeng Zhao, Yuqi Wang, Yuhao Shu, Weixiong Jiang, and Yajun Ha. 2022b. A Reliable 8T SRAM for High-Speed Searching and Logic-in-Memory Operations. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 30, 6 (2022), 769–780. https://doi.org/10.1109/TVLSI.2022.3164756

Lixia Han, Peng Huang, Yijiao Wang, Zheng Zhou, Yizhou Zhang, Xiaoyan Liu, and Jinfeng Kang. 2022. Efficient Discrete Temporal Coding Spike-Driven In-Memory Computing Macro for Deep Neural Network Based on Nonvolatile Memory. *IEEE Transactions on Circuits and Systems I: Regular Papers* 69, 11 (2022), 4487–4498. https://doi.org/10.1109/TCSI.2022.3194918

Wangxin He, Shihui Yin, Yulhwa Kim, Xiaoyu Sun, Jae-Joon Kim, Shimeng Yu, and Jae-Sun Seo. 2020. 2-Bit-Per-Cell RRAM-Based In-Memory Computing for Area-/Energy-Efficient Deep Learning. *IEEE Solid-State Circuits Letters* 3 (2020), 194–197. https://doi.org/10.1109/LSSC.2020.3010795

Supreet Jeloka, Naveen Bharathwaj Akesh, Dennis Sylvester, and David Blaauw. 2016. A 28 nm Configurable Memory (TCAM/BCAM/SRAM) Using Push-Rule 6T Bit Cell Enabling Logic-in-Memory. *IEEE Journal of Solid-State Circuits* 51, 4 (2016), 1009–1021. https://doi.org/10.1109/JSSC.2016.2515510

Zhewei Jiang, Shihui Yin, Jae-Sun Seo, and Mingoo Seok. 2020. C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism. *IEEE Journal of Solid-State Circuits* 55, 7 (2020), 1888–1897. https://doi.org/10.1109/JSSC.2020.2992886

Jaydeep P. Kulkarni, Ashish Goel, Patrick Ndai, and Kaushik Roy. 2011. A Read-Disturb-Free, Differential Sensing 1R/1W Port, 8T Bitcell Array. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 19, 9 (2011), 1727–1730. https://doi.org/10.1109/TVLSI.2010.2055169

Zhiting Lin, Shaoying Zhang, Qian Jin, Jianping Xia, Yunwei Liu, Kefeng Yu, Jian Zheng, Xiaoming Xu, Xing Fan, Ke Li, Zhongzhen Tong, Xiulong Wu, Wenjuan Lu, Chunyu Peng, and Qiang Zhao. 2023a. A Fully Digital SRAM-Based Four-Layer In-Memory Computing Unit Achieving Multiplication Operations and Results Store. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 31, 6 (2023), 776–788. https://doi.org/10.1109/TVLSI.2023.3266651

Zhiting Lin, Shaoying Zhang, Qian Jin, Jianping Xia, Yunwei Liu, Kefeng Yu, Jian Zheng, Xiaoming Xu, Xing Fan, Ke Li, Zhongzhen Tong, Xiulong Wu, Wenjuan Lu, Chunyu Peng, and Qiang Zhao. 2023b. A Fully Digital SRAM-Based Four-Layer In-Memory Computing Unit Achieving Multiplication Operations and Results Store. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 31, 6 (2023), 776–788. https://doi.org/10.1109/TVLSI.2023.3266651

Soumitra Pal and Aminul Islam. 2016. Variation Tolerant Differential 8T SRAM Cell for Ultralow Power Applications. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 35, 4 (2016), 549–558. https://doi.org/10.1109/TCAD.2015.2474408

Bhawna Rawat and Poornima Mittal. 2023. A Reconfigurable 7T SRAM Bit Cell for High Speed, Power Saving and Low Voltage Application. *ACM Trans. Des. Autom. Electron. Syst.* 28, 6, Article 96 (oct 2023), 14 pages. https://doi.org/10.1145/3616872

Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, H.-S. Philip Wong, and Gert Cauwenberghs. 2022. A compute-in-memory chip based on resistive random-access memory. *Nature* 608, 7923 (01 Aug 2022), 504–512. https://doi.org/10.1038/s41586-022-04992-8

H.-S. Philip Wong and Sayeef Salahuddin. 2015. Memory leads the way to better computing. *Nature Nanotechnology* 10, 3 (01 Mar 2015), 191–194. https://doi.org/10.1038/nnano.2015.29

Bi Wu, Kai Liu, Tianyang Yu, Haonan Zhu, Ke Chen, Chenggang Yan, Erya Deng, and Weiqiang Liu. 2023. High-Performance STT-MRAM-Based Computing-in-Memory Scheme Utilizing Data Read Feature. *IEEE Transactions on Nanotechnology* 22 (2023), 817–826. https://doi.org/10.1109/TNANO.2023.3336910

Jiayao Wu, Yijiao Wang, Pengxu Wang, Yiming Wang, and Weisheng Zhao. 2024. A STT-Assisted SOT MRAM-Based In-Memory Booth Multiplier for Neural Network Applications. *IEEE Transactions on Nanotechnology* 23 (2024), 29–34. https://doi.org/10.1109/TNANO.2023.

Just Accepted

3343834

Zihao Xuan, Chang Liu, Yue Zhang, Yuan Li, and Yi Kang. 2023. A Brain-Inspired ADC-Free SRAM-Based In-Memory Computing Macro With High-Precision MAC for AI Application. *IEEE Transactions on Circuits and Systems II: Express Briefs* 70, 4 (2023), 1276–1280. https://doi.org/10.1109/TCSII.2022.3224049

Shihui Yin, Zhewei Jiang, Jae-Sun Seo, and Mingoo Seok. 2020. XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks. *IEEE Journal of Solid-State Circuits* 55, 6 (2020), 1733–1743. https://doi.org/10.1109/JSSC.2019.2963616

Chengshuo Yu, Taegeun Yoo, Hyunjoon Kim, Tony Tae-Hyoung Kim, Kevin Chai Tshun Chuan, and Bongjin Kim. 2021. A Logic-Compatible eDRAM Compute-In-Memory With Embedded ADCs for Processing Neural Networks. *IEEE Transactions on Circuits and Systems I: Regular Papers* 68, 2 (2021), 667–679. https://doi.org/10.1109/TCSI.2020.3036209

Junwei Zeng, Nuo Xu, Yabo Chen, Chenglong Huang, Zhiwei Li, and Liang Fang. 2022. AIMCU-MESO: An In-Memory Computing Unit Constructed by MESO Device. *ACM Trans. Des. Autom. Electron. Syst.* 28, 1, Article 8 (dec 2022), 16 pages. https://doi.org/10.1145/3539575

Fan Zhang, Shaahin Angizi, Jiao Sun, Wei Zhang, and Deliang Fan. 2023. Aligner-D: Leveraging In-DRAM Computing to Accelerate DNA Short Read Alignment. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 13, 1 (2023), 332–343. https://doi.org/10.1109/JETCAS.2023.3241545

Received 19 May 2024; revised 20 August 2024; accepted 15 September 2024

Just Accepted

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.2024.3461588

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

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

Digital Object Identifier <doi>

Resp-BoostNet: Mental Stress Detection from Biomarkers Measurable by Smartwatches using Boosting Neural Network Technique

SANJAY KUMAR[1], ANSHUMAN RAJ CHAUHAN[2], AKHIL[3],AKSHI KUMAR[4], (Senior Member, IEEE) and GUANG YANG[5,6,7,8], (Senior Member, IEEE)

[1]Dept. of Computer Science and Engineering, Delhi Technological University, Delhi, India. (e-mail: sanjay.kumar@dtu.ac.in)
[2]Dept. of Computer Science and Engineering, Delhi Technological University, Delhi, India. (e-mail: its7arc@gmail.com)
[3]Dept. of Computer Science and Engineering, Delhi Technological University, Delhi, India. (e-mail: sah.akhil.01@gmail.com)
[4]Department of Computing, Goldsmiths, University of London, 8 Lewisham Way, London, United Kingdom. (e-mail: akshi.kumar@gold.ac.uk)
[5] Bioengineering Department and Imperial-X, Imperial College London, London W12 7SL, UK. (e-mail: g.yang@imperial.ac.uk)
[6]National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK.
[7] Cardiovascular Research Centre, Royal Brompton Hospital, London SW3 6NP, UK
[8] School of Biomedical Engineering & Imaging Sciences, King's College London, London WC2R 2LS, UK

Corresponding author: Guang Yang (e-mail: g.yang@imperial.ac.uk)

This study was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, Wellcome Leap Dynamic Resilience, and the UKRI Future Leaders Fellowship (MR/V023799/1).

ABSTRACT To maintain overall health and well-being, it is crucial to manage mental stress. This study focuses on developing a deep learning model for recognizing mental stress levels using the sensors of smartwatches. Most related research with notable performance has focused on mental stress detection using various physiological biomarkers obtained through sophisticated IoMT (Internet of Medical Things) devices. However, the ones using only the smartwatch's measurable physiological biomarkers, which do not include respiration rate, have comparatively lower performance because of a limited number of physiological biomarkers. In this paper, we introduce an improved model for mental stress detection using boosting neural network that can be integrated into a smartwatch. The proposed model consists of two phases. In the first phase, we introduce a boosting neural network technique that predicts the respiration rate by utilizing the biomarkers measurable by a smartwatch. The second phase uses the In the second phase, the modified set of biomarkers, which includes both the original biomarkers and the predicted respiration rate, is used for stress level classification via an artificial neural network. The necessary hyperparameter tuning is performed to obtain the optimal values of various model parameters. The training of the model is performed for fifteen different subjects of the publicly available multimodal WESAD (Wearable Stress and Affect Detection) dataset using various biomarkers measured by smartwatches. The proposed model predicts respiration rate with low error (0.035 MSE (Mean Squared Error)) and achieves high mental stress detection accuracy of 94% using smartwatch measurable biomarkers which is a $\sim$2% improvement over the current contemporary technique.

INDEX TERMS Artificial Neural Network, Boosting Neural Network, Deep Learning, Mental Stress Detection, Regression, Respiration rate

I. INTRODUCTION

MENTAL stress can be termed as psychological or emotional strain triggered by physical or mental discomfort felt while performing some demanding task or in some cases while performing some daily life routine. Mental stress can be caused by various factors such as work pressure, financial difficulties, relationship issues, traumatic events, or chronic health conditions. Mental stress is not necessarily 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.2024.3461588

negative thing and can be considered a normal component of life. It can be classified as positive mental stress which is generally acute mental stress and negative mental stress which is chronic mental stress or long-term mental stress [1]. In certain situations, to get optimal human output or to enter a state of motivation, short-term stress plays a vital role. But prolonged mental stress results in various health issues such as slower body recovery, raised blood pressure, bad sleep quality, increased vulnerability to infections, and decreased mental performance [2]. Prolonged or frequent exposure to stressors, which can be the result of performing vigorous activities, can lead to chronic mental stress. While performing vigorous activities the body undergoes a noticeable physical and psychological change that results in increased metabolic demand [3], [4]. To fulfill the increased demands, the body's sympathetic nervous system speeds up certain processes within the body such as heart rate, respiration rate, sweating rate, etc. The sympathetic nervous system slows down and the parasympathetic nervous system begins the rest and repair process [5] after the completion of intense activity. A balance between the two processes is desired for the normal functioning of the body. However, frequently working in a stressful environment can lead to higher responses from a sympathetic nervous system that may result in chronic mental stress. Chronic mental stress has an adverse effect on not just the mental but also the physical health of a person. Citing the various ill effects of mental stress and the increased mental stress levels across the population throughout the world due to the recent pandemic situation [6], a mental stress detection system has become a must for the self-management of mental stress.

There is a lack of awareness among people regarding mental stress which has made people unaware of their mental stress behavior. Due to this reason, it becomes difficult to collect stress-related data for research purposes. Often time researchers have to rely on self-assessments provided by volunteers [7]. Since stress condition is accompanied by some physiological and psychological changes, the usage of sensors to measure those changes provides another means to collect the data [8]. Surveys based on self-assessments often time provides biased data which makes the data quite unreliable. So, instead of relying on surveys for data collection, more accurate and reliable data can be collected easily using modern-day's advanced sensors that make measurements along three main modalities namely physiological changes, psychological changes, and behavioral changes. Mental stress detectors that can capture specific and effective physiological effects of mental stress are quite reliable in terms of physiological data collection. Physiological data is robust to distortions and it also does not get obscured. Biopacs, RespiBAN professional, Empatica, E4, MP150 etc can be used to measure physiological changes to produce good results. Few researchers have tried creating datasets for mental stress detection using these devices by measuring various physiological data. Healey et al. [9] published a dataset with features such as ECG (Electrocardiogram), EDA (Electrodermal Activity), RESP (Respiration Rate), and EMG (Electromyogram) data on driver stress. Koestra et al. [10] published DEAP (Dataset for Emotion Analysis using EEG, Physiological and video signal), a database for emotion analysis using physiological signals. The dataset contained EEG (Electroencephalogram) signals, facial videos, and peripheral physiological signals. More recently, Schmidt et. al. [11] published WESAD. This dataset has a rich set of features that includes signals such as EMG, EDA, ECG, RESP, Body Temperature, and three axes Acceleration. However, using sensor-based devices has its own limitations because of the subjective nature of mental stress. Different people perceive mental stress in different ways and their response (physiological changes) to mental stress also varies. Hence citing the reliability factor, a better alternative becomes combining sensor measurements with self-assessments. Wrist-worn devices can recognize mental stress quite reliably and many recent studies have been focussed on improving the reliability of predictions of these devices [12]. For instance, Gjoreski et. al. [13] used skin temperature, electrodermal activity, heart rate, blood volume pulse, and accelerometer signals to detect mental stress levels under laboratory conditions and were able to detect it with an accuracy of 83%. Remirez et al . [14] also used physiological biomarkers (heart rate, galvanic skin response and breath) for mental stress detection.On the other hand, Bannore et. al. [15] used facial expression recognition system to detect stress levels. Many of the studies related to mental stress detection use respiration rate in the recognition process [16] along with other general biosensors.

Many of the existing studies to detect mental stress use sophisticated chest-wearable IoMT devices for the measurement of physiological biomarkers [11]. However, wearing such devices is impractical for daily use. On the other hand, the studies utilizing physiological biomarkers collected through convenient devices such as a smartwatch, exhibit comparatively lower performance due to a limited range of physiological indicators. One such missing physiological biomarker is respiration which is an important indicator of mental stress as evidenced by multiple research showing a positive correlation between mental stress and respiration rate [17] [18] [19]. Our findings from a series of experiments also aligned with these researches and we also conclude that respiration rate is an important biomarker that can be used to detect mental stress levels with high accuracy (Tab. 5).

In this study, we intend to propose a method that utilizes a convenient wearable device (a smartwatch) for detecting mental stress with high accuracy. The proposed model for mental stress detection, employs a BNN (Boosting Neural Network) to predict respiration rate using smartwatch measurable physiological biomarker and combines it with the biomarker set to ultimately predict stress levels using an ANN (Artificial Neural Network). With the help of some compression techniques [20] [21] [22] [23] the proposed model can be easily integrated within smartwatches. The proposed model consists of two phases; the first one is a regression task for predicting respiration rate and the second

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.2024.3461588

one is a classification task for detecting the current mental stress level (affective state). Combining the predicted respiration rate as a feature in the existing data resulted in performance improvement of the classifier as compared to its performance in the absence of a respiration rate. We compare the performance of the proposed model with various baseline and contemporary methods of mental stress detection. The obtained results reveal the superiority of the proposed model over other comparing models. The major contributions of this work are as follows.

1) A deep learning-based improved model, named RespBoostNet, is introduced for detecting mental stress from physiological biomarkers measurable by smartwatches.
2) We effectively perform dataset aggregation to reduce the number of samples and dataset balancing, without any qualitative loss of data.
3) A Boosting Neural Network-based regression model is introduced to predict respiration rate using the physiological biomarker measured by a smartwatch and append that data to the feature vector of each sample from the dataset.
4) We perform intensive experimentation using the WESAD dataset and evaluate various performance metrics. The obtained results reveal the superior performance of the proposed model against the recently proposed model for mental stress detection.

The organization of the rest of the paper is as follows: Section II comprises the related work. A brief description of the dataset and evaluation metrics used is provided in Section III. Section IV contains details about the proposed methodology. Experimental analysis and performance evaluation are illustrated in Section V followed by concluding remarks and future scope in Section VI.

II. RELATED WORK

Due to the consequences of mental stress on health, there is extensive research on mental stress recognition. This section discusses some of the recent work in the domain of mental stress detection.

The frequency at which multi-modal devices take measurements is generally quite high which results in the generation of tons of data. Thus, the focus of some research pivots around finding efficient data clustering techniques to reduce the number of data points to make it easier for stress detection models. In this direction, Kumar et al. [24] proposed a genetically optimized Fuzzy C-means data clustering technique. The clustering helped in summarizing the data for each subject. They used a multi-level CNN (Convolutional Neural Network) architecture for the classification task. Divya et al. [25] proposed a similar method for clustering. They utilized time-frame restricted intra-similarity calculations to perform k-medoid data clustering to obtain summarized data. Hovsepian et al.'s [26] work primarily focused on the efficient pre-processing of data recorded by an AutoSense sensor. They performed interpolation and normalization on the ECG and respiration rate data and then used the refined data for mental stress prediction. They achieved an accuracy of 89% on the validation set and 72% on the test set using an SVM (Support Vector Machine) classifier. Sharma et al. [27] extracted entropy based features from EEG signals to detect the mental stress level using a support vector classifier. To enhance the performance of their SVM classifier they also used evolutionary optimizing techniques. Samarpita et al. [28] also worked with EEG signals and proposed an algorithm for stress level detection using EEG data. They used the bandpass filter method to separate EEG signals into EEG rhythms and normalized EEG signals using the k-mean clustering method.

Most of the research in the domain of mental stress detection combines the multi-modal data collected from wearable smart devices and sensors with machine learning techniques to estimate human affective state [29]–[34]. Sagbas et al. [35] and Garg et al. [36] leveraged smartphone and wearable devices for mental stress detection while performing daily routine activities. Arif et al. [37] also used smartphones to collect accelerometer and gyroscope sensor data of writing behavior for 46 subjects. They used several machine learning algorithms for classification and obtained the highest accuracy of 87.56% for the k-nearest Neighbour method. Gjoreski et al. [30] devised a method for continuous mental stress detection using data from wrist devices. Their method had three components: a laboratory mental stress detector, built using a random forest classifier, for detecting short-term mental stress; an activity recognizer to continuously monitor the user's activity; and a context-based mental stress detector that leverages the outputs of other two components to generate the final decision after every 20 minutes. This component was developed using an SVM classifier. Siirtola et al. [12] demonstrated the mental stress detection accuracy for various standard classifiers using smartwatch measurable features. The author also studied the importance of various physiological biomarkers for mental stress detection and concluded that EDA signals are unnecessary for recognizing mental stress when the window length is big enough. The window length here refers to the length of time considered for converting analog signals such as ECG to discrete data points. Zhu et al. [38] examined the classification performance of four different electrodermal activity databases for stress detection using several machine learning models. They achieved an accuracy of 92.9 for the approach using SVM.

Some recent researchers have tried to leverage the power of different deep-learning techniques for mental stress detection. Kumar et al. [39] proposed a multi-level deep neural network having hierarchical learning capabilities of CNN. To combine the high-level features into a single unified representation they used a model-level fusion strategy. The proposed model had an accuracy of 87.7% on the WESAD dataset. Yu et al. [40] proposed modality fusion and personalized attention mechanism for dealing with missing modalities and adapting the classification model for every user differently. Something along a similar line was done by Ahmed et

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.2024.3461588

al. [41]. They proposed a method that develops a multimodal diagnosis system that takes patient specific factors into consideration. Their model consists of an attention based multimodal classifier with selective dropout and normalization. Through rigorous experimentation on three different multimodal datasets, they achieved an F1 score of 0.945. Khan et al. [42] performed a comparison between supervised that used CNN combined with bi-directional LSTM (Long Short Term Memory) and semi-supervised (used Generative Adversarial Network) techniques to detect mental stress. Their study showed that semi-supervised techniques fairly outperformed fully supervised techniques.

Mou et al. [51] used an attention based CNN and LSTM model for driver stress detection. Iqbal et al. [52] tried to assess the relative sensitivity and specificity of common physiological biomarkers that are indicators of mental stress. They performed a statistical analysis of various physiological data gathered from healthy individuals and concluded that respiration rate and heart rate are the two best features for detecting the state of mental stress. Indikawati et al. [53] used standard machine learning classifiers to develop a personalized mental stress detection system using the WESAD dataset. They were able to achieve quite high accuracy for specific subjects. In summary, Tab. 1 lists the various research works studied in this article. The papers are listed in the order of appearance in the related work section.

All these works either used a limited number of physiological biomarkers, which limits their performance, or used physiological biomarkers that are measured by some sophisticated devices that are not suitable to be worn on a daily basis. This creates a trade-off between the performance of the stress detection model and the convenience of the user. The proposed work intends to bridge this gap by generating an additional physiological biomarker with the help of a strong deep learning based regressor, using only the smartwatch's measurable physiological biomarkers.

III. DATASETS AND EVALUATION METRICS

This section describes the details of the dataset used for training the proposed model and also the various evaluation metrics used to evaluate the performance of the proposed model.

A. DATASET

The model has been trained on the publicly available WESAD dataset. Schmidt et. al. [11] performed a lab study to collect the physiological and motion data of 17 subjects. During the experiments the devices attached to two of the subjects, for measuring their physiological changes, malfunctioned and hence the dataset contains data for only 15 subjects. The measurements were taken for a period of two hours through chest wearable (RespiBAN Professional) and wrist wearable (Empatica E4) IoMT device. A total of 12 biomarkers were measured out of which we used six in this work. The details of the same are listed in Tab. 2.

The dataset contains three different affective states (neutral, stress, and amusement) which is an improvement over previous datasets in this field which were majorly limited to two affective states, namely stress and no stress.

B. EVALUATION METRICS

Evaluation metrics are used to assess the performance of an algorithm. For assessing the performance of our proposed work we used various evaluation metrics. The usage of multiple evaluation metrics helps in better assessment of the performance of different techniques. For assessing the performance of classification tasks we used accuracy, precision, recall, and F1 score, while for regression tasks we used mean square error.

IV. PROPOSED WORK

In this section, we present our proposed model for mental stress detection using various physiological biomarkers measured by a smartwatch. First, we perform necessary data pre-processing, which consists of four tasks; (i) Data discretization to generate discrete samples from time series data, (ii) Data aggregation to reduce the number of samples for computational feasibility, (iii) Data set balancing to minimize chances of biasing and (iv) Data normalization for improving model's performance. The proposed model consists of two phases: (a) Regression task for predicting respiration rate for which we combined boosting technique with a neural network and (b) Classification task for detecting mental stress levels using an ANN.

Using the pre-processed data, we tried to ascertain the importance of each biomarker. For this task, we remove each biomarker, one at a time, from the feature vector of the samples of the dataset while keeping the rest biomarkers in the feature set and perform the classification task using an ANN. This analysis helped in assessing the effectiveness of all the biomarkers in detecting mental stress levels. We found that respiration rate, a biomarker not measured by currently available smartwatches (although in the WESAD dataset, ECG and EMG are measured by a chest-worn device, they can be measured by currently available smartwatches [54]), had a significant role in deciding the affective state of a person. Our analysis pointed out that the inclusion of respiration rate boosted the performance of the classifier. Hence, we propose to use a regression model to predict the data for respiration rate using other physiological biomarkers measured by smartwatches, append the predicted respiration rate to the feature vector, and then finally perform the classification task using the modified data. For the regression task, we introduce a BNN. Fig. 1 depicts the flowchart of our proposed work for mental stress detection. The successive sub-sections contain a detailed description of the various steps of the proposed method. Tab. 3 tabulates the various symbols that are used in the successive sub-sections.

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.2024.3461588

TABLE 1: A summary of various research papers studied by us in the related work section.

Ref No. & Year	Dataset used	Techniques used	Result
[12] 2019	WESAD	1. The importance of various physiological biomarkers was analyzed for mental stress detection 2. The accuracy of various standard machine learning classifiers for mental stress detection was demonstrated.	1. An accuracy of 87.4% was achieved. 2. Authors concluded that for larger window length (time interval used for discretization of analog signals) EDA signals are unnecessary for mental stress detection.
[24] 2021	WESAD	1. A genetically optimized Fuzzy C-means data clustering technique was used to generate summarized data. 2. A multi-level CNN was used for mental stress prediction	An average model accuracy of 87.7% was achieved.
[25] 2021	WESAD	1. Time-frame restricted intra-similarity calculations were used to perform k-medoid data clustering to obtain summarized data. 2. A CNN was trained on the summarized dataset for classification.	A reduction of 34% in the average execution time was achieved with comparable accuracy to the original WESAD paper.
[26] 2015	Collected data from an independent lab study with 21 participants	1. Interpolation and normalization on the ECG and respiration rate data were performed and then used the refined data for mental stress prediction using SVM classifier.	An accuracy of 72/% was achieved.
[27] 2022	EEG during mental arithmetic tasks dataset [43]	1. EEG signals were decomposed using stationary wavelet transform. 2. Entropy-based features were extracted from the decomposed signals. 3. SVM classifier was used for the classification of affective states. To improve the performance of a classifier, its parameters were optimized using different evolutionary-inspired approaches.	SVM optimized using whale optimization algorithm resulted in the best accuracy of 97.3%.
[30] 2017	Collected data from an independent lab study with 21 participants	1. This method had three components: - A laboratory mental stress detector, for detecting short-term mental stress. - An activity recognizer to continuously monitor the user's activity. - A context-based mental stress detector that leverages the outputs of other two components to generate the final prediction.	The method had a recall of 70% and a precision of 95%.
[38] 2023	CLAS [44], UTD [45], VerBIO [46], and WESAD	1. Different machine learning models were used for stress detection using four different electrodermal activity databases.	1. The SVM model outperformed other techniques and achieved an accuracy of 92.9% for binary classification. 2. The authors concluded that EDA signals significantly improve the model performance.
[39] 2021	WESAD	1. The authors combined the high-level features of data into a single unified representation using a model-level fusion strategy. 2. A multi-level deep neural network having hierarchical learning capabilities of CNN was used for classification.	A performance accuracy of 87.7% was achieved.
[40] 2021	Collected wearable sensor and self-report data from 41 participants	1. A modality fusion and personalized attention mechanism were proposed for dealing with missing modalities and adapting the classification model for every user individually.	The personalized attention strategy had a 77.4% f1-score.
[41] 2023	D-vlog [47], DAIC-WOZ [48] [49], MODMA [50]	1. This method uses a multimodal diagnosis system that takes patient specific factors into consideration. 2. The classifier model consists of an attention based multimodal classifier.	Through rigorous experimentation on three different multimodal datasets, they achieved an F1 score of 0.945.
[42] 2022	WESAD	1. The authors majorly performed a comparative study between supervised (used CNN combined with bi-directional LSTM) and semi-supervised (used Generative Adversarial Network) technique's performance for mental stress detection.	Their study showed that semi-supervised techniques fairly outperformed fully supervised techniques.
[51] 2021	Collected data using a driving simulator, sampled for 22 participants from multiple driving scenarios.	1. The authors used an attention based CNN and LSTM model to fuse non-invasive data that include data collected from driving simulator such as eye data, vehicle data, and environmental data. 2. Usage of self-attention mechanism in their work allows assigning different levels of attention to features from different modalities.	The authors achieved an accuracy of 95.5% for driver stress detection.

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.2024.3461588

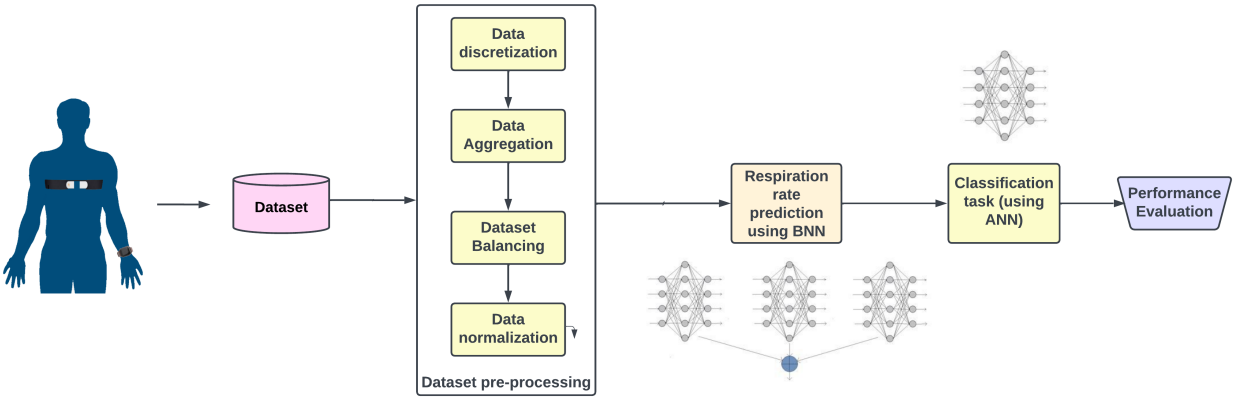


FIGURE 1: Flow diagram of the proposed Boosting Neural Network based framework for stress detection using wearable smartwatches

TABLE 2: Some of the physiological biomarkers present in the WESAD dataset

Biomarker	**Physiological changes measured**	**Measuring device**
Electrocardiogram	Frequency of cardiac cycle	RespiBAN
Electromyogram	Musculoskeletal movements Helps in recognizing hand and face gestures.	RespiBAN
Body Temperature	Skin temperature	RespiBAN, Empatica E4
Respiration	Rate of breathing and exhalation	RespiBAN
Electrodermal Activity	Skin conduction	RespiBAN, Empatica E4
Three Axis Acceleration	Motion change in three dimensions	RespiBAN, Empatica E4

TABLE 3: Table of description

Symbol	**Description**
X	Set of datapoints
X_i	i^{th} datapoint in the set of datapoints X
$X_{maximum}$	Maximum value in the set of datapoints X
$X_{minimum}$	Minimum value in the set of datapoints X
X_{in}	Normalised value of X_i
D	Dataset
n	Number of samples in dataset
d	Dimension (number) of feature in dataset
x_i	Feature set of i^{th} datapoint of dataset
y_i	Label of i^{th} datapoint of dataset
$\mathbb{R}$	Set of real numbers
F	The space of weak learners
f_k	k^th weak learner
β	Boosting rate
$(\bar{y_i})_m$	Output of m^th weak learner for i^th datapoint
$\mathcal{E}(x_i)$	Error for i^th datapoint

A. DATA PRE-PROCESSING

WESAD dataset is a huge dataset as each subject's data is recorded at a quite high frequency of 700 Hz. We convert the time series data to a discrete form for which we use a sliding window of 2 seconds. The window is shifted with steps of 0.25 seconds. A lot of samples with redundant values are generated as a result of this discretization step. Hence, for computational feasibility as well as removing the redundant data, the data is aggregated using a window of 5 seconds, and the values in that interval are replaced with their arithmetic mean. This step is unlikely to cause any significant data loss as mental stress levels will not see a drastic change in such a short interval. The next step in our work is to balance the training dataset as the original dataset is unbalanced with the three affective states in an approximate proportion of 1 (amusement) : 3 (neutral): 2 (stress). A model trained on such an unbalanced dataset is prone to some biasing. Hence the dataset is balanced by duplicating data corresponding to affective states having a lesser share in the dataset. We divide data points with labels neutral (baseline) and stress into three and two parts respectively so that each part almost consists of the same number of data points as the data points with amusement labels. Then we make three sets by combining one part from each label as depicted in Fig. 2.

Finally, we concatenate the three sets. We balance the dataset in this manner to increase its size as opposed to reducing it using some standard data clipping methods for dataset balancing. After these pre-processing steps, all the labels in the dataset are approximately in the same proportion.

The above steps is followed by dataset normalization to bring the values of biomarkers into the same range. We use min-max scaling for normalization to bring the value of each biomarker between 0 and 1. Mathematically normalisation is expressed as Eq. 1

$$X_{in} = \frac{X_i - X_{minimum}}{X_{maximum} - X_{minimum}} \tag{1}$$

where $X_{minimum}$ is the minimum of X, $X_{maximum}$ is 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.2024.3461588

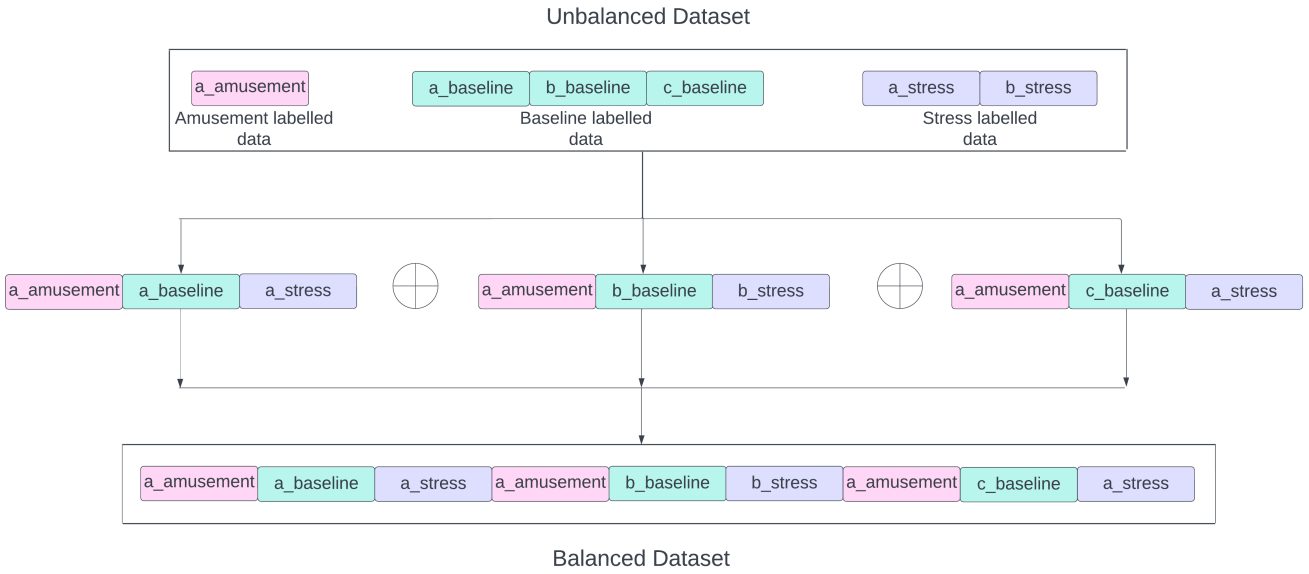


FIGURE 2: Segmenting dataset in multiple parts such that each part contains almost the same number of data points and combining them in a manner that balances the dataset

maximum of X and X_{in} is the normalised value of i^{th} data point X_i of X.

The pre-processed dataset after the above steps consists of 7695 samples in the train set and 1342 samples in the test set having an almost equal share of samples with the three affective states in the train set.

B. MODEL TRAINING

The goal of this work is to develop a model for mental stress detection that can be embedded within a smartwatch. Hence we use only the biomarkers that are measured by a smartwatch. Our proposed work consists of two phases. The first one is a regression model that predicts respiration rate using the physiological biomarkers measured by smartwatches and the second one uses a classifier that detects mental stress levels (three affective states viz amusement (assigned class 0), neutral (assigned class 1), stress (assigned class 2)). For the regression task, we introduce a BNN inspired by the works of Friedman et. al. [55] and Liu et. al. [56]. In their proposed boosting approach Liu et al. use an ANN as weak learners. To train a particular weak learner they combine the prediction results from the penultimate layer of the previous ANN (weak learner) with the original input features. The final output is calculated by combining output from each weak learner. In sub-section IV-B1 we explain our approach of creating a BNN. A simple ANN and an autoencoder were also considered for regression tasks but the proposed BNN outperformed both of them. The second sub-model is an ANN that is used as a classifier for detecting mental stress whose architectural details are included in section IV-B2.

1) Boosting Neural Network: For predicting respiration rate

Boosting algorithms and neural networks both provide good regression power individually. Traditional boosting techniques use a number of weak learners , typically decision trees, and leverages their combined power for regression and classification tasks. We use shallow neural networks instead of decision trees which paves the way for deep learning's modeling capabilities to be used with boosting techniques. The experimental results, discussed in section V-A, support the above claim by showing better results for BNN as compared to a simple ANN.

We use shallow neural networks having two layers as the weak learners for the boosting model. A series of 6 weak learners (the number of weak learners is a hyperparameter) are used for boosting. In the boosting algorithm, each weak learner predicts the error between the predicted value and the actual value of the target. We initialize the predicted values by assigning them a value equal to the mean of the actual values of respiration rates from the WESAD dataset. We train the model by predicting and eliminating the residual prediction error at each step using weak learners. The rate of learning is controlled by a user-defined hyperparameter, namely the boosting rate. We keep the boosting rate as the reciprocal of the number of weak learners to allow each weak learner to mitigate a fraction of the prediction error such that the overall model has minimal error. We formally present the mathematical formulation of our model below.

Considering a dataset D with n samples having d dimensional feature space

$$D = \{(x_i, y_i)_{i=1}^{n} \mid x_i \in \mathbb{R}^d,\ y_i \in \mathbb{R}\} \tag{2}$$

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.2024.3461588

Then the output of BNN will be given using K additive function given by Eq. 3

$$\bar{y_i} = \sum_{k=0}^{K} \beta_k f_k(x_i),\ f_k \in F \tag{3}$$

where F is the space of weak learners, the neural networks, β_k is the boosting rate K denotes the number of weak learners and each function f_k represents an independent, shallow neural network. The model's prediction is initialized with the mean of actual target values, i.e. $f_0 = \frac{\Sigma_{i=1}^{n}(x_i)}{n}$.

For each weak learner, the target value provided is the difference between the actual value y_i and the intermediate output of BNN given by eq 4. It can be viewed as the overall error value for the BNN model. Mathematically for the m^{th} weak learner the target value provided or alternatively the error function can be represented by eq 5

$$(\bar{y_i})_m = \sum_{k=0}^{m} \beta_k f_k(x_i) \tag{4}$$

$$(\mathcal{E}(x_i))_m = y_i - (\bar{y_i})_m \tag{5}$$

As the training progresses, with the aid of each weak learner the error $\mathcal{E}(x_i)$ is gradually minimized. Fig. 3 gives a diagrammatic representation of the BNN model adopted in this work. In the diagram, D represents the dataset, x represents the feature vector and y represents the target value. The diagram shows the k^{th} and $(k+1)^{th}$ weak learner of the model. The k^{th} layer accepts x as the input and $(y - \bar{y}_{k-1})$ as the target value. $f_k(x)$ denotes the output of the k^{th} weak learner which is then multiplied with boosting rate β_k for the k^{th} weak learner. Finally, $\bar{y}_{k-1}$ is added to that product, and the combined sum forms $\bar{y_k}$ which is passed on to the next weak learner, in the same manner, $\bar{y}_{k-1}$ was passed to k^{th} weak learner.

2) Artificial Neural network: For classification

The mental stress detection problem can be viewed as a multi-class or multi-label classification problem. Three different affective states viz. Amusement (class 0), Baseline or neutral (class 1), and Stress (class 2) constitute the target classes. For the classification task, we employ an ANN. For choosing the optimal architecture of the ANN we perform extensive hyperparameter tuning. The proposed ANN architecture is four layers deep with 64, 128, 64, and 1 being the number of neurons in each layer respectively. The ANN uses the ReLU (Rectified Linear Unit) activation function at each layer, Adam optimizer for optimized gradient descent, and a learning rate of 0.1.

C. OVERALL PROPOSED ALGORITHM

Algorithm 1 summarizes the steps of our proposed method. In the algorithm, the variables followed by parentheses denote a function and the ones without them represent data. Physiological biomarkers, namely ECG signals, EDA signals, EMG signals, body temperature, respiration rate, three-axis acceleration, and corresponding labels for affective state, namely amusement, neutral, and stress, form the input of our algorithm. The output of the algorithm consists of evaluated performance metrics. The algorithm starts with preprocessing of the dataset (lines 1 to 4 of algorithm 1). The time series dataset is discretized to generate discrete data samples, aggregated using a suitable window W, balanced to prevent biasing of the model, and finally normalized to improve model performance. Since the WESAD dataset already includes respiration rate, the respiration rate is extracted from the feature set of the pre-processed dataset. Lines 6 and 7 of algorithm 1, depict the regression task using BNN, details of which are presented in Section IV-B1. Boosting Neural Network() takes dataset ($wResp$; stands for dataset without respiration rate) as a parameter. Its other parameters include the number of weak learners(w), boosting rate(β), number of iterations for each weak learner(n), and model to be used as a weak learner (m). The BNN model is depicted as $BNNmodel()$ in the algorithm and produces the respiration rate denoted by $resp$. The predicted respiration rate is combined with the feature set of $wResp$ and the modified data set ($respD$) is fed as input to the classifier (line 9 of algorithm 1), for which we used a four-layer neural network. Finally, in line 10 of the algorithm, the various performance metrics are evaluated.

Algorithm 1 Algorithm for our proposed Resp-BoostNet

Input: ECG signals, EDA signals, EMG signals, body temperature, respiration rate and three-axis acceleration (from WESAD dataset), D; Labels (amusement, neutral, stress), L

Output: Evaluated performance metrics

1. $d \leftarrow$ Discretize the time series dataset D
2. $agg \leftarrow$ Aggregate the data points using a window W
3. $bal \leftarrow$ Bal(agg), to balance the dataset
4. $norm \leftarrow$ Norm(bal), to normalise the dataset
5. $wResp \leftarrow$ Extract respiration rate from feature set of dataset $norm$
6. $BNNmodel() \leftarrow$ Boosting Neural Network($wResp$, w, β, n, m)
7. $resp \leftarrow BNNmodel(wResp)$, To generate respiration rate
8. $respD \leftarrow$ To combine predicted respiration rate $resp$ within feature set of dataset $wResp$
9. $pred \leftarrow$ Classify($respD$), to predict stress level using $respD$
10. $perf \leftarrow$ Eval($pred$, L), to evaluate different performance metrics

return $perf$

D. HYPERPARAMETER TUNING

Hyperparameters are tunable parameters that have a significant impact on the performance of a deep learning 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.2024.3461588

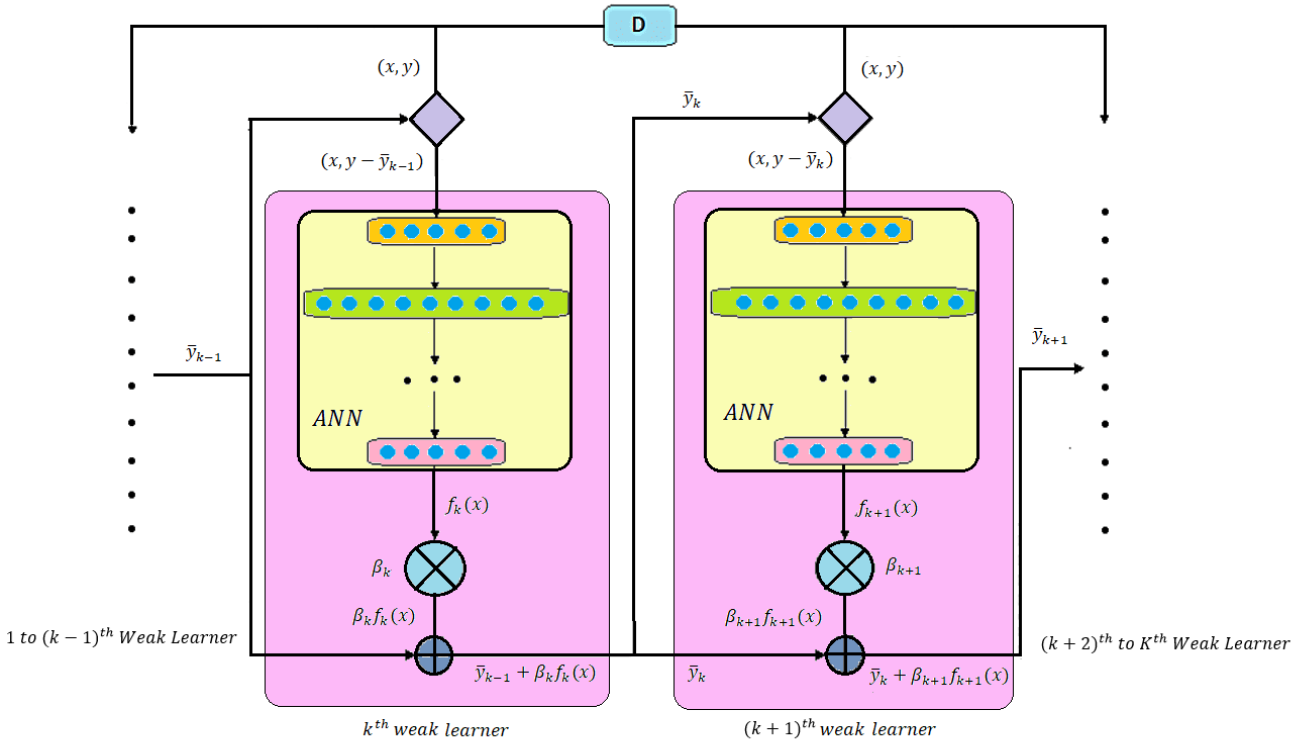


FIGURE 3: The architecture of Boosting Neural Network introduced in this study

model. Hence selection of suitable hyperparameters becomes an essential task. For this work, we used the Random Search technique to find the most appropriate hyperparameters. We iterated over different set of values for the hyperparameters and selected the ones that resulted in the highest accuracy of the model. For the regressor, i.e. the BNN, we tried different sets of number of weak learners (shallow neural network), boosting rate, number of iterations for each weak learner and optimizer to be used during gradient descent. For the classifier, i.e. the ANN, we tried different architectures (number of layers and number of neurons in each layer) and different learning rates. Tab. 4 lists the values of different hyperparameters chosen by us after performing extensive hyperparameter tuning for the regressor and classifier.

TABLE 4: The various hyperparameters for the regressor and the classifier of the proposed work

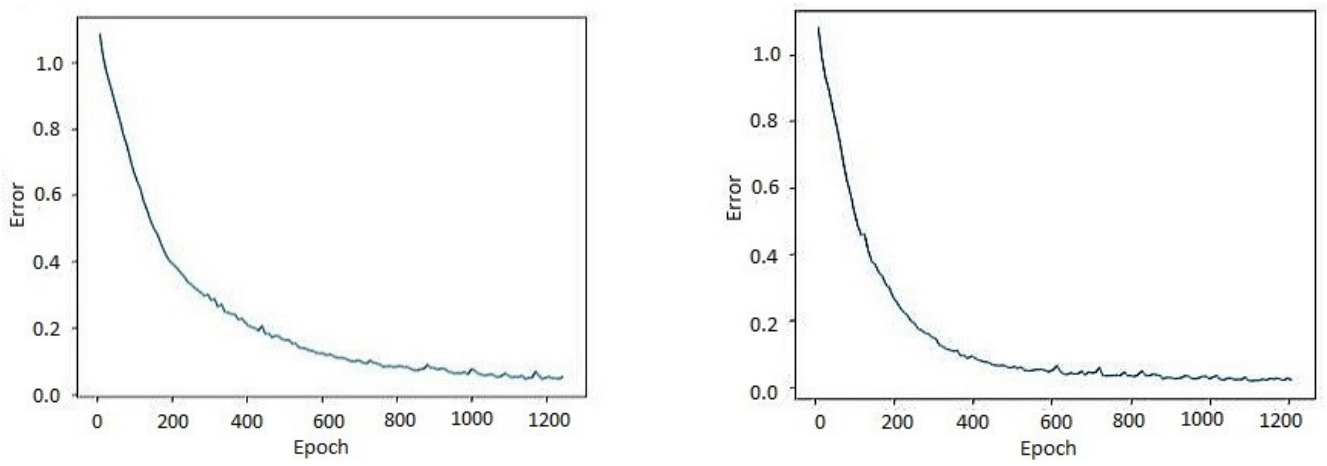
Regressor hyperparameters (BNN)	
Hyperparameter	**Description or value**
Number of weak learners (ANN)	6
Boosting rate	0.166
Iterations for each weak learner	100
Optimizer	Adam
Classifier hyperparameters (ANN)	
Hyperparameter	**Description or value**
Number of layers	4
Neurons in each layer	64, 128, 64, 3
Number of iterations	1250
Learning rate	0.001

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present a comprehensive set of experiments and the performance of the model to demonstrate the effectiveness of our proposed method for mental stress detection. This section also contains a comparative study involving various machine learning approaches namely, Linear and RBF, SVM, DT (Decision Tree), RF (Random Forest), AdaBoost, GNB (Gaussian Naive Bayes), and our proposed ANN as a classifier. We also made a performance comparison of our proposed method with different contemporary techniques that used the WESAD dataset, namely, Sirrtola et. al. [12], Kumar et. al. [24], Sharma et. al. [39], Tran et. al. [57], Iqbal et. al. [52], Indikawati et. al. [53] and Gjoreski et. al. [30] for mental stress detection. Section V-B deals with the first part of experimental analysis, i.e. the comparison with different baseline techniques, while section V-C deals with the latter one and presents a detailed performance comparison with various contemporary techniques listed above. We have used Python programming language for the implementation of the code for the proposed method. The work primarily uses Python's library for machine learning and deep learning in PyTorch, and some other libraries such as sklearn, pandas, and numpy. The experimental results are

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.2024.3461588



(a) (b)

FIGURE 4: Error vs Epoch plot for the proposed model. Figure (a) shows plot for the case when respiration rate is not present in the dataset and figure (b) shows plot for the case when predicted respiration rate is included in the dataset.

TABLE 5: Performance of neural network on WESAD dataset with different sets of biomarkers as feature

Biomarkers	Accuracy%
Acc, Temp, Resp, EDA, ECG, EMG	95.4
Temp, Resp, EDA, ECG, EMG	89.4
Acc, Resp, EDA, ECG, EMG	91.5
Acc, Temp, Resp, ECG, EMG	92.1
Acc, Temp, Resp, EDA, EMG	92.4
Acc, Temp, Resp, EDA, ECG	90.7
Acc, Temp, EDA, ECG, EMG	91.6

TABLE 6: MSE values for Autoencoder, Artificial Neural Network, and Boosting Neural Network when used for regression task using WESAD dataset

Regression model	MSE
Autoencoder	0.15
ANN	0.038
BNN	0.035

discussed below.

A. PERFORMANCE OF AUXILIARY TASKS

The first series of experiments involved in our work was assessing the importance of each biomarker in detecting mental stress levels. As discussed in section IV, we trained a neural network on the WESAD dataset multiple times, each time removing one of the biomarkers from the dataset to observe its impact on the model's performance. The accuracy of the neural network for each case is enlisted in Tab. 5. It is evident from this table that respiration plays a vital role in detecting mental stress.

For predicting respiration rate we proposed a BNN. Tab. 6 shows the MSE values for an autoencoder, an ANN, and the proposed BNN when trained on the WESAD dataset. It is evident from MSE values that the proposed BNN for regression outperforms the autoencoder and a simple ANN.

B. COMPARISON WITH VARIOUS BASELINE TECHNIQUES

This section contains the comparison of various baseline machine learning techniques and our proposed method, RespBoostNet on the WESAD dataset. We trained each model thrice to compare their performance in different scenarios which are; when the respiration rate is removed from the feature set of the WESAD dataset when the respiration rate is used from the WESAD dataset itself and finally when the WESAD dataset's respiration rate data is replaced by respiration rate predicted by the BNN based regressor (Tab. 7). Fig. 4 shows the Error versus Epoch plots for the proposed model. The plots indicate that the inclusion of respiration rate improved the model's performance which is evident from the fact that Fig 4(b) has a better learning curve.

To assess the generalizability of the models and more accurate performance estimation, we also performed k-fold cross-validation. The optimum value of k was found to be 8 and all the reported metrics are evaluated for this value of k. The data from Tab. 7 clearly shows that our proposed method has a better performance as compared to the various baseline methods including the traditional boosting method. Also, the accuracy of the model is significantly improved when the respiration rate is incorporated into the feature set of the WESAD dataset. Furthermore, the performance of the classifier when it is trained on a dataset containing originally measured respiration rate and when the dataset has predicted values is almost similar. Tab. 7 shows the performance of the proposed method, along with other baseline techniques, on the entire WESAD dataset, containing data from 15 subjects. The classifier was also trained on individual data of the 15 subjects and Fig. 5 shows the performance of the classifier

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.2024.3461588

TABLE 7: Performance comparison of our proposed Resp-BoostNet framework with different baseline machine learning techniques with and without using respiration rate

		Methods						
		Linear SVM	RBF SVM	DT	RF	AdaBoost	GNB	**Resp-BoostNet**
Without Respiration Rate	Accuracy (%)	41.4 ± 1.6	70.3 ± 2.0	71.6 ± 2.0	77.3 ± 1.6	76.9 ± 1.0	44.1 ± 0.7	88.5 ± 1.0
	Precision (%)	42.9 ± 1.6	78.4 ± 1.5	79.1 ± 1.6	80.9 ± 1.6	78.3 ± 1.2	49.6 ± 0.8	91.0 ± 0.6
	Recall (%)	41.5 ± 1.5	72.2 ± 1.8	71.5 ± 2.0	77.1 ± 1.6	77.3 ± 1.0	44.9 ± 0.7	88.6 ± 0.9
	F1 Score (%)	36.0 ± 1.5	64.9 ± 2.3	66.7 ± 2.3	74.1 ± 1.8	77.8 ± 1.3	39.3 ± 0.9	90.4 ± 0.7
With WESAD's Respiration Rate	Accuracy (%)	41.4 ± 1.6	70.0 ± 2.1	71.8 ± 2.1	78.6 ± 0.7	77.3 ± 1.2	44.1 ± 0.7	**94.1 ± 0.6**
	Precsion (%)	42.9 ± 1.6	76.9 ± 1.8	79.2 ± 1.6	78.1 ± 1.7	80.1 ± 1.7	49.6 ± 0.8	**95.3 ± 0.5**
	Recall (%)	41.5 ± 1.5	71.6 ± 1.9	72.4 ± 2.0	73.6 ± 1.7	79.7 ± 1.2	44.9 ± 0.7	**95.3 ± 0.5**
	F1 Score (%)	36.0 ± 1.5	63.3 ± 2.4	67.2 ± 2.3	76.8 ± 2.0	79.8 ± 0.5	39.3 ± 0.9	**94.9 ± 0.6**
With Predicted Respiration Rate	Accuracy (%)	41.4 ± 1.5	70.0 ± 2.1	69.9 ± 2.2	78.9 ± 0.5	77.3 ± 1.0	45.3 ± 0.7	**94.0 ± 0.3**
	Precision (%)	42.9 ± 1.6	76.9 ± 1.8	75.2 ± 2.0	76.1 ± 1.2	78.1 ± 1.2	46.1 ± 1.0	**95.5 ± 0.7**
	Recall (%)	41.5 ± 1.5	71.6 ± 1.9	70.7 ± 2.1	81.6 ± 1.7	82.4 ± 0.9	46.2 ± 0.7	**97.0 ± 0.3**
	F1 Score (%)	36.0 ± 1.5	66.3 ± 2.4	66.8 ± 2.4	78.8 ± 1.7	80.8 ± 0.7	40.0 ± 0.9	**96.9 ± 0.6**

TABLE 8: Performance comparison of the proposed Resp-BoostNet framework with different contemporary techniques for mental stress detection (Model's accuracy when trained on data for all 15 subjects of WESAD dataset)

Methods	**Accuracy%**
Gjoreski et al. [30]	85
Siirtola et al. [12]	87.4
Indikawati et al. [53]	92
Kumar et al. [24]	86.8
Sharma et. al. [39]	87.7
Iqbal et al. [52]	85.71
Tran et. al. [57]	87.5
Proposed Resp-BoostNet	**94.0**

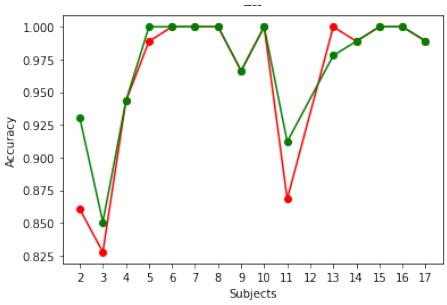


FIGURE 5: Accuracy for each subject for dataset without respiration rate (red plot) and with predicted respiration rate data (green plot)

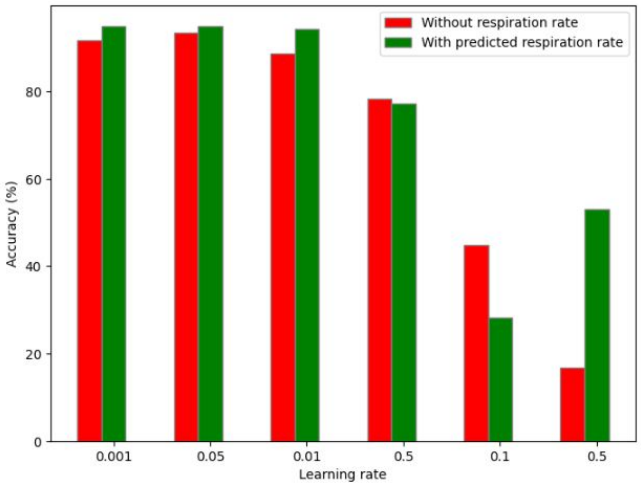


FIGURE 6: A comparison of classification accuracy of Artificial Neural Network for different learning rates with and without predicted respiration rate in the dataset

for each subject when the respiration rate is removed from the WESAD dataset (red plot) and when predicted respiration rate is added to feature set of WESAD dataset (green plot). The figure clearly shows that the inclusion of respiration rate in the feature set improves the classifier's performance for most of the subjects. The classifier used in the proposed work was selected after performing extensive hyperparameter tuning as discussed in IV-D. Fig. 6 shows a comparison of the performance of artificial neural networks with different learning rates on the WESAD dataset without the respiration rate and with the predicted respiration rate in its feature set. Smaller learning rates make the learning process very time-consuming and larger learning rates affect the model's performance hence, the search range was kept between 0.001 to 0.5. The optimal performance was obtained at a learning rate of 0.001

Hence, Tab. 7 clearly shows that our proposed model outperforms all other baseline machine-learning algorithms. AdaBoost-based classifier is the second best performer after the proposed model followed by the Random Forest classifier while Linear SVM is the worst-performing technique because of its limited ability to understand multi-dimensional data.

C. COMPARISON WITH VARIOUS CONTEMPORARY TECHNIQUES

The previous section contained the comparison between various baseline machine learning techniques and the proposed method Resp-BoostNet. This section compares the performance of our proposed method, Resp-BoostNet with several existing contemporary techniques for mental stress detection. Tab. 8 enlists the average accuracy of different techniques and our method for the 15 subjects in the WESAD dataset. 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.2024.3461588

results listed in Tab. 8 demonstrate that the proposed methodology outperforms other methods. The inclusion of predicted respiration rate in the biomarker set increases the number of biomarkers used which explains the improved performance of our method as compared to some of the works (for instance [53]). But some of the related works use the same set of biomarkers as our work but still have comparatively lower performance. The improved performance in those cases could be attributed to the efficient pre-processing of the dataset performed.

VI. CONCLUSION

Mental stress adversely affects the mental as well as physical health of a person and can lead to serious complications in the long run. This forms the pretext for developing an efficient technique for mental stress detection. However, predicting an individual's mental stress level from high-frequency, multivariate real-time biomarkers using multiple wearable devices is a complex process. The existing works either use sophisticated devices for measuring different physiological biomarkers to detect stress with higher accuracy or detect stress with comparatively lesser accuracy using limited physiological biomarkers measured with help of convenient measurable devices such as smartwatches. We tried to narrow this gap by providing a method that can detect stress with high accuracy using a smartwatch which can be conveniently worn. We observed that the respiration rate is an important biomarker for mental stress detection. The proposed model was able to achieve a high detection accuracy and outperform several contemporary methods using the limited biomarkers combined with respiration rate predicted using those limited biomarkers measured by a smartwatch. The study encountered a limitation due to the scarcity of publicly available datasets containing a comprehensive set of biomarkers. Moreover, the proposed model was trained on a single dataset containing data from 15 subjects and hence the trained model is prone to make generalized predictions. As each person possesses a distinct physiology, research focused on the development of personalized detection models is expected to produce significant improvements in results. A semi-supervised model for fine-tuning the proposed mental stress detection model, trained on data from multiple sources, in order to capture personalized physiological changes in response to stress conditions can be a potential direction for future research.

REFERENCES

[1] M. Chesnut, S. Harati, P. Paredes, Y. Khan, A. Foudeh, J. Kim, Z. Bao, and L. M. Williams, "Stress markers for mental states and biotypes of depression and anxiety: A scoping review and preliminary illustrative analysis," Chronic Stress, vol. 5, p. 24705470211000338, 2021, pMID: 33997582. [Online]. Available: https://doi.org/10.1177/24705470211000338

[2] S. Segerstrom and G. Miller, "Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry," Psychological bulletin, vol. 130, pp. 601–30, 08 2004.

[3] A. Angeli, M. A. Minetto, A. Dovio, and P. Paccotti, "The overtraining syndrome in athletes: A stress-related disorder," Journal of endocrinological investigation, vol. 27, pp. 603–12, 07 2004.

[4] A. R. Chauhan, Akhil, and S. Kumar, "Analysing effectiveness of different physiological biomarkers in detecting stress," in 2023 IEEE World Conference on Applied Intelligence and Computing (AIC), 2023, pp. 71–75.

[5] E. Oetter. (2011) Heart rate variability research review. [Online]. Available: http://www.8weeksout.com/2011/12/05/heart-rate-variability-research-review/

[6] B. Javed, A. Sarwer, E. B. Soto, and Z.-U.-R. Mashwani, "The coronavirus (covid-19) pandemic's impact on mental health," The International journal of health planning and management, vol. 35, no. 5, p. 993—996, September 2020.

[7] S. Cohen, T. P. Kamarck, and R. J. Mermelstein, "A global measure of perceived stress." Journal of health and social behavior, vol. 24 4, pp. 385–96, 1983.

[8] N. Parameswaran and D. Venkataraman, "A computer vision based image processing system for depression detection among students for counseling," Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, pp. 503–512, 04 2019.

[9] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 2, pp. 156–166, 2005.

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–31, 2012.

[11] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, ser. ICMI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 400–408. [Online]. Available: https://doi.org/10.1145/3242969.3242985

[12] P. Siirtola, "Continuous stress detection using the sensors of commercial smartwatch," in Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, ser. UbiComp/ISWC '19 Adjunct. New York, NY, USA: Association for Computing Machinery, 2019, p. 1198–1201. [Online]. Available: https://doi.org/10.1145/3341162.3344831

[13] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: In laboratory and real life," ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1185–1193. [Online]. Available: https://doi.org/10.1145/2968219.2968306

[14] A. Salazar-Ramirez, E. Irigoyen, R. Martinez, and U. Zalabarria, "An enhanced fuzzy algorithm based on advanced signal processing for identification of stress," Neurocomputing, vol. 271, pp. 48–57, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523121731216X

[15] A. Bannore, T. Gore, A. Raut, and K. Talele, "Mental stress detection using machine learning algorithm," in 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 2021, pp. 1–4.

[16] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," 2018. [Online]. Available: https://arxiv.org/abs/1811.08854

[17] D. Widjaja, M. Orini, E. Vlemincx, and S. Van Huffel, "Cardiorespiratory dynamic response to mental stress: A multivariate time-frequency analysis," Computational and Mathematical Methods in Medicine, vol. 2013, no. 1, p. 451857, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2013/451857

[18] I. Homma and Y. Masaoka, "Breathing rhythms and emotions," Experimental Physiology, vol. 93, no. 9, pp. 1011–1021, 2008. [Online]. Available: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/expphysiol.2008.042424

[19] R. Jerath, M. W. Crawford, V. A. Barnes, and K. Harden, "Self-regulation of breathing as a primary treatment for anxiety," Applied Psychophysiology and Biofeedback, vol. 40, no. 2, pp. 107–115, Jun 2015. [Online]. Available: https://doi.org/10.1007/s10484-015-9279-8

[20] H. Qin, X. Ma, Y. Ding, X. Li, Y. Zhang, Z. Ma, J. Wang, J. Luo, and X. Liu, "Bifsmnv2: Pushing binary neural networks for keyword spotting to real-network performance," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–13, 2023.

[21] H. Qin, Y. Ding, X. Zhang, J. Wang, X. Liu, and J. Lu, "Diverse sample generation: Pushing the limit of generative data-free quantization," IEEE

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.2024.3461588

Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 11 689–11 706, 2023.

[22] H. Qin, X. Zhang, R. Gong, Y. Ding, Y. Xu, and X. Liu, "Distribution-sensitive information retention for accurate binary neural network," Int. J. Comput. Vision, vol. 131, no. 1, p. 26–47, oct 2022. [Online]. Available: https://doi.org/10.1007/s11263-022-01687-5

[23] H. Qin, M. Zhang, Y. Ding, A. Li, Z. Cai, Z. Liu, F. Yu, and X. Liu, "Bibench: Benchmarking and analyzing network binarization," 2023. [Online]. Available: https://arxiv.org/abs/2301.11233

[24] A. Kumar, K. Sharma, and A. Sharma, "Genetically optimized fuzzy c-means data clustering of iomt-based biomarkers for fast affective state recognition in intelligent edge analytics," Applied Soft Computing, vol. 109, p. 107525, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494621004488

[25] D. Gupta, M. P. S. Bhatia, and A. Kumar, "Resolving data overload and latency issues in multivariate time-series iomt data for mental health monitoring," IEEE Sensors Journal, vol. 21, no. 22, pp. 25 421–25 428, 2021.

[26] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "Cstress: Towards a gold standard for continuous stress assessment in the mobile environment," in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ser. UbiComp '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 493–504. [Online]. Available: https://doi.org/10.1145/2750858.2807526

[27] L. D. Sharma, V. K. Bohat, M. Habib, A. M. Al-Zoubi, H. Faris, and I. Aljarah, "Evolutionary inspired approach for mental stress detection using eeg signal," Expert Systems with Applications, vol. 197, p. 116634, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422001233

[28] S. Samarpita and R. Satpathy, "Eeg-based stress detection using k-means clustering method," in Intelligent Systems and Machine Learning, S. Nandan Mohanty, V. Garcia Diaz, and G. A. E. Satish Kumar, Eds. Cham: Springer Nature Switzerland, 2023, pp. 35–43.

[29] W. Sanchez, A. Martínez-Rebollar, and M. Gonzalez-Mendoza, "Towards job stress recognition based on behavior and physiological features," 11 2017, pp. 311–322.

[30] M. Gjoreski, M. Lustrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," Journal of Biomedical Informatics, vol. 73, 08 2017.

[31] Z. Guo, Y.-J. Wang, B.-S. He, and J. Zhou, "Linc00312 single nucleotide polymorphism as biomarker for chemoradiotherapy induced hematotoxicity in nasopharyngeal carcinoma patients," Disease Markers, vol. 2022, 2022.

[32] H. Wang, T. Yang, J. Wu, D. Chen, and W. Wang, "Unveiling the mystery of sumo-activating enzyme subunit 1: A groundbreaking biomarker in the early detection and advancement of hepatocellular carcinoma," in Transplantation Proceedings, vol. 55, no. 4. Elsevier, 2023, pp. 945–951.

[33] M. Ciman and K. Wac, "Individuals' stress assessment using human-smartphone interaction analysis," IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 51–65, 2018.

[34] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition," IEEE Transactions on Affective Computing, 2021.

[35] E. Sağbaş, S. Korukoğlu, and S. Ballı, "Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques," Journal of Medical Systems, vol. 44, 02 2020.

[36] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress detection by machine learning and wearable sensors," in 26th International Conference on Intelligent User Interfaces - Companion, ser. IUI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 43–45. [Online]. Available: https://doi.org/10.1145/3397482.3450732

[37] E. A. Sağbaş, S. Korukoglu, and S. Balli, "Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques," Journal of Medical Systems, vol. 44, no. 4, p. 68, Feb 2020. [Online]. Available: https://doi.org/10.1007/s10916-020-1530-z

[38] L. Zhu, P. Spachos, P. C. Ng, Y. Yu, Y. Wang, K. Plataniotis, and D. Hatzinakos, "Stress detection through wrist-based electrodermal activity monitoring and machine learning," IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 5, pp. 2155–2165, 2023.

[39] A. Kumar, K. Sharma, and A. Sharma, "Hierarchical deep neural network for mental stress state detection using iot based biomarkers," Pattern Recognition Letters, vol. 145, pp. 81–87, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865521000490

[40] H. Yu, T. Vaessen, I. Myin-Germeys, and A. Sano, "Modality fusion network and personalized attention in momentary stress detection in the wild," in 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), 2021, pp. 1–8.

[41] S. Ahmed, M. Abu Yousuf, M. M. Monowar, A. Hamid, and M. O. Alassafi, "Taking all the factors we need: A multimodal depression classification with uncertainty approximation," IEEE Access, vol. 11, pp. 99 847–99 861, 2023.

[42] N. Khan, "Semi-supervised generative adversarial network for stress detection using partially labeled physiological data," ArXiv, vol. abs/2206.14976, 2022.

[43] I. Zyma, S. Tukaev, I. Seleznov, K. Kiyono, A. Popov, M. Chernykh, and O. Shpenkov, "Electroencephalograms during mental arithmetic task performance," Data, vol. 4, no. 1, 2019. [Online]. Available: http://www.mdpi.com/2306-5729/4/1/14

[44] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in 2019 International Conference on Biomedical Innovations and Applications (BIA), 2019, pp. 1–4.

[45] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-eeg biosignals dataset for assessment and visualization of neurological status," in 2016 IEEE International Workshop on Signal Processing Systems (SiPS), 2016, pp. 110–114.

[46] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. H. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1168–1182, 2022.

[47] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 11, pp. 12 226–12 234, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21483

[48] J. Gratch, R. Arstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L. Morency, "The distress analysis interview corpus of human and computer interviews." 01 2014.

[49] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," 2019.

[50] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, Z. Liu, Z. Yao, M. Yang, H. Peng, J. Zhu, X. Zhang, G. Gao, F. Zheng, R. Li, Z. Guo, R. Ma, J. Yang, L. Zhang, X. Hu, Y. Li, and B. Hu, "A multi-modal open dataset for mental-disorder analysis," Scientific Data, vol. 9, no. 1, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1038/s41597-022-01211-x

[51] L. Mou, C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, and W. Gao, "Driver stress detection via multimodal fusion using attention-based cnn-lstm," Expert Systems with Applications, vol. 173, p. 114693, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421001342

[52] T. Iqbal, P. Redon, A. Simpkin, A. Elahi, S. Ganly, W. Wijns, and A. Shahzad, "A sensitivity analysis of biophysiological responses of stress for wearable sensors in connected health," IEEE Access, vol. PP, pp. 1–1, 01 2021.

[53] F. I. Indikawati and S. Winiarti, "Stress detection from multimodal wearable sensor data," IOP Conference Series: Materials Science and Engineering, vol. 771, no. 1, p. 012028, mar 2020. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/771/1/012028

[54] V. Randazzo, J. Ferretti, and E. Pasero, "Ecg watch: a real time wireless wearable ecg," 06 2019, pp. 1–6.

[55] J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, pp. 1189–1232, 10 2001.

[56] S. Badirli, X. Liu, Z. Xing, A. Bhowmik, K. Doan, and S. S. Keerthi, "Gradient boosting neural networks: Grownet," 2020. [Online]. Available: https://arxiv.org/abs/2002.07971

[57] V.-T. Ninh, S. Smyth, M.-T. Tran, and C. Gurrin, "Analysing the performance of stress detection models on consumer-grade wearable devices," 2022. [Online]. Available: https://arxiv.org/abs/2203.09669

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.2024.3461588



DR. SANJAY KUMAR is an Assistant Professor in the Department of Computer Science and Engineering, Delhi Technological University (DTU), New Delhi, India. He earned a Ph.D. degree in Computer Applications from the Indian Institute of Technology (IIT) Delhi. He has completed an M.Tech in Computer Application from the Indian Institute of Technology (IIT) Delhi, India. Previously, he worked with the National Informatics Centre, Govt. of India as Scientist-B. His research interests include Machine Learning, AI, Social Network Analysis, NLP, and Design and Analysis of Algorithms. He has published numerous articles and proceeding papers in reputed journals and conferences. Notably, more than 25 of his papers have been published in top-rated SCI/SCIE journals, reflecting the quality and impact of his work. He is a recipient of the "Commendable Research Award for Excellence in Research" awarded by DTU for the last three consecutive years, namely, 2020, 2021, and 2022. His name has been included in the "Top 2% scientist of the world" list in the category of Image Processing and AI by Stanford University, USA in 2023. In addition to his research contributions, Kumar actively participates in the academic community as a reviewer for various journals associated with renowned publishers like Springer and Elsevier



ANSHUMAN RAJ CHAUHAN received his Bachelor of Technology in Computer Science and Engineering from Delhi Technological University in 2023, where he demonstrated a keen interest in the realm of deep learning, health informatics and NLP. He currently holds the position of Software Engineer at Google, where he contributes to the cutting-edge Cloud Spanner database within Google Cloud.



AKHIL received his Bachelor of Technology degree from Delhi Technological University in 2023, where he demonstrated a keen interest in the realm of deep learning. Currently, he serves as a Software Engineer at Chegg, a leading education technology company, where he has been actively contributing to projects involving large language models. His research interest mainly includes deep learning. Akhil's awards and honors include Excellence in Education Awards in 2018.



DR. AKSHI KUMAR is a Senior Lecturer and Director-Post Graduate Research (PGR) in the Department of Computing at the Goldsmiths, University of London, United Kingdom. Dr. Kumar has been endorsed by the Royal Academy of Engineering, United Kingdom as an Exceptional promise in the field of Artificial Intelligence/Data Science in 2022. She has received 9 research awards for Excellence in Research from various National and International organizations. Her name has been included in the "Top 2% scientist of the world" list by Stanford University, USA in 2023, 2022, and 2021. She has published more than 95 peer-reviewed journal papers, 65+ conference papers with 5 best paper awards, and 2 patents with the Indian Patent Office. She has been serving as an Associate editor and guest editor in various high-impact journals with reputed publishers. Her research interests are in affective computing, social network and media analytics, NLP, and AI for pervasive healthcare



DR. GUANG YANG is an Associate Professor (Senior Lecturer) in the Bioengineering Department and Imperial-X at Imperial College London. He holds a UKRI Future Leaders Fellowship and serves as an Honorary Senior Lecturer in the School of Biomedical Engineering & Imaging Sciences at King's College London. His research group is dedicated to developing novel and translational techniques for imaging and biomedical data analysis. The group's focus encompasses research and development in data-driven fast imaging, data harmonization, data synthesis, federated learning, explainable AI, and AI in drug discovery. Currently, his work spans a wide range of clinical applications in cardiovascular disease, lung disease, and oncology. For more information about Yang's Lab, visit: https://www.yanglab.fyi/

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

Check for updates

Received: 16 March 2024 | Revised: 8 July 2024 | Accepted: 20 August 2024

DOI: 10.1002/fld.5332

RESEARCH ARTICLE

WILEY

Response surface method-based hydraulic performance optimization of a single-stage centrifugal pump

Durvesh Yadav | **Raj Kumar Singh** | **K. Manjunath**

Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

Correspondence
Durvesh Yadav, Department of Mechanical Engineering, Delhi Technological University, New Delhi, India.
Email: durvesh_2k19phdme504@dtu.ac.in

Abstract

In this article, the response surface approach was employed to enhance the hydraulic performance of the pump at the rated point. Specifically, an approximate link between the design head and efficiency of the single-stage centrifugal pump and the parameters of the impeller's design was established. The first step in creating a one-factor experimental design involved selecting significant geometric variables as factors. Decision variables such as the number of blades, flow rate, and rotation were chosen due to their significant impact on hydraulic performance, while head and efficiency were considered as responses. Subsequently, the best-optimized values for each level of the parameters were identified using response surface analysis and a central composite design. The impeller schemes of the Design-Expert software were evaluated for head and efficiency using Computational fluid dynamics, and a total of 20 experiments were conducted. The simulated results were then validated with experimental data. Through the analysis of the individual parameters and the approximation model, the ideal parameter combination that increased head and efficiency by 7.90% and 2.06%, respectively, at the rated value was discovered. It is worth noting that in cases of a high rate of flow, the inner flow was also enhanced.

KEYWORDS

computational fluid dynamics, number of impeller blades, optimization technique, response surface methodology, rotation, single-stage centrifugal pump

1 | INTRODUCTION

A Centrifugal pump is a typical hydro-machinery with significant applications in a range of sectors including water treatment, chemical, oil and gas, and power generation in nuclear, thermal, and hydro plant.[1] This pump is a key member of the family of power-hungry turbo-machines and is the most popular and widely used pump to date. To boost fluid pressure, it is commonly employed in industrial sectors like the water, processing, and chemical industries. It consumes a significant portion of the total electricity produced worldwide. High pump efficiency is therefore crucial for cutting power consumption and realizing financial rewards. So one of the most important considerations when choosing and designing a pump is energy conservation.[2] With its rotor (impeller) and stator (volute), the centrifugal pump's constructive geometry is produced. While the volute performs energy transformation to further increase the fluid's pressure energy, the fluid's energy is enhanced by the conversion of mechanical work in the impeller. There are two ways to boost the overall efficiency of centrifugal pump.[3] The first step is to analyze how one variable or structural element affects the outcome.

 © 2024 John Wiley & Sons Ltd.

The second step entails applying mathematical formulas or computational techniques to investigate the combined causes of numerous factors on the pump's performance.

The several studies[4–6] focuses on understanding the dynamic behavior of the pump's energy conversion efficiency under varying operating conditions to optimize operational strategies for enhanced energy efficiency and reduced mechanical stress.

The optimization of hydraulic performance in centrifugal pumps is a critical area of research in fluid dynamics. There is a significant amount of literature related to the optimization of centrifugal pumps to improve their performance. Various studies have focused on utilizing different methodologies to enhance the efficiency of single-stage and multistage centrifugal pumps. For instance, the response surface methodology (RSM) has been applied to optimize the hydraulic performance of single-stage centrifugal pumps.[7–9] An important first step in the realm of turbomachinery optimization is the analysis of the key metrics for performance since it helps to make the pump design apparent.[10–13] Many studies concentrated on enhancing the efficiency and effectiveness of these pumps by optimizing various design modifications in the impeller, diffuser, and return passage.[14–18] Li et al.[19] presented a way for distributing the blade thickness by looking into secondary flow in centrifugal pumps. The findings demonstrate that the approach is effective at preventing secondary flow from developing and increasing performance. Results from Lin et al.[20] study, which looked at the impact of various bionic sinusoidal tubercle trailing edges (STTE) on a pump's operation and internal flow, demonstrate that the STTE has a significant impact on lowering energy usage. In a study by Wang et al.[21] utilizing numerical simulations, the impeller form of a centrifugal pump was optimized. They discovered that increasing blades while reducing blade angle improved the pump's efficiency and head performance. Matlakala et al.[22] investigated how the impeller diameter and blade count affected the performance of centrifugal pumps. How impeller blades and diffuser vanes fit have been investigated by Bai et al..[23] They have examined nine different impeller configurations using blades 5, 6, 7, and vanes 8, 9, and 10. For simulation, they employed the shear stress transfer SST model. The high-speed pump with $Q = 200\ m^3/h$, Ns = 71.6, and design head = 300 m was selected for numerical analysis. It has 7 impeller blades and 9 diffuser vanes. The combinations that were discovered to have the lowest head value were 5 + 8, 6 + 8, and 7 + 9. Keeping the impeller blades constant, the efficiency was optimum for 8 diffuser vanes and minimal for 10 diffuser vanes. For both steady and non-steady results, the pump with 7 blades and 8 vanes was the optimum option.

Other studies have prioritized improving operational parameters such as fluid viscosity, flow rate, and rotating speed. By using surrogate models and optimization techniques, it is also possible to practically determine the best combination of design parameters.[24–26] For instance, Li et al.[27] optimized the operating settings of a centrifugal pump using a RSM. They found that by increasing the discharge and reducing the rotation, the head of the pump and efficiency were improved.

According to extensive studies,[28–34] the impeller, a crucial part of the centrifugal pump, is highly influential to how well the same performs. Experimental and numerical methods are used to examine the effects of Bozorgasareh et al.'s[35] new impeller configurations with innovatively designed shrouds on the pressure head and the effectiveness of the semi-open centrifugal pump. Optimization techniques for impeller design that have been studied in recent years are Computational Fluid Dynamics (CFD), Genetic algorithms, Response surface methodology (RSM), and Multi-objective optimization. Additionally, researchers have explored the use of bionic optimization and neural networks to improve the energy consumption and efficiency of centrifugal pumps.[36,37] Collaborative optimization approaches have also been investigated to enhance the performance of multiphase pump impellers.[38,39]

In a study by Lee et al.,[40] a genetic algorithm was used to enhance the impeller design of a centrifugal pump. They found that increasing the number of blades while reducing the angle of the blades improved the pump's efficiency and head performance.

In several research Centrifugal pump design has been optimized using orthogonal optimization and RSM. In one example, study by Zhang et al.,[41] the impeller design of a centrifugal pump with low specific speed was optimized using orthogonal arrays, and it was investigated that increasing the blade numbers and decreasing the blade angle were among the most important factors for enhancing pump performance. Wang et al.[42] used RSM to optimize the impeller design of a centrifugal pump. They found that increasing the blade number and reducing the blade angle improved the pump's efficiency and head performance, which is consistent with the findings of their earlier study mentioned earlier. Design of experiments (DOE) and innovative algorithms can be used to categorize the techniques for multi-stage pump optimization. Although intelligent algorithms are superior at solving complex problems, the time required for optimization is heavily dependent on the speed at which the algorithms converge. For a pump, many parameters can be used as experimental design factors, making it difficult to select factors and maintain fewer designed schemes. To choose factors and reduce the number of experiments prior to the response surface technique, a sensitivity investigation and single factor

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

test were conducted in this research and less testing for more precise prediction is a benefit of response surface method experiment design.

1.1 | Novelty of this work

There are numerous works on the optimization of parameters that influence the efficiency and head of the pump. However, little attention has been given to optimizing factors such as the number of blades and rotation speed using RSM. In this study, we developed and validated a single-stage centrifugal pump through numerical simulation and experimental validation. The optimization of blade numbers and RPM in centrifugal pumps is a complex process that necessitates careful consideration of various design parameters. These factors play a crucial role in determining pump performance, including efficiency, pressure fluctuations, and overall reliability. This work uniquely focuses on optimizing number of blades and rotation speeds using RSM in a single-stage centrifugal pump, a novel approach not previously explored.

2 | MODEL DESCRIPTION

2.1 | Computational Modeling

A single-stage centrifugal pump having a specific speed of 21.54 is used as a pump model. The rated parameters used in designing the pump model are given as Discharge $Q_d = 464\ m^3/h$, Head $H_d = 40\ m$, and Rotation $N = 1750\ rpm$.

2.2 | Geometric and solid modeling

A three-dimensional single-stage centrifugal pump model mainly consisting of the impeller, diffuser, and return passage is modeled using Creo parametric software. Data as presented in Table 1 was used to generate the profile of the impeller, diffuser, and return passage referred from.[43]

Figure 1 depicts a complete model design of an impeller blade, from the development of the blade profile to the implementation of the blade, a complete model of an impeller, and a complete model of a diffuser. Figure 2 depicts a complete model of an inlet and an outlet of the assembly of a single-stage centrifugal pump.

2.3 | Mesh generation

The impeller geometry with a 2D domain generated triangular surface mesh whereas for the 3D flow domain, the unstructured tetrahedral mesh is generated because when there is a significant change in geometry or material properties within a model, using two different mesh types can help create smooth transition zones. In some cases, certain regions of a model may require a finer mesh to accurately capture complex geometry or high-stress concentrations, while other regions may not need such refinement as shown in Figure 3. Using two distinct mesh types allows for efficient use of computational resources by applying a finer mesh only where necessary. The impeller geometry consists of the impeller inlet, impeller outlet, hub, shroud, and blades. The outlet of the impeller is connected to the diffuser inlet with proper interfacing between them. Table 2 shows the total number of nodes, elements, and type of mesh generated. Figure 3 shows the three-dimensional view of the mesh generation of the impeller, diffuser, and return passage as well as the enlarged views of the same. The impeller is connected to a diffuser and the diffuser is linked to the return channel.

2.4 | Grid independence test

The impeller with five blades has undergone a grid independence test shown in Figure 4 at various element sizes, that is, 7, 5, 3, 2, and 1 mm. The velocity gradient generated has been taken as a criterion for independence for this analysis. Analysis based on the various grids showed that the findings at 5 mm and 3 mm are almost statistically the same with an error of 2.27%. It has been found that all the values after 4,041,112 cells, velocity gradient is going on decreasing. So, the

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 Design parameters for impeller, diffuser and return passage.[43]

Parts	Parameters	Values
Impeller	Impeller-outlet diameter	408 mm
	Impeller inlet diameter	195 mm
	Inlet angle β1	40°
	Outlet angle β2	20°
	Blade width at inlet and outlet	23 mm
	Speed	1600–1900 rpm
	Numbers of blades	5, 6, and 7
Diffuser	Diffuser inlet diameter	408 mm
	Diffuser outlet diameter	502 mm
	Diffuser inlet and outlet width	23 mm
	Number of vanes	8
	Inlet angle	5°
	Outlet angle	25°
Return passage	Inlet diameter	502 mm
	Casing diameter	570 mm
	Number of guide vanes	11
	Guide vane height	35 mm
	Casing height	58 mm

true maximum is 4,041,112 no. of cells. Furthermore, we got our solution converge at this maximum. After refining the grid 1,292,906 nodes and 4,041,112 elements have been used for further simulation.

2.5 | Boundary conditions

To obtain the result it is necessary first to apply boundary details, that is, inlet and outlet conditions. In this analysis, at the inlet, the static pressure is applied with reference pressure set as zero, and the mass flow rate of 128.416 kg/s is applied at outlet.[44] Figure 5 shows the boundary conditions at the inlet and outlet. Each domain is to be interfaced with each other because fluid must flow from one domain outlet to another domain inlet, that is, representing connectivity of the fluid domain shown in Table 3. Here there are some rotating parts, and some are stationary parts so in interference, the frame change model must be a frozen rotor having pitch change as none. Stationary wall motion is selected as the wall condition and no slip is chosen as a shear condition. The standard roughness model is selected with a roughness constant of 0.5.

2.6 | k-ε model

k-ε model is used as a turbulence model in this study. This is the simplest turbulence model, most widely used for solving two transport equations separately. The k-epsilon turbulence model was chosen because of the following points:

- The model is preferred in the scenario where a boundary layer under a strong adverse pressure gradient exists.
- The model is applicable where flow separation and recirculation persist.
- The model is also applicable where the mean flow of the complex structure is taken into account.
- The model exhibits superior performance for flows that involve the rotation of components.

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

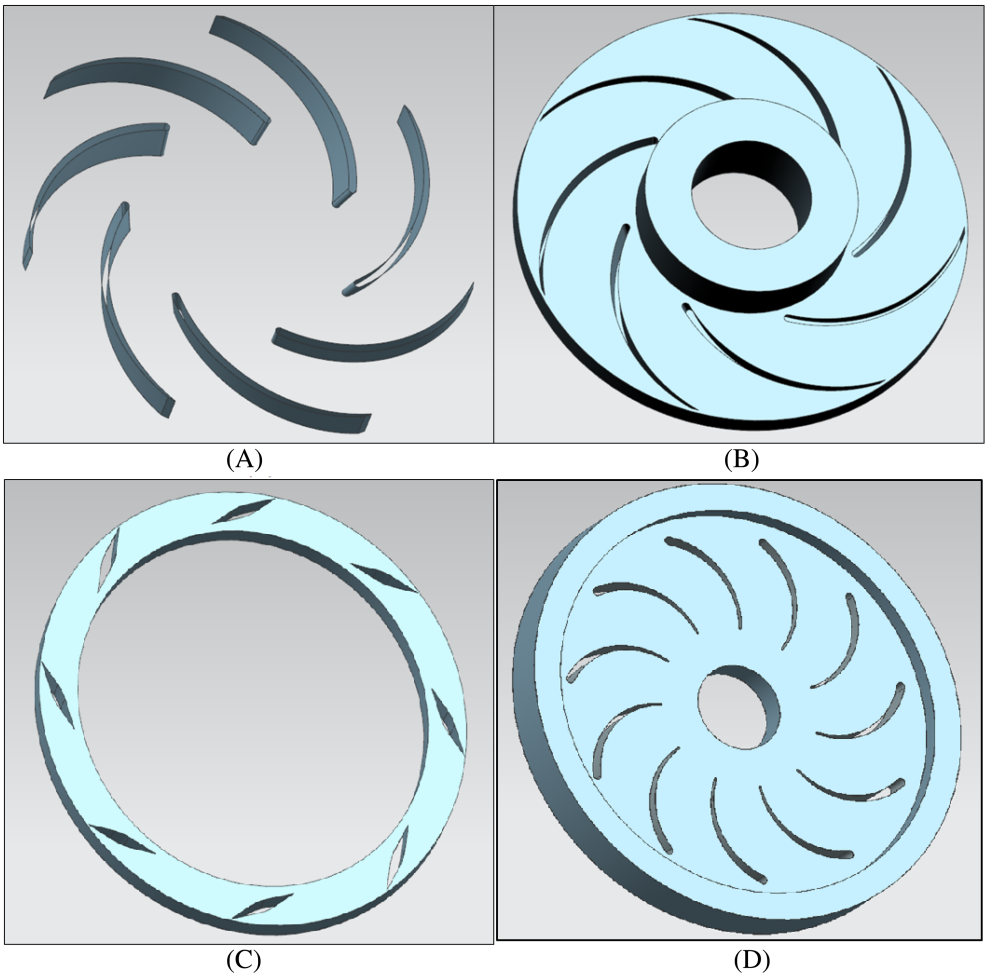


FIGURE 1 (A) Design of blades, (B) Complete impeller model, (C) diffuser model, and (D) return channel. [Colour figure can be viewed at wileyonlinelibrary.com]

2.7 | Solver

In the Ansys CFX, the solver control must be run after applying boundary details. A CFX solver involves the discretization of the flow domain into a differential equation. A certain number of iterations must be applied after the completion of all iterations a definition (.def) file will be generated.

Table 4 shows the general input data obtained from CFD software and has been used for the analysis. Whereas Table 5 shows the data used in the CFX solver. Here we have used a high-resolution for both advection scheme and turbulence numeric because it helps to accurately capture the transport and movement of fluid quantities, such as velocity, temperature, and pressure gradients. High-resolution advection schemes are designed to minimize numerical diffusion and dispersion, which can lead to smearing or damping of sharp gradients and discontinuities in the flow field. These schemes aim to preserve the sharpness and accuracy of these features by minimizing errors introduced during the discretization and numerical integration process. In computational fluid dynamics (CFD), a high-resolution turbulence scheme is used to accurately model and simulate turbulent flow. High-resolution turbulence schemes aim to capture the complex behavior of turbulence by minimizing numerical diffusion and dispersion, which can lead to the damping or smearing of turbulent fluctuations. The solver generates a result file (.res) at its conclusion, which will be examined in CFX-Post. The streamline pattern, pressure contour, blade-to-blade analysis, etc. are analyzed in CFX-The flow simulation. Using a function calculator, all the parameters, including different velocity components, static pressure, total pressure, torque, mass flow rate, etc., may be calculated in multiple domains and places.

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

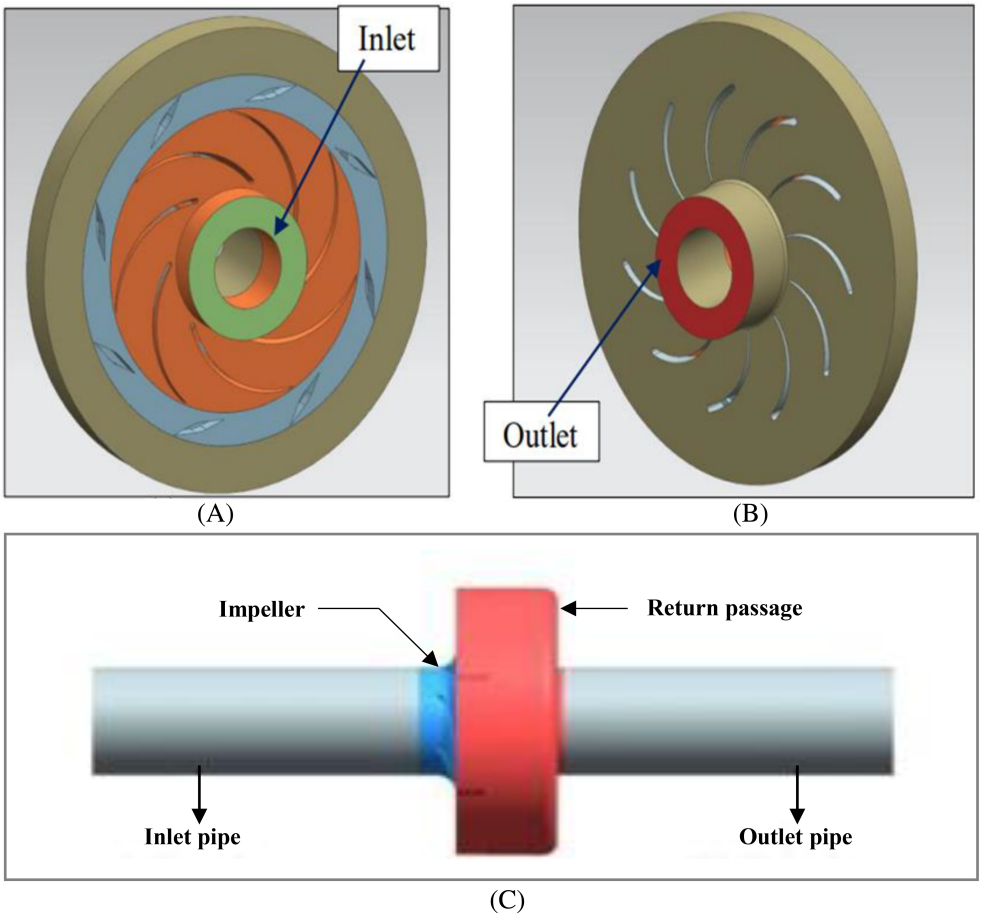


FIGURE 2 (A) Inlet, (B) outlet portion of the first stage of centrifugal pump, and (C) assembly of first stage centrifugal pump. [Colour figure can be viewed at wileyonlinelibrary.com]

3 | EXPERIMENTAL TEST RIG SETUP

Based on the aforementioned analysis, the test rig used for the experimental validation with simulation results is shown in Figures 6 and 7, comprises of 3 HP AC Drive motor that is connected to a pump with a stepped pulley arrangement, measuring tank of area $80 \times 80\,cm^2$ with a piezometer, Stainless Steel sump tank and, Galvanized Iron Pipe connections with the necessary control valve, and so on, all of which are mounted on a neatly painted Mild Steel structured frame manufactured by Aarvee Industries, made in India. The panel board has an energy meter to monitor the amount of power going into the motor, a vacuum gauge (0 to (−760) mm Hg) to measure the suction head, a pressure gauge (0–4.5 kg/cm^2) to measure the delivery head, and a dimmer starter to change the speed (0–5000 rpm). Figure 6 shows the line diagram of the experimental setup whereas Figure 7 shows the experimental test setup.

3.1 | Uncertainty analysis of experimental test rig

To accurately determine the standardized measurement uncertainty of various measuring systems and instruments, it is essential to calculate it using the relevant information provided by the manufacturers of these instruments. Additionally, it is crucial to calibrate these instruments as required, which helps to minimize systemic measurement uncertainty in the calibration units. When comparing test findings with stated results, it is important to consider the measuring uncertainty in a way that is not influenced by external factors such as the pump or the guaranteed values themselves. One effective

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

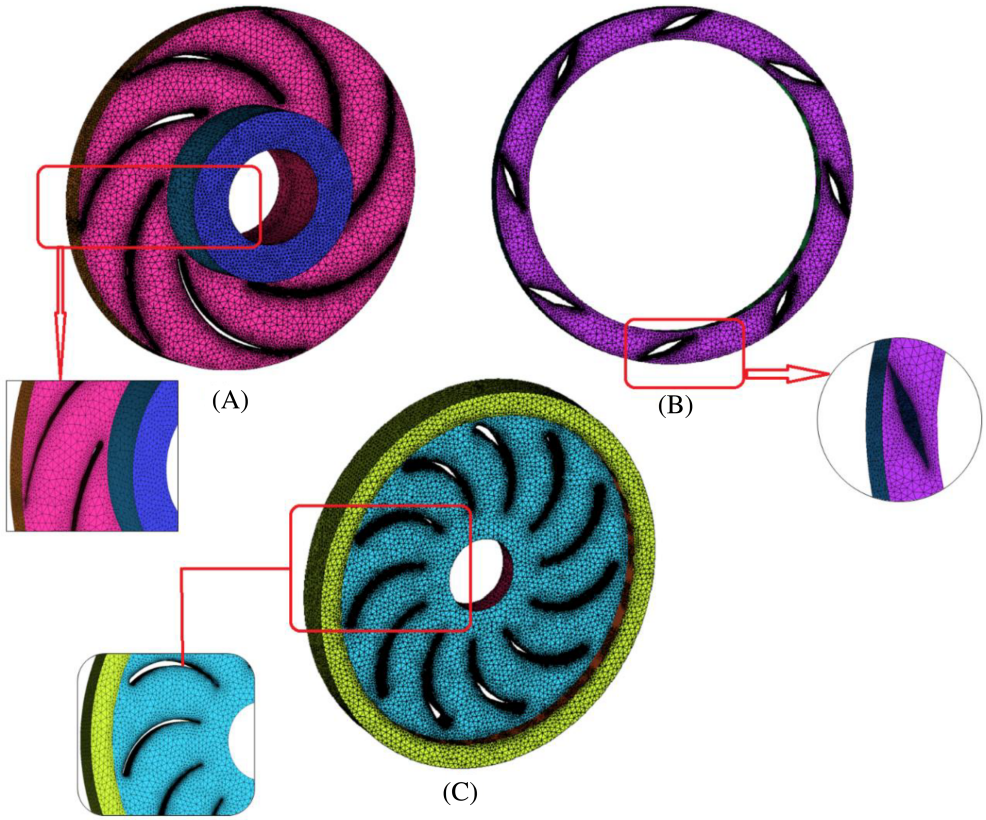


FIGURE 3 3D view and enlarged view of mesh generation of (A) impeller, (B) diffuser, and (C) Return passage. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Mesh data of each domain obtained from computational fluid dynamics software.[44]

Domain	No. of nodes	No. of elements		Element type
Impeller	398,099	Impeller inlet	7058	Triangle
		Hub	26,343	Triangle
		Shroud	26,078	Triangle
		Blades	227,181	Triangle
		Impeller outlet	7616	Triangle
		Shaft	2022	Triangle
		Flow domain	2,203,762	Tetrahedral
Diffuser	181,150	Diffuser inlet	6061	Triangle
		Diffuser outlet	8854	Triangle
		Vanes	91,357	Triangle
		Diffuser casing	26,461	Triangle
		Flow domain	1,002,296	Tetrahedral
Return passage of 1st stage	713,657	Inlet	6903	Triangle
		Outlet	3514	Triangle
		Casing	39,635	Triangle
		Guide vanes	361,971	Triangle

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

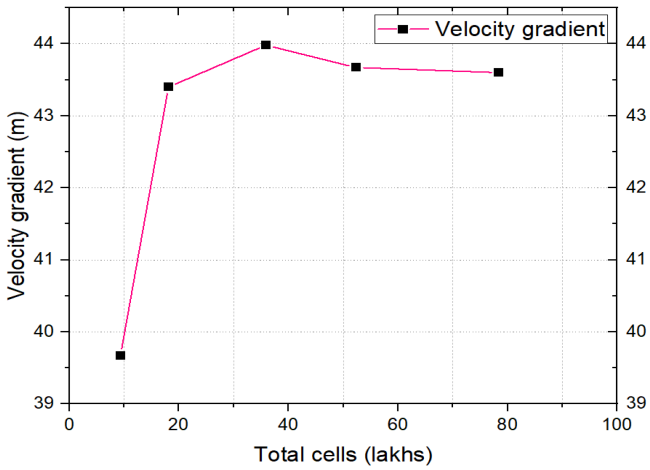


FIGURE 4 Grid independence test. [Colour figure can be viewed at wileyonlinelibrary.com]

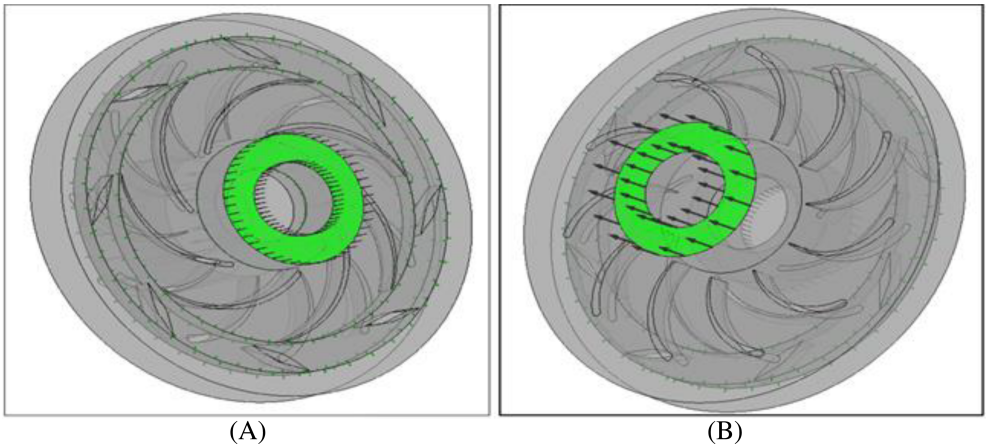


FIGURE 5 (A) Inlet boundary conditions and (B) outlet boundary conditions. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Domain interface data for single stage pump.[44]

Interface location	Interface type	Interface models	Frame change	Pitch change
Impeller–Diffuser	Fluid–Fluid	General connection	Frozen rotor	None
Diffuser–Return Passage	Fluid–Fluid	General connection	None	None

method for calculating uncertainty is through the propagation method.[45]

$$\Delta y = \sqrt{\left(\left(\frac{\partial y}{\partial x_1}\Delta x_1\right)^2 + \left(\frac{\partial y}{\partial x_2}\Delta x_2\right)^2 + \left(\frac{\partial y}{\partial x_3}\Delta x_3\right)^2 + \ldots + \left(\frac{\partial y}{\partial x_n}\Delta x_n\right)^2\right)}, \quad (1)$$

where y is the responses and x_1, x_2, $\ldots x_n$ are the independent variables. Table 6 displays the calculated and estimated levels of uncertainty for the responses and independent variables. The outcome shows that the uncertainty error is less than 6%. As per BS EN ISO 9906:2000, that much unpredictability in an engineering analysis is permissible.

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 General input data obtained from computational fluid dynamics software.[44]

Domain type	Fluid domain	
Fluid	Water at 25°C	
Density of fluid	997 kg/m^3	
Reference pressure	0 atm	
Domain motion	Impeller	Rotating
		1. 1600 rpm
		2. 1750 rpm
		3. 1900 rpm
	Diffuser	Stationary
	Return passage	Stationary
Rotation axis	*Z*-axis	
Turbulence model	k–ε	

TABLE 5 Data used in CFX solver.[44]

Advection scheme	High-resolution
Turbulence numeric	High-resolution
Number of iterations	500
Length scale control	Conservative
Time-scale control	Auto time-scale
Time-scale factor	1
Residual-type	Root mean square
Residual-target	10^{-06}

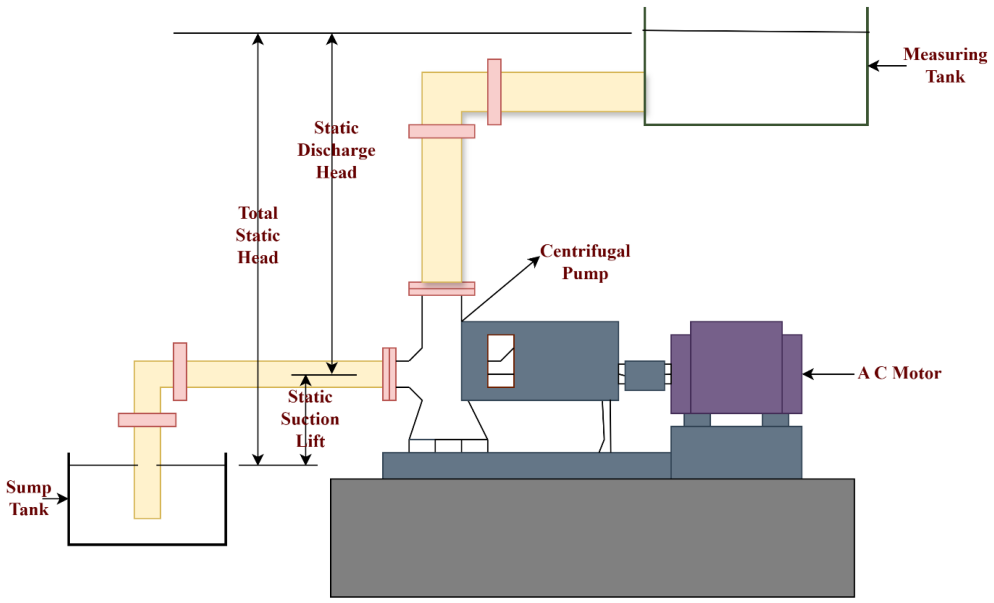


FIGURE 6 Schematic diagram of test rig. [Colour figure can be viewed at wileyonlinelibrary.com]

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

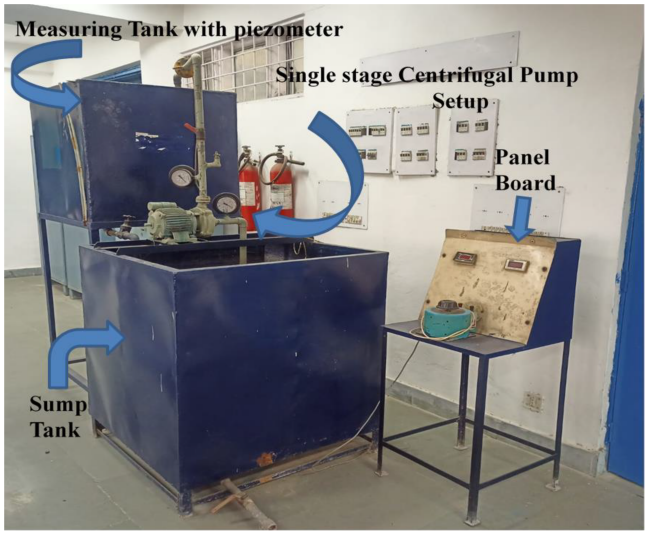


FIGURE 7 Experimental test rig setup. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Calculation of uncertainty of instruments.

	Range	Uncertainty (%)
Non-dependent variables		
Discharge pressure (kg/cm^2)	0–4.5	±1.8
Suction pressure (mm Hg)	0 to (−760)	±1
Flow rate (m^3/h)	0–850	±0.5
Torque (N m)	0–2500	±0.06
Rotational speed (RPM)	0–5000	±0.1
Dependent variables		
Head, H (m)	±5.64%	
Efficiency, η (%)	±5.9%	

3.2 | Validation of experimental with simulation

The steady-state computational study performed for 5, 6, and 7-bladed impeller centrifugal pumps in this paper is comparable to the experimental analysis performed using 5-bladed impeller centrifugal pumps. Figure 8 presents the validation of experimental and simulated results. The differences between experimental results and CFD results in the head are 1.92%, 4.7%, 3.2%, 7.10%, 7.02%, 6.1%, 10.6%, and 9.09%, respectively. The differences between efficiency results and CFD results in efficiency are 8.4%, 5.24%, 5.01%, 5.25%, 7.4%, 7.5%, and 7.3%, respectively. The overall average error in experimental results and CFD results for head and efficiency is 6.21% and 5.77%, respectively. Due to a lack of consideration during the test for the energy loss caused by the pump itself, the test outcomes for the head and efficiency were lower than those of the numerical calculations. As a result, the statistical results produced in this study and from other work have been gathered with reliability. As a result, the computational results reported in the research can be trusted. Figure 8 shows a comparison of the characteristics curves for experimental results tests with simulation results.

4 | OPTIMIZATION OF SINGLE-STAGE CENTRIFUGAL PUMP USING RESPONSE SURFACE METHODOLOGY

Response surface methodology (RSM) is a statistical and mathematical technique employed to simplify complex systems by examining the connection between input-independent variables and response patterns at the output. RSM is a used

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

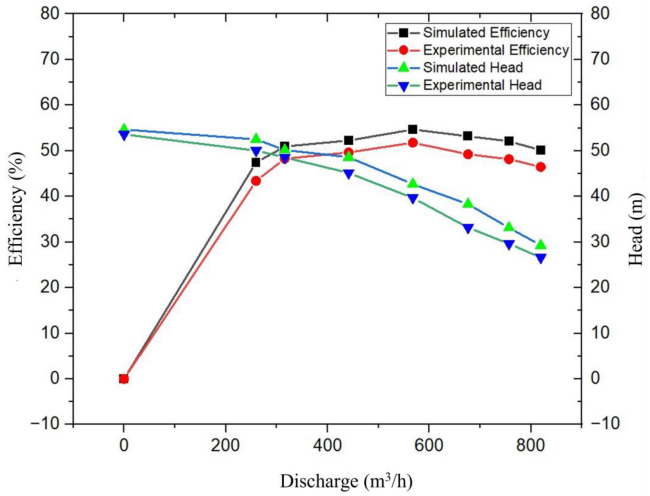


FIGURE 8 Validation of experimental results with computational fluid dynamics results. [Colour figure can be viewed at wileyonlinelibrary.com]

to optimize responses and improve product quality by modeling the relationship between input variables and output responses.

The RSM employs a variety of experiment design techniques. Box–Behnken Design (BBD) and Central Composite Design (CCD) are the two that are used the most frequently. The main difference between the two is the number of levels for each input variable. BBD uses three levels (−1, 0, and 1) for each input variable, resulting in a total of 15 experimental runs. This design is useful for identifying quadratic effects and interactions between variables. CCD, on the other hand, uses more levels (−1.68, −1, 0, 1, and 1.68) for each input variable, resulting in a total of 20–30 experimental runs. This design is useful for identifying both linear and quadratic effects, as well as interactions between variables. Use a suitable experimental design, such as a factorial design, CCD, or Box–Behnken design, to plan the experiments that will be conducted. The design should allow for the estimation of main effects, interactions, and quadratic effects. In this study face centered CCD is used with $\alpha = 1$. In the literature, you can find more information about the RSM and CCD. Figure 9 shows the flowchart of the RSM.

The effects of independent variables, that is, rotational speed (N), the blade numbers (z), and the discharge (Q) on the response (head and efficiency) are examined by employing the Response surface analysis with the CCD method. According to Table 7, an L9 orthogonal array was used in this experiment, has three control factors and three levels. In this study, the maximum and lowest values were used to examine the head and efficiency, respectively. Table 7 shows the factor levels of CCD. Table 8 shows the scheme and results of the CCD where no. of experiments performed was 20. These tables are obtained in Design Expert using RSM.

5 | RESULTS AND DISCUSSIONS

The RSM results and numerically simulated results are computed and analyzed. The hydraulic performance of the pump is also calculated using equations provided in this section. The effect of input parameters on internal flow has been presented and discussed.

5.1 | RSM results

To identify the responses such as head and efficiency which are a function of the independent variables (rotation, number of blades, and flow rate), second-order polynomial Equations (2) and (3) obtained from RSM analysis were used as shown

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

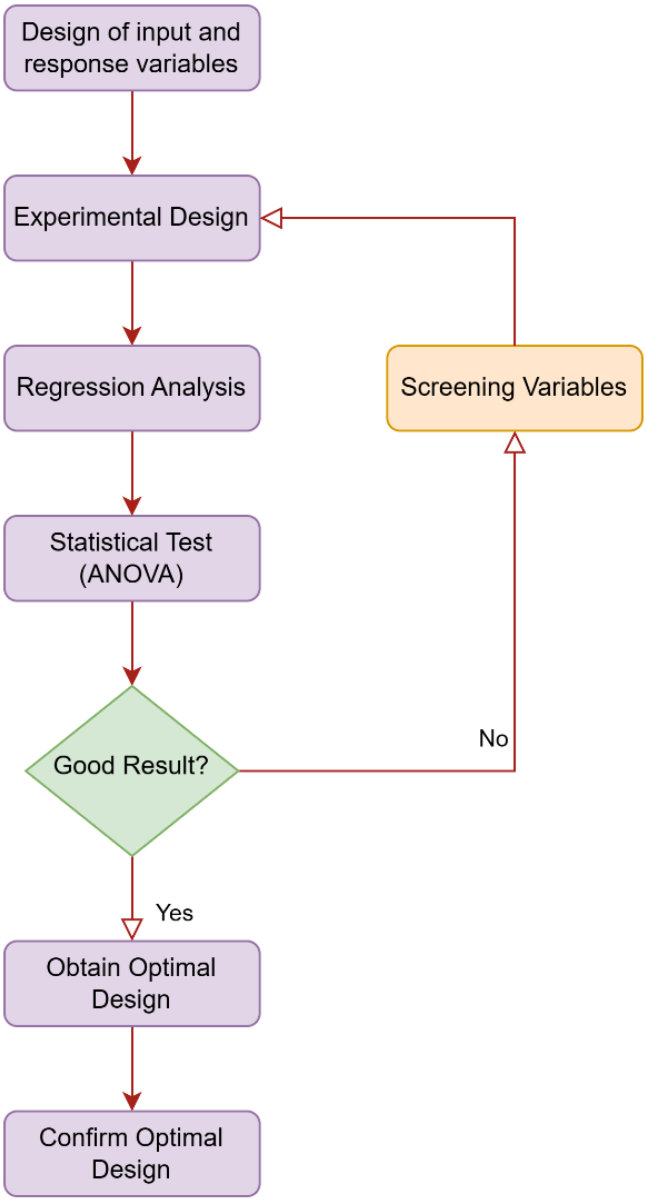


FIGURE 9 Response surface methodology flowchart. [Colour figure can be viewed at wileyonlinelibrary.com]

below:

$$H = 9.97863 - 1.50117z - 0.006372N - 0.004486Q + 0.000810zN + 0.000750zQ - 4.19689 * 10^{-6}NQ - 0.011611z^2 + 1.52680 * 10^{-6}N^2 - 5.37287 * 10^{-6}Q^2, \quad (2)$$

$$\eta = -18.78094 + 1.25178z + 0.021236N + 0.000228Q + 0.000011zN + 0.000255zQ - 1.00031 * 10^{-6}NQ - 0.086717z^2 - 5.30595 * 10^{-6}N^2 + 1.21688 * 10^{-6}Q^2. \quad (3)$$

Validation of the response surface model has been done to ensure that it accurately represents the true relationship between the input factors and the response. This involve checking for model adequacy, lack of fit, and statistical significance of model terms. The findings demonstrated that the linear effect of rotational speed ($p \leq 0.05$) followed by the linear effect of blade numbers and discharge and the interaction effect between several blades and rotational speed followed by blade numbers and discharge, and rotation per minute and discharge, and the quadratic effect of blade numbers were the variables with largest to smallest effects on the head. The remaining quadratic effects had no impact on the head. Thus,

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 7 Factor levels of centre composite design.[46]

	Code level		
Factors	**−1**	**0**	**1**
Blade numbers z	5	6	7
Rotation per minute N (rpm)	1600	1750	1900
Flow rate Q (m^3/h)	350	450	550

TABLE 8 Scheme and results of central composite design.[46]

Run	Number of blades z	Rotational speed N rpm	Flow rate Q m^3/h	Head H(m)	Efficiency %
1	6	1750	450	32.88	47.43
2	5	1900	550	38.21	45.72
3	6	1750	450	32.88	47.43
4	5	1900	350	33.01	43.6
5	5	1750	450	26.92	41.11
6	6	1750	550	35.43	48.64
7	6	1900	450	44.01	50.93
8	7	1900	350	43.71	52.39
9	6	1750	450	27.02	47.43
10	5	1600	550	21.36	37.49
11	6	1750	450	32.88	47.43
12	7	1750	450	37.12	51.26
13	6	1750	450	32.88	47.43
14	6	1750	450	32.88	47.43
15	7	1600	350	23.21	42.68
16	6	1750	350	27.02	46.17
17	7	1600	550	28.11	46.97
18	7	1900	550	53.46	55.95
19	5	1600	350	20.01	35.08
20	6	1600	450	22.02	40.57

the final models incorporated the significant terms. The model fits well, as shown by the lack of fit test ($p > 0.05$) being insignificant for the head. Table 9 shows that the variables with the largest to smallest effects on the efficiency were the linear effect of blade numbers, rotation per minute, and discharge, and the interaction effect between number of blades and flow rate and quadratic effect of blade numbers and rotation per minute respectively ($p \leq 0.01$). The interaction effect between the blade numbers and rotation per minute, rotation per speed and discharge, and the quadratic effect of the discharge had no impact on the efficiency. The findings demonstrated that the model fits well and that the lack of fit test ($p > 0.05$) was not relevant for efficiency. In Design-Expert software under the Analysis of Variance (ANOVA) function, the approximate model was examined by using R-squared ($R^2 = 0.9681$) shown by Equation 4, Adjusted R-squared (Adj.$R^2 = 0.9394$) shown by Equation 5 and Predicted R-squared (Pred. $R^2 = 0.9060$).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (y - y_i)^2}. \tag{4}$$

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 9 Analysis of variance ANOVA for quadratic model.[46]

	Head			Efficiency		
Variable	**mean square**	***F*-values**	***p*-values**	**Mean square**	***F*-values**	***p*-values**
Intercept	167.42	880.29	<0.0001	52.42	308.70	<0.0001
Main effect						
Z	227.15	1194.30	<0.0001	213.91	1259.77	<0.0001
N	1167.05	6136.15	<0.0001	209.76	1235.37	<0.0001
Q	69.91	367.56	<0.0001	22.05	129.87	<0.0001
Interaction effect						
z*N*	34.44	181.11	<0.0001			
z*Q*	1.19	6.23	0.0316	0.0052	8.11	0.0173
NQ	1.58	8.33	0.0162			
Quadratic effect						
z^2	2.68	14.07	0.0038	0.0207	18.24	0.0016
N^2				0.0392	36.26	0.0001
Q^2						
Lack of fit	0.0162	0.3393	0.8697	0.0008		
R^2	0.9681			0.9968		
Adjusted R^2	0.9394			0.9939		

$$Adj.R^2 = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right). \tag{5}$$

According to Equation (5) n denotes the size of the sample; k denotes the number of independent factors. The Predicted R^2 of 0.9060 and the Adjusted R^2 of 0.9394 are reasonably in accord, that is, the difference is < 0.2 which shows that there is a very close range between actual optimization objectives and parameters & the regression Equations (2) and (3) are important to anticipate the head and efficiency. The Design Expert software's optimization module was used for predicting optimum impeller parameters. The head and efficiency of the centrifugal pump were evaluated in this investigation using the greatest and lowest values, respectively. To get a high-efficiency impeller with a suitable head, set both the efficiency and head optimization conditions to "maximize."

Figure 10 shows the plot between predicted versus actual values of experiments for head and efficiency. To comprehend the dominant influencing parameters, the effect of variables on the responses is investigated. Figure 10 shows the variation in predicted and actual values of experiments for head and efficiency for different number of blades respectively. Its shows the dominating influence of input parameter, that is, number of blades on the responses (head and efficiency). Figure 11 represents the response surface and contour plots of the interaction between rotational speed and number of blades for (A) Head and (B) Efficiency of single stage centrifugal pump. It shows that as we increase the number of blades from 5 to 7 and rotational speed from 1600 to 1900 rpm it increases the head and efficiency increase significantly. But the effect of number of blades is lower than that of rotational speed. Thus, rotational speed has the dominant effect on the head and efficiency of the pump. The validated response surface model was used to identify the optimal factor levels that maximize or minimize the response variable. This involve finding the settings that result in maximum yield, minimum cost, or optimal product quality.

A confirmation experiment was conducted using the optimal factor levels identified by the response surface model to validate the predicted results and ensure that the optimum has been achieved.

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

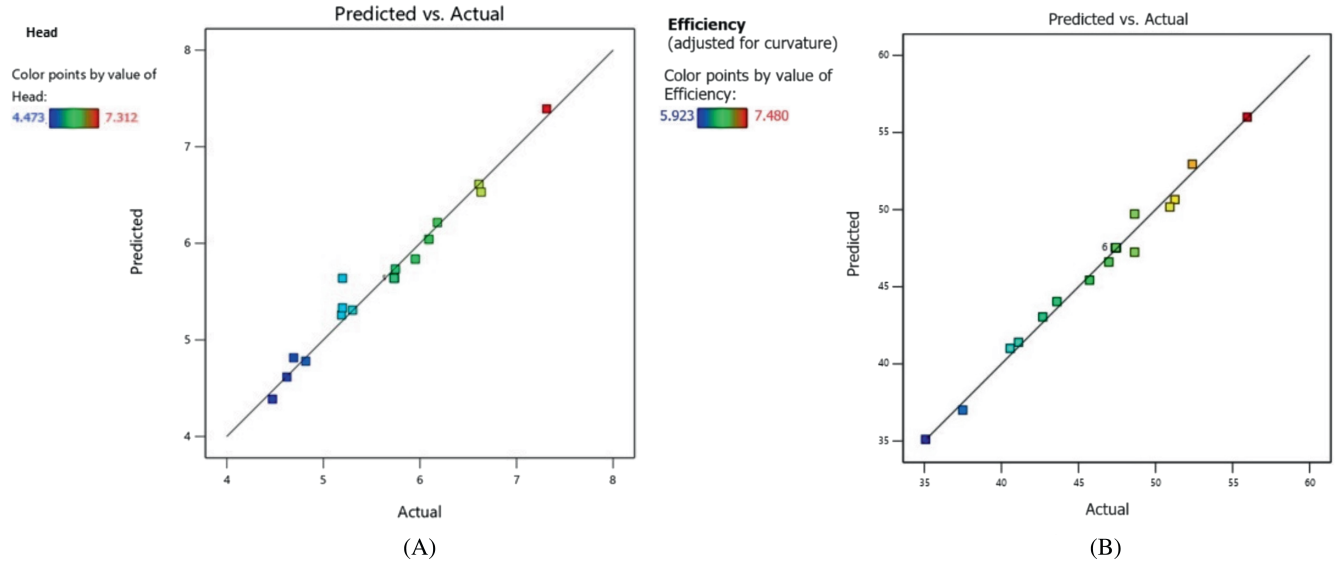


FIGURE 10 Predicted versus. actual values plot for (A) head, and (B) efficiency using response surface methodology. [Colour figure can be viewed at wileyonlinelibrary.com]

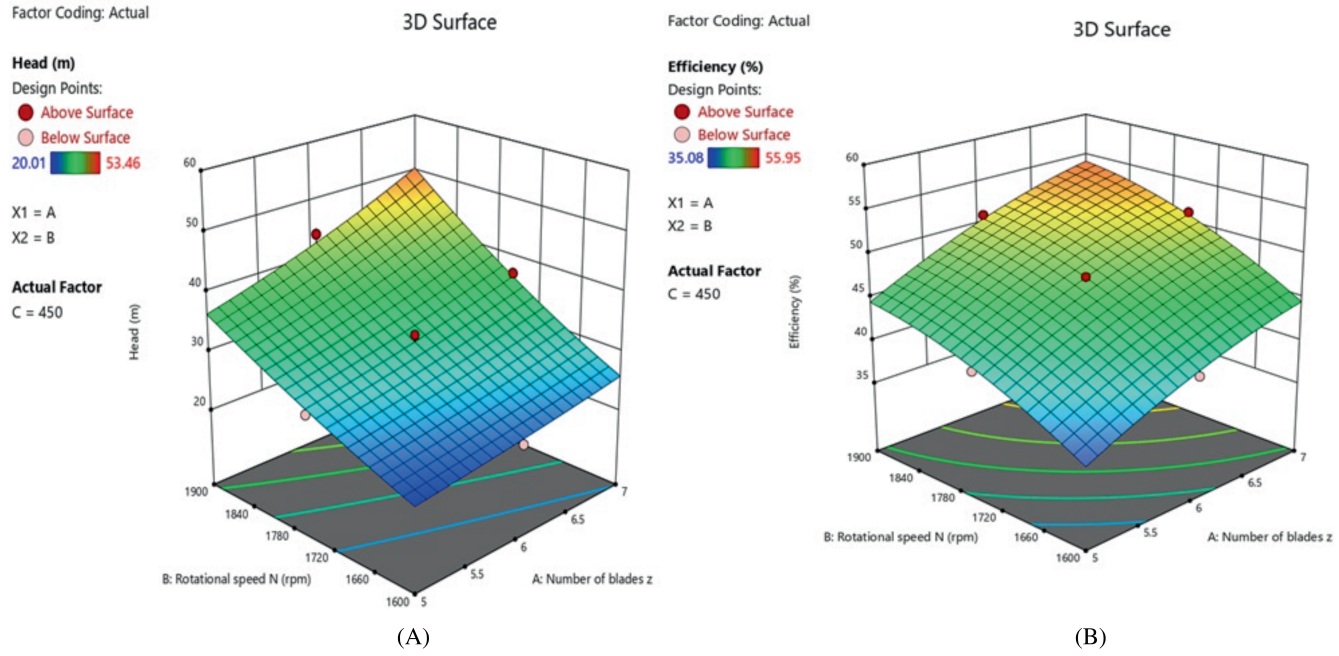


FIGURE 11 Response surface and contour plots of the interaction between rotational speed and number of blades for (A) head and (B) efficiency of single stage centrifugal pump. [Colour figure can be viewed at wileyonlinelibrary.com]

5.2 | Hydraulic performance using CFD

The single stage has been analyzed by varying blades at different rotation speeds. The CFD simulation was carried out for a single-stage pump which has an impeller with 5 blades, 6 blades, and 7 blades, a diffuser with 8 vanes, and a return passage with 11 return guide vanes. The investigation is carried out by varying the speed of the impeller to 1600, 1750, and 1900 rpm. The performance of the pump is represented by plotting graphs for parameters such as head, and efficiency

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

w.r.t. discharge. The parameters are calculated[43] as follows:

$$\text{Head coefficient} \quad H^* = \frac{gH}{u_2^2}, \tag{6}$$

$$\text{Power coefficient} \quad P^* = \frac{P_{in}}{\rho\omega^3 D_2^5}, \tag{7}$$

$$\text{Head generated} \quad H = \frac{(P_2 - P_1)\ stan.\ frame}{\gamma}, \tag{8}$$

$$\text{Flow coefficient} \quad \Psi = \frac{W_2}{\sqrt{2gH}}, \tag{9}$$

$$\text{Speed coefficient} \quad \phi = \frac{u_2}{\sqrt{2gH}}, \tag{10}$$

$$\text{Specific speed} \quad N_s = \frac{N\sqrt{Q}}{H^{3/4}}, \tag{11}$$

$$\text{Input power} \quad P_{in} = \frac{2\pi NT}{60000}, \tag{12}$$

$$\text{Output power} \quad P_{out} = \frac{\gamma QH}{1000}, \tag{13}$$

$$\text{Efficiency} \quad \eta = \frac{P_{out}}{P_{in}}. \tag{14}$$

The following table, Table 10 illustrates how the flow characteristics, including pressure at the inlet and outlet, head, output power, input power, efficiency, specific speed, head coefficient, power coefficient, flow coefficient, and speed coefficient, change with speed for impellers with five, six, and seven blades.

Above shown figures, Figure 12 shows the performance curves of (A) head (H) versus discharge (Q), and (B) efficiency (η) versus discharge (Q) for 5, 6, and 7 blades. Moreover, Figure 12A shows that the efficiency curve for 5 blades is less than the efficiency curve for 6 blades and 7 blades. It demonstrates that as blade numbers increase, efficiency increases to a certain discharge and then decreases beyond that discharge. This is because more blades can create more turbulence and resistance, which can lead to a decrease in efficiency. Figure 12B shows that as discharge increases head or pressure produced by the pump decreases because the pump is intended to run at a certain discharge and any deviation from this flow rate will cause a change in the head produced. As the flow rate increases, the fluid experiences more friction as it passes through pipes, fittings, and other components of the pumping system. This friction causes a loss in energy, resulting in a decrease in head. The suction pressure of the fluid drops below its vapor pressure, causing cavitation. Higher flow rates can increase the likelihood of cavitation, which can result in a decrease in head.

5.3 | Internal flow

This analysis is carried out at different speeds by varying numbers of impeller blades for a single stage. The objective is to numerically simulate the initial model and optimized model of the first stage of the multistage pump and compare its performance parameters. The results are shown in streamline and pressure contour graphs.

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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 10 The variations in flow characteristics for impellers with five, six, and seven blades at various speeds.

No. of blades	Speed (rpm)	1600	1750	1900
5	P1 (pa)	246,788	251,630	263,691
	P2 (pa)	429,140	515,758	615,133
	H (m)	18.588	26.924	35.824
	Pout (W)	23486.93	34019.68	45265.72
	Pin(W)	64719.90	82740.80	101537.48
	η (%)	36.29	41.11	44.58
	Ns	64.142	53.135	46.566
	H^*	0.1560	0.1889	0.2133
	P^*	0.2039	0.2179	0.2268
	Ψ	0.4185	0.3459	0.2973
	Φ	1.7897	1.6265	1.5309
6	P1 (pa)	246,000	231,586	231,963
	P2 (pa)	462,076	554,234	663,765
	H (m)	22.02	32.88	44.01
	Pout (W)	27830.58	41557.06	55616.09
	Pin (W)	68588.31	87604.5	109200.8
	η (%)	40.57	47.43	50.93
	Ns	56.477	45.729	39.90247
	H^*	0.1849	0.2308	0.262111
	P^*	0.2161	0.2307	0.243998
	Ψ	0.3846	0.3100	0.2615
	Φ	1.6441	1.4716	1.3811
7	P1 (pa)	232,711	224,983	223,596
	P2 (pa)	481,350	588,991	707,657
	H (m)	25.34	37.10	49.34
	Pout (W)	32024.70	46884.2	62347.0
	Pin (W)	71436.42	91462.60	114066.9
	η (%)	44.82	51.26	54.65
	Ns	50.83	41.77	36.62
	H^*	0.2128	0.2604	0.2938
	P^*	0.2250	0.2408	0.2548
	Ψ	0.3515	0.2935	0.2524
	Φ	1.5327	1.3855	1.3044

5.3.1 | Velocity streamline

All cases' streamlined velocities have been prepared. If we look at the streamlined profile, we can see that the fluid particle velocity at the inlet of the pump is higher than it is at the outlet because the fluid is being drawn in at a high velocity at the inlet due to suction. Additionally, the impact of the impeller's rotation-which forces the fluid outside of its domain and causes it to flow at its fastest near the impeller outlet-was examined.

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

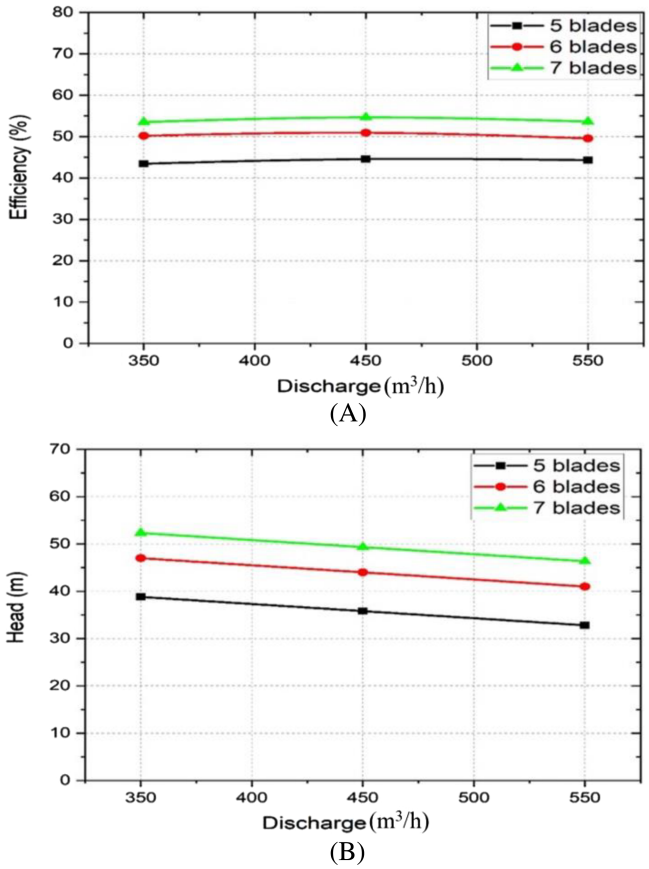


FIGURE 12 The performance curves of (A) head (H) versus discharge (Q), and (B) efficiency (η) versus discharge (Q) for 5 blades, 6 blades, and 7 blades. [Colour figure can be viewed at wileyonlinelibrary.com]

Figures 13 and 14 show the variation of velocity streamline pattern for a single-stage Centrifugal pump with 5, 6, and 7 blades at (A) 1600 rpm, (B) 1750 rpm, (C) 1900 rpm, and Streamline pattern for a single-stage optimal pump with 7 blades at 1900 rpm respectively. According to Figure 13, there is maximum velocity is found near the walls of impeller blades because, near the wall of the impeller blades, there is a sudden change in the shape of the blade causing turbulence to increase the velocity of the fluid. Figure 14 also shows the behavior of the streamlined distribution of velocity in the return channel with a 3D enlarged view. These figures show that velocity increases as rotational speed increases because it causes faster rotation of impeller blades creating a greater centrifugal force, which pushes the fluid outwards and increases velocity increasing the area of contact in between, and the amount of liquid it sucked also increases. But when the blade numbers increase it causes a reduction in velocity because the fluid has to pass through more blades which increases the resistance to flow. From Figure 13B we can interpret that the function of the return guide vanes is to deswirl the fluid before moving on to the next phase which results in an undistorted flow into the following impeller and can draw more fluid in it.

5.3.2 | Pressure distribution

The pressure distribution of a complete single-stage centrifugal pump with an impeller of 5, 6, and 7 blades as an initial model with a speed range of 1600–1900 rpm and optimized model with 7 blades, 1900 rpm, and 550 m^3/h are shown below for a pump. According to research, as the blade numbers and rotation rise, the head and hydraulic efficiency do as well. This is due to the pre-rotation of the water induced by the rotation of the impeller enhancing the water flow's energy.

Figures 15 and 16 shows the pressure contour for a single-stage pump with 5, 6, and 7 blades at (A) 1600 rpm, (B) 1750 rpm, and (C) 1900 rpm, and the Pressure contour for a single-stage optimal pump with 7 blades at 1900 rpm respectively. Figure 15 shows that a vacuum is there at the impeller's inlet due to the suction of fluid at the inlet which satisfies the theory of pump. The maximum pressure is found near diffuser vanes. The impeller is in charge of supplying the fluid

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

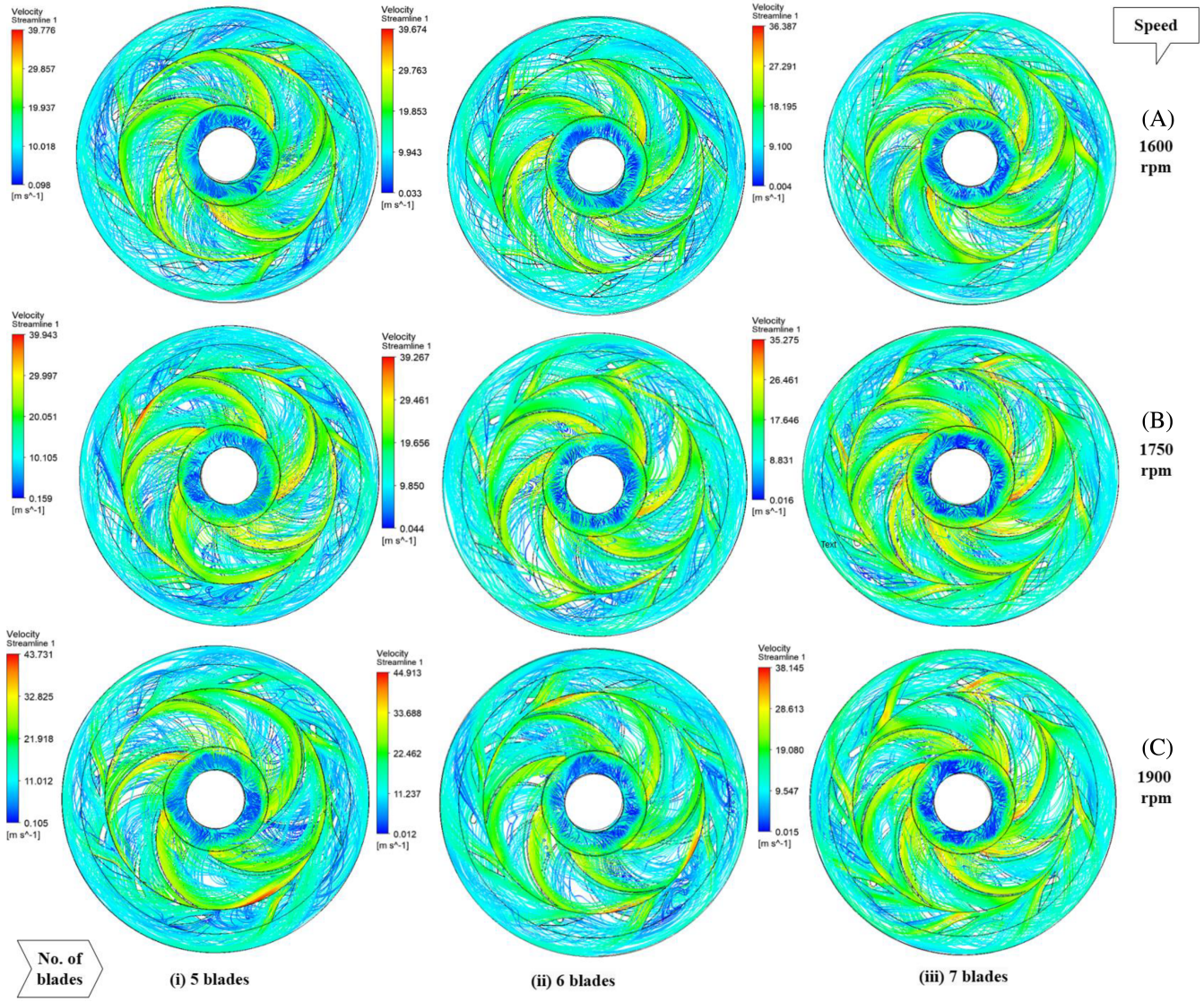


FIGURE 13 Streamlines pattern for single stage with 5, 6, and 7 blades at (A) 1600 rpm, (B) 1750 rpm, and (C) 1900 rpm. [Colour figure can be viewed at wileyonlinelibrary.com]

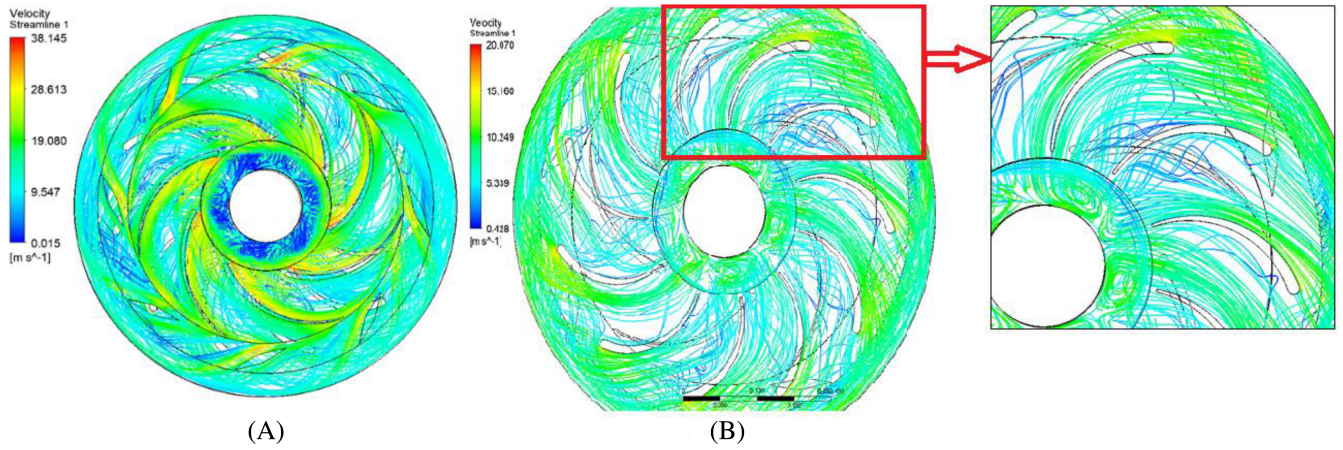


FIGURE 14 (A) Streamline pattern for optimized single-stage pump in impeller and diffuser with 7 blades, and (B) return passage at 1900 rpm. [Colour figure can be viewed at wileyonlinelibrary.com]

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

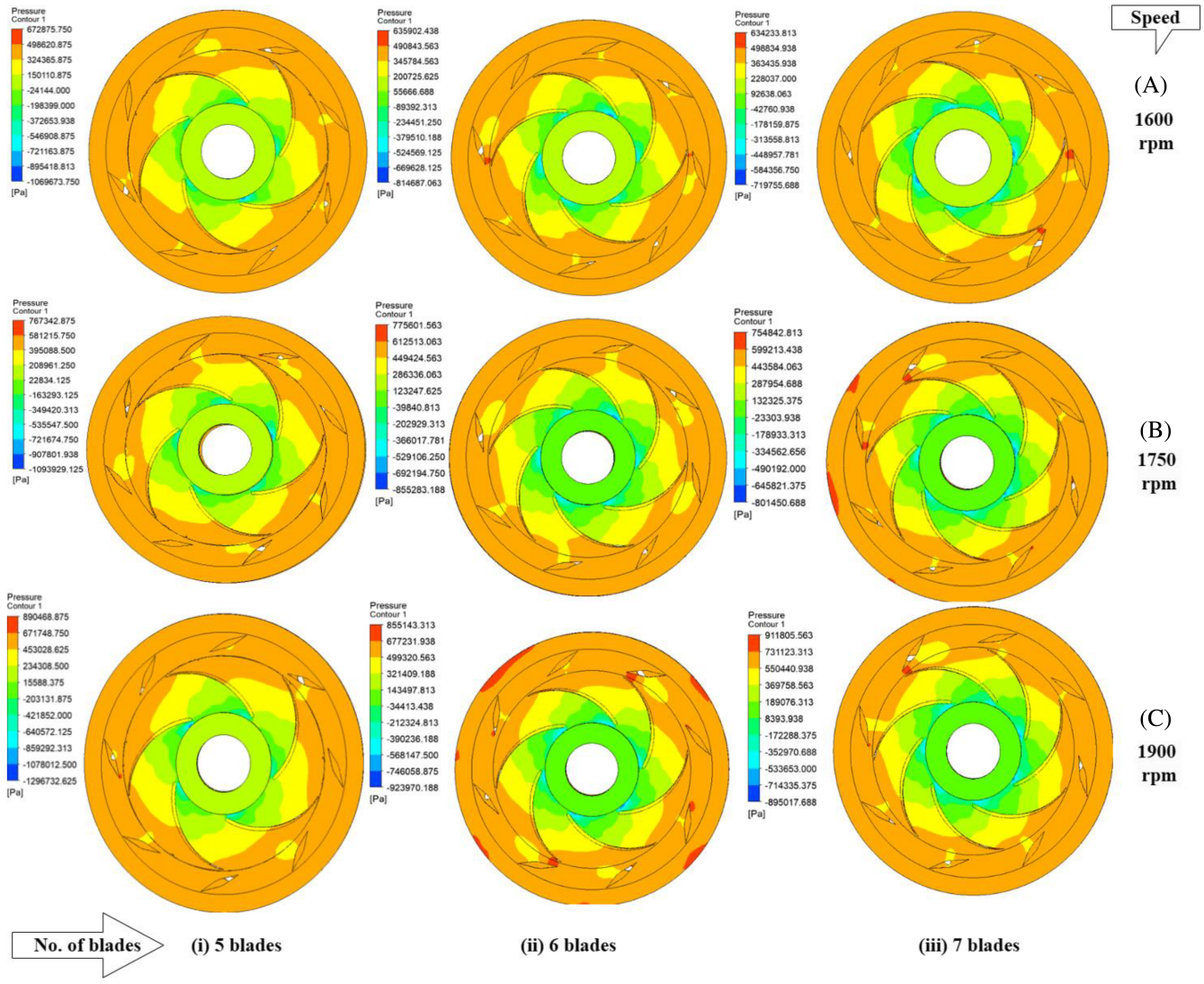


FIGURE 15 Pressure contour for a single-stage pump with 5, 6, and 7 blades at (A) 1600 rpm, (B) 1750 rpm, and (C) 1900 rpm. [Colour figure can be viewed at wileyonlinelibrary.com]

with energy by rotating and creating a centrifugal force. The fluid accumulates kinetic energy and velocity as it passes through the impeller, which causes the pressure to drop. Contrarily, the diffuser's purpose is to transform the energy of the fluid's motion into the energy of pressure. It is a stationary component with a gradually expanding cross-sectional area. An increase in the area of the cross-section results in a drop in velocity and a rise in pressure when the fluid exits the impeller and moves into the diffuser. This design efficiently transforms the impeller's kinetic energy into pressure energy, enabling the pump to produce more significant pressure. By having more pressure in the diffuser than in the impeller, the pump can achieve a higher overall discharge pressure. Figure 16 shows higher head is generated with the optimized single-stage pump with seven impeller blades at 1900 rpm.

5.4 | Validation of results

To validate the results obtained from RSM we create the 3D model using the parameters from Table 1 to obtain the optimized impeller for the entire computational domain. The optimal pump model is then estimated in CFX 2022R1 to confirm the optimization findings. The head and efficiency calculated from the numerical simulation are 49.34 m and 54.65% with errors of 7.90% and 2.06% respectively as shown in Table 11. This is because RSM takes into account a wider range of operating parameters and their interactions. Numerical simulations often focus on a limited set of parameters

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

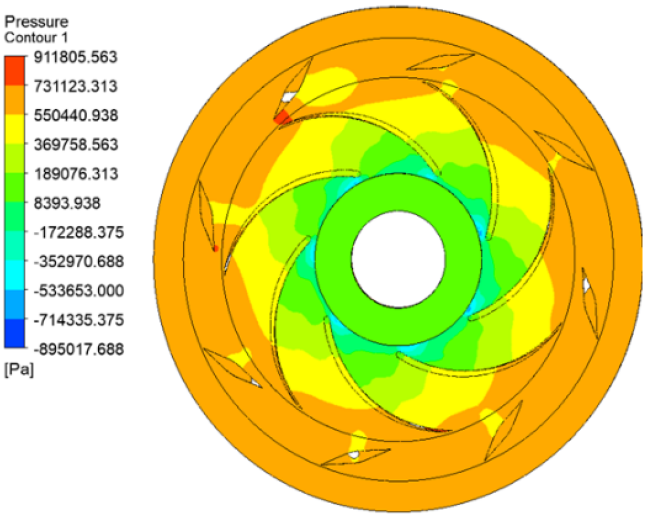


FIGURE 16 Pressure contour for the optimized single-stage pump with 7 blades at 1900 rpm. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 11 Comparison of numerically simulated results and selected solution from response surface methodology.

Results	Number of blades z	Rotation N (rpm)	Flow rate Q (m^3/h)	Head H (m)	Efficiency η (%)
Simulated results	7	1900	450	49.34	54.65
RSM results	7	1900	450	53.24	55.78

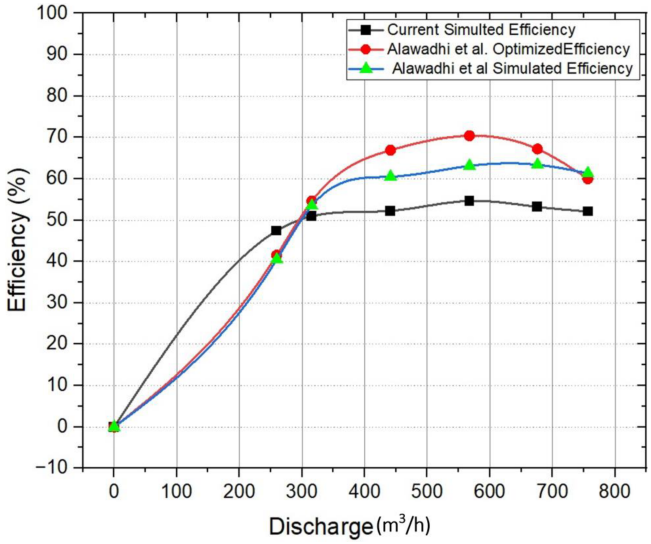


FIGURE 17 Comparison of Alawadhi et al. results and Current simulated results.[47] [Colour figure can be viewed at wileyonlinelibrary.com]

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

and assume linear relationships between them, which may not accurately capture the complex interactions that occur in a real pump system. RSM, on the other hand, uses statistical techniques to model the relationships between multiple variables and optimize the design based on these interactions. Numerical simulations typically assume idealized conditions and do not consider factors such as fluctuations in fluid properties, changes in system resistance, or variations in operating temperatures. RSM, however, can incorporate these uncertainties into the optimization process, leading to a design that performs well under a wider range of real-world conditions. Table 11 shows the numerical simulation results of the optimized single-stage pump. Figure 17 shows the comparison of Alawadhi et al.[47] work results with current simulated results. It shows that the current simulated work follows the trend of the characteristics curve of Alawadhi et al.[47] work with a reasonable error. This presents that the efficiency goes on increasing as we increase the flow rate and after a certain discharge, it decreases as we increase discharge. Because of the following high discharge rates, the pump may operate at a lower head and flow rate than its design point, leading to reduced efficiency. This is because the pump may be operating away from its optimal operating conditions, where it is designed to perform most efficiently.

6 | CONCLUSIONS

The following conclusions could be derived from the investigations and discussions of these findings:

- Three distinct speeds were used in this study to simulate the flow of a single-stage centrifugal pump with three different impellers that had 5, 6, and 7 blades.
- A model of a centrifugal pump with an impeller, diffuser, and return channel is constructed, and the stage head and efficiency of the pump are investigated regarding the impacts of increasing the number of impeller blades and rotational speed.
- From the above performance charts, velocity streamlines, and pressure contour, it has been observed that as the rotation and blade numbers increase, the performance parameters go on increasing.
- It is found that the variation of efficiency with discharge first increases, and then decreases with larger discharge after rated discharge because the pre-rotation of the water caused by the impeller's spin increases the flow energy of the water. As the number of blades increases the surface in contact with also increases which causes more turbulence and resistance.
- The head generated exhibits a diminishing pattern with discharge which follows the empirical trend of the H-Q curve. It is investigated that the single-stage pump running at a speed of 1900 rpm gives greater performance among all remaining speeds which can be validated with an optimization technique, that is, RSM.
- Using RSM with the face-centred CCD method an optimized pump model is obtained with 7 blades, 1900 rpm, and 450 m^3/h which resulted in a head and efficiency improvement of 7.90% and 2.06% respectively.
- In all, the basic pattern of the experimental results is fit by the statistical curves for head and efficiency.

NOMENCLATURE

P_1	pressure at inlet
P_{in}	input power
P_{out}	output power
H	hydraulic efficiency
N	the rotational speed of the impeller
Ω	angular velocity of the impeller
N_S	specific speed of the centrifugal pump
Q	flow rate
η	efficiency
z	blade number
P_2	pressure at outlet
H_e	Euler's head
H^*	head coefficient
P^*	power coefficient
H	head generated

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

Ψ	flow coefficient
Φ	speed coefficient
g	gravity
γ	the specific weight of water

ACKNOWLEDGMENTS

The authors want to take this opportunity to convey their gratitude to Delhi Technological University for helping to provide the DTU Fellowship.

FUNDING INFORMATION

No financial assistance is available for this work.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

1. Huang R, Zhang Z, Zhang W, Mou J, Zhou P, Wang Y. Energy performance prediction of the centrifugal pumps by using a hybrid neural network. *Energy*. 2020;213:119005. doi:10.1016/J.ENERGY.2020.119005
2. Arun Shankar VK, Umashankar S, Paramasivam S, Hanigovszki N. A comprehensive review on energy efficiency enhancement initiatives in centrifugal pumping system. *Appl Energy*. 2016;181:495-513. doi:10.1016/J.APENERGY.2016.08.070
3. Wu C, Pu K, Li C, Wu P, Huang B, Wu D. Blade redesign based on secondary flow suppression to improve energy efficiency of a centrifugal pump. *Energy*. 2022;246:123394. doi:10.1016/J.ENERGY.2022.123394
4. Li W, Li H, Liu M, Ji L, Agarwal RK, Jin S. Energy dissipation mechanism of tip-leakage cavitation in mixed-flow pump blades. *Phys Fluids*. 2024;36(1):015115. doi:10.1063/5.0183540
5. Li W, Long Y, Ji L, et al. Effect of circumferential spokes on the rotating stall flow field of mixed-flow pump. *Energy*. 2024;290:130260. doi:10.1016/j.energy.2024.130260
6. Li W, Huang Y, Ji L, Ma L, Agarwal RK, Awais M. Prediction model for energy conversion characteristics during transient processes in a mixed-flow pump. *Energy*. 2023;271:127082. doi:10.1016/j.energy.2023.127082
7. Zhao H. Neural network with data augmentation in multi-objective prediction of multi-stage pump. 2020 http://arxiv.org/abs/2002.02402
8. Jiang W, Pei J, Wang W, Yuan S. Optimization for first stage of multistage pump based on response surface methodology. *IOP Conf Ser Earth Environ Sci*. 2021;627(1):012017. doi:10.1088/1755-1315/627/1/012017
9. Zhang L, Wang X, Wu P, Huang B, Wu D. Optimization of a centrifugal pump to improve hydraulic efficiency and reduce hydro-induced vibration. *Energy*. 2023;268:126677. doi:10.1016/j.energy.2023.126677
10. Kim JH, Lee HC, Kim JH, et al. Improvement of hydrodynamic performance of a multiphase pump using design of experiment techniques. *J Fluids Eng Trans ASME*. 2015;137(8):081301. doi:10.1115/1.4029890
11. Zhou L, Shi W, Wu S. Performance optimization in a centrifugal pump impeller by orthogonal experiment and numerical simulation. *Adv Mech Eng*. 2013;5:385809. doi:10.1155/2013/385809
12. Kim S, Lee KY, Kim JH, Kim JH, Jung UH, Choi YS. High performance hydraulic design techniques of mixed-flow pump impeller and diffuser. *J Mech Sci Technol*. 2015;29(1):227-240. doi:10.1007/s12206-014-1229-5
13. Wang W, Yuan S, Pei J, Zhang J, Yuan J, Mao J. Optimum hydraulic design for a radial diffuser pump using orthogonal experimental method based on CFD. *Am Soc Mech Eng Fluids Eng Div FEDSM*. 2014;1A:21038. doi:10.1115/FEDSM2014-21038
14. Zhou L, Shi W, Lu W, Hu B, Wu S. Numerical investigations and performance experiments of a deep-well centrifugal pump with different diffusers. *J Fluids Eng Trans ASME*. 2012;134(7):071102. doi:10.1115/1.4006676
15. San AT, Oo MM, Yee Zaw CW. Design calculation and performance analysis of single suction centrifugal pump. *Int J Sci Res Publ*. 2019;9(8):p9298. doi:10.29322/ijsrp.9.08.2019.p9298
16. Gölcü M, Pancar Y. Investigation of performance characteristics in a pump impeller with low blade discharge angle. *World Pumps*. 2005;2005(468):32-40. doi:10.1016/S0262-1762(05)70749-5
17. Bai L, Yang Y, Zhou L, Li Y, Xiao Y, Shi W. Optimal design and performance improvement of an electric submersible pump impeller based on Taguchi approach. *Energy*. 2022;252:124032. doi:10.1016/J.ENERGY.2022.124032
18. Khoeini D, Shirani E. Enhancement of a centrifugal pump performance by simultaneous use of splitter blades and angular impeller diffuser. *Int J Fluid Mach Syst*. 2018;11(2):191-204. doi:10.5293/IJFMS.2018.11.2.191
19. Ji L, Li W, Shi W, Tian F, Agarwal R. Effect of blade thickness on rotating stall of mixed-flow pump using entropy generation analysis. *Energy*. 2021;236:121381. doi:10.1016/J.ENERGY.2021.121381
20. Lin Y, Li X, Zhu Z, Wang X, Lin T, Cao H. An energy consumption improvement method for centrifugal pump based on bionic optimization of blade trailing edge. *Energy*. 2022;246:123323. doi:10.1016/j.energy.2022.123323

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

21. Wang Y, Huo X. Multiobjective optimization design and performance prediction of centrifugal pump based on orthogonal test. *Adv Mater Sci Eng*. 2018;2018:1-10. doi:10.1155/2018/6218178
22. Matlakala ME, Kallon DVV, Mogapi KE, Mabelane IM, Makgopa DM. Influence of impeller diameter on the performance of centrifugal pumps. *IOP Conf Ser Mater Sci Eng*. 2019;655(1):012009. doi:10.1088/1757-899X/655/1/012009
23. Bai Y, Kong F, Xia B, Liu Y. Effect of blade number matching of impeller and diffuser in high-speed rescue pump. *Adv Mech Eng*. 2017;9(5):168781401770359. doi:10.1177/1687814017703595
24. Zhang Y, Hu S, Wu J, Zhang Y, Chen L. Modeling and multi-objective optimization of double suction centrifugal pump based on kriging meta-models. *Springer Proc Math Stat*. 2015;95:251-261. doi:10.1007/978-3-319-08377-3_25
25. Goel T, Dorney DJ, Haftka RT, Shyy W. Improving the hydrodynamic performance of diffuser vanes via shape optimization. *Comput Fluids*. 2008;37(6):705-723. doi:10.1016/j.compfluid.2007.10.002
26. Derakhshan S, Pourmahdavi M, Abdolahnejad E, Reihani A, Ojaghi A. Numerical shape optimization of a centrifugal pump impeller using artificial bee colony algorithm. *Comput Fluids*. 2013;81:145-151. doi:10.1016/j.compfluid.2013.04.018
27. Wang C, Shi W, Wang X, et al. Optimal design of multistage centrifugal pump based on the combined energy loss model and computational fluid dynamics. *Appl Energy*. 2017;187:10-26. doi:10.1016/j.apenergy.2016.11.046
28. Al-Obaidi AR. Monitoring the performance of centrifugal pump under single-phase and cavitation condition: a CFD analysis of the number of impeller blades. *J Appl Fluid Mech*. 2019;12(2):445-459. doi:10.29252/jafm.12.02.29303
29. Feng J, Benra F k, Dohmen HJ. Numerical study on impeller-diffuser interactions with radial gap variation chair of turbomachinery Vol. 2006. 2006.
30. Cheah KW, Lee TS, Winoto SH, Zhao ZM. Numerical flow simulation in a centrifugal pump at design and off-design conditions. *Int J Rotating Mach*. 2007;2007:1-8. doi:10.1155/2007/83641
31. Tong Z m, Xin J g, Tong S g, Yang Z q, Zhao J y, Mao J h. Internal flow structure, fault detection, and performance optimization of centrifugal pumps. *J Zhejiang Univ Sci A*. 2020;21(2):85-117. doi:10.1631/jzus.A1900608
32. Mendiratta K, Yadav KK, Gahlot VK. Performance optimization of radial flow centrifugal pump impeller using Cfd. *IJRET Int J Res Eng Technol*. 2016;5:134-138. doi:10.15623/ijret.2016.0511025
33. Babayigit O, Kocaaslan O, Aksoy MH, Guleren KM, Ozgoren M. Numerical identification of blade exit angle effect on the performance for a multistage centrifugal pump impeller. *EPJ Web Conf*. 2015;92:02003. doi:10.1051/epjconf/20159202003
34. Mentzos MD, Margaris DP, Papanikas DG. A numerical simulation of the impeller volute interaction in a centrifugal pump. 2004.
35. Ali M. The effect of rotational speed variation on the static pressure in the centrifugal pump (part 1). *IOSR J Mech Civ Eng*. 2013;8(6):83-94. doi:10.9790/1684-0868394
36. Lin Y, Li L, Yang S, Chen X, Li X, Zhu Z. Performance prediction and optimization of hydrogenation feed pump based on particle swarm optimization–least squares support vector regression surrogate model. *Eng Appl Comput Fluid Mech*. 2024;18(1):2315985. doi:10.1080/19942060.2024.2315985
37. Liu H, Cheng Z, Ge Z, dong l, dai c. Collaborative improvement of efficiency and noise of bionic vane centrifugal pump based on multi-objective optimization. *Adv Mech Eng*. 2021;13(2):168781402199497. doi:10.1177/1687814021994976
38. Cancan P, Xiaodong Z, Zhiguang G, Ju W, Yan G. Research on cooperative optimization of multiphase pump impeller and diffuser based on adaptive refined response surface method. *Adv Mech Eng*. 2022;14(1):168781402110729. doi:10.1177/16878140211072944
39. Wang CN, Yang FC, Nguyen VTT, Vo NTM. CFD analysis and optimum design for a centrifugal pump using an effectively artificial intelligent algorithm. *Micromachines*. 2022;13(8):1208. doi:10.3390/mi13081208
40. Lee YT, Ahuja V, Hosangadi A, et al. Impeller design of a centrifugal fan with blade optimization. *Int J Rotating Mach*. 2011;2011:1-16. doi:10.1155/2011/537824
41. Zhang N, Liu X, Gao B, Wang X, Xia B. Effects of modifying the blade trailing edge profile on unsteady pressure pulsations and flow structures in a centrifugal pump. *Int J Heat Fluid Flow*. 2019;75:227-238. doi:10.1016/J.IJHEATFLUIDFLOW.2019.01.009
42. Yu-qin W, Ze-wen D. The effect of rotational speed variation on the performance in the centrifugal pump. *Vacuum*. 2020;172:109058. doi:10.1016/j.vacuum.2019.109058
43. La Roche-Carrier N, Dituba Ngoma G, Ghie W. Numerical investigation of a first stage of a multistage centrifugal pump: impeller, diffuser with return vanes, and casing. *ISRN Mech Eng*. 2013;2013:1-15. doi:10.1155/2013/578072
44. ANSYS. CFX-Solver Theory Guide. 2009.
45. Moffat RJ. Describing the uncertainties in experimental results. *Exp Therm Fluid Sci*. 1988;1:3-17.
46. Experiments D of. response surface methodology. 2018.
47. Alawadhi K, Alzuwayer B, Mohammad TA, Buhemdi MH. Design and optimization of a centrifugal pump for slurry transport using the response surface method. *Machines*. 2021;9(3):60. doi:10.3390/machines9030060

How to cite this article: Yadav D, Singh RK, Manjunath K. Response surface method-based hydraulic performance optimization of a single-stage centrifugal pump. *Int J Numer Meth Fluids*. 2024;1-24. doi: 10.1002/fld.5332

10970363, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/fld.5332 by Delhi Technological University, Wiley Online Library on [11/09/2024]. 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

Rapid #: -23205804

CROSS REF ID: **6420163**

LENDER: **GHENT (Ghent University Library) :: Ejournals**

BORROWER: **HUL (Harvard University) :: Widener Library**

TYPE:	Article CC:CCG
JOURNAL TITLE:	International journal of intelligent robotics and applications.
USER JOURNAL TITLE:	International journal of intelligent robotics and applications.
ARTICLE TITLE:	Review on design, development, and implementation of an unmanned aerial vehicle for various applications
ARTICLE AUTHOR:	Shekh, Mubina
VOLUME:	
ISSUE:	
MONTH:	
YEAR:	2024
PAGES:	NA
ISSN:	2366-5971
OCLC #:	

Processed by RapidX: 9/27/2024 2:08:13 AM

This material may be protected by copyright law (Title 17 U.S. Code)

International Journal of Intelligent Robotics and Applications
https://doi.org/10.1007/s41315-024-00359-6

REVIEW

Review on design, development, and implementation of an unmanned aerial vehicle for various applications

Mubina Shekh[1] · **Sushila Rani**[1] · **Rituparna Datta**[2]

Received: 30 December 2022 / Accepted: 27 June 2024
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2024

Abstract
Drones are technically denoted as unmanned aerial vehicles (UAVs) or unmanned aerial systems (UAS). These are remotely or autonomously piloted aircraft and can fly without an onboard human pilot while being controlled from the ground. The need for UAVs is commonly found in wireless communication beyond human trajectory. Recently, drones have been used for a wide variety of tasks, including military operations, mining, agriculture, and general-purpose applications such as surveillance, exploration, rescue, arms and product delivery, mapping and surveying, entertainment, photography, etc. These UAVs are equipped with a range of Micro-electromechanical Systems (MEMS), including sensors, cameras, controllers, etc. In this manuscript, a novel taxonomy of flying drones with their explicitly defined applications, encompassing UAVs to smart sensors, has been reviewed. The paper also consists of a quick overview of various designing, controlling, and intelligent controller and navigation approaches of UAVs with possible applications. The goal of this study is to offer a review of the best techniques that the researcher can employ or suggest for various UAV-based applications. The paper highlights several challenges that are associated with the modeling, building, controlling, and navigating of UAVs. The article also attempts to suggest some lesser-explored areas that have the potential for future research work, such as battery refilling management, anti-theft devices, sensors to avoid collision, etc.

Keywords Drone · UAVs · Design · Analysis · Controller approach · Potential applications

1 Introduction

UAVs are wireless communication devices that are used in various fields like agriculture for stockpile estimation, wildlife for rescuing & monitoring, forest for protection from fire, mining for surveying, mapping & exploration, riverside in the flooded region for rescuing lives, haul road optimization, tailing dam monitoring, firefighting in multi-storey buildings, etc. The foremost motive of this research work is to provide a technical reference and enhance awareness and recognition of UAV monitoring in mining, military, and several regions. UAVs finds wide applications in place where the human trajectory is complex. Other advantages associated with UAVs lie in their landing, which requires less space, compact size, less cost, and no skilled operator is required as UAVs are crewless vehicles. With the help of UAVs, real-time data is collected for further planning and execution, which is further helpful in reducing the site's hazards, clearing the path, surveillance, safety, and rescuing the miners, to name a few. Winkelmaier et al. (2021) have introduced a new approach to ensure mine safety by using both aerial photography and visual analytics to monitor stress fractures. It was suggested that the mines' safety operations may be predicted by the size and position of stress fractures, as well as their progression according to a CAD model by developing tension fractures on work- or catch-benches dug. Chen et al. (2017) have suggested that computational analysis of UAVs is required to be able to reduce the plastic deformation and breakage of the product. The study has recommended the use of various software such as ANSYS, CATIA, NASTRAN, SolidWorks, Creo, etc., for the simulation of UAVs. The author used PTC Creo to generate a geometric model and to establish simulation parameters, such as materials, loads, limitations, mesh, etc.

✉ Mubina Shekh
mubina_2k21phdme502@dtu.ac.in

[1] Department of Mechanical, Production, Industrial and Automobile Engineering, Delhi Technological University, Delhi 110042, India

[2] AI/ML, Accenture, Bangalore 560066, India

Published online: 04 September 2024

for the study of non-rotor aerodynamic models in place of multi-rotor UAVs.

Zhong et al. (2021) have studied transition characteristics for tail-sitter UAVs, which have a wide angle of attack, high climbing altitude, and lack of control. Tail-sitter is a small dual-rotor UAV. It provides better aeroplane control during flight. It has been concluded that the static and dynamic analysis of tail sitter UAVs is required for a detailed transition analysis of UAVs. Genetic Algorithm was used to address the static trimming issue and the steady state during transition. Lahmeri et al. (2021) have studied the future of wireless communication networks in UAVs. They have studied, various perspectives of AI usage in UAV-based networks. Several supervised (Regression Algorithms, Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and unsupervised ML (Clustering Algorithms, Dimensionality Reduction Algorithms Machine Learning) methods have been studied. Beretta et al. (2019a) have studied the use of Drones for real-time monitoring, mapping, etc. They have suggested the use of photogrammetry software for the analysis of surface mining operations. Compared to the manual classified model with the automated lithological classification using machine learning (ML) methods found accuracy above 90% of the contact zones.

Elmeseiry et al. (2021) classified UAVs based on their wings and landing type. Based on wings, UAVs are classified as fixed-wing type and multirotor wing type. Based on landing, UAVs are classified as Horizontal take-off landing (HTOL) and Vertical take-off landing (VTOL). Fixed-wing type UAVs are similar to aircraft, having a Horizontal take-off landing (HTOL), and requiring large space for landing. Fixed-wing-type UAVs are used for large path planning because of their high battery life. Fixed-wing type UAVs are used where stationary in the air is not required as in cases i.e. for surveillance, in the military for the attack, to rescue the civilians, etc. On the other hand, rotary wing-type UAVs are similar to helicopters, quadcopters, etc. having a Vertical take-off landing (VTOL). And it requires less space for landing. It cannot cover a large space due to its battery life. Rotary wing type UAVs where stationery in the air is required as in cases i.e. for surveillance, in the military for the attack, to rescue the civilians, etc.

Hassanalian et al. (2017) have studied the classification of drones based on their size and weight. The authors have reviewed a wide range of drones ranging from UAVs with a maximum wing span of 61 m and weight of fifteen thousand kg to smart dust (SD) with a minimum size of 1 mm and weight of 5 mg. At both ends of the defined spectrum, between UAV and SD, there are various types of drones known as micro drones. The demand for intelligence missions has increased as a result of the creation of smaller air drones known as Micro Air Vehicles (MAV), Nano Air Vehicles (NAV), and Pico Air Vehicles (PAV). The study also has discussed various problems that was faced in the design of UAVs. The drones were differentiated on the basis of material used in its construction, its complexity and expenses on the control system.

In Sect. 1, the authors have reviewed research work focused on design parameters, static and dynamic analysis, control approaches, and applications of UAVs. The classification of UAVs in different categories is shown in Fig. 1

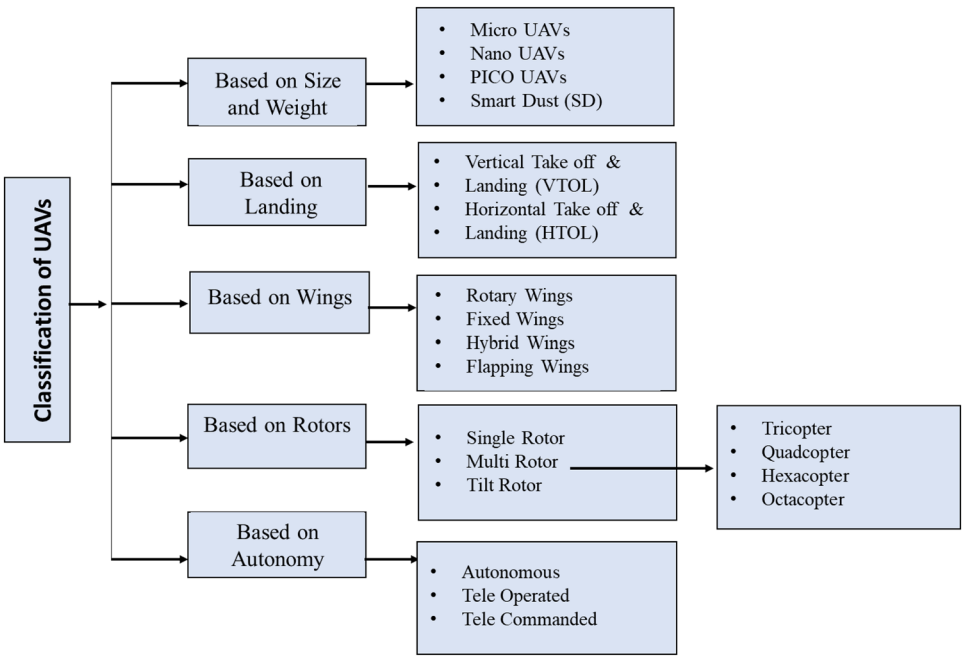


Fig. 1 Classification of UAVs

and the actual view of the drones/UAVs is shown in Fig. 2. Further research work in the field of design and development of UAVs is to be reviewed to understand in-depth research in this area.

2 Design and Development of UAVs

The designing of UAVs includes three considerations i.e. static, dynamic, and control systems. Also, it includes mathematical calculations to evaluate parameters such as the path of trajectory, lift and drag forces, orientation, shape, and size of UAVs. Papageorgiou (2018) has focused on the design optimization of UAVs by enhancing the current MDO (Multidisciplinary Design Optimization) capabilities and by exploring the use of SoS (System of Systems) models. MDO is a technique that can be used to examine design alternatives through the concurrent study of various engineering disciplines in the development of complex products. Modeling at a higher level, known as an SoS formulation, tries to provide new and enhanced capabilities that go beyond those of a single system. Rajagopal et al. (2009) have designed UAVs created using multi-objective genetic algorithms (MOGA) based on the kriging method. The design problem is expressed as a multidisciplinary design optimization (MDO) problem in conjunction with aerodynamic and structural analysis. The optimization findings support the Kriging-based MOGA's viability for UAV design optimization and offer helpful launching points for the detailed structural design of wings. Arfaoui (2017) has focused on UAVs used for wireless communications using electronic devices like sensors, cameras, etc. These UAVs are associated with a Global Positioning System (GPS) and an inertial navigation system (INS), conventional cameras, hyperspectral cameras, and multispectral cameras are used for a wide range of high-performance applications. Mallick et al. (2017) have shown that UAV is Controlled through Graphical User Interface (GUI). They proposed a mathematical model and artificial algorithm control technique for autonomous quadrotor flight. Quadrotors find applications such as trajectory tracking, graceful motion, and accurate altitude hold performance.

UAV is designed with the concept of generating ideas or usefulness and evaluating the benefits and limitations of each idea when it comes to implementation. To create an acceptable phase to produce a Drone, the design morphology depicted in Fig. 3 is employed. In Sect. 2, the literature available for designing and modeling of UAVs is reviewed.

2.1 Analysis of UAVs

Static and dynamic analysis is performed to evaluate structural analysis and dynamic behavior of the components. The structural analysis includes the evaluation of stress, strain, deformations on the body, etc. Dynamic analysis involves the evaluation of natural frequencies, nodes, aerodynamics, and stability of the systems. UAV stability is required to complete its task without any kind of hindrance. Various accessories, like sensors, cameras, batteries etc., are

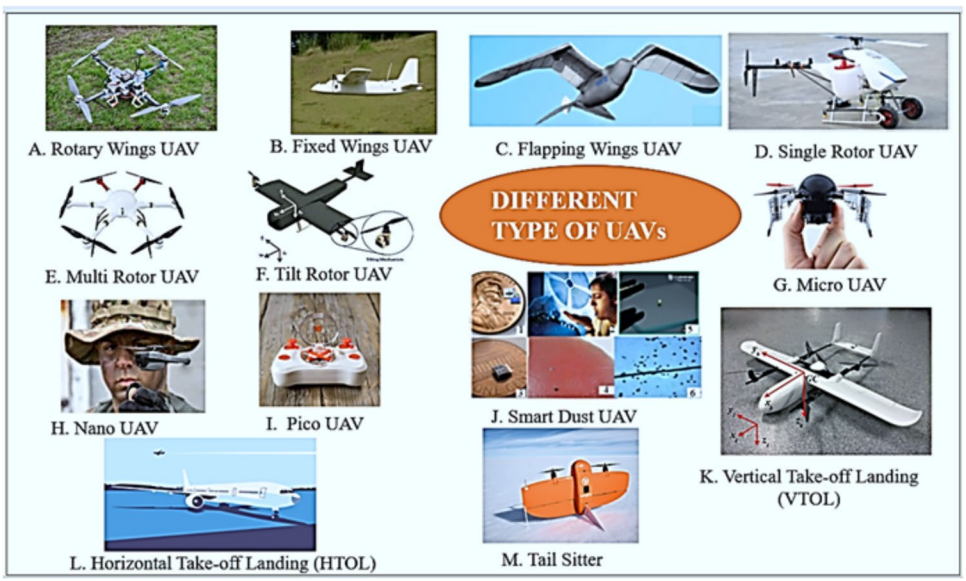


Fig. 2 Types of UAVs ((**A**–**C**, **E**–**M** (Bin Junaid et al. 2018; https://www.geeksforgeeks.org/comparison-centralized-decentralized-and-distributed-systems/; Yang et al. 2018; Zhang et al. 2019; Panigrahi et al. 2021; https://medium.com/@thegeospatialnews/why-are-micro-drones-important-107d1244dc30; https://www.militaryaerospace.com/unmanned/article/14200780/unmanned-miniature-surveillance; https://odditymall.com/skeye-pico-drone-worlds-smallest-drone; Gu et al. 2017; https://bagsid.com/2021/10/04/plane-talking-time-to-board/; https://docs.px4.io/main/en/frames_vtol/tailsitter.html), **D** (Zhu et al. 2020), **J** (Hassanalian and Abdelkefi 2017))

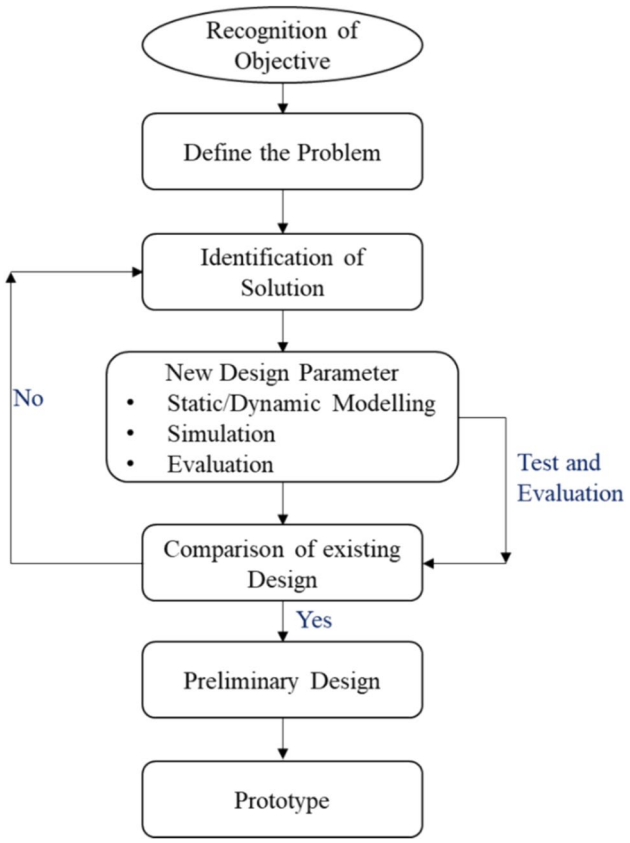


Fig. 3 Morphology of the design process

mounted on it for different applications. Therefore, UAV dynamic analysis is required to check its performance. A significant part of designing UAVs involves the use of computer-aided engineering (CAE). CAE is used for (1) stress analysis, (2) thermal and fluid dynamics analysis, (3) motion and power analysis, (4) In acoustic analyses, finite element analysis and the boundary element approach are used. (5) Mechanical operations simulation in UAV manufacturing execution systems (MES). Software such as SolidWorks, CATIA, ANSYS, NASTRAN, and PTC Creo (Pro/Engineer) enable the application of UAVs (Chen et al. 2017). Kontogiannis et al. (2013) have analyzed UAV wings. They have evaluated the efficacy and aerodynamic characteristics of the wing and the airfoil section using CFD. Further, they have optimized the results of CFD to increase overall design aerodynamic efficiency. They proposed the compact unmanned aerial vehicles are designed to perform the role of a surveillance plane equipped with a camera for photography and video. The aircraft design is completed through the use of linear aerodynamic analysis and linear stability analysis.

Li et al. (2021), have provided solutions to the problem of vibration that occurred during the take-off of UAVs. They created a single-axis and multi-axis vibration resistance module to minimize data jitter caused by the accelerometer and gyroscope sensor in the model. Singh et al. (2019), have carried out the static analysis of quadcopters using the finite elements method (FEM) based on software ANSYS 18.0 to evaluate deformation and stress contours. They have found that maximum stress is recorded at the inner edge of the quadcopter's arms, where maximum deformation is visible around the edges. Bhandari et al. (nd), have performed vibrational analysis to evaluate the dynamic behavior of quadcopters. The three-dimensional (3-D) model of the quadcopter is created using Creo 2.0. Further, the 3-D model is imported into ANSYS 16.2 for static and dynamic analysis. The result depicts the area that is most likely to fail by considering boundary conditions as fixed propeller arms are fixed bodies. Creo 2.0 was used to create the 3-D model of the quadcopter body frame, and Ansys 16.2 was used for analysis. Daud et al. (Daud et al. 2018), have carried out SolidWorks software for structural analysis of UAVs. They created various designs by reverse engineering techniques of the FURIA helicopter's design. For static and dynamic loadings, the critical components of the chosen design were analyzed using ABAQUS software. Furthermore, the simulation was carried out using MATLAB/Simulink to evaluate the values of the spring and damping constants used in the UAV suspension system. Idrissi et al. (2021) have used MATLAB/Simulink software for the in-depth analysis of UAV dynamics. The ADAMS environment is employed to justify the execution of mathematical dynamics in Simulink. The quadrotor frame is created in SolidWorks to conform to the Simulink-proposed system's requirements for its attributes. From the result, it has been concluded that the performance from the ADAMS environment has similar results to the mathematical model. Idrissi et al. (2022) have reviewed several vertical take-off and landing (VTOL) formations. They presented various strategies for UAV dynamics, simulation tools, and control. They have explained various design approaches such as Dynamic simulation, aerodynamic effect on the propeller, etc., and controller approaches such as Backstepping, Gain Scheduling, Adaptive Control, Neural Networks, Fuzzy Logic, PID, Sliding mode controller (SMC), etc. Nigro et al. (2019) have proposed a fully actuated ODQuad (OmniDirectional Quadrotor). These UAVs may simultaneously change the tilting angles of all the propellers by utilizing two extra actuators, which allow position and attitude motions to be independent of one another. Due to the non-parallel axes of the propellers, this method avoids internal forces and energy dissipation, unlike other fully actuated UAVs with tilted propellers. The kinematic and dynamic models are created with a preliminary mechanical design of UAVs. Additionally, they have developed a hierarchical two-loop-based motion control technique. Further, a simulation of the designed UAV has been performed to determine its feasibility.

Zhu et al. (2020) have designed a compact drone for a certain airframe size which is aerodynamically efficient.

This method involved drone modeling configuration, CFD analysis, and flight test operations. They suggested a new technique for an octocopter drone that uses a hybrid method to produce 41.5% more thrust than a coaxially configured model with the same rotor blade size and 71% more thrust than a traditional model with the same length of the airframe arm. Panagiotou et al. (2020) have conducted the aerodynamic analysis and drag breakdown analysis for a realistic UAV wing platform. In this research, the authors investigated the enhancement of efficiency/performance for fixed-wing UAVs that operate in the low-speed subsonic regime. Omotuyi et al. (2021) have an Unscented Kalman Filter (UKF) based method for designing UAVs. The authors proposed the use of an Unscented Kalman Filter (UKF) to combine data from the quadrotor's dynamics with visual-inertial measurements to estimate the UAV's 6-DOF orientation. The author estimated that a visual-inertial sensor system can be improved by information from UAV dynamics. To precisely estimate the vehicle's state, they offered a UKF-based method. It reduced translational error by 2–5 percent and rotational error by 37%. Hossain and Krouglicof (2016) have proposed bond graph modeling in their research for a multi-body simulation and the stabilization control law for quadrotors (UAVs.) This study showed that using a bond graph and multi-body modeling approach together makes it easier to extend or simplify the system model or its controllers. It described the complicated and coupled dynamics of robotics systems with more accuracy. Fernando et al. (2013) has modeled the dynamic model of Unmanned aerial vehicles using the Newton–Euler (NE) method. The model suggests the consequences of the forces and torques derived by the quadcopter's four propellers on its motion. Based on the model, a quadrotor simulator in MATLAB® Simulink was created, which can be used to develop and test various control algorithms.

There are a variety of aerial vehicles available. Most of the time, we employ rotary-wing aircraft since they need less space for landing and are more convenient to operate than fixed-wing planes. Additionally, this vehicle is capable of soaring at great altitudes. Designing an unmanned aerial vehicle (UAV) is difficult because of its form and significant power consumption while in flight. For the UAVs, Intelligence is also required because it is an autopilot (pilotless) device, it make decisions by its intelligence therefore AI is required. For the successful implementation of AI, some electronic components are also required like sensors, controllers, etc.

2.2 Challenges during the designing of the UAVs

Various challenges occur during the design and development of Unmanned Aerial Vehicles (UAVs). In drone technology, payload capacity and flying time are always compromises (Elmeseiry et al. 2021). UAVs are typically powered by onboard, small, lightweight lithium-ion batteries, although their power backup is inferior to that of other batteries. The mission may not be completed due to a decrease in endurance caused by an increase in shipment. The important challenge in designing of UAV is to ensure its stability i.e. to keep it stable and quiet while maintaining a modest weight. In addition, UAVs are unable to carry heavy loads because of their small weight (Saha et al. 2011). To overcome this challenge, UAVs should be designed by the type of application performed by UAVs, and the design needs to be improved. Generally, the key UAV challenges such as 3D deployment, performance analysis, channel modeling, and energy efficiency are explored along with representative results.

1. Battery capacity is another challenge while designing UAVs as the length of a UAV mission depends upon battery capacity. Researchers have proposed numerous solutions for charging batteries such as a battery charge prediction model based on the Particle Filter algorithm (Wang and Ma 2016). Moreover, an automatic charging station system can be integrated into the UAVs. The station system is made up of four components: a battery, a power converter, a wireless charging pad, and a solar panel (Zuo et al. 2022).
2. The conventional control design system is adequate for general activities. However, when external disturbances such as wind and clouds occur, it is inadequate. At that point, learning-based control techniques are needed to solve the control and navigation issues. As a result, the combination of data-driven and learning-based methodologies with the theory of dynamical systems creates a new potential for tackling hitherto insurmountable challenges in the modeling and control of unmanned aerial vehicles (UAVs) (Blakeney 2021).
3. UAV collision avoidance is a significant difficulty because it is crucial to prevent mishaps with any obstruction. Obstacles might be moving or still, and UAVs are susceptible to colliding with them. The collision avoidance system automatically detects and avoids collisions and quickly alerts the UAV operator, who may decide to assume control to resolve the situation. This control action can be performed by a sequence of to sense, detect, and avoid the collision (Deng et al. 2021).
4. The safety and security of UAVs for each essential component required to be ensured. The UAV systems face cybersecurity difficulties since they could be targeted by hackers. The communication channels that connect the system's different parts of the equipment are vulnerable to numerous attacks. To avoid the attack of intruders UAVs' self-control is required. It can be performed with the help of Data analytics and machine learning which

can detect the attacks and take action accordingly (Elmeseiry et al. 2021).

5. The issue of privacy poses another hurdle in the advancement of UAVs, posing risks to both public and private property. While various criteria and certifications exist, they do not specifically address this concern. To address privacy issues, it is suggested that individually identifying numbers for UAVs and operator licenses be employed as potential solutions (Elmeseiry et al. 2021).

In summary, the challenges inherent in UAV design encompass a broad spectrum of considerations, ranging from power sources and control strategies to safety, security, and privacy concerns. A summary of the challenges and proposed solutions is provided in Table 1. Addressing these challenges is pivotal to the continued development and safe integration of UAV technology into various domains.

3 Flight control system and computational modelling

For flight control of UAVs, hardware, and software components are required as shown in Figs. 4 and 5. It will let the aircraft be remotely piloted either manually by the pilot or automatically by an onboard computer. Since UAV flight dynamics are very unpredictable and non-linear, maintaining attitude and stability could necessitate ongoing calculations and flight system adjustments. Control system analysis is required to control the flight of the aerial vehicle in dynamic conditions. Pu et al. (2013) have proposed a Brain Emotional Learning (BEL) intelligent control system for simulation over traditional flight control systems. The Lyapunov theorem for stability is first used to evaluate the learning algorithm of BEL. Yakovlev et al. (2015) have proposed preparing a multirotor unmanned aerial vehicle investigator's duty for path planning. The author suggested a method for predicting path geometry constraints based on the UAV flight dynamics model and control constraints, the author suggested a method for predicting path geometry constraints.

Cledat et al. (2020) proposed a method for forecasting mapping quality based on data available before the flight, such as the flight plan, expected flight time, rough digital terrain model (DTM), prevalent surface roughness, and embedded sensor characteristics. In a scenario where satellite signals are frequently blocked due to the drone's motion or its surroundings. Here, the PPK (post-processed kinematic) technique is preferred over the RTK (real-time kinematic) technology. The technique of satellite positioning quality prediction is worth taking into account for navigation safety not only at camera stations but also (potentially with various criteria) over the full drone trajectory, including the take-off

Table 1 Summary of the challenges (Singhal et al. 2018)

Challenge	Concern	Present Scenario	Proposed Solution
Design and Modelling	Payload and Endurance Limit	The tradeoff between payload and Flight time	Design for standard improvement
Control and Navigation	Windy and adverse climate change	Emergency Landing	Learning-based control techniques are needed
Battery Capacity	Charging of battery to accomplish the task	Planning scheduling and replacement of battery	Solar-powered UAVs, Wireless Charging or Battery swapping
Collision Avoidance	Avoid crashing	Emergency landing or change the path	Learning-based control techniques
Safety and Security	Parts of the UAVs	Cyber Security Challenge	Data Analytics and Machine learning
Privacy	Exposure to public and private property	Legal action is taken	Proper rules and regulations are formed by the government

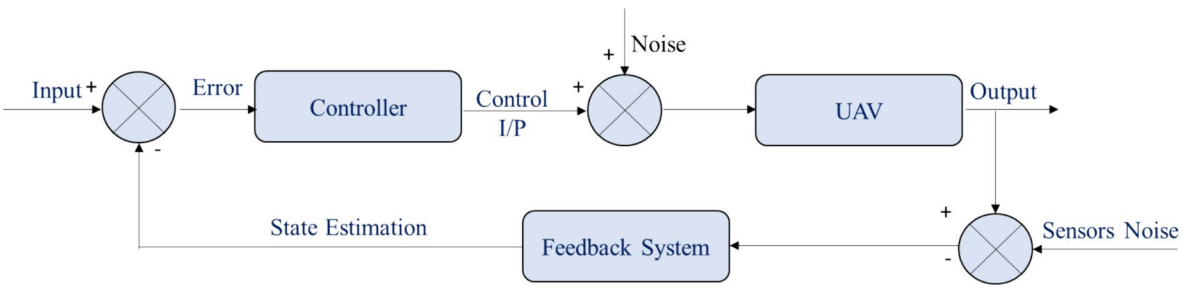


Fig. 4 UAV Control system (Elmeseiry et al. 2021)

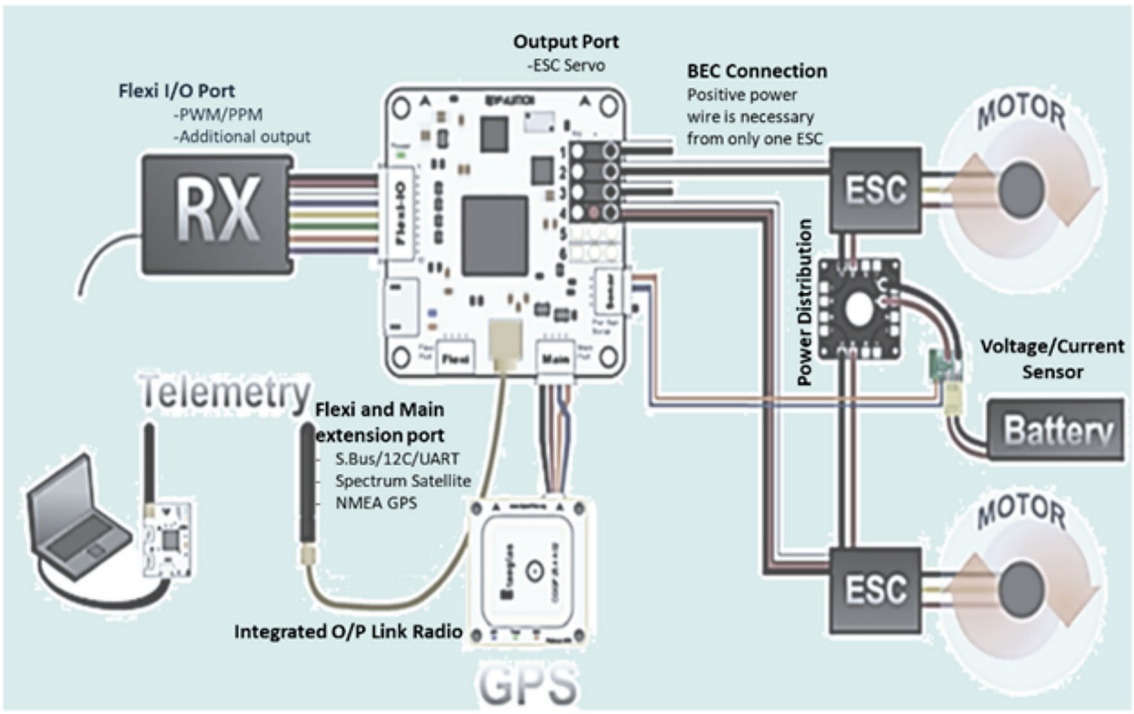


Fig. 5 Components of UAVs (Kangunde et al. 2021)

and landing zones. Karimi and Pourtakdoust (2013) resolved the motion planning issue in real time using a brand-new dynamic hybrid approach. With less computational cost, this approach integrates desired characteristics including quick convergence to an optimal path. Tang et al. (2020) suggested an active and passive feedback system-based trajectory fusion approach for unmanned aerial vehicles. The passive surveillance data came from the unmanned aerial vehicle ground control station, while the active surveillance data came from autonomous dependent surveillance broadcast. The UAV was tracked using the IMM (interacting multiple-model) approach to produce local trajectories. The simulation results demonstrated that the local trajectory is produced by the interaction of multiple model techniques. The study also showed that multiple model techniques were rougher than the fusion trajectory produced using the convex combination fusion algorithm. Chen et al. (2018) have proposed a modeling and simulation-based development process for drones. Autonomous features using open-source hardware and software frameworks are developed. Based on open-source hardware platforms like CC3D and Raspberry Pi, a generic drone structure is being developed. The planned structure is supported by Libre Pilot, an open-source software package to control multirotors. Using model-based design in Scilab, the autonomous feature design is produced.

Schmidt et al. (2019) showed two flexible flying-wing UAVs' flight dynamics, flutter, and control algorithms in their research. The UAVs are being created for flight study into the modeling and control of incredibly elastic aircraft. They will have wing spans of 10–14 feet, aspect ratios of over 10, and control surfaces along the whole wing trailing edges. An autopilot and many active flutter reduction systems are among the control systems created for this aircraft. Pitch and roll attitude control are also included in the inner loops of the autopilot together with an airspeed control. Santoso et al. (2020), have suggested a control system for Drones, that uses fuzzy feedback control autopilot to conduct precise trajectory tracking in the presence of uncertainty. The authors suggested a hybrid feedback and feedback control autopilot that effectively eliminates both the external instabilities and the cross-coupling disturbance between the adjacent and longitudinal loops about the vertical loop. Akhtar et al. (2017), proposed a control system, and haptic feedback used to operate the UAV in real-time monitoring and digital simulations. In modeling, the control configuration for the UAV's motion and altitude controllers, PID & PD control techniques were used. Java Net Beans are utilized for haptic control, MATLAB/Simulink, PID & PD Controller, Solid Works for CAD, and a virtual simulation model for controlling a UAV system. Haptic technology will substantially improve the operator's and UAV's functional performance, reducing accidents.

The term "UAV control" refers to a broad range of methodologies and tools. The design and analysis of UAV control must adhere to a precise methodology since UAV designs are inherently unstable. To ensure stable behavior

and the required performance of the UAV, which may be impacted by external disturbances, the mission of the flight control system of the UAV is crucial. The controller strives to minimize the error between the desired state and the estimated state, which could be the position, velocity, or attitude of the UAV with disturbances and activities in various environmental conditions (Elmeseiry et al. 2021).

An unmanned aerial vehicle (UAV) can be made to fly in a specific direction and at an appropriate height above ground level by adjusting its altitude. It can also be guided along a predetermined course by adjusting its velocity and orientation. For UAVs, many control laws can be applied. As shown in Fig. 6, there are four different types of control systems i.e. linear control, nonlinear control, learning-based control, and swarm control systems. Additional categories for linear controllers include PID, PD, LQR, and gain scheduling control systems. Nonlinear controllers, such as Pixhawk open-source controller, Backstepping, Sliding Mode, and Feedback Linearization. MATLAB, fuzzy logic, neural systems, and Swarm control systems, which include centralized, decentralized, and distributed control systems, are examples of learning-based control systems (Deng et al. 2021). Early attempts at autonomous flight employed linear controllers, including:

- A Proportional–derivative PD controller, The extended state observer (ESO)–based proportional derivative (PD) controller is developed to build the inner-loop attitude control, which improves the UAV system's capacity to block interference from both internal and external uncertainty (Melo et al. 2022).
- A Proportional–Integral–Derivative PID controller, This is one of the most popular UAV controllers. The controller gains are chosen to align the UAV's present state with a reference state. They are very common for UAVs because of how simple they are to use. However, their robustness and optimality are also constrained (Melo et al. 2022).
- A Linear Quadratic Regulator LQR controller, This controller minimizes a suitable cost function in dynamic systems. With a substantial stability buffer to loop errors, it is regarded as robust about process uncertainty. However, it necessitates knowing the entire system's state, which is not always possible (Melo et al. 2022).
- Gain Scheduling Controller, The PID controller can be modified using a fuzzy-gain scheduling approach to stabilize both position and altitude. This control technique needs to be efficient, straightforward, and resistant to UAV uncertainties and outside disturbances (Melo et al. 2022).

System non-linearity and coupling elements of UAV state variables are dealt with by non-linear controllers. A discussion of these controllers is provided below:

- The Backstepping Controller, functions by dividing the controller work into numerous steps so that each sub-system can be processed and gradually stabilized. The technique quickly converges, uses fewer computer resources, and handles uncertainty from the outside world (Jiang et al. 2019).
- The Sliding Mode Controller (SMC), depends on the concepts of Lyapunov stability. It employs a discontin-

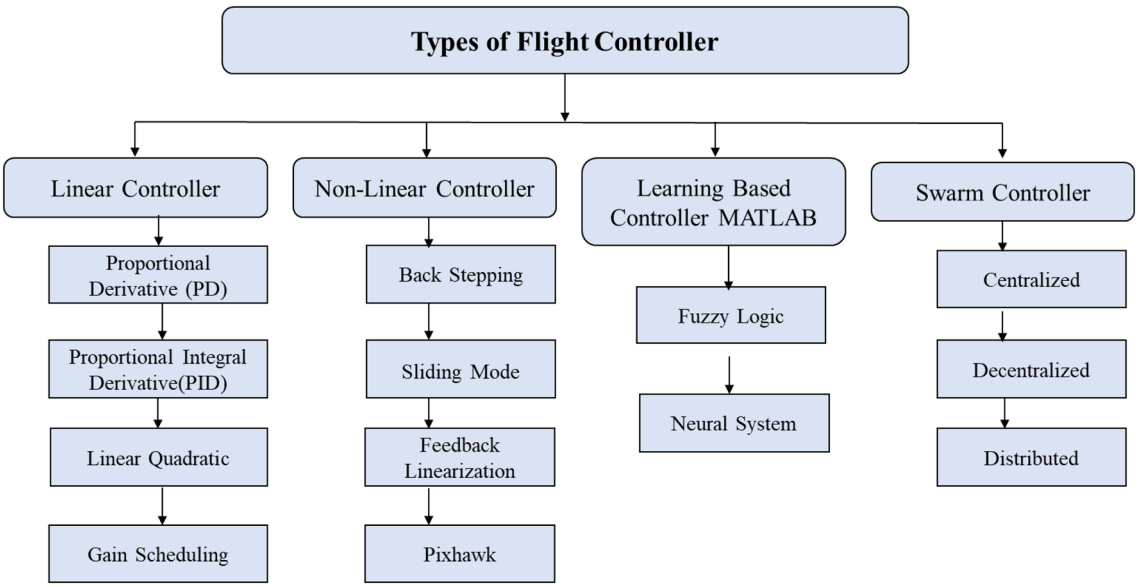


Fig. 6 Flight control systems (Idrissi et al. 2022; Kim et al. 2019)

uous control signal to compel the system to slide along a recommended path. It is not affected by perturbations, modeling mistakes, or parametric uncertainties (Jiang et al. 2019).

- Feedback Linearization, This approach is suggested to construct attitude tracking control of UAV, utilizing linear extended state observer (LESO) to estimate and account for the disturbances. As the position controller, a saturation control law with arc tangent functions is used (Jiang et al. 2019).
- Pixhawk open-source controller, The open-source Pixhawk flight controller platform is frequently used by UAVs and autonomous vehicles. Its hardware customization options and PX4 software offer flexibility across a wide range of applications. Pixhawk facilitates the integration of numerous sensors and mission planning via ground control stations (Priandana et al. 2020).

Robust Control Algorithms, to make sure the controller performs effectively and within the allowable disturbance range, are used. To deal with the system for self-assessment of the disturbance and identify the solution while applying UAVs, learning-based controllers using MATLAB are employed.

- Fuzzy Logic: Precise battery charge levels must be taken into account so that the system doesn't become unstable before it allows control of height and orientation angles (Pitch, Roll, and Yaw). A MATLAB interface is created and put into use for the visualization and communication with the drone. It enables interaction between the user and every system feature so that the mode of execution can be selected (Betancourt et al. 2019).
- Nural network: To forecast the linear accelerations from the flight record, a neural network model is trained. To accomplish real-time online optimization, the neural network-based model predictive controller is then put into use with the automatic control and dynamic optimization toolkit (ACADO) (Jiang et al. 2022).
-
-

Swarm control: The following systems, including centralized, decentralized, and distributed control systems, were employed in the early years to achieve autonomous flight:

- Centralized: Centralized Systems that use client/server architecture and include one or more client nodes that are directly connected to a central server. In many firms, this is the type of system that is most frequently used, where a client submits a request to a company server and then receives the response (https://www.geeksforgeeks.org/comparison-centralized-decentralized-and-distributed-systems/).
- Decentralized: Every node in decentralized systems makes its own decisions. The sum of the decisions made by the individual nodes determines how the system ultimately behaves. It should be noted that the request is not handled by a single entity (https://www.geeksforgeeks.org/comparison-centralized-decentralized-and-distributed-systems/).
- Distributed: A distributed system is a group of software applications that pool processing resources from various independent computation nodes to accomplish a single, shared objective. It utilizes different nodes to interact and synchronize over a shared network and is also known as distributed computing or distributed databases (https://www.geeksforgeeks.org/comparison-centralized-decentralized-and-distributed-systems/).

Khan and Al-Mulla (2019) have suggested in their research, that the design requirements for drones that can fly longer and more autonomously are continually evolving as a result of the advancements in UAV technology. This study is based on the utilization of UAVs along with ML and its algorithms in various locations. The application of ML has improved UAV capabilities and opened doors to new areas. The combination of ML and UAVs has produced quick and accurate outputs. The UAVs' ability to fly at low altitudes, tiny size, high resolution, lightweight, and likelihood are advantages. Shi et al. (2019), showed that a new robust nonlinear controller (Neural-Lander) based on deep learning enhanced the control performance of a quadrotor during landing. They compared to theoretical ground-effect models that can learn from coupled unsteady aerodynamics and vehicle dynamics to produce more precise estimates. This model outperformed the standard controller in all directions and can account for both the ground effect and other non-dominant aerodynamics.

Beretta et al. (2019b) described the use of ML algorithms (Support Vector Machine (SVM), k-nearest Neighbour (kNN), Random Forest (RF), and Gradient Tree Boost (GTB)) to classify a complete, in-depth geological model by automated detection of surface features. Flying over an opencast was done with a DJI Phantom 4 Pro data entry for classifying rocks. The UAV training was operated through a machine-learning model with appropriate metrics. It has been proven successful in automating the classification of lithological groupings using visible light components. Four classes that had regions that could be easily distinguished were manually chosen and used as training data Beretta et al. (2019a). The use of photogrammetry software to keep up with surface mining activities onboard UAVs is a reality. Using ML algorithms for automated lithological classification is a low-cost

advancement. When compared to lithologies based on automatic dense point cloud classification and manual (or reality) categorized models, it may display accuracy levels above 90% of the contact zones. The Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel is the default library setting with a sample matrix size factor that exhibits the maximum accuracy.

In this section, the literature available in the control system is reviewed. Many researchers are working on UAVs' flight control autonomously. There are several types of controllers discussed in this. UAV flight modes must be controlled using algorithms and a complicated logic system in a variety of applications. There are various algorithms are used by researchers. Some of them are discussed in this review paper. As discussed, UAVs are unmanned vehicles used for wireless communication. Therefore, to make communication easy, various artificial intelligence techniques are used as shown in Figs. 6 and 7.

A comparison of computation modeling approaches, which are also quite advanced methodologies, is shown in Fig. 7. For UAV autonomous flight controllers, it illustrates many sorts of computational modeling methodologies. We shall first talk about these methods. As a whole, artificial intelligence constitutes the system. AI includes machine learning as a subset. Neural networks are the foundation of deep learning algorithms, which are a branch of machine learning (https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/). A single neural network differs from a deep learning procedure, which requires more than three node layers, or depth, in neural networks.

3.1 Comparison of controller algorithms

Table 2 shows different types of control algorithms along with their technology and methods.

4 Mapping and surveying analysis

As per the requirement, we are using UAVs for the real-time monitoring process. It is used to gather real-time data using a camera and various sensors and other accessories required for the same. Map analysis is required to clear understanding of the path and safety of the miners in the mining industry. It is also required for stockpile estimation in the case of mining as well as for agriculture. There are various types of Maps used as shown in Fig. 8.

By using UAV mapping or aerial surveying the final deliverables 2D and 3D maps are created. Photogrammetry and LiDAR are the two commonly employed UAV mapping techniques. Fig 8 forms of maps that are frequently produced using UAV mapping:

- Ortho-mosaic maps: It is a two-dimensional top to down map. The camera took a composite image by looking down directly at a survey site or item. An Ortho-mosaic map is produced by stitching together digital photos captured by UAVs (https://theconstructor.org/surveying/types-maps-drone-mapping/562980/).
- 3D point clouds: The densified point cloud is created from the UAV photos as a 3D point cloud. It is composed of numerous points that constitute a geometric coordinate. Geographical (x, y, z) and color information are provided for each point. A three-dimensional model is

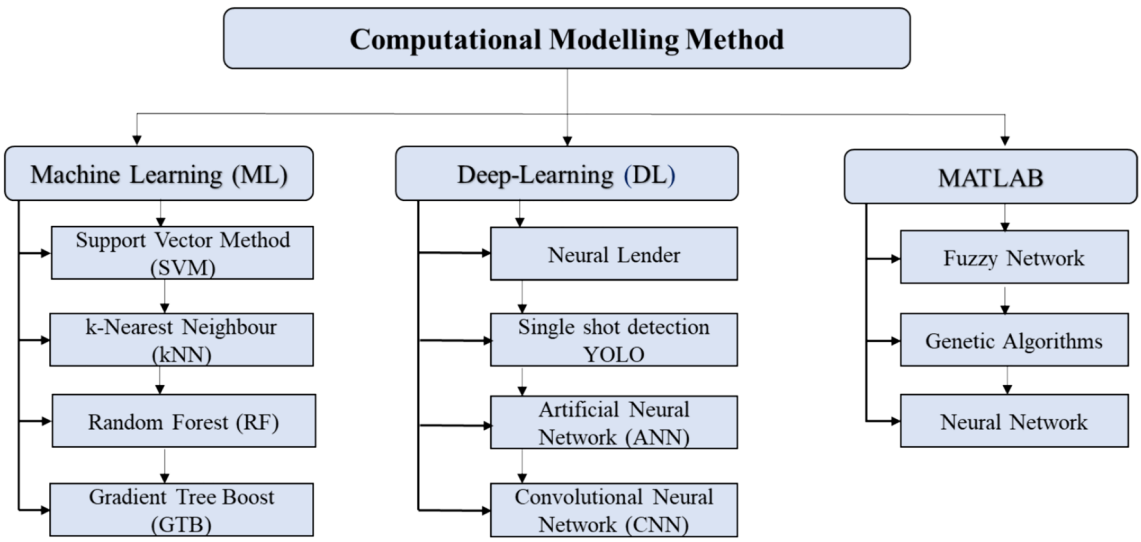


Fig. 7 Various computational modelling techniques

Table 2 Comparision of controller Algorithms

S.N	Name of controller Algo		Technology & Methods
	1	Machine Learning	Various classification models were put to the test. These tests were conducted using the Scikit-Learn Python module. According to the following descriptions, various algorithms produced representative models (Beretta et al. 2019b)
	1a	Support Vector Method (SVM)	It uses the border points between classes to classify the model (Beretta et al. 2019b)
	1b	k-Nearest Neighbour (KNN)	It predicts the value from the evaluated point's first k neighbors (Beretta et al. 2019b)
	1c	Random Forest (RF)	Based on the training samples, it generates several decision trees, and then develops a classification by voting among them (Beretta et al. 2019b)
	1d	Gradient Tree Boost (GTB)	It develops a decision tree through training that regularizes a second tree based on its flaws. This is repeated as often as the user specifies (Beretta et al. 2019b)
2	Deep Learning		Despite their intricate features, varied architectural patterns, and unique designs, buildings can be accurately spotted in drone photos (Ramachandran and Sangaiah Jun. 2021)
	2a	Neural Lender	Conventional control techniques frequently fall short of achieving a smooth landing because they are unable to adequately take into consideration these complicated dynamics. A novel deep learning-based robust nonlinear controller, called Neural-Lander, is employed in this scenario to enhance the control performance of a quadrotor during landing (Shi et al. 2019)
	2b	Single Shot Detection, YOLO	When compared to region-based detection systems, single-shot detection methods are faster and need less memory. The multi-box in an image is used by algorithms based on the single-shot detection approach to identify numerous items in a single shot. Furthermore, the idea behind YOLO is that there aren't any independent classification or detection modules that need to be synchronized with one another, and there aren't any repeated region suggestions like there are in region-based detectors. There is only one output layer, but it has several features (Ramachandran and Sangaiah 2021)
	2c	Artificial Neural Network (ANN)	To learn and complete the task, a computer system uses animal brain simulations. It is an effective computing technique for creating intricate relationships between sets of data. Because of this benefit, it is possible to anticipate the performance of UAVs' wings and airfoils using artificial neural networks (Khan and Al-Mulla 2019)
	2d	Convolutional Neural Network (CNN)	It is a feature extraction network. There are two trainable subnetworks. It's employed to shorten training's computational time and with good performance (Khan and Al-Mulla 2019)
3	MATLAB (Thesis and Akhtar 2017)		It is employed to organize and carry out UAV missions for both fixed-wing and multirotor UAVs. It is a fourth-generation programming language and multi-paradigm numerical computation environment. The MathWorks-developed proprietary programming language MATLAB enables matrix manipulation, function and data visualization, algorithm implementation, user interface construction, and interaction with other programs written in languages such as C, C++, Java, Fortran, and Python (Thesis and Akhtar 2017; https: https://www.mathworks.com/help/uav/planning-and-control.html)
	3a	Fuzzy Network	Fuzzy logic control has been preferred over traditional control techniques. Because it can handle highly nonlinear systems successfully, allows for imperfect data (https://www.mathworks.com/help/uav/planning-and-control.html; Hammoodi et al. 2020)
	3b	Genetic Algorithms	A system of genetically based algorithms is employed to resolve the UAV scheduling issue. It directly designs the initial population generation process, genetic procedures, and encoding/decoding techniques (https://www.mathworks.com/help/uav/planning-and-control.html)

Table 2 (continued)

S.N	Name of controller Algo	Technology & Methods
3c	Neural Network	The effectiveness of the controller was assessed by simulations using the MATLAB program, which exhibits improved speed and stability. The NN is effectively learning from a series of training patterns. An NN is used in system identification to record system dynamics and thus trains the NN to act as a controller (https://www.mathworks.com/help/uav/planning-and-control.html; Hammoodi et al. 2020)

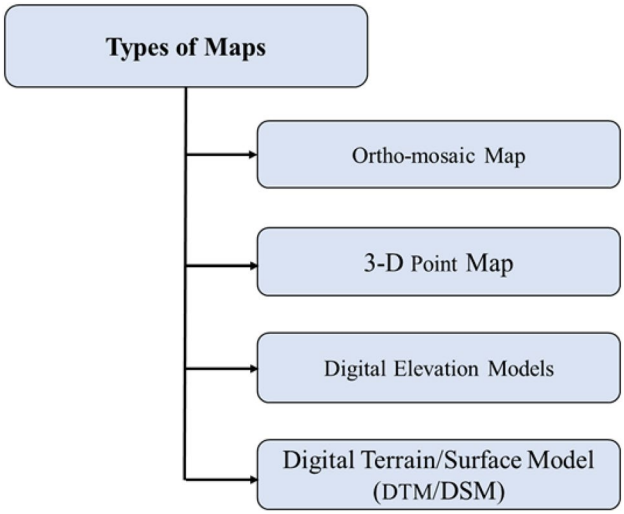


Fig. 8 Classification of maps

created by combining all of these coordinates. The model becomes denser with the increase in coordinates (https://theconstructor.org/surveying/types-maps-drone-mapping/562980/).

- Digital Elevation Models (DEM): Every point that the drones recorded on the site are represented visually in a DEM. DEMs employ colors to show the elevation difference rather than an accurate representation of the site. The figure represents the difference between DEM, DTM, and DSM UAV's maps (https://theconstructor.org/surveying/types-maps-drone-mapping/562980/; Modeling and - the differences between DTM, DSM DEM – Plex-Earth Support Desk (plexearth.com).
- Digital Terrain/Surface Model (DTM/DSM): A subset of DEM is the digital terrain model (DTM) and digital surface model (DSM). While DSM includes all the obtained points and represents both the man-made and natural features on the earth's surface, DTM offers a bare-earth surface model with just the ground and the terrain. Figure 9 shows the difference between DSM and DTM (Modeling and - the differences between DTM, DSM DEM – Plex-Earth Support Desk (plexearth.com).
- Sensor fusion technology uses UAVs to carry many sensors and software that combines the data for improved results. These systems automatically combine input data from numerous sensors, such as a conventional RGB camera and a thermal camera, to enhance the functionality of the system and its application. It is required to use top photogrammetry software to convert high-resolution photographs into accurate maps and models (Chaurasia and Mohindru 2021). Some of the best UAV mapping programs are as follows:
- Pix4D-Mapper Photogrammetry Software;
- Drone-Deploy 3D Mapping Solutions;
- Autodesk ReCap Photogrammetry Software.

These Software are used for photogrammetry that converts 2D pictures or images into 3D models or measurements of objects and scenes.

Puniach et al. (2021) have shown the effect of Land surface displacement and deformation in underground mining. Aerial photogrammetry is used as one approach to their



Fig. 9 Difference between DTM and DSM, (Source: Elevation modeling—the differences between DTM, DSM & DEM – Plex-Earth Support Desk (plexearth.com))

observation to reduce their negative consequences. Some methods make it possible to automate the measurement of horizontal displacement using photogrammetric data. Using ultra-high resolution orthomosaics, underground mining-related horizontal displacements were measured. The study compared the performance of picture registration techniques for multi-temporal orthomosaic matching. Ramachandran and Sangaiah (2021) recommended employing region-based detection algorithms such as for image detection using the Deep Learning approach, Single shot detection-YOLO, and faster R-CNN. Object detection is essential for agricultural security. Drone intelligence needs a deep learning model for autonomous action to be done. Deep learning algorithms outperform more conventional image processing approaches, according to research.

Langan et al. (2020) described that the distributed strain and pressure sensor signals were used to teach an ANN-based airspeed controller to behave like a traditional airspeed controller. Wu et al. (2022) explored the identification of objects in photos and videos taken by unmanned aerial vehicles. A picture captured by an unmanned aerial system (UAS) may need a variety of post-processing techniques such as data processing to extract useful information from the image. Additionally, networks based on optical flow, memory networks, and tracking are also options. It is possible to track several objects in UAV-borne movies using various methods, including detection tracking, single object tracking, multiple object tracking assisted by multiple object tracking, and multiple object tracking using memory networks.

Climent et al. (2012) have modelled a UAV for the accurate mapping of agricultural sites. To better understand the relationships between canopy architecture and the fraction of intercepted photosynthetically active radiation, land surface bidirectional reflectance simulations were performed (fIPAR). Three-dimensional radiative transfer FLIGHT model and measurements with ground and remote instruments on board UAV were used to investigate the relationship between canopy reflectance and instantaneous fIPAR in peach and citrus orchards. Multi-spectral cameras are installed on UAV platforms to produce high-resolution radiation maps. As depicted in Fig. 10, it might be helpful for precision farming approaches including agricultural management of cohesive zones and the identification of possible fruit quality areas for harvest.

Karim et al. (2018) proposed shape identification techniques for storing the geometrical shape of weapons for detection, and the system compares the shape of a fresh input image to a predefined database to recognize the specific weapon. It is simpler to identify illicit guns, and this will help to lower crime rates if the same-shaped weapons are legalized with the appropriate permitted licenses for the citizens. Two computing units were used in the suggested drone plane's design to perform the

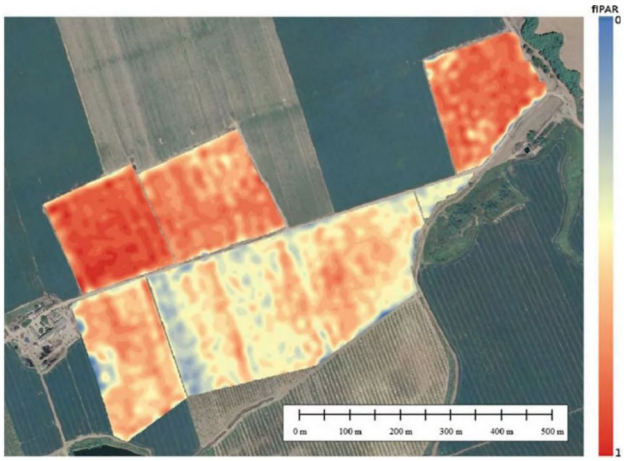


Fig. 10 Maps obtained for citrus and peach orchards showing the variation of vegetation cover (Guillen-Climent et al. 2012)

controlling and image processing functions. Weapon detection, in this case, is accomplished using a Histogram of oriented gradient (HOG), while classification is accomplished using a Support Vector Machine (SVM) (Beretta et al. 2019a). The photogrammetry technique is used to maintain surface mining activities i.e. mounted on UAVs. It shows 90% accuracy in the contact zones. As shown in Fig. 10, It showed the Point cloud representation of the mining area after topographic reconstitution using photogrammetry techniques.

The advantages of the UAV include its ability to fly at low altitudes, small size, excellent resolution, lightweight design, and portability. To assist us in making better and more informed decisions for the surveying, the UAV can be equipped with sensors, stereo cameras, and Lidars. This gives us a wider and more accurate view of all surrounding regions. Therefore, the path planning for UAVs should be done in a way that the entire area can be covered in the camera's field of view with the shortest possible flying path. To create a larger image with a wider field of view, numerous images with small but overlapping fields of view are combined in a process called mapping. The RGB camera's data is fed into a mapping algorithm that identifies critical spots before retrieving local invariant descriptors from the first two consecutive input frames using the Scale-Invariant Features Transform (SIFT) and Speeded Robust Features (SURF) methods. This model was created as a point cloud with geometric accuracy, as shown in Figs. 10 and 11. A portion of the resulting point cloud is manually characterized as a training and testing dataset using ML algorithms.

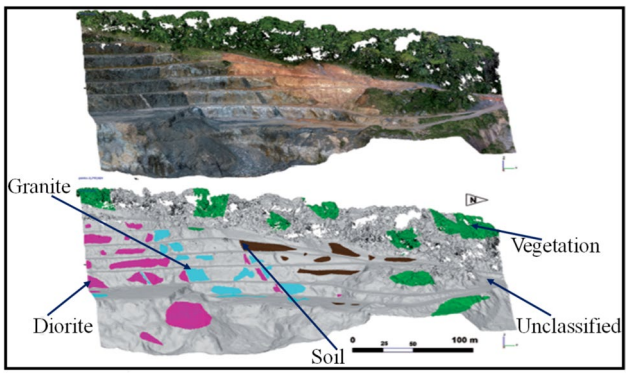


Fig.11 Area represented by point cloud using photogrammetry software (Beretta et al. 2019a)

5 Applications of UAVs

In the last two decades, several applications have recommended the use of UAVs with high-resolution cameras and satellite locating systems as shown in Fig. 12. For instance, they enable higher-resolution images to be captured at longer ranges and also provide an affordable, accurate fix for many different issues.

Park and Choi (2020) explored the possible application of UAVs for mining sites through remote sensing such as topographic surveys in open-pit mines, rock slope analysis, airborne geophysical surveys at the investigation/exploration phase, working environment analysis, underground surveying at the exploitation phase, and ground subsidence monitoring through the soil, water, ecological restoration, and ground subsidence monitoring.

Shahmoradi et al. (2020) discussed that the mines are often found in isolated and rugged highland regions. Therefore, Drones may be used to keep records on mining operations and the geography of the surrounding region, providing valuable information for better mine planning and safety. To ensure the safety of the production pit, a UAV equipped with a hyperspectral frame camera was deployed. In the case of open pit mining, slope angle optimization is critical by traditional methods. But 3D mapping of open-pit slope zones using UAVs equipped with terrestrial laser scanning and aerial photography cameras will be easy. The authors proposed the use of UAVs for reducing production costs, improving mine efficiency, and recycling resources. Backes et al. (2019) examined that the UAVs would be very helpful for implementing a new imagery approach to flooded areas as an application so that image data can be effectively used to train and guide the drone. The developed algorithm also requires validation of the technical approach on hardware platforms performing navigation, inspection, mapping, etc. Flood mitigation may be improved by using customized, accurate, and trustworthy flood models and forecasts. On the other hand, flooding in urban areas is the most difficult to deal with. High-resolution surface models may be generated from images obtained by a variety of space and aerial platforms by using improved sensor capabilities and automated post-processing techniques. Topographic mapping of small local regions has been made possible by drones flying at low altitudes, such as miniature Unmanned Aerial Systems (UAS) or "drones". Nicolás et al. (2021), and Poudel (2021) showed two distinct disciplines such as floodplain modeling, hydrology, and hydraulics (H&H) in their manuscript. Both peak flood flows and peak water surface levels were calculated through hydraulic analysis in the context of a flood region's analysis. When floodplain computer models are used in conjunction with a GIS, productivity increases. The authors discussed three methods for it i.e. the Interchange method, the Interface method, or the Integration technique. Velusamy et al. (2022), and kim et al. (2019) explored that the agriculture industry is now confronted with several issues, including labor shortages,

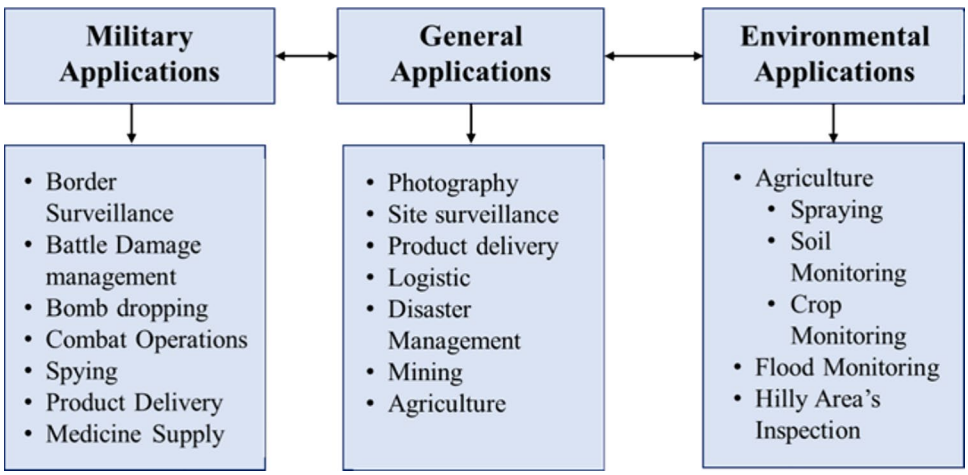


Fig. 12 Potential applications of UAVs

rising production costs, environmental deterioration, and so on. One of the most important challenges researchers face is crop loss due to illness. The authors mainly focused on pesticide and fertilizer spraying, seed sowing, growth assessment and mapping, crop monitoring, and pest management using remote sensing techniques of UAVs. It is still difficult to keep tabs on crop development and spot insect infestations in their early stages. 3D camera stockpile estimate is also utilized to obtain volume estimation of the land. Sherstjuk et al. (2018) suggested that UAVs and remote sensing techniques can be used for forest firefighting operations. This manuscript showed methods for image processing and remote sensing, a real-time decision support system for forest fire response. Researchers demonstrated the outcome of a developed system for the detection of forest fires (92%) in close to real-time settings with a processing time of fewer than 2 min, 10 m cell size, and 16 levels of the colour discretization. Alsammak et al. (2022) showed a systematic review of the use of swarm UAVs to cover the area of firefighting. This assessment showed the environment, detection of the fire's location, and efficient suppression. An innovative approach for remotely managing the firefighting effort has been the subject of several investigations. As a result, they found the following challenges, the model should be designed as Swarm robots as a distributed system, self-detecting and covering are the affected areas.

Rango et al. (2017) explored that drones are associated with various specialized sensors, cameras, and fertilizers, as well as other emerging technologies, which can help farmers combat risks and, in one example, eradicate parasites that could damage plants in cultivated fields. To inform neighboring UAVs of the existence of parasites or previously visited areas of land. There have been two coordinating strategies suggested and investigated, i.e., the first one relies on a local field map distribution. The second relies on the extension of a link-state routing protocol to the agriculture application domain to distribute all topology information among all UAVs. It is demonstrated that the dispersed link-state distribution is more appropriate since it enables lower energy usage and more parasites to be exterminated. Figure 12, represents a line diagram of potential applications of UAVs. This line diagram represents broadly three categories of UAV applications i.e. military applications, general applications, and environmental applications. Beyond these three categories, UAVs found several applications in other areas too. Moreover, UAVs have a lot of scope in agriculture as shown in Fig. 13. UAV technology has been successfully deployed in a wide spectrum of agricultural applications such as crop yield prediction, seed sowing, fertilization, pest detection, spaying of insecticides and pesticides, and crop damage assessment by environmental conditions

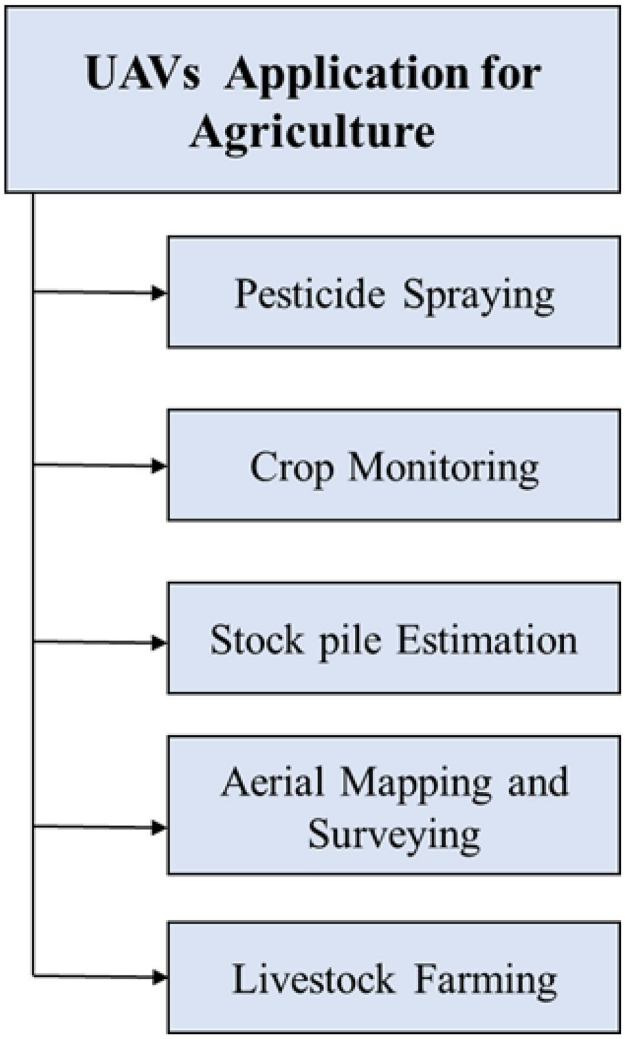


Fig. 13 UAVs agricultural control system

such as heavy rain, no rain, hail, and storm (Rahman et al. 2021). Figure 13 presents a line diagram of the potential application of UAVs in agriculture. All these potential applications are discussed in detail as:

Pesticide spraying: UAVs are used for spraying pesticides in agricultural fields. The spraying of pesticides using UAVs has the advantage over traditional spraying techniques. Using traditional spraying techniques, people are exposed to pesticides while spraying, which results in loss of pesticide from adequately constructed nozzles and is hazardous for human health. While using UAVs which are fitted with equipment that allows to estimated UAVs distance from ground (Rakesh et al. 2021).

Crop monitoring: Crop monitoring is one of the most important aspects of farming. UAVs are equipped with equipment such as a Camera (RGB, Multispectral, Thermal), and sensors (Near-infrared (NIR) that help to monitor crops precisely (Maddikunta et al. 2021).

Weed Control: To control weeds (unwanted plants) in the fields, UAVs help the farmer spray the weedicides (chemicals) (Rakesh et al. 2021).

Stock-Pile estimation: For the management of bulk grains and crops, stockpile volume estimation is needed. A multirotor UAV is used to gather data using light detection and ranging (LIDAR) point cloud data while it autonomously navigates through the warehouse (Gago et al. 2021).

Aerial mapping and surveying: UAVs can map and survey the soil and terrain in three dimensions. After creating a 3D map and seeding, UAVs are used to provide information for an irrigation system (Rakesh et al. 2021).

Livestock Farming: UAVs with real-time streaming capability are being employed, to monitor the large-scale movement of animals and their behaviors (Rakesh et al. 2021).

Future aspects for farming using UAVs: These drones are equipped with cutting-edge technologies, including image processing methods, visual systems, data analytics tools, Artificial Intelligence Algorithms, and effective spraying techniques, which enable farmers to increase productivity and improve the accuracy and efficacy of agriculture. There are the following aspects of future parameters (Khan et al. 2021).

- Increase productivity of food grains: UAVs are helpful to increase the production of food grains to meet high demands due to the increase in population globally. UAVs will be helpful in soil monitoring, improving irrigation systems, controlling pesticides/weedicides, and stockpiling from the warehouse.
- Reduction of human effort: When a human trajectory is impossible or the operator's life is in danger, UAVs are useful. The skin of a person may be affected by substances like pesticides. As a result, we can use UAVs to carry out this kind of operation. UAVs can be used for seed sowing, spraying, monitoring, and other tasks to lessen the need for human labor.
- Product delivery at the doorstep: UAVs will also assist in delivering grains, insecticides, and other items to fields where transportation is difficult to come by. This technology will be used for precision farming.

UAV technology is used for monitoring where a human approach is not possible. These are the main areas such as aerospace, boilers' inspection, drainage systems, etc. Further improvement such as increasing the payload capacity, and the endurance limit of UAVs is required to improve the quality of UAVs and further, it can help to enhance the productivity of farming.

UAVs have a huge potential for usage in both military and civilian applications due to their economic feasibility, safety advantages, and great mobility. Unmanned aerial vehicles (UAVs) are also being studied for their usage on building sites, for various activities such as inspection, tracking the progress of the job, and safety visualization (Hubbard et al. 2015). Radio and video-based electronic performance and tracking systems (EPTS) for position detection are widely used in a variety of monitoring in buildings (Rachmawati and Kim 2022). The drone-based video system not only identifies and tracks workers' positions and motions but also offers performance evaluation measures for use in training and competitive environments (Russomanno et al. 2022). UAVs help in the creation of objects including 3D modeling to simulate buildings. Regarding time and space, there are numerous approaches of UAVs to conduct monitoring (Russomanno et al. 2022). Figure 14 shows the life cycle stages of building monitoring:

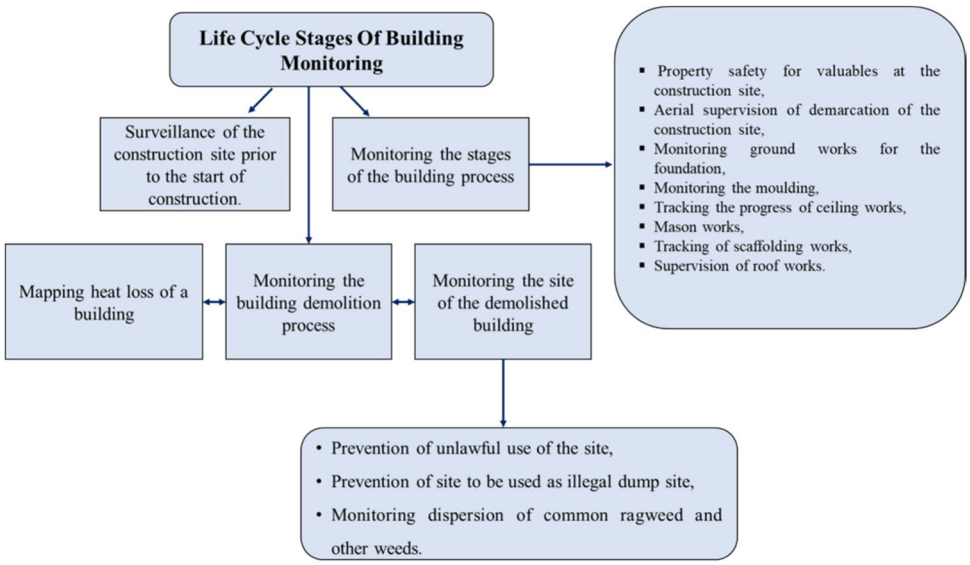


Fig. 14 Life cycle stages of building monitoring (Russomanno et al. 2022)

Supervisors and performance analysts will be able to see and quantify the X and Y positions of every individual on the field by implementing a drone-based video system (Russomanno et al. 2022). Unmanned aerial vehicles (UAVs) have drawn a lot of attention because of their wide range of capabilities for both military and general uses. Figure 12 shows a list of some applications for the various attributes that are appropriate for them. These potential applications chosen for each feature have some relationship to other features. UAVs in the military and agricultural sectors have limitless potential. There remain numerous limitations and challenges in such preliminary stages of study and innovation.

6 Future scope

Nowadays, UAVs are used in both the commercial and humanitarian sectors. These will be used even more frequently soon due to easy access. Various UAV-related opportunities are coming in several areas. There are different unsolved objectives of the work to be done in the future, like Stability and controller strategies, that are required to be solved. UAV mission paths should be optimized for energy efficiency with collision avoidance, and tracking and path planning algorithms also should be improved. For various applications, self-learning UAV systems need high-level support like ML/DL. It is used for scheduling batteries, tracking, path planning, resource allocation, and collision avoidance. Wireless System for Haptic Device Think Film Technology Solar-Powered UAVs are required to conduct a smooth operation. An actual implementation framework is necessary to fulfill the requirements of the applications in various fields.

7 Conclusion

In this review paper, an overall analysis, control system, and application of aerial vehicles are discussed. The design and deployment of UAV control and navigation systems are also discussed. Here, we identify the various parameters that occur in static and dynamic modeling. The degree of autonomy is crucial for deploying UAVs in different challenging application domains and unfavorable environmental conditions. A UAV is considered fully autonomous if it can make decisions and react to unforeseen circumstances in real time without any direct human input. Many algorithmic and technological developments are still necessary for this level of autonomy. In addition to a list of exciting open problems, we also presented a series of concluding observations that explain the present challenges, limitations, and future scope. Numerous implementations could be made in the future to support this research. The study will serve as an introductory overview of UAVs, which will be beneficial not only for novice researchers but also for experienced ones who want to explore the field of UAVs.

Acknowledgements I was delighted when I got the option to write a review paper on this significant and valuable subject. It is a complex undertaking, but I have been working continuously to complete it. I am pleased to submit my review paper on "Review on Design, Development, and Implementation of an Unmanned Aerial Vehicle for various Applications." I want to express my sincere respect and warm regard to the late Prof. Vikas Rastogi for providing continuous guidance and support during his presence. I dedicated this review paper to my supervisor, who sadly died before publication.

Author contributions Discussed several Methodologies, Design software, Control systems, and potential applications of the UAVs. This manuscript has been read and approved by all authors.

Funding This research was not funded by any other organizations.

Declarations

Conflict of interest In this manuscript, authors show their interest in the UAV's design, control, and implementation of UAVs for various applications. The authors declare no conflict of interest.

References

Alsammak, I.L.H., Mahmoud, M.A., Aris, H., Alkilabi, M., Mahdi, M.N.: The use of swarms of unmanned aerial vehicles in mitigating area coverage challenges of forest-fire-extinguishing activities: a systematic literature review. Forests **13**(5), 811 (2022). https://doi.org/10.3390/f13050811

Arfaoui, A.: Unmanned aerial vehicle: review of onboard sensors, application fields, open problems and research issues. Aymen Arfaoui Int. J. Image Process. **11**, 12 (2017)

Backes, D., Schumann, G., Teferele, F.N., Boehm, J.: Towards a high-resolution drone-based 3D mapping dataset to optimise flood hazard modelling. Int. Arch. Photogr. Remote Sens. Spatial Inf. Sci. **42**, 181–187 (2019). https://doi.org/10.5194/isprs-archives-XLII-2-W13-181-2019

Beretta, F., Rodrigues, Á.L., de Peroni, R.L., Costa, J.F.C.L.: Using UAV for automatic lithological classification of open pit mining front. Rev. Esc. Minas **72**(1), 17–23 (2019a). https://doi.org/10.1590/0370-44672018720122

Beretta, F., Rodrigues, A.L., Peroni, R.L., Costa, J.F.C.L.: Automated lithological classification using UAV and machine learning on an open cast mine. Appl. Earth Sci. Trans. Inst. Min. Metall. **128**(3), 79–88 (2019b). https://doi.org/10.1080/25726838.2019.1578031

Betancourt, F.R., Campos F.A., Pinzón, V.S. Control of UAV based on fuzzy logic in a controlled environment. Arpn J. Eng. Appl. Sci. **14**(24), (2019)

Bhandari, A., Ahmad, F., Kumar, P., Patil, P.P.: Design and vibration characteristics analysis of quadcopter body frame. http://www.ripublication.com

Bin Junaid, A., Sanchez, A.D.D.C., Bosch, J.B., Vitzilaios, N., Zweiri, Y.: Design and implementation of a dual-axis tilting quadcopter. Robotics **7**(4), 1–20 (2018). https://doi.org/10.3390/robotics7040065

Blakeney, R.A. Air to air sense and avoid on unmanned aircraft systems. https://ryanblakeney.com/air-toair-sense-and-avoid-on-uas/. Accessed 25 Oct 2021.

Chaurasia, R., Mohindru, V. Unmanned aerial vehicle (UAV): a comprehensive survey. Unmanned Aerial Vehicles for Internet of Things (IoT) Concepts, Techniques, and Applications, 1–27 (2021)

Chen, Y.L., Fang, J., Shiau, Y.R., Hung, W.C.: A study on static structural of non-rotor unmanned aerial vehicle. MATEC Web Conf. **119**, 01053 (2017). https://doi.org/10.1051/matecconf/201711901053

Chen, S., Durak, U., Hartmann, S.: Modeling and simulation-based development of autonomy features for drones. SNE Simul. Notes Eur. **28**(2), 55–60 (2018). https://doi.org/10.11128/sne.28.tn.10413

Cledat, E., Jospin, L.V., Cucci, D.A., Skaloud, J.: Mapping quality prediction for RTK/PPK-equipped micro-drones operating in complex natural environment. ISPRS J. Photogramm. Remote Sens. **167**(July), 24–38 (2020). https://doi.org/10.1016/j.isprsjprs.2020.05.015

Daud, M.E.H., Ammoo, M.S., Ishak, I.S., Mat, S.: Design and structural analysis of a quadrotor. Jurnal Mekanikal. **41**, 46–58 (2018)

De Rango, F., Palmieri, N., Tropea, M., Potrino, G. UAVs team and its application in agriculture: A simulation environment. SIMULTECH 2017 - Proc. 7th Int. Conf. Simul. Model. Methodol. Technol. Appl., no. Simultech, pp. 374–379, 2017, https://doi.org/10.5220/0006466303740379.

Deng, Z., Wu, L., You, Y.: Modeling and design of an aircraft-mode controller for a fixed-wing VTOL UAV. Math. Probl. Eng. **2021**, 1–17 (2021)

Elevation modeling - the differences between DTM, DSM & DEM – Plex-Earth Support Desk (plexearth.com)

Elmeseiry, N., Alshaer, N., Ismail, T.: A detailed survey and future directions of unmanned aerial vehicles (Uavs) with potential applications. Aerospace **8**(12), 363 (2021). https://doi.org/10.3390/aerospace8120363

Fernando, H.C.T.E., De Silva, A.T.A., De Zoysa, M.D.C,K. Dilshan, A.D.C., Munasinghe, S.R.: Modelling, simulation and implementation of a quadrotor UAV. 2013 IEEE 8th Int. Conf. Ind. Inf. Syst. ICIIS 2013 - Conf. Proc., pp. 207–212, (2013). https://doi.org/10.1109/ICIInfS.2013.6731982.

Gago, R.M., Pereira, M.Y., Pereira, G.A.: An aerial robotic system for inventory of stockpile warehouses. Eng. Rep. **3**(9), e12396 (2021)

Gu, H., Lyu, X., Li, Z., Shen, S., Zhang, F.: Development and experimental verification of a hybrid vertical take-off and landing (VTOL) unmanned aerial vehicle(UAV). 2017 Conf. Unmanned Aircr. Syst. ICUAS **2017**, 160–169 (2017). https://doi.org/10.1109/ICUAS.2017.7991420

Guerra-Langan, A., Araujo-Estrada, S., Richards, A., Windsor, S. Simulation of a machine learning based controller for a fixed-wing uav with distributed sensors. AIAA Scitech 2020 Forum, vol. 1 PartF, no. January, pp. 1–18, (2020). https://doi.org/10.2514/6.2020-1239.

Guillen-Climent, M.L., Zarco-Tejada, P.J., Berni, J.A.J., North, P.R.J., Villalobos, F.J.: Mapping radiation interception in row-structured orchards using 3D simulation and high-resolution airborne imagery acquired from a UAV. Precis. Agric. **13**(4), 473–500 (2012). https://doi.org/10.1007/s11119-012-9263-8

Hammoodi, S.J., Flayyih, K.S., Hamad, A.R.: Design and implementation speed control system of DC motor based on PID control and matlab simulink. Int. J. Power Electron. Drive Syst. **11**(1), 127–134 (2020)

Hassanalian, M., Abdelkefi, A.: Classifications, applications, and design challenges of drones: a review. Prog. Aerosp. Sci. **91**(April), 99–131 (2017). https://doi.org/10.1016/j.paerosci.2017.04.003

Hossain, M.R., Krouglicof, N.: Multi-body dynamics modeling & control of Quadrotor helicopter using bond graph. Simul. Ser. **48**(10), 102–111 (2016). https://doi.org/10.22360/summersim.2016.icbgm.012

https://bagsid.com/2021/10/04/plane-talking-time-to-board/

https://docs.px4.io/main/en/frames_vtol/tailsitter.html

https://medium.com/@thegeospatialnews/why-are-micro-drones-important-107d1244dc30

https://odditymall.com/skeye-pico-drone-worlds-smallest-drone

https://theconstructor.org/surveying/types-maps-drone-mapping/562980/

https://www.911security.com/learn/airspace-security/drone-fundamentals/types-of-drones-fixed-wing

https://www.geeksforgeeks.org/comparison-centralized-decentralized-and-distributed-systems/

https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/

https://www.mathworks.com/help/uav/planning-and-control.html

https://www.militaryaerospace.com/unmanned/article/14200780/unmanned-miniature-surveillance

Hubbard, B., Wang, H., Leasure, M., Ropp, T., Lofton, T., Hubbard, S., Lin, S. Feasibility study of UAV use for RFID material tracking on construction sites. In 51st ASC Annual international conference proceedings (pp. 669–676). (2015)

Idrissi, M., Annaz, F., Salami, M.: Mathematical physical modelling of a quadrotor UAV. 2021 7th Int. Conf. Control. Autom. Robot. ICCAR 2021, pp. 206–212, (2021). https://doi.org/10.1109/ICCAR52225.2021.9463447

Idrissi, M., Salami, M., Annaz, F.: A review of quadrotor unmanned aerial vehicles: applications, architectural design and control algorithms. J. Intell. Robot. Syst. Theory Appl. **104**(2), 1–53 (2022). https://doi.org/10.1007/s10846-021-01527-7

Jiang, H., Xia, Y., Hu, R., Ma D., Hao, C. A Feedback linearization and saturated control structure for quadrotor UAV; 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 8277–8282, https://doi.org/10.23919/ChiCC.2019.8866172

Jiang, B., Li, B., Zhou, W., Lo, L.Y., Chen, C.K., Wen, C.Y.: Neural network-based model predictive control for a quadrotor UAV. Aerospace **9**(8), 460 (2022)

Kangunde, V., Jamisola, R.S., Theophilus, E.K.: A review on drones controlled in real-time. Int. J. Dyn. Control **9**(4), 1832–1846 (2021). https://doi.org/10.1007/s40435-020-00737-5

Karimi, J., Pourtakdoust, S.H.: Optimal maneuver-based motion planning over terrain and threats using a dynamic hybrid PSO algorithm. Aerosp. Sci. Technol. **26**(1), 60–71 (2013). https://doi.org/10.1016/j.ast.2012.02.014

Karim, S., Zhang, Y,. Laghari, A., Asif, M.R. Image processing based proposed drone for detecting and controlling street crimes. Int. Conf. Commun. Technol. Proceedings, ICCT, vol. 2017-Octob, pp. 1725–1730, (2018). https://doi.org/10.1109/ICCT.2017.8359925.

Khan, A.I., Al-Mulla, Y.: Unmanned aerial vehicle in the machine learning environment. Proc. Comput. Sci. **160**, 46–53 (2019). https://doi.org/10.1016/j.procs.2019.09.442

Khan, N., Ray, R.L., Sargani, G.R., Ihtisham, M., Khayyam, M., Ismail, S.: Current progress and future prospects of agriculture technology: gateway to sustainable agriculture. Sustainability **13**(9), 4883 (2021)

Kim, J., Kim, S., Ju, C., IlSon, H.: Unmanned aerial vehicles in agriculture: a review of perspective of platform, control, and applications. IEEE Access **7**, 105100–105115 (2019). https://doi.org/10.1109/ACCESS.2019.2932119

Kontogiannis, S.G., Ekaterinaris, J.A.: Design, performance evaluation and optimization of a UAV. Aerosp. Sci. Technol. **29**(1), 339–350 (2013). https://doi.org/10.1016/j.ast.2013.04.005

Lahmeri, M.A., Kishk, M.A., Alouini, M.S.: Artificial intelligence for UAV-enabled wireless networks: a survey. IEEE Open J.

Commun. Soc. **2**(April), 1015–1040 (2021). https://doi.org/10.1109/OJCOMS.2021.3075201

Li, J., Lei, Z., Gao, D., Xu, K., Shi G.: Design and research on UAV vibration resistance module. In Proceedings - 2021 7th Annual International Conference on Network and Information Systems for Computers, ICNISC 2021, 2021, pp. 550–553. https://doi.org/10.1109/ICNISC54316.2021.00104.

Maddikunta, P.K.R., Hakak, S., Alazab, M., Bhattacharya, S., Gadekallu, T.R., Khan, W.Z., Pham, Q.V.: Unmanned aerial vehicles in smart agriculture: applications, requirements, and challenges. IEEE Sens. J. **21**(16), 17608–17619 (2021)

Mallick, T. C., Bhuyan, M.A.I., Munna, M.S. Design & implementation of an UAV (Drone) with flight data record. 2016 Int. Conf. Innov. Sci. Eng. Technol. ICISET 2016, pp. 1–6, (2017). https://doi.org/10.1109/ICISET.2016.7856519

Melo, A.G., Andrade, F.A., Guedes, I.P., Carvalho, G.F., Zachi, A.R., Pinto, M.F.: Fuzzy gain-scheduling PID for UAV position and altitude controllers. Sensors **22**(6), 2173 (2022)

Nigro, M., Pierri, F., Caccavale, F.: Preliminary design, modeling and control of a fully actuated quadrotor UAV, 2019 Int. Conf. Unmanned Aircr. Syst. ICUAS 2019, pp. 1108–1116, 2019. https://doi.org/10.1109/ICUAS.2019.8798092

Omotuyi, O., Kumar, M.: UAV Visual-Inertial Dynamics (VI-D) odometry using unscented Kalman Filter. IFAC-PapersOnLine **54**(20), 814–819 (2021). https://doi.org/10.1016/j.ifacol.2021.11.272

Panagiotou, P., Yakinthos, K.: Aerodynamic efficiency and performance enhancement of fixed-wing UAVs. Aerosp. Sci. Technol. **99**, 105575 (2020). https://doi.org/10.1016/j.ast.2019.105575

Panigrahi, S., Krishna, Y.S.S., Thondiyath, A.: Design, analysis, and testing of a hybrid vtol tilt-rotor uav for increased endurance. Sensors **21**(18), 1–21 (2021). https://doi.org/10.3390/s21185987

Papageorgiou, A.: Design optimization of unmanned aerial vehicles design optimization of. 2018.

Park, S., Choi, Y.: Applications of unmanned aerial vehicles in mining from exploration to reclamation: a review. Minerals **10**(8), 1–32 (2020). https://doi.org/10.3390/min10080663

Poudel, K.: Hydrological and hydraulic modeling for flood analysis : a case study for modi catchment. Int. J. Eng. Res. Technol. **10**(08), 534–544 (2021)

Priandana, K., Hazim, M., Kusumoputro, B. Development of autonomous UAV quadcopters using pixhawk controller and its flight data acquisition. In 2020 International Conference on Computer Science and Its Application in Agriculture (ICOSICA) (pp. 1–6). IEEE. (2020)

Pu, H., Zhen, Z., Jiang, J., Wang, D.: UAV flight control system based on an intelligent BEL algorithm. Int. J. Adv. Robot. Syst. **10**, 1–8 (2013). https://doi.org/10.5772/53746

Puniach, E., Gruszczyński, W., Ćwiąkała, P., Matwij, W.: Application of UAV-based orthomosaics for determination of horizontal displacement caused by underground mining. ISPRS J. Photogramm. Remote Sens. **174**, 282–303 (2021). https://doi.org/10.1016/j.isprsjprs.2021.02.006

Rachmawati, T.S.N., Kim, S.: Unmanned Aerial Vehicles (UAV) integration with digital technologies toward construction 4.0: a systematic literature review. Sustainability **14**(9), 5708 (2022)

Rahman, M.F.F., Fan, S., Zhang, Y., Chen, L.: A comparative study on application of unmanned aerial vehicle systems in agriculture. Agriculture **11**(1), 22 (2021)

Rajagopal, S., Ganguli, R..: Multidisciplinary design optimization of a uav wing using kriging based multi-objective genetic algorithm. Collect. Tech. Pap. - AIAA/ASME/ASCE/AHS/ASC Struct. Struct. Dyn. Mater. Conf., no. May, pp. 1–18, 2009, https://doi.org/10.2514/6.2009-2219.

Rakesh, D., Kumar, N.A., Sivaguru, M., Keerthivaasan, K.V.R., Janaki, B.R., Raffik, R. Role of UAVs in innovating agriculture with future applications: a review. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) (pp. 1–6). IEEE. (2021)

Ramachandran, A., Sangaiah, A.K.: A review on object detection in unmanned aerial vehicle surveillance. Int. J. Cogn. Comput. Eng. **2**, 215–228 (2021). https://doi.org/10.1016/j.ijcce.2021.11.005

Russomanno, T.G., Blauberger, P., Kolbinger, O., Lam, H., Schmid, M., Lames, M.: Drone-based position detection in sports—validation and applications. Front. Physiol. **13**, 850512 (2022)

Saha, B., Koshimoto, E., Quach, C.C., Hogge, E.F., Strom, T.H., Hill, B.L., Vazquez, S.L., Goebel, K. Battery health management system for electric UAVs. In Proceedings of the 2011 Aerospace Conference, Hangzhou, China, 12–14 June 2011; pp. 1–9.

Santoso, F., Garratt, M.A., Anavatti, S.G., Petersen, I.: Robust hybrid nonlinear control systems for the dynamics of a quadcopter drone. IEEE Trans. Syst. Man. Cybern. Syst. **50**(8), 3059–3071 (2020). https://doi.org/10.1109/TSMC.2018.2836922

Schmidt, D.K, Danowsky, B.P.. Kotikalpudi, A., Seiler, P.J., Kapania R.K., Flight-dynamics, flutter analysis, and control of mdao-designed flying-wing research drones. AIAA Scitech 2019 Forum, 1–29, (2019). https://doi.org/10.2514/6.2019-1816.

Shahmoradi, J., Talebi, E., Roghanchi, P., Hassanalian, M.: A comprehensive review of applications of drone technology in the mining industry. Drones **4**(3), 1–25 (2020). https://doi.org/10.3390/drones4030034

Sherstjuk, V., Zharikova, M., Sokol, I. Forest fire-fighting monitoring system based on UAV team and remote sensing. 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO). pp. 663–668, (2018). https://doi.org/10.1109/ELNANO.2018.8477527.

Shi, G. et al., Neural lander: Stable drone landing control using learned dynamics. Proc. - IEEE Int. Conf. Robot. Autom., vol. 2019-May, pp. 9784–9790, 2019, https://doi.org/10.1109/ICRA.2019.8794351.

Singh, R., Kumar, R., Mishra, A., Agarwal, A.: Structural analysis of quadcopter frame. Mater. Today Proc. **22**, 3320–3329 (2019). https://doi.org/10.1016/j.matpr.2020.03.295

Singhal, G., Bansod, B., Mathew, L. Unmanned aerial vehicle classification, applications and challenges: a review. Preprints 2018, 2018110601. https://doi.org/10.20944/preprints201811.0601.v1

Tang, X.M., De Chen, J., Li, T.: Unmanned aerial vehicle trajectory data fusion based on an active and passive feedback system. Proc. Inst. Mech. Eng. Part GJ. Aerosp. Eng. **234**(3), 887–895 (2020). https://doi.org/10.1177/0954410019890796

Thesis M., Akhtar, S. N.: The use of modern tools for modelling and simulation of UAV with Haptic. 2017.

Vélez-Nicolás, M., García-López, S., Barbero, L., Ruiz-Ortiz, V., Sánchez-Bellón, Á.: Applications of unmanned aerial systems (UASs) in hydrology: a review. Remote Sens. (2021). https://doi.org/10.3390/rs13071359

Velusamy, P., Rajendran, S., Mahendran, R.K., Naseer, S., Shafiq, M., Choi, J.G.: Unmanned aerial vehicles (Uav) in precision agriculture: applications and challenges. Energies (2022). https://doi.org/10.3390/en15010217

Wang, C.; Ma, Z. Design of wireless power transfer device for UAV. In Proceedings of the 2016 IEEE International Conference on Mechatronics and Automation, Harbin, China, 7–10 August 2016; pp. 2449–2454.

Winkelmaier, G., Battulwar, R., Khoshdeli, M., Valencia, J., Sattarvand, J., Parvin, B.: Topographically guided UAV for identifying tension cracks using image-based analytics in open-pit mines. IEEE Trans. Ind. Electron. **68**(6), 5415–5424 (2021). https://doi.org/10.1109/TIE.2020.2992011

Wu, X., Li, W., Hong, D., Tao, R., Du, Q.: Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey. IEEE Geosci. Remote Sens. Mag. **10**(1), 91–124 (2022). https://doi.org/10.1109/MGRS.2021.3115137

Yakovlev, K.S., Makarov, D.A., Baskin, E.S.: Automatic path planning for an unmanned drone with constrained flight dynamics. Sci. Tech. Inf. Process. **42**(5), 347–358 (2015). https://doi.org/10.3103/S0147688215050093

Yang, W., Wang, L., Song, B.: Dove: a biomimetic flapping-wing micro air vehicle. Int. J. Micro Air Veh. **10**(1), 70–84 (2018). https://doi.org/10.1177/1756829317734837

Zhang, K.F., Zhang, Z., Zhang, Y.H., Li, H.: Experimental study of single-rotor UAV on droplet deposition distribution in soybean field. Appl. Ecol. Environ. Res. (2019). https://doi.org/10.15666/aeer/1706_1383313844

Zhong, J., Wang, C.: Transition characteristics for a small tail-sitter unmanned aerial vehicle. Chin. J. Aeronaut. **34**(10), 220–236 (2021). https://doi.org/10.1016/j.cja.2020.12.021

Zhu, H., Nie, H., Zhang, L., Wei, X., Zhang, M.: Design and assessment of octocopter drones with improved aerodynamic efficiency and performance. Aerosp. Sci. Technol. **106**(2020), 106206 (2020). https://doi.org/10.1016/j.ast.2020.106206

Zuo, Z., Liu, C., Han, Q.L., Song, J.: Unmanned aerial vehicles: control methods and future challenges. IEEE/CAA J. Autom. Sinica **9**(4), 601–614 (2022)

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.

Mubina Shekh is pursuing her Ph.D. in the area of unmanned aerial vehicles (UAVs) in Delhi Technological University, Delhi, India. She is mostly interested in design, modelling, and simulation. She has a 12-year teaching tenure at esteemed engineering institutes.

Sushila Rani has her Ph.D. in Machine Design (Vibrations) from Delhi Technological University, Delhi, India. Presently, she is working as an assistant professor in the Mechanical, Production, Industrial, and Automobile Engineering Department at Delhi Technological University, Delhi, India. Her research interest is fault diagnosis in machinery.

Rituparna Datta is a distinguished data scientist and researcher with extensive experience in optimization, artificial intelligence, and machine learning. Currently, he serves as the AI/ML Engineering Manager at Accenture, India. He has also held the position of Computer Research Scientist and Adjunct Faculty in the Department of Computer Science at the University of South Alabama, USA, and has worked as an Operations Research Scientist with Boeing Research and Technology.

Journal of Electronic Materials
https://doi.org/10.1007/s11664-024-11426-w

28TH INTERNATIONAL CONFERENCE ON NUCLEAR TRACKS AND RADIATION MEASUREMENTS

rGO-Embedded Polymer Nanocomposite Layer for Improved Performance of Triboelectric Nanogenerator

Shilpa Rana[1] · **Bharti Singh**[1]

Received: 7 January 2024 / Accepted: 29 August 2024
© The Minerals, Metals & Materials Society 2024

Abstract
A triboelectric nanogenerator (TENG) working on a contact electrification and electrostatic induction principle is a promising energy source for fulfilling the energy demand of low power electronic devices by converting the ambient mechanical energy to useful electrical energy. Here, a polymer nanocomposite film-based triboelectric nanogenerator has been designed by embedding reduced graphene oxide (rGO) nanosheets in a polyvinylidene fluoride (PVDF) matrix as one of the friction layers. The PVDF nanocomposite film-based TENG was constructed and examined for structural, electrical, and surface properties with varied weight percentages of rGO nanofillers (0.0 wt%, 0.5 wt%, 1.0 wt%, 1.5 wt%, and 2.0 wt%). The experimental results demonstrate that the addition of rGO in a PVDF matrix considerably increased the output performance of the TENG device. The TENG device with 1.5 wt% of rGO can deliver the maximum output voltage and current of 95.9 V, and 16.8 μA, respectively, which are ~ 3 and ~ 7 times the voltage and current produced by pristine PVDF film-based TENG. The enhanced performance of the nanogenerator is attributed to the addition of conductive nanofillers in the polymer matrix which improves the surface charge density of polymer nanocomposite films by forming a conduction network, resulting in more effective charge transfer. Moreover, the output of the nanogenerator is stored in the capacitor and used to drive commercial LEDs, revealing the TENGs' potential applications for designing self-powered electronic devices.

Keywords Triboelectrification · Nanogenerator · Energy harvesting · rGO

Introduction

The global energy crisis with the technological advancement and increase in the human population has become one of the major problems that has motivated researchers all over the world to actively seek for innovative solutions to generate electricity from unconventional renewable energy sources. In light of this, various technologies have been developed and studies are still ongoing to harvest energy from sustainable energy sources, such as solar, thermal, vibration, wind, etc.[1–4] Among these technologies, triboelectric nanogenerators (TENGs) operating on the principle of triboelectrification and electrostatic induction have garnered considerable attention as an innovative technology which converts mechanical energy from diverse sources into electrical energy. Because of their working principle, TENGs possess several advantages over other energy-harvesting devices due to their outstanding performance in energy harvesting with strong output signals, cost-effectiveness, robustness, simple design, and abundant choice of materials.[5–7] Also, it is well known that the surface charge density of triboelectric materials has a significant impact on TENGs' output performance. Therefore, various methods have been adopted by researchers to improve the TENG performance, among which two major approaches are (1) to select a suitable friction material with markedly different polarities and structural optimization, as the performance of the device greatly depends on the structure and material selection, and (2) surface modification with micro-/nanostructure formation, doping composition, and charge injection to improve and regulate the friction ignition effect.[8–10] Many materials, organic, inorganic, polymer, etc., have been explored since the invention of the TENG in 2012, and their triboelectric series have been made in

✉ Bharti Singh
bhartisingh@dtu.ac.in

[1] Department of Applied Physics, Delhi Technological University, Main Bawana Road, Delhi 110042, India

Published online: 12 September 2024

which polymers currently dominate the triboelectric series due to their high charge transfer capacity.[11,12] PVDF is one of the most versatile polymer materials which serves as the negative triboelectric layer in TENGs due to its high electronegativity and flexibility, which keeps PVDF in the frontline of polymers to integrate with TENGs.[13–15] Up to now, several methods have been used to boost the performance of PVDF-based TENGs by electrospinning, microstructure formation, and porous structure formation which increase the effective contact surface area of the PVDF.[16] Another possible way to enhance the output performance of PVDF-based TENGs is the addition of optimized concentrations of nanofillers in the polymer matrix.[17–19] It has been observed that reduced graphene oxide (rGO) is preferred over other materials because of its electron-trapping capabilities, good conductivity, and higher electronegativity which help in efficiently capturing external electrons.

Hence, in this study, we propose a polymer composite film-based TENG by coupling of PVDF nanocomposite film with nylon film. rGO nanosheets with different weight percentages (0.5 wt%, 1.0 wt%, 1.5 wt%, 2.0 wt%) were incorporated into a PVDF polymer matrix and subjected to structural and morphological characterization to confirm successful doping of the nanofiller into the polymer matrix. The TENG's electrical output performance was optimized by subjecting it to tapping via a dynamic shaker at different force and frequency parameters. A maximum voltage and current of 95.9 V and 16.8 μA, respectively, was produced by a device with 1.5 wt% of rGO. Furthermore, the device's ability to harvest energy has been shown through capacitor charging and powering of LEDs. Thus, the present studies demonstrate an easy and efficient approach for boosting TENGs' performance by the addition of nanofillers, thus extending its potential application in self-powered and wearable devices.

2. Materials and Methods

First, graphene oxide (GO) was synthesized via a modified Hummers' method,[20] and rGO was synthesized using a hydrothermal method, as reported in our previous study.[21] Initially, an aqueous solution of GO (1.5 mg/mL) was subjected to ultrasonication to achieve a homogeneous dispersion of the GO. Subsequently, this prepared solution was transferred to an autoclave and heated to 180°C for 12 h. The resulting solution was then washed with distilled water and ethanol and kept in oven at 60°C overnight to obtain rGO powder. For the synthesis of the PVDF polymer nanocomposite films, various weight percentages of rGO (0.5 wt%, 1.0 wt%, 1.5 wt% and 2.0 wt%) were mixed in 10 ml of N,N-dimethylformamide (DMF) and stirred vigorously until the rGO completely dispersed in the solvent. Then, 1 g of the PVDF powder was slowly added to the prepared solution of rGO-dispersed DMF and stirred for 30 min to form a homogenous composite solution. Using a micropipette, the prepared solutions were drop-casted onto glass substrates and dried in an oven at 90ºC. After natural cooling of the oven, the films were immersed in distilled water to obtain free-standing composite films (Fig. 1). The same procedure was adopted for synthesis of the PVDF film without addition of rGO.

The nylon films were synthesized by phase-inversion processes reported by Soin et al.[22]

Furthermore, a TENG device was fabricated by assembling the PVDF and the PVDF/rGO nanocomposite with the nylon film in vertical contact separation mode. For this, the PVDF/rGO nanocomposite and nylon films were cut into 2 × 2 cm^2 and used as the two active triboelectric layers. Aluminum tape was applied to the back side of each film to act as electrodes, while copper wire was utilized to establish electrical connections.

Characterization

The crystallinity of the synthesized powder and thin films were analyzed using x-ray diffraction (XRD; DY 3733 Analytical Xpert pro) with the Cu Kα x-ray source of 1.54 Å wavelength. Fourier-transform infrared (FTIR; PerkinElmer spectrum-II) spectroscopy of the thin films was carried out to ascertain the amount of β-phase present in the polymer nanocomposite films and to verify the existence of various functional groups. Field-emission scanning electron microscopy (FESEM; Zeiss GeminiSEM 500) was used to analyze the surface morphology of the samples. Raman spectroscopy of the rGO was carried out using a WiTec alpha 300 RA spectrometer with a 514-nm wavelength laser. The dielectric measurements of the samples in the frequency range of 20 Hz- 10^4 Hz were performed using an LCR meter. The output performance of the TENGs were analyzed using an MDO34 Tektronix Oscilloscope by tapping the nanogenerator with the Micron MV-0025 electrodynamic shaker with an MPA-0200 function generator. A Keithley 6514 electrometer was used to carry out the short-circuit current and charge measurements. A force sensor was used to measure the force applied on the TENGs.

4. Results and Discussions

Several characterizations techniques have been used to confirm successful synthesis of the rGO. Figure 2a illustrates the synthesized rGO sample's XRD pattern, which is dominated by two broad peaks at 25.4° and 43.5°

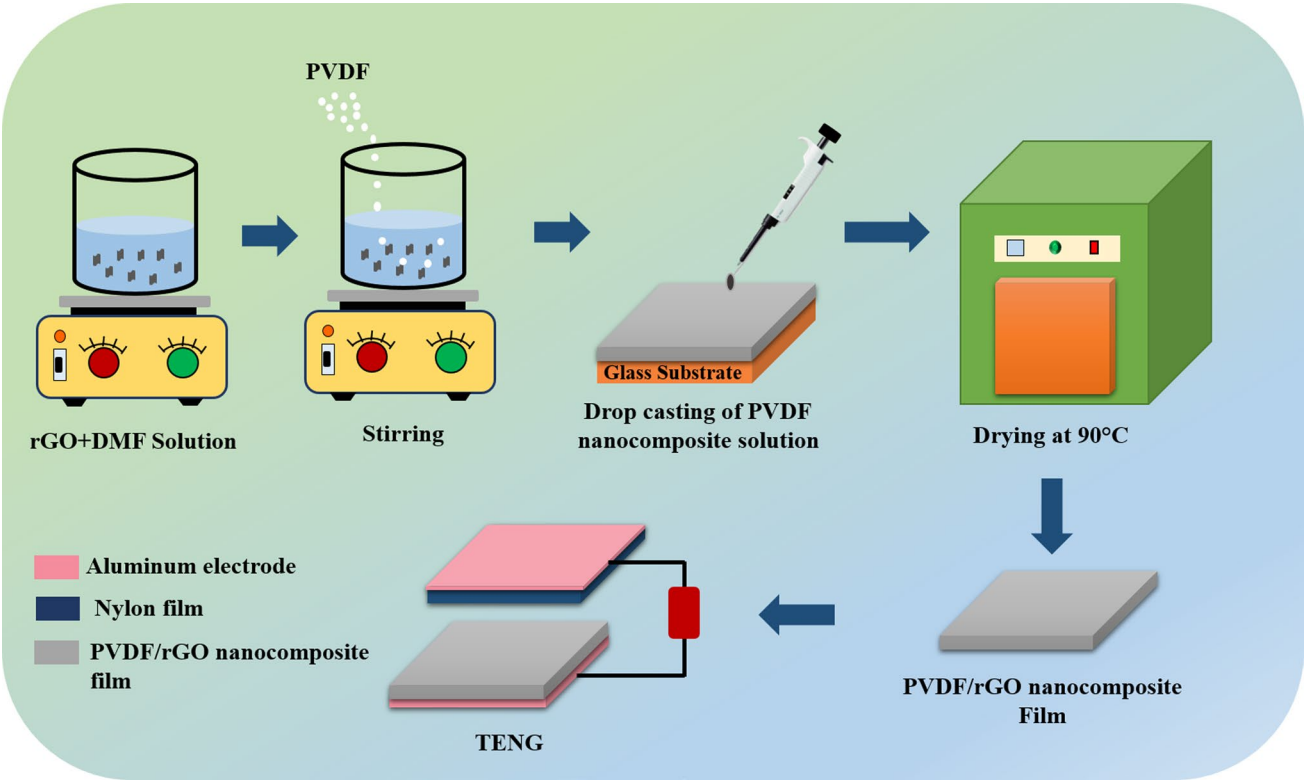


Fig. 1 Schematic showing the synthesis of PVDF/rGO nanocomposite film via drop-casting and fabrication of a TENG.

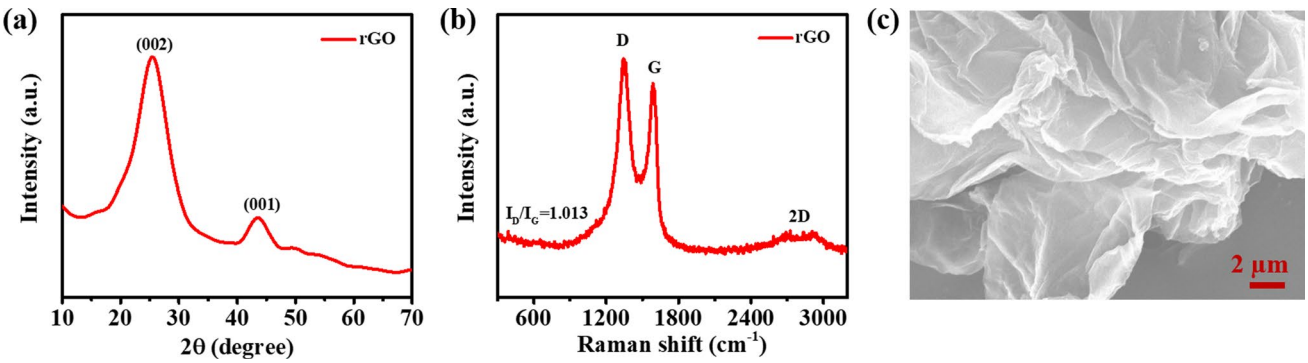


Fig. 2 (a) XRD pattern, (b) Raman spectra, and(c) FESEM micrograph of rGO.

corresponding to the (002) and (001) planes, respectively. The peak at 25.4° is attributed to the 2D structure of rGO, indicating that the sp^2 network structure of graphene is formed during the reduction process. The peak's position and width indicate that GO was successfully reduced to rGO, which is consistent with the rGO's defective and amorphous structure. Additionally, a less intense peak at 43.5° can be observed, and this peak is associated with the turbostratic band of disordered carbon material.[23,24]

Raman spectroscopy is an effective technique that has been used widely to identify the structure, disorders, defects, and the number of graphitic layers in carbon-based materials.[25] Raman spectra of rGO which consists mainly of two characteristic bands D and G band located at ~1347 cm^{-1}, and ~1590 cm^{-1} , respectively, are shown in Fig. 2b. The G band emerged from the in-plane vibrations of the sp^2 -bonded carbon atoms, whereas the D band also referred to as the defective band is linked to structural defects. Furthermore, the intensity ratio of the D and G bands, i.e., I_D/I_G, respectively, has been calculated to determine the degree of structural defects and disorder in the rGO, and has a value of 1.01 indicating that more defects are introduced

in the carbon structure and the effective reduction of GO to form rGO. In addition to two major peaks corresponding to D and G bands, a less intense broad 2D peak around 2700 cm^{-1} is also observed, indicating the formation of multilayered graphene material.

FESEM was utilized to investigate the surface morphology of the rGO, and the corresponding FESEM micrograph is shown in Fig. 2c, demonstrating that rGO exhibits a laminated structure, comprising multiple thin sheets with distinct edges and a wrinkled texture spanning across the entire surface. These thin and semi-transparent layers revealed that rGO nanosheets are composed of multiple layers rather than a single layer, as in case of graphene.

Furthermore, synthesized rGO nanosheets with different concentrations of 0.0 wt%, 0.5 wt%, 1.0 wt%, 1.5 wt%, and 2.0 wt% rGO were mixed with PVDF to form the polymer nanocomposite films and were labeled P, PR1, PR2, PR3, and PR4, respectively. The crystallinity of PVDF and its composite films with varied concentrations of rGO (0 wt%, 0.5 wt%, 1.0 wt%, 1.5 wt%, and 2.0 wt%) were analyzed by XRD (Fig. 3a). All the films exhibit two characteristic peaks at diffraction angles of $2\theta = 18.3°$ and 20.4°, respectively. The peak at 18.3° corresponds to the (020) reflection plane and is associated with the α phase, whereas the peak at 20.4° is attributed to the β phase of the (110/200) reflection plane.[25,26] As we increased the concentration of rGO, the peak intensity associated with the β phase rose, but fell after 1.5 wt%. Furthermore, the peak conforming to rGO was not observed in the XRD spectra, which may be due to a relatively lower diffraction intensity and concentration of rGO in the PVDF nanocomposite film. For further better understanding of the crystallographic phase in the polymer nanocomposite films, FTIR spectroscopy was further used to quantify the rise in the content of the β phase of PVDF after doping it with rGO. Figure 3b shows that several vibrational groups are present in the FTIR spectra of the films, while absorbance peaks corresponding to different phases (α, β, and γ phases) are indicated in the spectra.[27] Moreover, when we doped rGO in the polymer matrix, the intensity of the peak corresponding to non-polar α phase was diminished and the relative fraction of electroactive polar β phase was determined using:

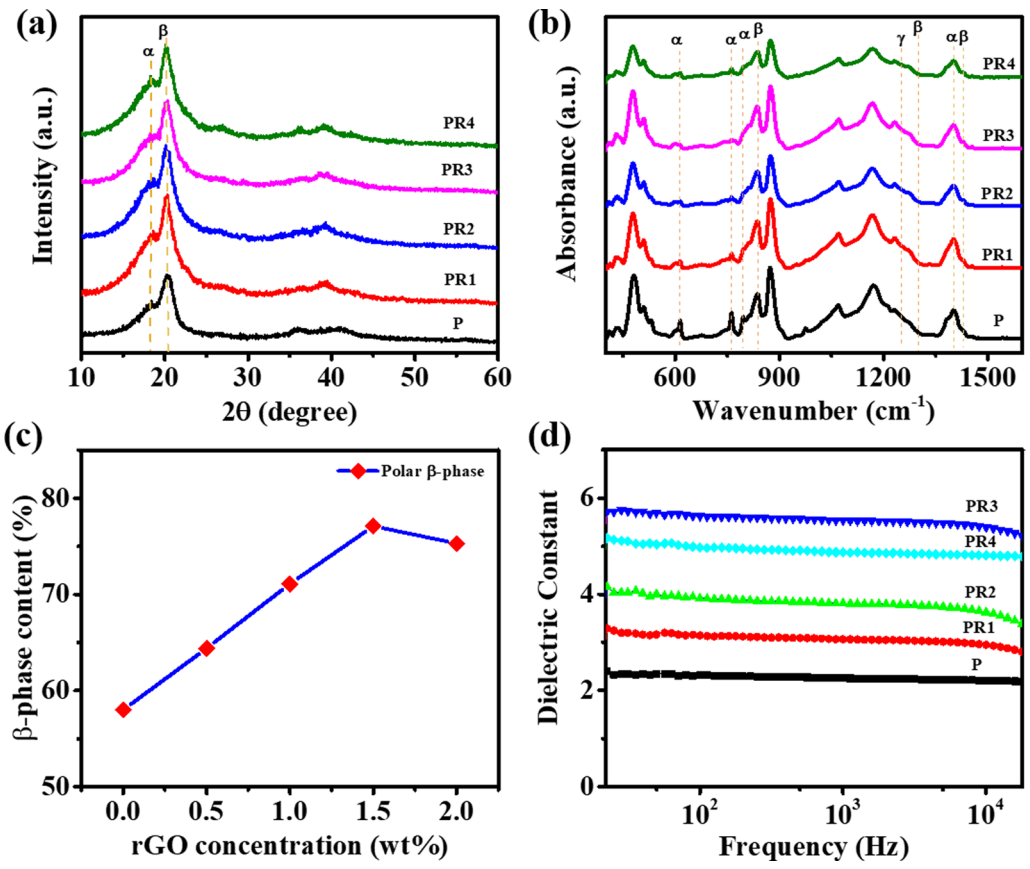


Fig. 3 (a) XRD patterns, (b) FTIR spectra, (c) content of electroactive polar β phase, and (d) dielectric constant of PVDF nanocomposite films with different weight percentages of rGO.

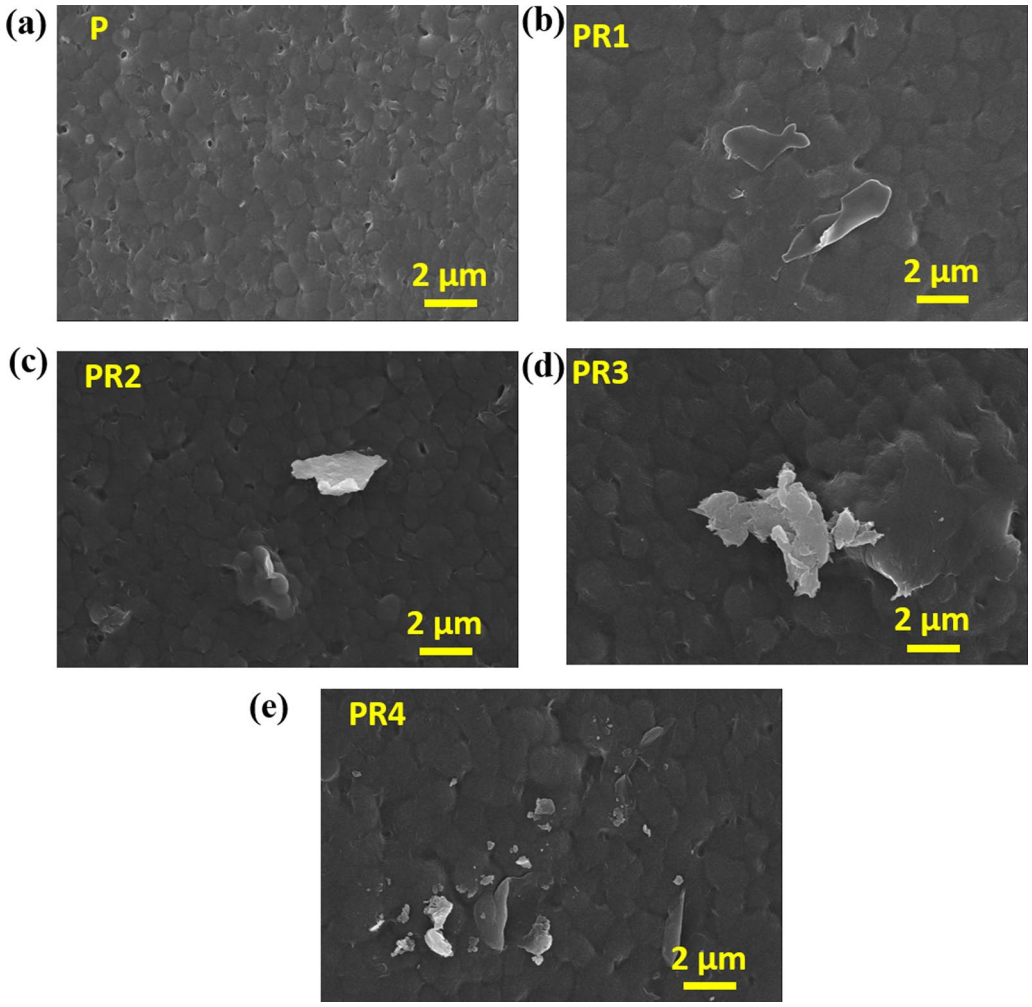


Fig. 4 FESEM micrographs of (a–e) PVDF and PVDF nanocomposite films with varied concentrations of rGO.

$$F(\beta) = \frac{A_\beta}{1.26A_\alpha + A_\beta} \times 100\% \quad (1)$$

where A_α and A_β are the absorption intensities at 762 and 840 cm^{-1}, respectively. The calculated values of the relative fractions of the electroactive polar phase $F(\beta)$ of P, PR1, PR2, PR3, and PR4 were 58%, 64.4%, 71.1%, 77.1%, and 75.3%, respectively, and the related results are illustrated in Fig. 3c. The percentage of the β phase content in the PVDF nanocomposite films was significantly enhanced with the increase in rGO concentration up to 1.5 wt%, while further addition of the nanofillers reduced its value. The possible reason behind the initial increase in the value of $F(\beta)$ up to a certain concentration and thereafter decreasing can be understood by the fact that, when we add rGO in the PVDF matrix, the crystallinity of the nanocomposites first increases due to the interaction between surface charges (i.e., oxygen-containing functional group and delocalized π electrons) of rGO with the CH_2 and CF_2 dipoles of PVDF which help in the gradual enhancement in the nucleation of the polar β phase by maximizing all trans molecular conformations.[21,28] However, after a certain concentration of rGO nanosheets (i.e., 1.5 wt% in our case), the alignment of the polymer chain became disturbed, resulting in suppressing the tendency of β phase formation arising because of a poor dipole arrangement in the PVDF chain.[26,29] Also, it has been reported in the literature that the dielectric constant of the triboelectric material is an important property which helps in improving the TENG device performance. Therefore, the dielectric constant of PVDF and its composite films were measured as a function of frequency with the help of the LCR meter (Fig. 3d). With increases in the rGO concentration, the dielectric constant of the PVDF film increases from 2.4 to 3.3, 4.1, 5.7, and 5.2 corresponding to the 0.5 wt%, 1.0 wt%, 1.5 wt%, and 2.0 wt%, respectively, of the rGO nanofillers. As expected, 1.5 wt% (i.e., PR3) has the maximum dielectric constant which is consistent with the XRD and FTIR studies. This increase

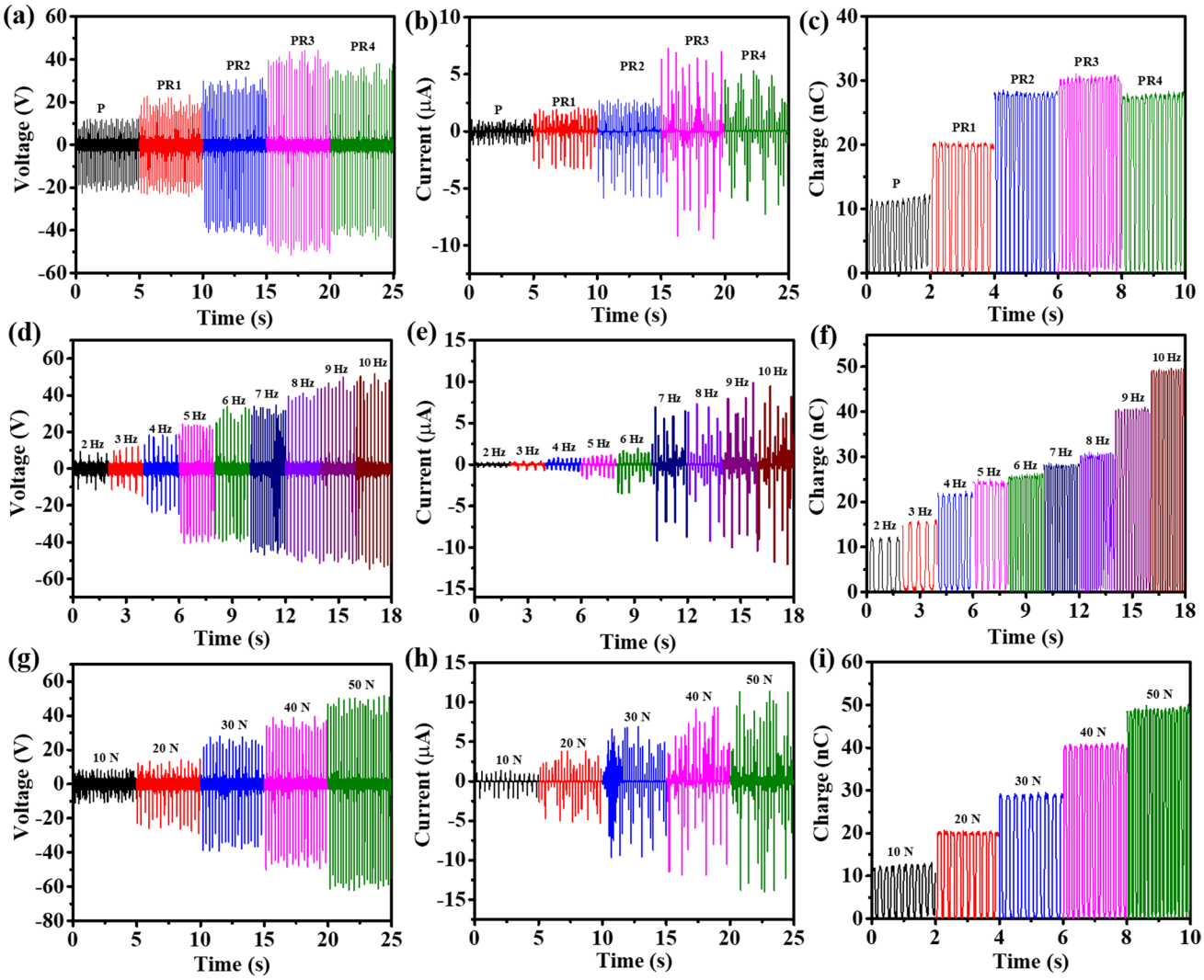


Fig. 5 Measured open-circuit voltage, short-circuit current, and charge waveforms of the TENG obtained by (a–c) varying the concentration of rGO in the PVDF, (d–f) varying the input frequency, and (g–I) the impact force on the TENG with 1.5 wt% of rGO.

Table I Comparison of PVDF/rGO-nylon-based TENG performance with various earlier reported TENGs.

Materials	Device area (cm^2)	V_{oc} (V)	I_{sc} (μA)	Power density ($\mu W/cm^2$)	Reference
PVC-rGO, aluminum	4.8 × 5	42.3	4.56	2.04	18
GO:PPy-PDMS, copper	–	27	0.359	–	32
CF-CNT, PTFE	2.5 × 2.5	60	1.8	11.06	33
SC-rGO, PTFE	3 × 3	81.5	–	255	34
PI/rGO, nylon	10 × 1.5	130	7.5	–	35
PDMS-rGO, water droplets	2 × 2	2	0.002	–	36
GO-AgNPs cellulose paper, PTFE	7.5 × 7.5	4	400	0.71	37
Ti@MoS_2/PP, nylon	–	29.3	–	–	38
ZnONRs@CB/NF, PTFE	–	28	4.5	80	39
PVDF/rGO, nylon	2 × 2	95.9	16.8	95	This work

in dielectric constant value after the addition of nanofillers is due to the formation of microscopic dipole and micro capacitor network formation which increases the capacitance of the nanocomposite film. However, the subsequent

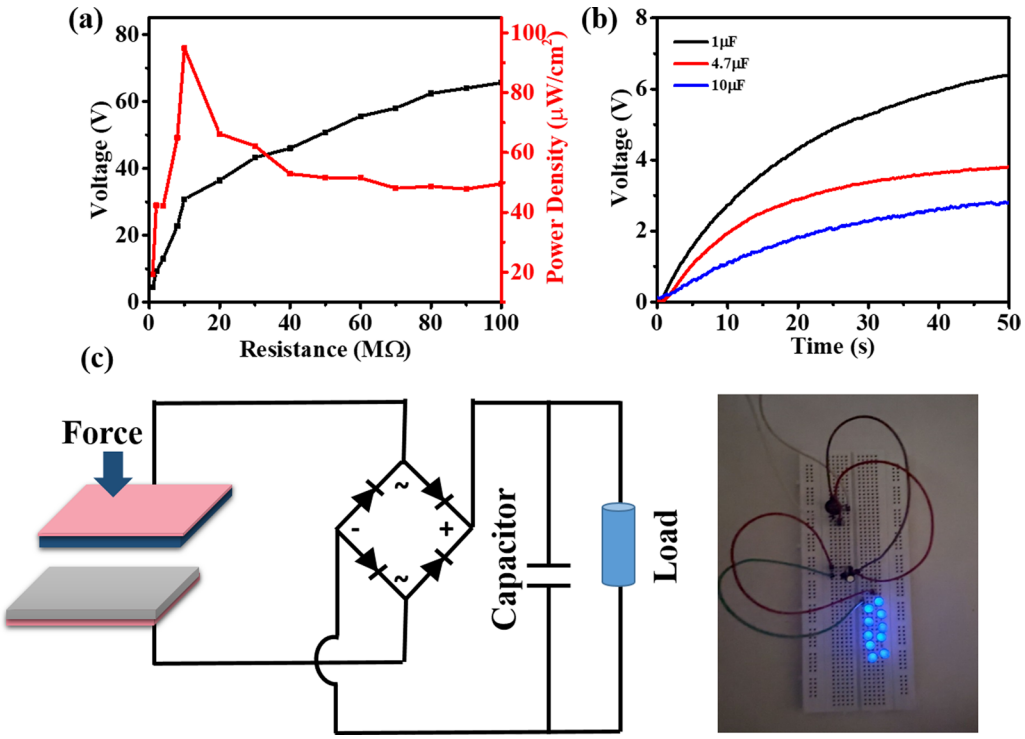


Fig. 6 (a) Output voltage and power density as a function of load resistance, (b) charging of capacitors by a nanocomposite film-based TENG with 1.5 wt% of rGO , and (c) images of LEDs powered by TENGs with a schematic circuit diagram.

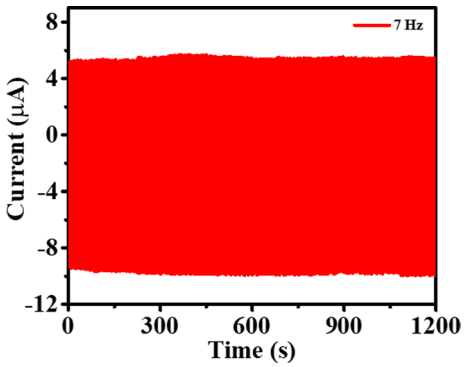


Fig. 7 The durability of the PVDF/rGO TENG with 1.5 wt% of rGO.

decrease in value from 1.5 wt% to 2.0 wt% occurs due to the establishment of a conductive pathway, leading to a reduction in the film's capacitance. Hence, 1.5 wt% of the nanofillers is the optimum amount for the generation and retention of a significant quantity of triboelectric charges on the dielectric layer. The surface morphologies of PVDF and its composite films with varied concentrations of rGO are illustrated in Fig. 4a–e. It can be clearly seen from the figures that the rGO nanosheets are uniformly dispersed within the PVDF matrix without any agglomeration.

Figure 1 shows a schematic of the as-fabricated TENG formed by the coupling of PVDF and PVDF-rGO nanocomposite films with nylon as two highly negative and positive triboelectric layers operating in a simple vertical contact separation mode. In order to assess the TENG's performance, a series of tests were carried out by tapping nanogenerators with the help of an electrodynamic shaker. Initially, the electrical outputs including open-circuit voltage (V_{OC}), short-circuit current (I_{sc}) and transferred charges were measured with different concentrations of rGO, as depicted in Fig. 5a–c. A TENG with 1.5 wt% loading amount of rGO (i.e., the PR3-based TENG) generates the highest V_{OC}, I_{SC}, and charges of 95.9 V, 16.8 μA, and 28.1 nC in contrast to a pristine PVDF-based TENG which can generate V_{OC}, I_{SC}, and charges of 35.4 V, 2.3 μA, and 11.6 nC, while PR1, PR2, and PR4 film-based nanogenerators can produce V_{OC} and I_{SC} of 48.1 V, 74.9 V, 81.9 V, 5.4 μA, 8.9 μA, and12.7 μA, respectively. The maximum output performance of 1.5 wt% of rGO is also supported by the FTIR and dielectric studies which show the maximum content of the β phase and dielectric constant corresponding to 1.5 wt% of rGO doping in the PVDF matrix. Moreover, the impact of the input excitation force and frequency on the TENG output performance with 1.5 wt% of rGO (PR3) was also investigated and the corresponding V_{OC}, I_{SC}, and charge waveforms are shown in Fig. 5d–I. The frequency of tapping

was also varied from 2 Hz to 10 Hz, with a highest value of voltage and current of ~ 107 V and 22 μA being attained at 10 Hz frequency. In the same vein, the impact of the applied force on the output capabilities of the TENG with 1.5 wt% rGO was explored through force variations ranging from 10 N to 50 N (Fig. 5g–I). At a forceful impact of 50 N, the TENG yielded a maximum voltage and current of around 113 V and 20 μA, respectively. It is clearly illustrated from Fig. 5d–I that both the frequency and force follow similar rising patterns for voltage, current, and charge. The enhanced open-circuit voltage and short-circuit current in relation to the operating frequency and force may be attributed to the following reasons. First, the faster charge transfer at higher tapping frequency, and second, when we increase the impact force, there is an increase in the elastic deformation and effective contact area in the polymer nanocomposite film causing a rise in TENG capacitance which may be helpful in storing more triboelectric charges, and as a result improved the TENG performance.[30,31] Table I shows the comparison between the effect of nanofillers on the output performance of the TENG. To investigate the ideal output condition of the as-fabricated TENG, the generated voltage was also measured as a function of load resistance ranging from 1 MΩ to 100 MΩ, as shown in Fig. 6a. The optimal TENG can deliver a maximum power density of ~95 μW/cm^2 across a matching load resistance of 10 MΩ. Furthermore, to confirm the practical energy-harvesting potential of the fabricated TENG, capacitors of 1.0 μF, 4.7 μF, and 10 μF were charged by the PR3 film-based TENG after rectification. The corresponding voltage charging curves of capacitors are illustrated in Fig. 6b, showing the voltage buildup over a brief period during their charging process. According to the results, the 1 μF capacitor demonstrated swift charging, reaching up to 6.4 V within 50 s. In contrast, the 4.7 μF and 10 μF capacitors only attained charges of 2.8 V and 3.8 V, respectively, within the same timeframe. Apart from charging the capacitior, the energy-harvesting performance of the TENG is illustrated via lighting of commercial LEDs and corresponding circuit diagram with an image of the LEDs shown in Fig. 6c. The video demonstration of the LEDs lighting with the help of the electrodynamic shaker is shown in Movie 1. The durability of the as-fabricated TENG is illustrated in Fig. 7, which shows excellent stable cyclic operation of TENG over 1200 cycles with negligible change in the value of the current. These results demonstrate the good durability of the TENG for long-term applications.

Conclusions

We successfully fabricated PVDF/rGO nanocomposite film-based TENGs and examined the impact of rGO doping on TENG performance by coupling of PVDF nanocomposite films with nylon film. The results showed that the addition of rGO in the PVDF matrix greatly improved the nucleation of crystalline β phase, dielectric, and charge transfer properties of PVDF film, resulting in the improved performance of the TENG device. A maximum voltage and current of 95.9 V, and 16.8 μA, respectively, was obtained by the PR3 film-based TENG (i.e., 1.5 wt% of rGO), whereas a pristine PVDF film-based TENG could only generate 35.4 V voltage and 2.4 μA current. Also, this device produced the highest power density of ~ 95 μW/cm^2 when connected to a matching load resistance. The generated power was stored within the capacitor and employed to illuminate commercial LEDs, showing the potential application of TENGs in future low-power portable and wearable electronic devices.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11664-024-11426-w.

Acknowledgments The authors express their gratitude to CSIR (Award No. 08/133(042)/2019-EMR-I) for financial support that enabled the execution of this research endeavor.

Conflict of interest Authors asserts that they have no conflict of interest.

References

1. Z. Kınas, A. Karabiber, A. Yar, A. Ozen, F. Ozel, M. Ersöz, and A. Okbaz, High-performance triboelectric nanogenerator based on carbon nanomaterials functionalized polyacrylonitrile nanofibers. *Energy* 239, 122369 (2022). https://doi.org/10.1016/j.energy.2021.122369.
2. S. Sripadmanabhan Indira, C. Aravind Vaithilingam, K.S.P. Oruganti, F. Mohd, and S.J.N. Rahman, Nanogenerators as a sustainable power source: state of art, applications, and challenges. *Nanomaterials* 9(5), 773 (2019). https://doi.org/10.3390/nano9050773.
3. C. Wu, A.C. Wang, W. Ding, H. Guo, and Z.L. Wang, Triboelectric nanogenerator: a foundation of the energy for the new era. *Adv. Energy Mater.* 9(1), 1802906 (2019). https://doi.org/10.1002/aenm.201802906.
4. H. Li, C.S.L. Koh, Y.H. Lee, Y. Zhang, G.C. Phan-Quang, C. Zhu, Z. Liu, Z. Chen, H.Y.F. Sim and C.L.J.N.E. Lay, A wearable solar-thermal-pyroelectric harvester: achieving high power output using modified rGO-PEI and polarized PVDF. *Nano Energy* 73, 104723 (2020). https://doi.org/10.1016/j.nanoen.2020.104723
5. C. Wu, T.W. Kim, and H.Y. Choi, Reduced graphene-oxide acting as electron-trapping sites in the friction layer for giant triboelectric enhancement. *Nano Energy* 32, 542 (2017). https://doi.org/10.1016/j.nanoen.2016.12.035.
6. D. Zhang, Z. Xu, Z. Yang, and X. Song, High-performance flexible self-powered tin disulfide nanoflowers/reduced graphene oxide nanohybrid-based humidity sensor driven by triboelectric nanogenerator. *Nano Energy* 67, 104251 (2020). https://doi.org/10.1016/j.nanoen.2019.104251.

7. K. Cheng, S. Wallaert, H. Ardebili, and A. Karim, Advanced triboelectric nanogenerators based on low-dimension carbon materials: a review. *Carbon* 194, 81 (2022). https://doi.org/10.1016/j.carbon.2022.03.037.
8. R. Mao, D. Zhang, Z. Wang, H. Zhang, D. Wang, M. Tang, L. Zhou, H. Cai, and H. Xia, Deep-learning-assisted low-cost flexible cotton yarn-based triboelectric nanogenerator for ultra-sensitive human-computer merging interfaces. *Nano Energy* 111, 108418 (2023). https://doi.org/10.1016/j.nanoen.2023.108418.
9. S.A. Lone, K.C. Lim, K. Kaswan, S. Chatterjee, K.-P. Fan, D. Choi, S. Lee, H. Zhang, J. Cheng, and Z.-H. Lin, Recent advancements for improving the performance of triboelectric nanogenerator devices. *Nano Energy* 99, 107318 (2022). https://doi.org/10.1016/j.nanoen.2022.107318.
10. Y. Zou, J. Xu, K. Chen, and J. Chen, Advances in nanostructures for high-performance triboelectric nanogenerators. *Adv. Mater. Technol.* 6(3), 2000916 (2021). https://doi.org/10.1002/admt.202000916.
11. H. Zou, Y. Zhang, L. Guo, P. Wang, X. He, G. Dai, H. Zheng, C. Chen, A.C. Wang, and C. Xu, Quantifying the triboelectric series. *Nat. Commun.* 10(1), 1427 (2019). https://doi.org/10.1038/s41467-019-09461-x.
12. G. Khandelwal, N.P. Maria Joseph Raj, and S.J. Kim, Materials beyond conventional triboelectric series for fabrication and applications of triboelectric nanogenerators. *Adv. Energy Mater.* 11(33), 2101170 (2021). https://doi.org/10.1002/aenm.202101170.
13. P. Pandey, D.-H. Jung, G.-J. Choi, M.-K. Seo, S. Lee, J.M. Kim, I.-K. Park, and J.I. Sohn, Nafion-mediated barium titanate-polymer composite nanofibers-based triboelectric nanogenerator for self-powered smart street and home control system. *Nano Energy* 107, 108134 (2023). https://doi.org/10.1016/j.nanoen.2022.108134.
14. M. Tayyab, J. Wang, J. Wang, M. Maksutoglu, H. Yu, G. Sun, F. Yildiz, M. Eginligil, and W. Huang, Enhanced output in polyvinylidene fluoride nanofibers based triboelectric nanogenerator by using printer ink as nano-fillers. *Nano Energy* 77, 105178 (2020). https://doi.org/10.1016/j.nanoen.2020.105178.
15. H.H. Singh and N. Khare, Flexible ZnO-PVDF/PTFE based piezo-tribo hybrid nanogenerator. *Nano Energy* 51, 216 (2018). https://doi.org/10.1016/j.nanoen.2018.06.055.
16. Y. Wang, Y. Yang, and Z.L. Wang, Triboelectric nanogenerators as flexible power sources. *npj Flex. Electron.* 1(1), 10 (2017). https://doi.org/10.1038/s41528-017-0007-8.
17. S. Rana, V. Singh, and B. Singh, Recent trends in 2D materials and their polymer composites for effectively harnessing mechanical energy. *iScience* (2022). https://doi.org/10.1016/j.isci.2022.103748.
18. R.F.S. Muktar Ahmed, S.B. Mohan, S.M. Ankanathappa, M. Shivanna, S.A. Basith, M.H. Chandrashekara Shastry, A. Chandrasekhar, and K. Sannathammegowda, Sucrose assisted chemical-free synthesis of rGO for triboelectric nanogenerator: Green energy source for smart-water dispenser. *Nano Energy* 106, 108085 (2023). https://doi.org/10.1016/j.nanoen.2022.108085.
19. G.-Z. Li, G.-G. Wang, Y.-W. Cai, N. Sun, F. Li, H.-L. Zhou, H.-X. Zhao, X.-N. Zhang, J.-C. Han, and Y. Yang, A high-performance transparent and flexible triboelectric nanogenerator based on hydrophobic composite films. *Nano Energy* 75, 104918 (2020). https://doi.org/10.1016/j.nanoen.2020.104918.
20. D.C. Marcano, D.V. Kosynkin, J.M. Berlin, A. Sinitskii, Z. Sun, A. Slesarev, L.B. Alemany, W. Lu, and J.M. Tour, Improved synthesis of graphene oxide. *ACS Nano* 4(8), 4806 (2010). https://doi.org/10.1021/nn1006368.
21. S. Rana, V. Singh, and B. Singh, Tailoring the output performance of PVDF-based piezo-tribo hybridized nanogenerators via B, N-codoped reduced graphene oxide. *ACS Appl. Electron. Mater.* 4(12), 5893 (2022). https://doi.org/10.1021/acsaelm.2c01085.
22. N. Soin, P. Zhao, K. Prashanthi, J. Chen, P. Ding, E. Zhou, T. Shah, S.C. Ray, C. Tsonos, T. Thundat, E. Siores, and J. Luo, High performance triboelectric nanogenerators based on phase-inversion piezoelectric membranes of poly(vinylidene fluoride)-zinc stannate (PVDF-ZnSnO3) and polyamide-6 (PA6). *Nano Energy* 30, 470 (2016). https://doi.org/10.1016/j.nanoen.2016.10.040.
23. N. Díez, A. Śliwak, S. Gryglewicz, B. Grzyb, and G. Gryglewicz, Enhanced reduction of graphene oxide by high-pressure hydrothermal treatment. *RSC Adv.* 5(100), 81831 (2015). https://doi.org/10.1039/C5RA14461B.
24. S. Shukla, I. Khan, V.K. Bajpai, H. Lee, T. Kim, A. Upadhyay, Y.S. Huh, Y.-K. Han, and K.M. Tripathi, Sustainable graphene aerogel as an ecofriendly cell growth promoter and highly efficient adsorbent for histamine from red wine. *ACS Appl. Mater. Interfaces* 11(20), 18165 (2019). https://doi.org/10.1021/acsami.9b02857.
25. A. Anand, D. Meena, K.K. Dey, and M.C. Bhatnagar, Enhanced piezoelectricity properties of reduced graphene oxide (RGO) loaded polyvinylidene fluoride (PVDF) nanocomposite films for nanogenerator application. *J. Polym. Res.* 27(12), 1 (2020). https://doi.org/10.1007/s10965-020-02323-x.
26. S.K. Karan, D. Mandal, and B.B. Khatua, Self-powered flexible Fe-doped RGO/PVDF nanocomposite: an excellent material for a piezoelectric energy harvester. *Nanoscale* 7(24), 10655 (2015). https://doi.org/10.1039/C5NR02067K.
27. X. Cai, T. Lei, D. Sun, and L. Lin, A critical analysis of the α, β and γ phases in poly (vinylidene fluoride) using FTIR. *RSC Adv.* 7(25), 15382 (2017). https://doi.org/10.1039/C7RA01267E.
28. S. Rana, and B. Singh, Polymer nanocomposite film based piezoelectric nanogenerator for biomechanical energy harvesting and motion monitoring. *J. Mater. Sci. Mater. Electron.* 34(25), 1764 (2023). https://doi.org/10.1007/s10854-023-11207-x.
29. R. Bhunia, S. Gupta, B. Fatma, Prateek, R.K. Gupta and A. Garg, Milli-Watt Power harvesting from dual triboelectric and piezoelectric effects of multifunctional green and robust reduced graphene oxide/P(VDF-TrFE) composite flexible films. *ACS Appl. Mater. Interfaces* 11(41), 38177 (2019). https://doi.org/10.1021/acsami.9b13360
30. L. Shi, H. Jin, S. Dong, S. Huang, H. Kuang, H. Xu, J. Chen, W. Xuan, S. Zhang, and S. Li, High-performance triboelectric nanogenerator based on electrospun PVDF-graphene nanosheet composite nanofibers for energy harvesting. *Nano Energy* 80, 105599 (2021). https://doi.org/10.1016/j.nanoen.2020.105599.
31. T. Bhatta, P. Maharjan, H. Cho, C. Park, S.H. Yoon, S. Sharma, M. Salauddin, M.T. Rahman, S.M.S. Rana, and J.Y. Park, High-performance triboelectric nanogenerator based on MXene functionalized polyvinylidene fluoride composite nanofibers. *Nano Energy* 81, 105670 (2021). https://doi.org/10.1016/j.nanoen.2020.105670.
32. E.J. Jelmy, D. Jose, K.V. Vijoy, K.J. Saji, and H. John, Enhanced triboelectric performance of graphene oxide-conducting polymer hybrid modified polydimethylsiloxane composites. *Mater. Adv.* 3(17), 6897 (2022). https://doi.org/10.1039/D2MA00771A.
33. G.M. Rani, C.-M. Wu, K.G. Motora, R. Umapathi, and C.R.M. Jose, Acoustic-electric conversion and triboelectric properties of nature-driven CF-CNT based triboelectric nanogenerator for mechanical and sound energy harvesting. *Nano Energy* 108, 108211 (2023). https://doi.org/10.1016/j.nanoen.2023.108211.
34. H.J. Hwang, J.S. Yeon, Y. Jung, H.S. Park, and D. Choi, Extremely foldable and highly porous reduced graphene oxide films for shape-adaptive triboelectric nanogenerators. *Small* 17(9), 1903089 (2021). https://doi.org/10.1002/smll.201903089.

35. X. Zhao, B. Chen, G. Wei, J.M. Wu, W. Han, and Y. Yang, Polyimide/graphene nanocomposite foam-based wind-driven triboelectric nanogenerator for self-powered pressure sensor. *Adv. Mater. Technol.* 4(5), 1800723 (2019). https://doi.org/10.1002/admt.201800723.
36. A.N. Parvez, M.H. Rahaman, H.C. Kim, and K.K. Ahn, Optimization of triboelectric energy harvesting from falling water droplet onto wrinkled polydimethylsiloxane-reduced graphene oxide nanocomposite surface. *Compos. B Eng.* 174, 106923 (2019). https://doi.org/10.1016/j.compositesb.2019.106923.
37. A.A. Pillai, J. John, R.M. Mathew, J. Jose, E.S. Zacharia, R. Abraham, N.M. Johnson, R.L. Thomas, and V. Thomas, Enhanced output performance of graphene oxide based triboelectric nanogenerators via plasmon coupling. *Mater. Sci. Eng. B* 296, 116637 (2023). https://doi.org/10.1016/j.mseb.2023.116637.
38. S. Veeralingam and S. Badhulika, Ti@MoS2 incorporated Polypropylene/Nylon fabric-based porous, breathable triboelectric nanogenerator as respiration sensor and ammonia gas sensor applications. *Sens. Actuators B Chem.* 380, 133346 (2023). https://doi.org/10.1016/j.snb.2023.133346.
39. N. Jayababu, and D. Kim, ZnO nanorods@conductive carbon black nanocomposite based flexible integrated system for energy conversion and storage through triboelectric nanogenerator and supercapacitor. *Nano Energy* 82, 105726 (2021). https://doi.org/10.1016/j.nanoen.2020.105726.

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 Asia Pacific Conference on Innovation in Technology (APCIT)
Mysuru, India. Jul 26-27, 2024

Sentiment Classification on Suicide Notes Using GPT, Bi-LSTM and CNN

Akash Bansal
Department of Computer Science & Engineering
Delhi Technological University
New Delhi, India
akashbansalhry@gmail.com

Rohit Beniwal
Department of Computer Science & Engineering
Delhi Technological University
New Delhi, Inida
dr.rohitbeniwal@gmail.com

2024 Asia Pacific Conference on Innovation in Technology (APCIT) | 979-8-3503-6153-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/APCIT62007.2024.10673647

***Abstract*—In the modern day, where sentiments and feelings are important in every element of our daily lives, from personal to professional, sentiment classification has become a crucial field of research. This entails learning what people think about different products and how they feel about a specific topic. Relevant research has demonstrated that several models and techniques have been applied to the sentiment classification of suicide notes. Still, the state-of-the-art in this field was restricted to lower accuracy datasets. As a result, for the purpose of this study, we taken dataset for sentiment classification tasks on suicide notes utilizing information from Twitter, Reddit, blogs , and various Infomedia websites.Sentences are divided into six categories based on suicidality, including Proud, Love, Neutral, Sad, Hate and Happy.According to our results, the suggested model has achieved a 77% F1 score, a 78% recall, and a 77% precision.**

***Index Terms*—GPT2, Bi-LSTM, CNN, Classification, Sentiment, Sentiment Classification, Suicide notes**

I. Introduction

Sentiment analysis is a term used to define the emotions carried by particular sentences [1]. In brief, we can say that we can find the emotions present in the sentences. Opinion mining is the term used for this purpose. A sentence can be majorly three major types of emotions positive, negative, and neutral then we further divide it into categories positive we have happiness, love, etc, negative we have anger, and hate. This field of study comes into a trend in the 1960s. When in the 2000s social media started gaining popularity again it trends because on social media one is expressing feelings and thoughts in the form of text. And then machine learning comes into the field. We have huge data on social media we can analyse this by using social media. We know that data is the fuel of today's world, and sentiment analysis is an important feature of this data. We can determine on a particular topic what the views of the people or what they feel as emotions about a particular news or any incident. There is huge work in this field on different datasets, peoples use many machine learning algorithms and deep learning algorithms for sentiment analysis [2]. Accuracy also matters for the analysis. Hence we have different accuracies on using different algorithms because there are advantages and disadvantages to different algorithms. In today's scenario transfer-based learning models are on trends for sentiment analysis. So we use a combined model of transfer-based learning, deep learning, and convolutional neural networks. From the serval areas, we use the suicide notes for sentiment analysis purposes. Suicide notes are pieces of a letter written by an individual before he commits suicide, hence this is a very important topic. By analyzing these data we can understand the suicidal minds, and it is important to analyze because it is a very huge problem in our society, especially for particular categories like the students who are preparing for competitive exams or college students who committed suicide under pressure of studies. So by analyzing suicide notes sentiment we can understand the suicidal minds and help them further and prevent the suicide. By understanding their emotions, after identifying the suicidal minds we can help them by escaping these difficult situations. The literature review will represent the method used in this field of study and till now how much accuracy we have achieved by using different models. By analyzing this model we decide the combination of models we will use on our dataset. We used a dataset of suicide notes having six types of categories those are hate, love, proud, neutral, sad, and happy. We manually label this dataset one by one in these mentioned categories. The model we trained is a combination of GPT2, Bi-LSTM, and CNN. We have 7 layers used in our model these layers are tokenization using GPT2, then we have an embedding layer, Bi-LSTM, a 1D convolutional layer, a 1D max pool layer, a flattened layer, a dropout layer, an output layer with softmax function. Our results are better than the previous results in this field. we achieved a precision value is 77%, recall value is recall value is 78% and an f1-score is also 77%. Further, our paper is divided into different sections, The next section is the Literature review in this section we describe the previous work that happened in this field, the next section is the methodology and we have a further subsection on methodology. Data collection shows how we collected and have some random samples in this section and a graphical representation of the number of sentences in each category. Data preprocessing explains the tokenization and conversion of datasets in vectors. Then in the next section, we describe each layer and their working.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:27 UTC from IEEE Xplore. Restrictions apply.

II. Literature Review

A hybrid system is used by Wang, Chen and others including a machine learning classifier and a rule-based rule-based classifier, was used. The result is shown as F-measure 0.5038 [3]. Desmet, B and other used a standard SVM model was used on the ib2b challenge dataset and having a set of 15 emotions, and the result was presented as F-score is above 50 bootstrap with 40 annotations [4].Beniwal and Dobhal used the Bi-LSTM model on a dataset of suicide notes with multiple layers of Bi-LSTM, the F1 score is 0.72 [5].In a study by Ghosh, S, and others Classification of suicide notes by professionals is 63%, and by machine learning algorithms is 78% [6]. A private attentive model is used with the multi-network cascades and the MR score is 60.25% and overall F1-score is 87.5% by single task system [7].Burnap, P, and others stated Analysis of suicide-related conversations on Twitter by a combined model with a Rotation Forest algorithm and maximum probability voting classification, which has seven labels in the dataset, and the F-measure is 0.728 [8].Sarsam, Alzahrani and others used YATSI classifier is used to analyze suicide-related tweets having emotions of fear, sadness, and negative sentiments [9].Cherry,Mohammad and others used Two different large-margin models; they can handle all the labels effectively, and the F-measure is 55.22% [10]. The result is from 1200 papers over a period of eight years using different accuracies, and the range by quality work is 71.5% to 88.2% on MIMIC and ib2b datasets [11]. Analysis of Punjabi text by using a deep neural network model obtained after 10-fold cross-validations is 83.3% [12]. A 2-layered attention mechanism is used on suicide-related lexicons and tree hole contents, and the accuracy was 91% [13]. Suicide-related data is collected from Kazakhstan, and NLP and Machine learning techniques are used to us analyse this collected data related to suicide [14]. Supervised Learning is used to detect suicdial ideation in the text post by user on the social media platform [15].Desmet, B and others purposed a genetic algorithm is used to detect the suicidality performed on Dutch language text extracted from different social media platforms [16]. RNN was used to prevent suicide, detection is important, and the accuracy achieved for the detection process using RNN is 74% [17].Long, S, and others approaches RNN, CNN, and attention-based models are used on two datasets; 28 different combinations are provided with seven inputs, and five pre-trained language models are used [18]. Random forest, SVM, and logistic regression are used on the datasets collected from Reddit, having 10000 random posts and detection. The accuracy achieved is 50%, and the F1 score is 66% [19].Wang, R and others performs survey on novel medical-level research achieved an accuracy of 56% in the prediction of 10 labels [20].

III. Methodology

This section contains a few subsections, which include the data collection phase, data preprocessing, an explanation of what was used for the analysis of the dataset, and the result section.

A. Data Collection Phase

Our dataset contains sentences extracted from the suicide notes. Around 250 suicide notes further slpit and manually labeled into six categories thos are love, hate, sad, neutral, happy, and proud. After splitting, we have 2600 labeled samples. TABLE I shows some random samples from our dataset that have two columns of text and labels. A bar graph in Fig. 1. presents a number of samples in each category.

TABLE I
Some random samples from the dataset along with their labels

Text	Labels
I love you	love
As for my family, I am sorry I let you down	sad
Take me home Jesus, take me home Lord, take me home Lord	proud
Dell, you take care of Gaylene always	love
Be strong in the Lord. I love you sister	love
I didn't know what happened	neutral
You know how we grew up in the same house	neutral
You have to live and die by the choices that we make	neutral

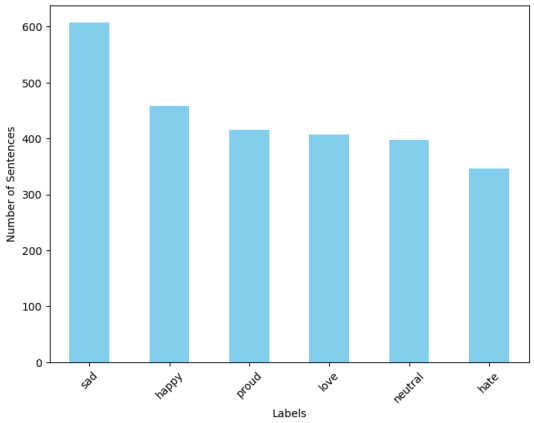


Fig. 1. Number of samples for each class of dataset

B. Data Preprocessing Phase

Now, our dataset is ready, we need to convert this dataset into tokens using GPT2, and then we split the dataset into two parts, one for training purposes and the other for testing purposes at 80% and 20%, respectively. Then, we do padding on the vectors generated and fill the space with zeroes till a particular length. Now we use encoding and then convert our padded dataset in the form of vector for passing it in our model and making it model friendly.

C. Model

We have used a combination of models having seven layers in the model and use this model for training and testing on our dataset for classification purposes. firstly GPT a transform learning model is used for tokenization and converts that dataset into to vector and then uses this as input for the next

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:27 UTC from IEEE Xplore. Restrictions apply.

layer. We have an embedding layer, which we use to convert the input vector into the dense vectors and convert the input layer into the output layer with some specific dimension, the value of dimension for the input layer is 60, and then in the output layer, this is 128.

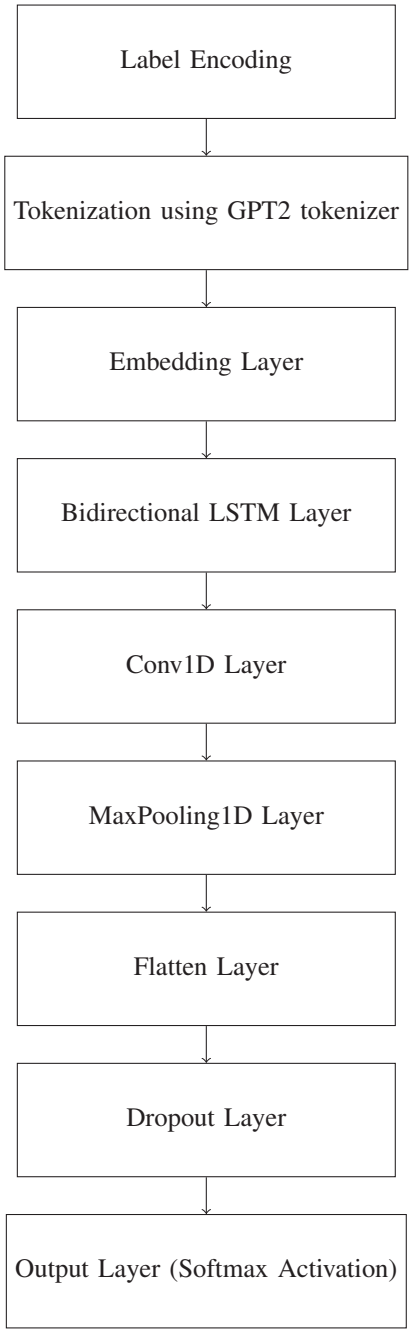


Fig. 2. Model Architecture

In the next layer, we have Bi-LSTM which stands for Bidirectional Long-Short Term Memory Network. It is a type of RNN which we used in our model.

It is different from standard RNNs because it works in two directions forward and backward at a time. It includes summation operation which works in both directions and the output is combined of both the RNNs. In the forward pass, it works from left to right, and in the backward pass it works from right to left, it takes the input and also the previous input as a hidden state, and in the output provides a hidden state for the current input and cell state. Then it takes the output of forward and backward passes and combines them and this combined value is the output of the Bi-LSTM model. There is one characteristic of the output by Bi-LSTM is that the length of the output is the same as the input and it has information about both past and future contexts at a time because of bidirectional working. It also efficiently handles long-term dependencies of our dataset because it can easily reduce the problem because of vanishing gradients. Also the reason for choosing this for our dataset is it effectively works on sequence labeling. The next layer we have is a standard convolutional layer. We have many layers of CNN used in our model these are, the 1D max pool layer for dimensionality reduction, the flattened layer to convert the pooled input into a 1D array, the Dropout layer for avoiding overfitting in our model and the final layer is the output layer. Convolutional layers are designed in a way that these are learnable layers that can easily identify local patterns and this is the purpose we use this in our. Each filter is efficient and uses its advantages to learn different patterns. There are some spatial dimensions produced, which can be handled by the pooling layer including max pooling and average pooling. In the dropout layer convert some input vectors to zero to avoid any kind of overfitting because it can be a huge disadvantage for our model that is why we used it. The final layer of our model is the dense output layer and we use the softmax function to calculate probabilities.

IV. Result and Analysis

In this implementation, we used many Python-based libraries for data handling and model building. Pandas and Numpy we use for calculations and handling arrays. For the label encoder and train test split we used Sklearn. From transformers, we import GPT2 as a part of our model. Then we used Keras from TensorFlow, this play a very important role in our implementation because we have done major work by using this. For padding sequences in the dataset, we use the Keras preprocessing sequence, we use Keras's layers for the import input layer, dense layer, dropout layer, Bi-LSTM, conv1D, maxpool1D, embeddings, and flatten layer, adam from Keras optimizers, sklearn for confusion matrix and classification report, and matplotlib for graphical representations.

we have various parameters, hyperparameters, and their values as shown in TABLE II and TABLE III. These parameters are related to the layers of our model. The first layer we have is embeddings and for that, we have three parameters input_dimension, output_dimension, and input_length. For input_dimension the parameter value is several tokens in GPT2 vocabulary, for output_dimension value is 128, and input_length is 60. the second layer name is Bidirectional LSTM and the parameter name is units and under_sequences, the values for these parameters are '64' and 'true'. The third

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:27 UTC from IEEE Xplore. Restrictions apply.

TABLE II
LAYER-WISE PARAMETERS AND THEIR VALUES

Layer Name	Parameter Name	Parameter Value
Embedding	input_dim output_dim input_length	# of Tokens in GPT2 Vocabulary 128 60
Bidirectional LSTM	units return_sequences	64 True
Conv1D	filters kernel_size activation	64 3 relu
MaxPooling1D	pool_size	2
Flatten	-	-
Dropout	rate	0.5
Dense	units activation	num_classes softmax

TABLE III
HYPERPARAMETERS

Hyperparameter Name	Value
Pre-trained model	'gpt2'
Input length	60
Embedding output dimension	128
Bi-LSTM units	64
Conv1D filters	64
Conv1D kernel size	3
Conv1D activation	relu
MaxPooling1D pool size	2
Dropout rate	0.5
Learning rate	0.001

layer is Conv1D and there are three parameters for this layer filter, kernel_size, and activation and the values are 64, 3, and relu respectively. The fourth layer is Maxpooling1D and the parameter name is pool_size and the value is 2. The next layer is flattened and for this layer, there is no parameter. the next layer is Dropout and the parameter name is rate and the value for the parameter is 0.5. The next and last layer is the dense layer and the parameters are units and activation and their respective values are num_classes and softmax.

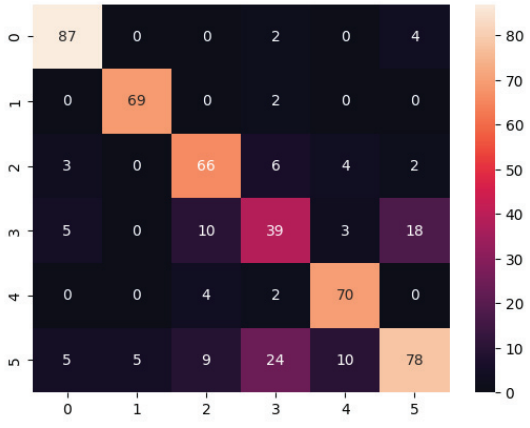


Fig. 3. Predicted label

Then we have Fig. 3. which shows the confusion matrix and Fig. 4. which is the classifications report containing the precision, recall, f1-score, and support. The result is from the training dataset, we have 660 samples in the training dataset divided into 6 categories. these categories are hate, love, proud, neutral, sad, and happy. As we can analyze from the table we have recalled more than 90% for happy, hate, and proud. precision is more than 90% for love, and the f1-score is also more than 90% for hate. We achieved amazing results. Overall we have a precision value is 78%, recall value is recall value is 73% and an f1-score is also 73%.

```
Classification Report:
              precision    recall  f1-score   support

       happy       0.87      0.94      0.90        93
        hate       0.93      0.97      0.95        71
        love       0.74      0.81      0.78        81
     neutral       0.52      0.52      0.52        75
       proud       0.80      0.92      0.86        76
         sad       0.76      0.60      0.67       131

    accuracy                           0.78       527
   macro avg       0.77      0.79      0.78       527
weighted avg       0.77      0.78      0.77       527
```

Fig. 4. Result Table

Beniwal and Dobhal used a model having two layers of Bi-LSTM, and they achieved precision 73%, recall 72%, and F1-score 72% [5]. We are using a hybrid approach on the same dataset and proposed a hybrid model having the first layer as GPT2, then the second layer as Bi-LSTM, and then a few convolutional layers and achieved outstanding results is shown in Fig. 5. the comparison of the results of Bi-LSTM and our hybrid model. We can clearly analyze that our hybrid model outperformed the available dataset with higher accuracies.In this Bar graph we have compared Bi-Lstm and our hybrid model accuracies and analyzed the differences.

Model	Precision	Recall	F1-Score
Bi-LSTM	73%	72%	72%
GPT, Bi-LSTM, CNN	77%	78%	77%

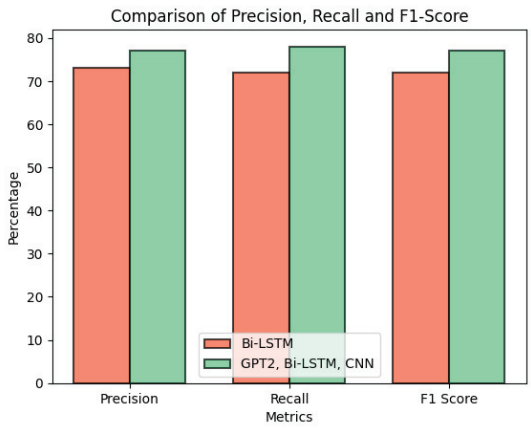


Fig. 5. Comparison of the Bi-LSTM and GPT2, Bi-LSTM, CNN

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:27 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion and future scope

In this study, our purpose is to detect suicidal minds and prevent them from committing suicide. For that, we have done an analysis of suicide note data using deep learning and transfer learning-based models. The datase contains data from 250 suicide notes and is labeled manually into six categories: love, hate, sad, happy, neutral, and proud. We used a hybrid model containing GPT2, Bi-LSTM, and some convolutional layers for sentiment analysis of our dataset. Our result is better than the previous work in this field, and we achieved a 77% F1 score, a 78% recall, and a 77% precision.

We see several opportunities to direct our study on suicide note analysis in the future. A significant obstacle to reaching the required accuracy levels is the lack of large-scale datasets, even with the variety of machine learning models we use in our work, such as CNN, BiLSTM, and GPT. Nevertheless, this is a chance for creativity and teamwork. We plan to establish collaborations and investigate cutting-edge data augmentation methods to broaden the scope of our collection and improve our comprehension of the intricate linguistic patterns seen in suicide notes.

Future research on hybrid technique integration is an exciting area of study. We want to improve suicide note analysis effectiveness and uncover new insights by integrating the advantages of many models and honing our techniques. Our study highlights the importance of tackling the complex problems of mental health and suicide prevention in addition to technological improvements. We aspire to have a major impact on this crucial sector by raising awareness of the needs of individuals in need and providing support for them through interdisciplinary cooperation and a strong commitment to excellence

References

[1] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113

[2] Souma, W., Vodenska, I., & Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods. Journal of Computational Social Science, 2(1), 33-46.

[3] Wang, W., Chen, L., Tan, M., Wang, S., & Sheth, A. P. (2012). Discovering fine-grained sentiment in suicide notes. Biomedical informatics insights, 5, BII-S8963.

[4] Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. Expert Systems with Applications, 40(16), 6351-6358.

[5] Beniwal, R., & Dobhal, A. (2023, June). Sentiment Classification on Suicide Notes Using Bi-LSTM Model. In International Conference on Data Analytics & Management (pp. 1-10). Singapore: Springer Nature Singapore.

[6] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. Biomedical informatics insights, 3, BII-S4706.

[7] Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). Deep cascaded multitask framework for detection of temporal orientation, sentiment and emotion from suicide notes. Scientific reports, 12(1), 4457.

[8] Burnap, P., Colombo, W., & Scourfield, J. (2015, August). Machine classification and analysis of suicide-related communication on twitter. In Proceedings of the 26th ACM conference on hypertext & social media (pp. 75-84).

[9] Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., Alnumay, W., & Smith, A. P. (2021). A lexicon-based approach to detecting suicide-related messages on Twitter. Biomedical Signal Processing and Control, 65, 102355.

[10] Cherry, C., Mohammad, S. M., & De Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. Biomedical informatics insights, 5, BII-S8933.

[11] Denecke, K., & Reichenpfader, D. (2023). Sentiment analysis of clinical narratives: a scoping review. Journal of Biomedical Informatics, 104336.

[12] Singh, J., Singh, G., Singh, R., & Singh, P. (2021). Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification. Journal of King Saud University-Computer and Information Sciences, 33(5), 508-517.

[13] Cao, L., Zhang, H., Feng, L., Wei, Z., Wang, X., Li, N., & He, X. (2019). Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. arXiv preprint arXiv:1910.12038.

[14] Meraliyev, B., Kongratbayev, K., & Sultanova, N. (2021, April). Content Analysis of Extracted Suicide Texts From Social Media Networks by Using Natural Language Processing and Machine Learning Techniques. In 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-6). IEEE.

[15] Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. Complexity, 2018.

[16] Desmet, B., & Hoste, V. (2018). Online suicide prevention through optimised text classification. Information Sciences, 439, 61-78.

[17] Singh, S., & Chandra, S. K. (2022, October). Sentiment analysis for depression detection and suicide prevention using machine learning models. In International Conference on Information Systems and Management Science (pp. 452-460). Cham: Springer International Publishing.

[18] Long, S., Cabral, R., Poon, J., & Han, S. C. (2022). A quantitative and qualitative analysis of suicide ideation detection using deep learning. arXiv preprint arXiv:2206.08673.

[19] Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O., & Bingol, H. O. (2018). Detecting suicidal ideation on forums: proof-of-concept study. Journal of medical Internet research, 20(6), e9840.

[20] Wang, R., Yang, B. X., Ma, Y., Wang, P., Yu, Q., Zong, X., ... & Liu, Z. (2021). Medical-level suicide risk analysis: A novel standard and evaluation model. IEEE Internet of Things Journal, 8(23), 16825-16834.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:27 UTC from IEEE Xplore. Restrictions apply.

SOME CHALLENGES AND ISSUES FOR WATER TRANSFER FROM GANGA-BRAHMAPUTRA BASINS TO CAUVERY-PENNAR BASINS IN INDIA

S.K. MAZUMDER

Former AICTE Emeritus Professor of Civil Engineering
Delhi College of Engineering (Now Delhi Technology University)
E-mail:somendrak64@gmail.com

Abstract

Some eminent engineers of India proposed different plans for transfer of surplus water of Ganga Brahmaputra basins to water deficit Cauvery-Pennar basins in the past for management of recurring floods and droughts in India. These proposals were, however, not accepted due to techno-economic reasons. National Perspective plan drawn by National Water Development Agency envisages the same objective by interlinking rivers through 30 link canals. Some of the challenges and issues involved in execution of the scheme have been critically examined and discussed with particular reference to views of Bangladesh and India in this regard.

Keywords: Water Transfer, River Links, Brahmaputra-Ganga, Cauvery- Pennar, Issues & Challenges

1.0 INTRODUCTION

Govt. of India, proposes increasing water storage and accelerated irrigation by building dams and link canals to transfer water from water surplus to water deficit basins. to address the recurring floods and droughts and to ensure water security in India

Concept of national water grid for effective management of flood and drought situations in India was introduced by a number of eminent engineers

National Water Development Agency (NWDA) studied the possibility of water transfer from surplus basins to deficit basins and proposed the National Perspective Plan (NPP) consisting of 14 river links under the Himalayan component and 16 river links under the peninsular component (Fig.1) involving 12,500 km of canal length and 32 dams

India advocates inter-basin water transfer from the Ganga-Brahmaputra basin in the north to Cauvery-Pennar basin in the south through a series of link canals to transfer the surplus water of Ganga-Brahmaputra basin to address the severe water scarcity in the Cauvery-Pennar basin (Verghese,1999). India's proposal for Brahmaputra-Ganga link through a feeder canal across Bangladesh was the shortest link without any lift component. It was, however, rejected by the Banglsadesh Govt. for political reasons. Apart from its claim for more water downstream of Farakka barrage, Bangladesh also claims water share of Tista river. Keeping in view the Bangladesh demand for more water from Ganga, Brahmaputra and Tista basins, Govt of India has now changed the earlier plan and the new Brahamaputra-Tista-Ganga link -1 is shown in Fig.2(Full green line) which passes entirely through India. The excess water upstream of Farakka barrage will be transported through links10, 11, 15, 18, 22 and 23 to the water deficit Cauvery-Pennar basins in the south. It will also satisfy Bangladesh demand for more water downstream of the proposed barrage at Yogigopa on Brahmaputra river and existing barrages at Gazoldoba on Tista river and Farakka barrage on Ganga river in India (Fig.1 &2). The paper is intended to discuss the various issues and challenges related to implementation of NPP for water transfer of surplus water in Ganga-Brahmaputra basins to water deficit Cauvery- Pennar basins.

2.0 HISTORICAL DEVELOPMENT

Concept of national water grid for effective management of flood and drought situations in India was introduced by a number of eminent engineers in the past, like Sir Arthur Cotton, Dr. K.L. Rao, Captain M. N. Dastur and many others (IWRS-1996). But all the earlier proposals were rejected as

they were found to be techno-economically not viable. It was Indira Gandhi who set up NWDA in 1982 to study the possibility of water transfer from surplus basins in the north to water deficit areas in the south and north-west (Table-1).

Table-1: Water Surplus and Water Deficit Basins in India

Surplus Basins		Scarce Basins	
Basins	Per Capita Availability in M3 Per Year	Basins	Per Capita Availability in m3 Per Year
BrahmaputraBasin	18,417	East flowing Rivers between Mahanadi and Pennar	919
BarakBasin	7,646	Cauvery	666
estflowingRiversbetween TadriandKanyakumari	3,538	Pennar	648
estflowingRiversbetween TapiandTadri	3,194	West flowing River Basin of Kutch and Saurashtra including Luni	631
Narmada	2,855		
Brahmani-Baitarni	2,696		
Mahanadi	2,546	East flowing River Basins between Pennar and Kanyakumari	383
Godavari	2,026		
Indus	1,757	between Pennar and	
Ganga	1,473	Kanyakurnari	

NWDA proposed the National Perspective Plan (NPP) consisting of 14 river links under the Himalayan component and 16 river links (Mazumder,2011) under the peninsular component as shown in Fig.1 (NWDA,2005). A Task Force (2003) was appointed by the late Vajpayi led NDA Govt. National Commission of Integrated Water Resources Development (NCIWRD,1999) was not in favor of long-distance river links and suggested to execute short links initially. The past UPA Govt. wanted to hear the views of all the stakeholders and experts before taking a final decision. A standing committee was formed and the committee invited suggestions /opinions of public and experts in the subject. The committee examined the representations and finally submitted its recommendation to the Govt. regarding implementation of the proposed NPP.

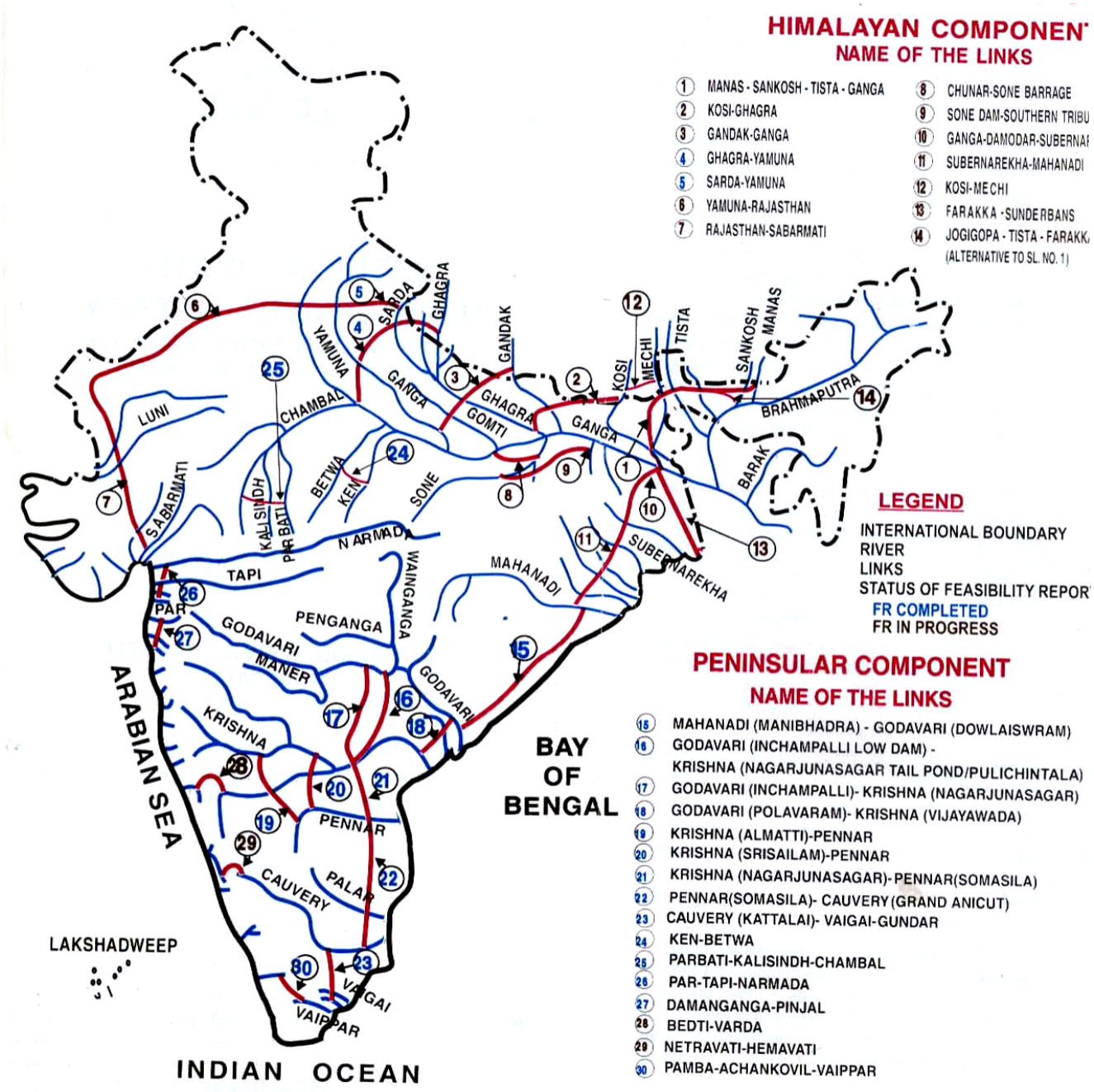


Fig.1 National Perspective Plan for Linking Indian Rivers

Iyer (2003) remarked that NPP scheme should be undertaken with full recognition of the serious ecological damages that may be caused by interlinking rivers and that the benefits should outweigh the costs. Goyal (2003) was highly critical about viability of NPP. Reddy (2002) said Interlinking rivers will be a disaster,

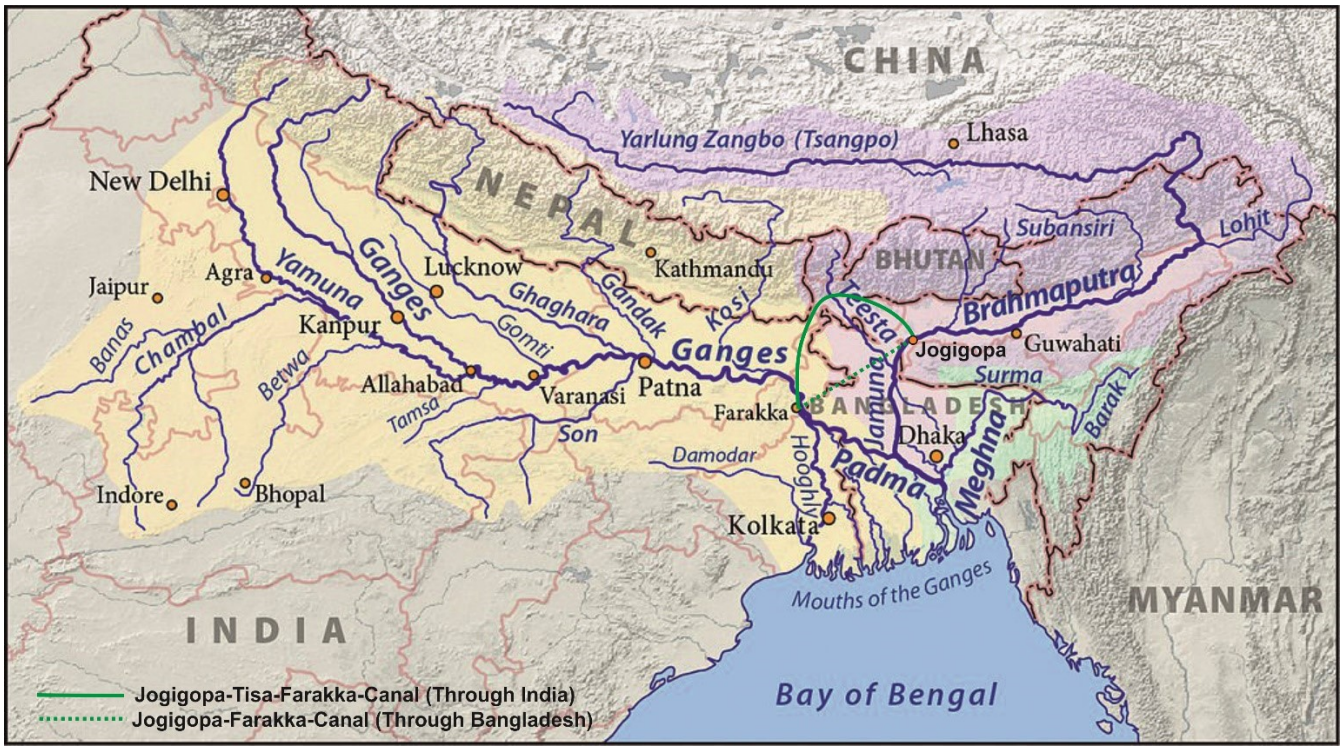


Fig. 2 Revised Plan of Linking Ganga-Brahmaputra via Tista River

3.0 NECESSITY OF INTERLINKING INDIAN RIVERS

Areas having water less than 1000 m3 per capita per year are designated as scarcity areas. All India having an average 1345m3 /capita/year (as on 2025) indicates that India is not water deficit right now. But the spatial distribution of water varies widely due to extreme non-uniform rainfall distribution over the country. Whereas per capita water availability in Brahmaputra basin is 18,417 m3/year, the same for Cauvery-Pennar basin is about 450m^3 only. Due to rapid rise in population and ever increasing water pollution, many surplus basins are likely to be water deficit basins soon. Water transfer from surplus to scarce basins has been found necessary to fight recurring floods and droughts in many parts of the country. It is estimated that an additional area of 35 mha of land can be brought under irrigation and 50,000 MW Hydro-power can be generated through river linking project.

4.0 PRIORITY OF RIVER- LINKING

To meet the growing requirements of water for various uses and to be counted as a developed nation, it is imperative not only to develop the new water sources but to conserve, recycle and reuse water wherever possible. For the food security of our growing population, various options have been considered in quantitative terms as possible sources to augment the anticipated deficit. It is also important to note that only through projects of this nature, we can generate the hydro-power required for groundwater pumping, provide water for inland navigation and meet industrial, municipal and ecological requirements (INAE, 2008). The priority of action, however, must be for increasing productivity of land, choice of proper crops as per prevailing climatic conditions, increase in irrigation efficiency of existing projects through improved water management, rainwater harvesting and groundwater recharge, recycling and reuse of wastewater, preservation of water in tanks to be followed by intra basin and inter basin transfers. The National Commission on Integrated Water Resources Development (NCIWRD,1999) has recognized that long distance inter-basin transfer of water is an outstandingly large complex program of water management. Studies have to be done with the help of computer simulation models and systems analysis.

5.0 PRIORITY OF GANGA–BRAHMAPUTRA LINK

Average annual water availability of India is 1869 billion cubic meters (BCM) from different river basins in the country (Iyer,1989, CWC,1993). The utilizable water with conventional approach is 1121 BCM which comprises of 690 BCM of surface water and 431 BCM of replenish able ground water. Our meagre storage capacity built so far is only 305 BCM which corresponds to 33.88% of the utilizable water resources and 23% of the potential water resources of India which is inadequate to meet the agricultural and other demands. India badly needs more storage since monsoon lasts only about 2 to 3 months in a year. Water resources potential of Ganga-Brahmaputra basin is 1162 BCM i.e. 62% of the country (IWRS, 2007). Hydro-power potential of Brahmaputra-Ganga basin is 1,12,366 MW i.e. 76% of India's total hydro-power potential of 1,48,701 MW (CBIP, 2022). Untapped hydro-power in Ganga and Brahmaputra basins are 26,859MW and 54,329 MW respectively which corresponds to 51.4% and 93.1% of the potential available in the respective basins. Major and minor dams can be built in the northern Himalayan basins of Ganga and Brahmaputra both for storage of water and hydro-power development.

6.0 SOME CHALLENGES AND ISSUES INVOLVED

There are several challenges and issues in water transfer. Several difficulties likely to be faced in its implementation are summarized below.

6.1 Opposition by NGOs

A group of people, especially the NGOs, the environmental group (Iyer,2003) are strongly against the inter-link. They apprehend that such a massive inter- basin transfer of water will result in environmental degradation (Khitolya et al,2005), loss of aquatic eco-system, loss of land, forests, fisheries and the livelihood of the poor people, massive displacement, evaporation losses, water logging and salinity and possible change in the climate (Desai et.al,2005). The massive investment of Rs.5.6 lakh crores (at 2003 price index) which is likely to further increase manifold due to cost and time overrun, is going to deprive other important projects for our socio-economic development. The loans from financing agencies (World Bank, 2006) may subject the country to a permanent vicious debt trap where the country is likely to lose its economic sovereignty and dictated by foreign powers

6.2 Change in Constitution

Water is a state subject under entry 17 of state list- II subject to entry 56 of central list- I at present. Even if the project is found to be techno-economically feasible, implementation of the same will be a herculean task. It needs constitutional amendment. Most of the donor states, even though surplus, is reluctant to part with its resources free of cost and are trying to project their future demands arguing that their surpluses are owing to lack of sufficient storage and due to lack of investment in their states.

6.3 Interstate Dispute

In almost all the river valley projects executed in the country so far, water of a river basin has been shared only by the riparian/cobasin states. The proposed scheme envisages transfer of water from surplus basins to drought prone basins irrespective of whether they are riparian/co-basin or not. Recently, the country has witnessed bitter quarrel and animosity amongst the states of Karnataka and Tamilnadu over the sharing of water from rivers Cauveri and Krishna, even though both the states are riparian. One can well imagine the degree of complexity and the dispute that will arise over sharing of water for the proposed scheme where a large numbers of states are involved, resulting in tension and rivalry.

6.4 Resistance of People in the East and North-East

Most of the surplus water lies in the East and North-East states where people are economically backward mainly due to lack of investment. People may resist inter- basin transfer of their water resources free of cost to the beneficiary states unless the Govt. of India bring their economic condition at par with other developed states of the country. There is a massive unemployment and unrest amongst the unemployed youths in these states. Unless the Govt. addresses these problems first, it may be almost impossible for the project authorities to implement the proposed river linking scheme, however well planned it may be.

6.5 Poor Performance of Many Exiting Projects

Performances of many of the projects are not satisfactory (Mazumder,2002,1984). About 70% of irrigation water is currently being wasted (CWC,2010) due to improper maintenance, lack of proper co-ordination between users and controlling authorities, mismanagement of water at farm level, wrong and unrevised pricing policy (INAE,2008). Farmers of Punjab state are not allowing completion of Sutlej-Yamuna river link while half of the project is already completed by the Haryana Govt. long back and the entire investment is lying idle. One of the DVC main canals which was designed for inland navigation has not carried a single vessel so far.

6.6 Poor Economic Return & Faulty Pricing Policy

Unlike other commodities, water for irrigation is currently distributed almost free of cost. The present irrigation water rates are extremely low and the realization of even those low rates is difficult. Whereas during the British days, 87% of the rates used to be realized from users, today the revenue receipt has come down to 15% only (Mohile,2000). As a result, there is hardly any maintenance of the irrigation projects after their completion. Presently, the overall irrigation efficiency of most of the surface irrigation schemes for agriculture, which consume about 80% of available water, is about 30% only.

6.7 Poor Yield of Farms

Existing yield of cereals per hectare of farms is very poor compared with other developed countries in the world.
All alternatives to river linking suggested by NGOs e.g. increasing productivity of land which is far below other developed countries (Fig.3) should be examined.

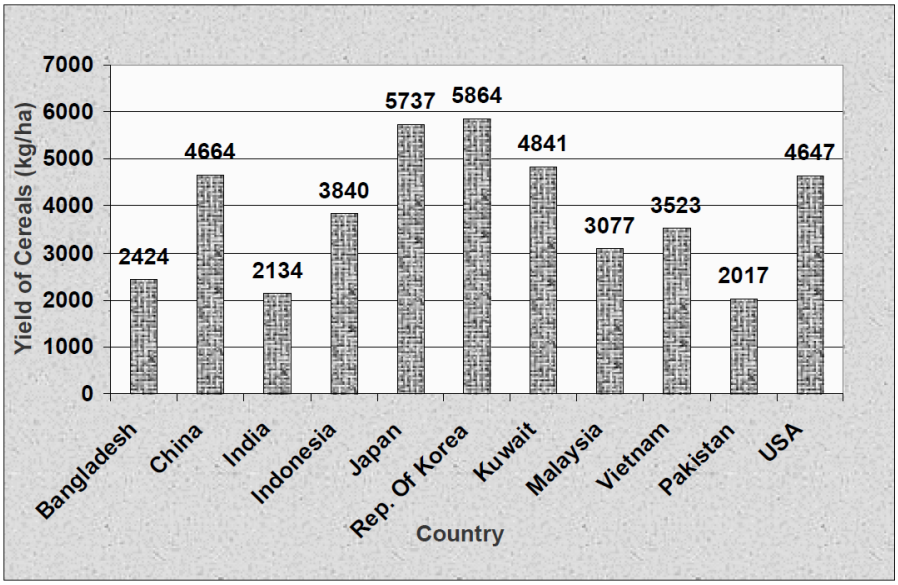


Fig. 3 Showing Yield of Cereals in India Visa Vis Other countries

6.8 Possible Climatic Changes

In an article published in Hindu, 26th August, 2005, consequences of the proposed interlinking on south-west monsoon were discussed. In the article, the inter-relation between the reduction in fresh run-off to Bay of Bengal (BoB) and consequent climate change have been examined by the earth scientists. Once fresh water is transferred (estimated as 200 Km3 annually), scientists are of the view that it is likely to increase salinity of top layer of BoB resulting in decrease of water circulation from BoB to the Indian ocean affecting the south-west monsoon.

7.0 VIEWS OF BANGLADESH OVER BRAHMAPUTRA-GANGA LINK

Bramaputra- Ganga link is essentially needed for transfer of excess water of Brahmaputra-Ganga to Cauvery-Pennar basins. However, Bangladesh opposes Brahmaputra flow diversion on the ground that similar problems may arise in Bangladesh due to Brahmaputra flow diversion as presently being experienced due to Ganga flow diversion from Farakka barrage. Bangladesh is in favour of Ganga flow augmentation to the extent 3,300 to 5,300 m3/sec (Rahaman, 2009) during the lean season by building several storage reservoirs in Nepal. These incremental flows alone are over four times the present lean season flows in the Ganges at Farakka. At the 2005 Joint Rivers Commission (JRC) meeting, Bangladesh proposed holding tripartite talks between Nepal, India and Bangladesh to discuss the construction of several water reservoirs in Nepal. Bangladesh seeks water-sharing agreements also for 53 other common rivers flowing from the Himalayas through the Indian states of Bihar, Assam and West Bengal to Bangladesh basins located downstream of all these rivers (Daily Star, 2005). Bangladesh argues that the augmentation of the Ganga water should be solved through construction of multiple storage reservoirs in Nepal as there is enough water in the Ganga basin. It wants the water managed in such a way as to minimize flooding during monsoon months and water shortage during lean months in Bangladesh. Such a scheme presents less of a threat to Bangladesh sovereignty and also avoids India's claim to the Brahmaputra waters (Crow et al., 1995). It advocates a tri-partite agreement between China, India and Bangladesh for any flow diversion from Brahmaputra river. China has agreed to invest huge amount money for construction of two long barrages inside Bangladesh - one across river Brahmaputra (at Bahadurabad, Jamalpur) and the other across the Ganga (at Pangsha on Tista) without consulting India. India has already constructed a barrage on Tista river at Gajoldoba (near Shiliguri) in west Bengal for diversion of flow of Tista into two major irrigation canal systems for agriculture in the states of Jharkhand and West Bengal.

8.0 INDIA'S STAND OVER BRAHMAPUTRA-GANGA LINK

India wants to divert water from the Brahmaputra basin for augmenting the dry season flow of the Ganga and Tista basins through the proposed Brahmaputra-Tista -Ganga link for addressing the immediate problems being faced by Bangladesh during lean season downstream of Farakka and Tista barrages. It will help in transfer of excess of combined flow of Ganga and Brahmaputra basins from north to south of India in the long run by interlinking Farakka (Link-1) with other rivers (Links 10, 11, 15, 18, 22 and 23 in Fig.1). The proposed link will help generating hydropower using God given terrain head of north-eastern states south of Himalayas. Several major storage reservoirs on tributaries of Brahmaputra e.g. Sankosh, Manash, Subansiri, Dihang etc. are proposed besides minor storages on innumerable branches and sub-branches of Brahmaptra and Barak basins discharging into Brahmaputra river. It will be highly beneficial for water transport for navigational use (National Waterways-1 &2).The proposed link is exclusively through Indian Territory and will divert flow generated in Indian catchments south of Tibet.

Large numbers of dams and reservoirs have been already built by China in Brahmaputra basin (called Tsangpo river basin in Tibet) for storage of water of Tsangpo river and its tributaries and diverting flow to dry areas in the north of China without entering into any agreement with either India and Bangladesh.

Although 90% of Ganga and Brahmaputra catchments south of Himalayas belong to India, it has agreed for bilateral agreement with Bangladesh for water sharing of Ganga and Brahmaputra rivers through joint rivers commission (JRC). Put simply, Bangladesh wants to share water of Brahmaputra and Ganga multilaterally by involving China, Nepal, India and Bagladesh. But India wants to share water bilaterally with Bangladesh since the proposed Brahmaputra-Ganga link has nothing to do with either Nepal or China who has violated international norms of consulting riparian states while constructing multiple reservoirs on Tsangpo river and its tributaries in Tibet. In short, there is an unresolved dilemma between the proposals of Bangladesh and India over Brahmaputra –Ganga link for water sharing over time and space (Rao, et al, 2012). At the 36th JRC meeting, India, however, assured Bangladesh that it would implement the Ganga- Brahmaputra link after consultation with Bangladesh (Daily Star, 2005)

REFERENCES

CBIP (2022), "Water and Energy International" Journal *pub.by Central Board of Irrigation & Power, MalchaMarg, Chanakyapury, Vol. 64, No.11, Feb. 2022*

CWC (2010) "Water Use Efficiency for Irrigation Projects" Pub. by central Water Commission, Min. of water Res., River Dev. And Ganga Rejuvenation, Govt. of India, April

CWC(1993) "Water Resources Potential In The River Basins Of India" chapter-1 : *Publication by Central Water Commission, Min. of Water Resources, River Development and Ganga Rejuvenation, Govt. of India*

Crow, B., Lindquist, A. & Wilson, D. (1995) "Sharing the Ganges: The Politics and Technology of River Development", *Dhaka University Press Limited, Dhaka, Bangladesh*

Daily Star (2005) "JRC Meet Ends on Good Note", The Daily Star, 5(471), 22 September Available at: *http://www.thedailystar.net/2005/09/22/d5092201011.htmxx*

Desai V.R.,Ghosh S.N and Patro J.(2005), "National Water Grid- its Objectives and Challenges", *ISH Journal of Hyd. Engineering., Vol. 11, No.1, March*

Goyal, J. (2003) "Is interlinking of rivers viable?" The Tribune (March 13). *http://tribuneindia.com/2003/20030313/science.htm#1*

INAE (2008) "Water Resources Management-Role of Water Sector in India", *Pub. by Indian National Academy of Engineers, IIT(Delhi), Feb.21-22*

IWRS (2007) "Role of Water Resources Development & Management in Bharat Nirman", Theme paper presented on Water Resources Day observed by *Indian Water Resources Society (IWRS) at ICID, New Delhi, May 9th*

IWRS (1996) "Inter-basin Transfer of Water for National Development- Problems and Prospects" Theme paper *pub. on Water Resources Day by IWRS, New Delhi*

Iyer, S.S. (1989) "Water Resources of India – An Overview", *Journal of Indian Water Resources Society, Vol. 9, No. 2, April*

Iyer,R.R(2003),"River Link To disaster" *Pub. in The Statesman, April,5th*

Khitolya A.K, Goyal A.K and. Kumar D. (2005) "Environmental and other Aspects of Interlinking Rivers", *Proceedings of Recent Advances in Water Resources Development and Management, organised by Deptt. of Water Resources Development and Management (WRDM) earlier known as WRDTC), IIT, Roorkee, 23-25 Nov*

Mazumder, S. K. (2011) "Interlinking Indian Rivers – Merits, Demerits and Difficulties in Implementation" *published in the Journal of Bharati Vidyapeeth, Pune*

Mazumder S.K (2002) "Performance of Surface Irrigation Schemes in India", keynote speech pub. in the *proceedings of All India seminar on water and environment-issues and challenges, organized by IWRS and WRDTC, IIT (Roorkee), Oct.*

Mazumder, S.K. (1984) "Efficiency of Irrigation in the Command Areas of DVC, Mayurakshi and Kangsabati projects in West Bengal", *Proc. National Seminar on Water Resources Management for Rural Development, KNIT, Sultanpur, Oct. 18-19*

Mohile A.D (2000) "River Valley Development- a Comprehensive Approach", *proceedings of International conference on hydro-power and river valley development, Oxford & IBH Pub. Co. Pvt. Ltd, New Delhi, Dec.*

NCIWRD (1999), "Annual Report of National Commission for Integrated Water Resources Development", *Min. of Water Resources, Govt. of India*

NWDA (2005), "Water for life", *11th National Water Convention, org. by National Water Development Agency, Govt. of India Delhi, May 11th*

Rao S. V. N., S. V. Vijaya Kumar, K. S. Ramasastri & P. R. S. Rao (2012) "Optimal Inter Basin Transfer in Space and Time : A Case Study of Interlinking of River Systems in South India", *ISH Journal of Hydraulic Engineering, Published online: 07 June*

Rahaman Muhammad Mizanur (2009) "Integrated Ganges Basin Management: Conflict and Hope for Regional Development", *Water Policy, January,www.researchgate.net/publication*

Reddy, C.R. (2002). "Interlinking Rivers will be a Disaster". Chennai: *The Hindu Sunday Magazine, October 20*

Task Force (2003), "Inter-Linking of Rivers "-Agenda for the meeting of *IITs and IISC under the chairmanship of Suresh Prabhu, task force, August*

TFIR (Task Force on Interlinking of Rivers) (2005). Official Website of the Task Force on Interlinking of Rivers, *http://nwda.gov.in/(accessed 12 May 2005)*

World Bank (2006), 'World Bank and Swiss Agency for International Co-operation, Report-, '*India's Water Economy Bracing for a Turbulent Future', World Bank Report, 2006*

Verghese, B. G. (1999). "Waters of Hope: From Vision to Reality in Himalaya-Ganga Development Cooperation", *Dhaka University Press Limited, Dhaka, Bangladesh*

ExLibris RapidILL

Rapid #: -23205871

CROSS REF ID: **6420166**

LENDER: **HNK (Hong Kong Univ. of Science & Technology) :: Main Library**

BORROWER: **HUL (Harvard University) :: Widener Library**

TYPE:	Article CC:CCG
JOURNAL TITLE:	Electrical & computer engineering
USER JOURNAL TITLE:	Iranian journal of science and technology.
ARTICLE TITLE:	Stacked Denoising Variational Auto Encoder Model for Extractive Web Text Summarization
ARTICLE AUTHOR:	Yadav, Madhuri
VOLUME:	
ISSUE:	
MONTH:	
YEAR:	2024
PAGES:	NA
ISSN:	2228-6179
OCLC #:	

Processed by RapidX: 9/26/2024 7:58:40 PM

Copy supplied for research or private use only. Not for further reproduction or distribution.

The Hong Kong University of Science & Technology Library Document Supply Service

ILLiad TN: 332053

ARTICLE SLIP

Borrower: RAPID:HUL

Lending String:

Journal Title: Iranian journal of science and technology

Volume: **Issue:**
Month/Year: 2024
Pages: NA 18p.

Article Author: Yadav, Madhuri

Article Title: Stacked Denoising Variational Auto Encoder Model for Extractive Web Text Summarization

Imprint:

ILL Number: -23205871

Call #:
http://rapid.exlibrisgroup.com/redirect?thid=NDU3NTUwNjA=
Location: SpringerLink Journals

ILLiad TN: 332053

Charge
Maxcost:

Shipping Address:
NEW: Widener Library

Fax:
Ariel:

Hong Kong University of Science & Technology Library

Document Supply Service

Odyssey: 143.89.104.29
Email: lbill@ust.hk
Tel: (852)2358-6766

Copyright Notice

Copy supplied for research or private use only. Not for further reproduction or distribution.

Iranian Journal of Science and Technology, Transactions of Electrical Engineering
https://doi.org/10.1007/s40998-024-00751-9

RESEARCH PAPER

Stacked Denoising Variational Auto Encoder Model for Extractive Web Text Summarization

Madhuri Yadav[1] · Rahul Katarya[1]

Received: 18 May 2023 / Accepted: 28 August 2024
© The Author(s), under exclusive licence to Shiraz University 2024

Abstract
Extracting and concatenating distilled content from a corpus into a summary is a technique known as extractive summarization. In recent days, extractive summarization of web text has become popular due to the wide usage of social media. Hence various researches have been conducted on extractive summarization of web text, but the processing of huge amounts of web text and understanding the context is difficult due to the requirement of a lot of storage and time. To solve this issue, the continuous bag of words text vectorization model has been used that reduce the processing time by producing a distributed combination of words in vector arrangement. Moreover, the polysemous words are unable to be captured, which makes extraction difficult. Hence a novel Hierarchical Attention pointer Stacked Denoising Variational Autoencoder Model has been proposed in which the SDVAE model forms latent distribution for contextualized words and bidirectional attention mechanism extracts keywords and features from sentences thereby capturing polysemic words. Furthermore, the summary is obtained with dangling anaphora whereas antecedent morphological expression and verb referents are not considered in the summary. Hence a novel Multilayered Competitive Probable Modular Perception Model has been proposed in which the competitive layer scores the sentence and the scored sentences are ranked using string kernel and class conditional probability thereby considering the antecedent morphological expression and then, Graph based Quadruplicate Lexicon Summarization is used that forms quadruplicate lexicon chain in graphical format to eliminate dangling anaphoric expressions. The experimental results obtained show that the proposed model has achieved a comparatively high accuracy of 98.3% and recall, precision, and F-measure of 98%.

Keywords Continuous bag of words · Extractive summarization · Hierarchical Attention · Modular perception model · Stacked denoising variational autoencoder model

1 Introduction

Text resources have proliferated on the web as a result of the cyclopean use of digital devices combined with Internet accessibility. The most common large collections of freely accessible information are different versions of websites and online platforms. On the internet, there is a multitude of textual information. Web text encompasses a wide range of contexts/domains and is constantly updated by emerging multi-dimensional information (Abualigah et al. 2020; Hossain and Hoque 2020; El-Kassas, et al. 2021). People are not interested in reading a long piece of text and as a result, typically skip crucial sections of it. This has boosted the need for text summarization which is a method that makes key information extraction easier from a document shortly and straightforwardly.

In the current emergent information era, text summarization has evolved into a critical and relevant engine for supporting and illustrating text material. People find it much more challenging to physically summarize lengthy texts (Al-Maleh and Desouki 2020; Kumar et al. 2021; Madhuri and Ganesh Kumar 2019). Data mining is used to describe the process of dealing with large amounts of raw data and extracting meaningful information from it using

✉ Rahul Katarya
rahulkatarya@dtu.ac.in

Madhuri Yadav
madhuriyadav.me@gmail.com

[1] Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Published online: 13 September 2024

an algorithm (Ghodratnama, et al. 2020; Kanapala et al. 2019). The methods that fall under the area of text summarization are extractive approaches and abstractive approaches. The extractive procedures choose a few lines or phrases from the original text, whereas the abstractive ways create a summary by building a semantic structure or semantic tree and then using the natural language generation methodology (Shirwandkar and Kulkarni 2018; Song et al. 2019).

The extractive text summarization technique is further separated into two types: supervised and unsupervised learning. Supervised learning requires a person to identify the sentences in the original training text and a learning data set while the unsupervised technique does not necessitate the need for a human for summary generation (Mao, et al. 2019; Alami et al. 2019). To construct a summary, text summarization techniques employ natural language-generating techniques. Extraction-based summarization is used by the majority of text synthesis tools. Topic identification, interpretation, summary creation, and summary evaluation are the primary challenges in text summarization. Important tasks in extraction-based summarizing include locating relevant phrases and using them to select sentences for the summary. All extraction-based summarizers do three distinct activities, (i) gathering key text components and saving them as an intermediate representation, (ii) grading text sentences based on that representation, and (iii) creating a summary by selecting several phrases.

Traditional scoring methods, such as sentence length, sentence position, and TF-IDF-based features, incorporate feature engineering as a necessary and labor-intensive task. Best n, maximal marginal relevance, and global selection are some of the methods for selecting sentences for the summary (Mishra et al. 2019; Siautama and Suhartono 2021; Mao, et al. 2010). However, these existing techniques for extractive text summarization have issues in the extraction of features due to the consideration of decontextualized sentences with the core meaning of the word only, lack coverage, and have redundancy issues (Wang, et al. 2013). Text summarization methods based on neural networks have improved substantially in recent years. Extracting semantic characteristics with neural networks for extractive summarization has gotten more attention. In the realm of natural language processing, these semantic latent characteristics are beneficial(Zhou, et al. 2020). Deep learning-based approaches (Suleiman et al. 2019; Doğan and Kaya 2019; Magdum and Rathi 2021; Shini et al. 2021) have been used to tackle the issues in traditional text summarization techniques. However, during extractive text summarization, some issues have been noticed such as the dangling anaphora problem and the inability to capture polysemy (Steinberger et al. 2007; Li et al. 2021). Hence, it is necessary to develop a novel solution to tackle the aforementioned issues. This work focuses on filling in the gaps in text summarization, with particular attention on web-specific feature extraction, web-specific data processing, sentence scoring, ranking difficulty, redundancy, and dangling anaphora resolution in web material.

1.1 Major Contribution

The following are the major contributions provided by this paper:

- In the proposed work, the CBoW text vectorization model has been introduced to eliminate the delay in processing large-scale data from websites.
- We propose a Hierarchical Attention pointer Stacked Denoising Variational Auto Encoder Model for efficient feature extraction.
- We propose a Multilayered Competitive Probable Modular Perception Model for scoring and ranking sentences.
- A Graph-based Quadruplicate Lexicon Summarization has been proposed to produce a summary by resolving the dangling anaphora issue.

Our research hypothesizes that the proposed model will eliminate processing delays, improve feature extraction, enhance sentence scoring and ranking in text summarization, and effectively address dangling anaphora issues. We aim to empirically verify these hypotheses, anticipating significant advancements in the field of text summarization.

1.2 Paper Structure

The content of the paper is structured as follows: In the next section the literature survey has been discussed, Sect. 3 discusses the process of extractive summarization, Sect. 4 discusses the proposed approach, Sect. 5 discusses experimental results and their comparison; finally, Sect. 6 concludes the paper.

2 Literature Survey

Extractive summarization is a process in which we try a sentence using a strategy and then select the most important sentences to form the output sentence. Our primary goal when summarizing a document is to produce a summary that encapsulates the document's overall conclusion. In this endeavor, extractive summarization identifies the phrases or paragraphs that accurately and precisely convey the significance of the content. Elbarougy et al. (Elbarougy et al. 2020) suggested a graph-based system in which the

text is represented as a graph, with the sentences as the vertices. A modified PageRank algorithm is used for each node with a starting score equal to the number of nouns in this phrase. The cosine similarity between phrases is used to create a final summary that incorporates sentences with more information and is well related to one another. The three main stages of text summarization are pre-processing, feature extraction, and graph formation, followed by summary extraction using the Modified PageRank approach. The Modified PageRank method uses a variable number of iterations to discover the number that produces the best summary results, and the extracted summary is based on compression ratio, which takes into consideration reducing redundancy based on sentence overlapping. However, ranking score determination takes huge time and becomes expensive. El-Kassas et al. (El-Kassas, et al. 2020) presented "EdgeSumm," revolutionary graph-based architecture based on four described algorithms. The first technique builds a new text graph model representation from the supplied content. The next two algorithms search the constructed text graph for sentences to include in the proposed summary. When the final candidate summary exceeds the user-specified threshold, the fourth strategy is utilized to select the most important sentences. EdgeSumm combines a variety of extractive ATS techniques (including graph-based, statistical-based, semantic-based, and centrality-based methods) to maximize their benefits and minimize their drawbacks. EdgeSumm is unsupervised, global (not limited to a particular topic), and does not need any training data. However, dangling anaphora problems are not focused on graph-based approaches hence additional resolution techniques are required. Patel et al. (Patel et al. 2019) provided a multi-document summary to ensure enough content coverage and information variety. The fuzzy model is used by the statistical feature-based approach to deal with the erroneous and ambiguous feature weight. In addition to this technique, redundancy elimination using cosine similarity is offered. On the DUC 2004 dataset, the proposed technique is compared to DUC participant systems and other summarizing systems such as TexLexAn, ItemSum, Yago Summarizer, MSSF, and PatSum using the ROUGE measure. However, sentence ordering is difficult using this model. Manjari et al. (Manjari, et al. 2020) employ a method that creates an extractive summary of the information depending on the user's query, data was collected from numerous websites all over the internet. Selenium web scraping is also covered. For text summarization, the Term Frequency–Inverse Document Frequency (TF-IDF) method is used. This method is original and effective for producing summaries in response to user requests. However, other web-scraping techniques and summarization techniques are required to enhance the performance of text summarization.

Chatterjee et al. (Chatterjee and Yadav 2019) presented a text-summarizing approach that combines latent semantic analysis with random indexing. In addition, tests to compare the results to several relevant baseline approaches. A hybrid strategy based on random indexing and latent semantic analysis known as LSA-RI tries to reduce the amount of time necessary for SVD method matrix computations. The relative improvement in outcomes above the baseline LSA-based strategy demonstrates the usefulness of the hybrid method devised. However, this technique is unable to determine multiple meanings of a word and hence it is challenging to compare documents. Hernández-Castañeda et al. (Hernández-Castañeda, et al. 2020) provided a new keyword detection strategy for the ATS task that takes advantage of semantic information. This method (Uçkan and Karcı 2020) not only improves coverage by clustering sentences to identify the primary subjects in the source document, but it also improves precision by finding keywords within the clusters. The solution offered performed better than earlier methods using a common collection, according to the findings of this study's experiments. However, the representation of the produced summary is redundant without any verb referents.

This paper (Gangathimmappa et al. 2023) proposed a DLCLS–MQO model for cross-lingual multi-document summarization. it enables queries in one language summarization to another and uses deep learning and metaheuristics for superior results. The author (Joshi et al. 2023) proposed a deep Summ novel extractive summarization method using topic modeling and word embedding. It combines topic vectors and sequence networks to enhance summary quality and accuracy. Authors in Ma et al, (2304) proposed an impression GPT that improves radiology report generation by levering large language models (LLMs) within dynamic prompt an iterative optimization algorithm. It achieves state-of-the-art results on medical datasets, binding the gaps between LLMs and domain-specific language processing. Similarly, this paper provides an alternative approach (Luo 2303), which explores ChatGPT's effectiveness in evaluating factual consistency in text summarization. In (Ghadimi et al. 2023) the author focused on sentence embedding and feature learning using a submodular convolution network.

Authors of Mao, et al. (2010) a maximum marginal relevance-based pre-trained BERT model is used which tackles the redundancy however dangling anaphora issue is still not focused. Authors of Elbarougy et al. (2020) take huge time for ranking score determination and in El-Kassas, et al. (2020), the dangling anaphora problem is not focused on graph-based approaches hence additional resolution techniques are required. Sentence ordering is difficult in Patel et al. (2019) and (Manjari, et al. 2020) requiring other web-scraping techniques and

summarization techniques to enhance the performance of text summarization. (Chatterjee and Yadav 2019) is unable to determine multiple meanings of a word and (Hernández-Castañeda, et al. 2020) forms a summary with redundant sentences. In (Gangathimmappa et al. 2023) model focused on multi-document and multi-lingual summary generation, but did not focus on short-length text. In(Joshi et al. 2023), the author emphasized redundancy but did not address the dangling anaphoric issue. The authors of Ma et al, (2023) concentrated on domain-specific summarization using large language models but did not give attention to concerns such as redundancy and dangling anaphora. (Luo 2023) exhibits limitations such as favoring lexically similar options false reasoning, and incomplete instruction understanding. In (Ghadimi et al. 2023) the author focused on the issues of redundancy but did not address the dangling anaphora issue. Hence, there is a need to develop a novel model to solve the issues in the aforementioned techniques.

3 Summarization Methodology

Extractive summarization is the process of finding the important and most significant sentences of the text. However, the existing techniques are having issues in extracting features from text and in producing relevant summaries. The following phases are involved in the extractive summarization approach:

3.1 Preprocessing

This is a step that plays an important role in summarization. The following tasks are done that preprocess the original web text by performing tokenization, stop word removal, and lemmatization((Abualigah et al. 2020; El-Kassas, et al. 2021)). To provide an efficient processing of web text, a novel CBoW Text Vectorization Model has been introduced. This model converts the tokens formed from the Tokenization process into text vectors to form a combination of distributed representations of words which makes the processing of huge web text easier in minimum time.

3.2 Processing

3.2.1 Feature Extraction

To identify the key features of data we need to do some processing on the data (Abualigah et al. 2020; El-Kassas, et al. 2020). At the stage of extracting features from web text, a Hierarchical Attention pointer Stacked Denoising Variational Autoencoder Model has been proposed that undergoes two phases. In the first phase, SDVAE encodes word and sentences, and generate contextualized sentence representation using deep stacked layers that capture all possible word meanings based on context from preprocessed text and produce a smooth latent representation of sentences via a Denoised latent distribution layer. In the second phase, the Bidirectional Attentive Pointer Mechanism points its attention on the words from contextualized sentences in a bidirectional routine that extracts features including polysemy and thereby contextualized feature extraction is achieved.

3.2.2 Scoring and Ranking the Sentences

For scoring and ranking in summary generation, the Multilayered Competitive Probable Modular Perception Model has been proposed in which sentences are scored based on proximity and singularity of words using competitive probability. Then, scored sentences were ranked with the optimal percentage of string kernels based on conditional probabilistic rules which avoid irrelevant cataphoric expressions in ranked sentences.

3.3 Post-Processing

Additionally, Graph-based Quadruplicate Lexicon Summarization has been proposed that form quadruplicates of nouns, adjectives, determiners, and predicate lexicon chains in a graphical structure to consider verb referents thereby eliminating dangling anaphoric expressions. Hence, the proposed model increases the accuracy and precision of text summarization with consideration of polysemy, antecedent morphological expression, and verb referents.

4 The Proposed Work

The proposed model for web text summarization has been shown in Fig. 1 in which the preprocessing is enhanced by using CBoW-based vectorization and then features were extracted using a hierarchical attention pointer-based SDVAE model with bidirectional connection.

The SDVAE incorporates the variational inference principle which makes it suitable for dealing with uncertainty and task of data summarization. SDVAE has structured latent space which makes it beneficial for feature extraction. Denoising is the key component of SDVAE which helps it to capture noise. Then, the summary is generated based on the ranking and scoring process in the Multilayered Competitive Probable Modular Perception Model also, to generate the summary with verb referents and without sentence repetition, Graph based Quadruplicate Lexicon Summarization has been used.

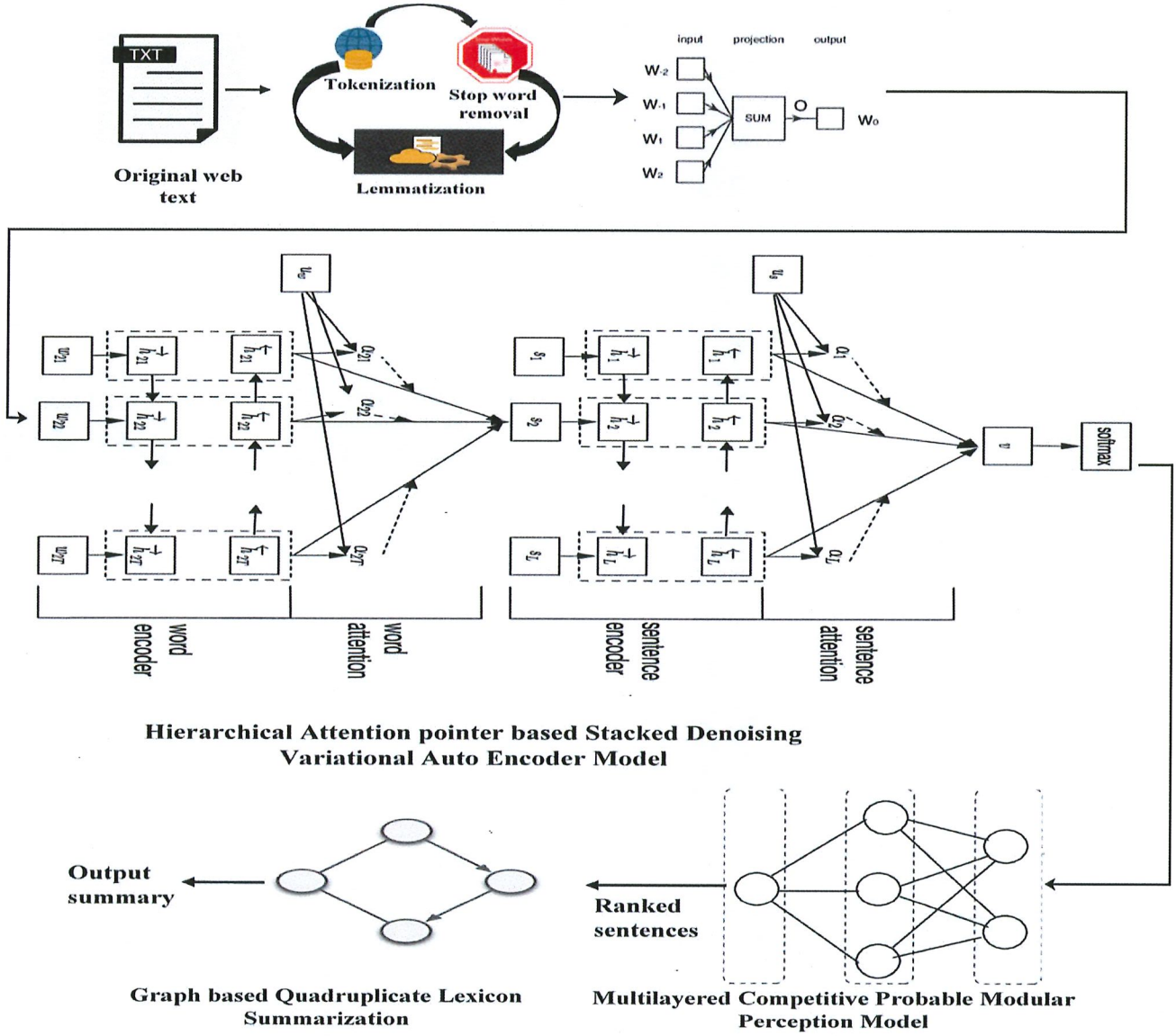


Fig. 1 Architecture of proposed web text summarization

4.1 CBoW Text Vectorization Model

Web text from Twitter has been taken as an input and this web text data is preprocessed by removing stop words from the sentences and grouping different forms of the same word using lemmatization. Then, tokenization was performed to divide a chunk of text into distinct words using a delimiter, which forms word tokens. These word tokens were given as the input to the CBoW Text Vectorization model. The process takes place on the CBoW text vectorization model which is shown in Fig. 2.

The CBOW model tries to understand the context of the words and uses that information as input. It then tries to predict the center word on the basis of the context of surrounding terms. If four context words are used to forecast one target word, the input layer will have four $1XW$ input vectors where W is the number of words in the vocabulary. The hidden layer receives these input vectors and multiplies them by a WXN matrix, where N size of the hidden layer. Finally, the $1XN$ output matrix from the hidden layer enters the sum layer, which performs an element-wise summing on the vectors before performing a final activation and obtaining the output. These preprocessed and vectorized words were given to the hierarchical model for feature extraction which is explained in the next subsection.

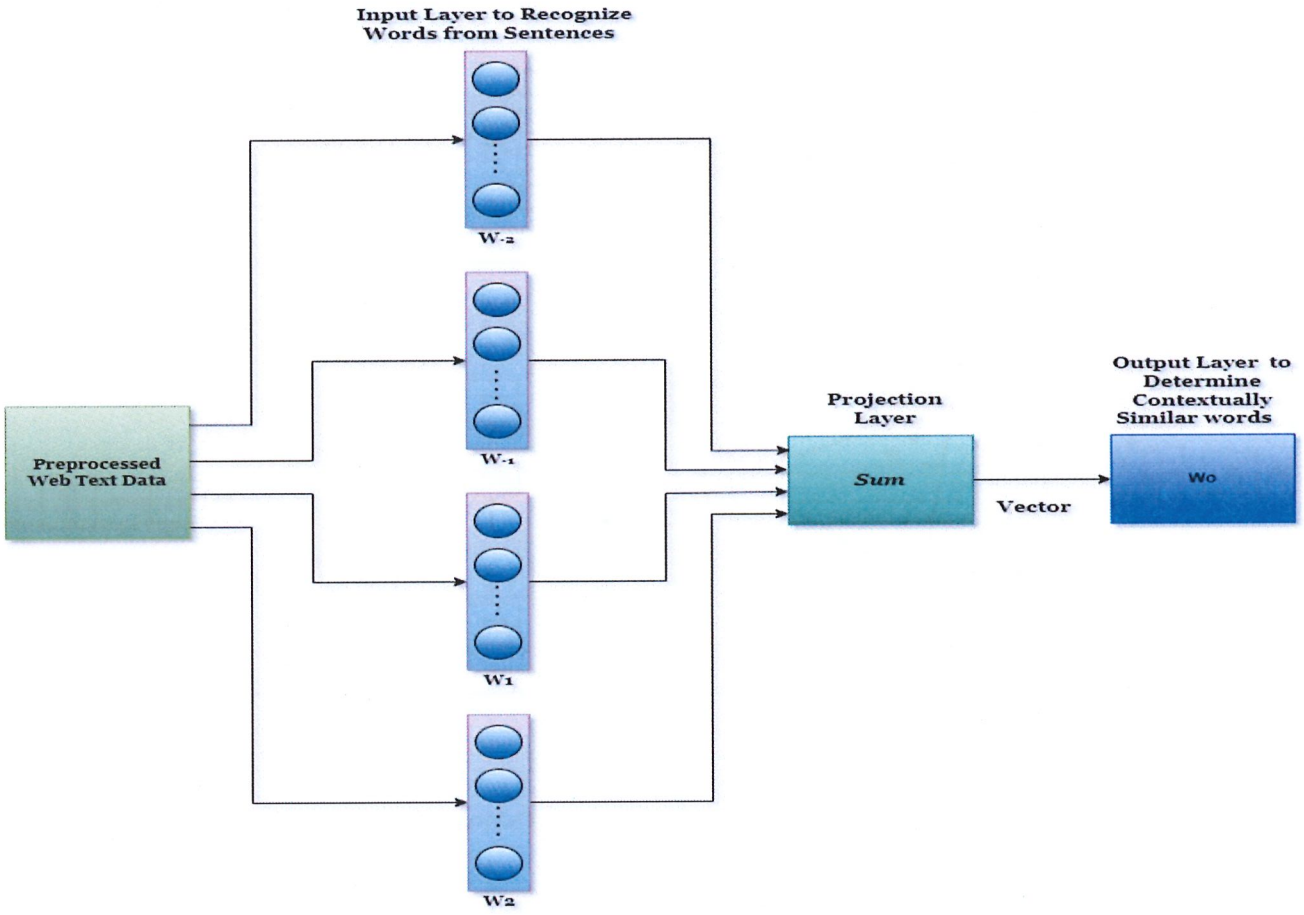


Fig. 2 CBoW text vectorization model

4.2 Hierarchical Attention Pointer Based Stacked Denoising Variational Autoencoder Model

Contextualized feature extraction with the consideration of polysemy words has been performed using the Hierarchical Attention pointer-based Stacked Denoising Variational Autoencoder (SDVAE) Model. This model undergoes a two-phase process to extract the features efficiently. In the first phase, words were encoded using SDVAE and in the second phase, an attention mechanism was utilized to extract polysemy, linguistic, and semantic features with the consideration of contextualized words. These two phases, encoding and attention pointing have been performed for sentences also. The SDVAE architecture for performing word and sentence encoding has been shown in Fig. 3.

The preprocessed vectorized web text data has been given as input to the SDVAE model. The noise in the input layer is reduced by denoising the variational auto encoder which improves the robustness and generalization ability in the extraction of more useful feature representation. Only a small amount of label noise is introduced into the original input, allowing the model to rebuild "pure" data from "polluted" input. The logarithm of the variance of the latent variables in an SDVAE parameters that is learned during the training process, and its value will vary depending on the specific input data. The purpose is to capture the uncertainty or variance in the latent space, which allows the SDVAE to generate data with varying degrees of randomness during the decoding process. The parameter settings are eliminated into enter the local optimum space. The smooth latent representation of words was obtained after performing denoising and latent state distribution in hidden layers of the encoder. Latent vectors are low-level data representations that are created by the encoder from high-level data distribution representations. After that, the decoder takes in low-level data representation and produces high-level data representation. Latent contextualized distribution relates the observable vector y to the low-dimensional latent variable x. SDVAE calculates the probability of observable vector as follows:

$$P_{\varphi}(y) = \int P_{\varphi}(y|x)P_{\mu}(x)dx \tag{1}$$

In Eq. (1), $P_{\mu}(x)$ is the prior probability of latent variables modeled by SDVAE model with parameter μ and

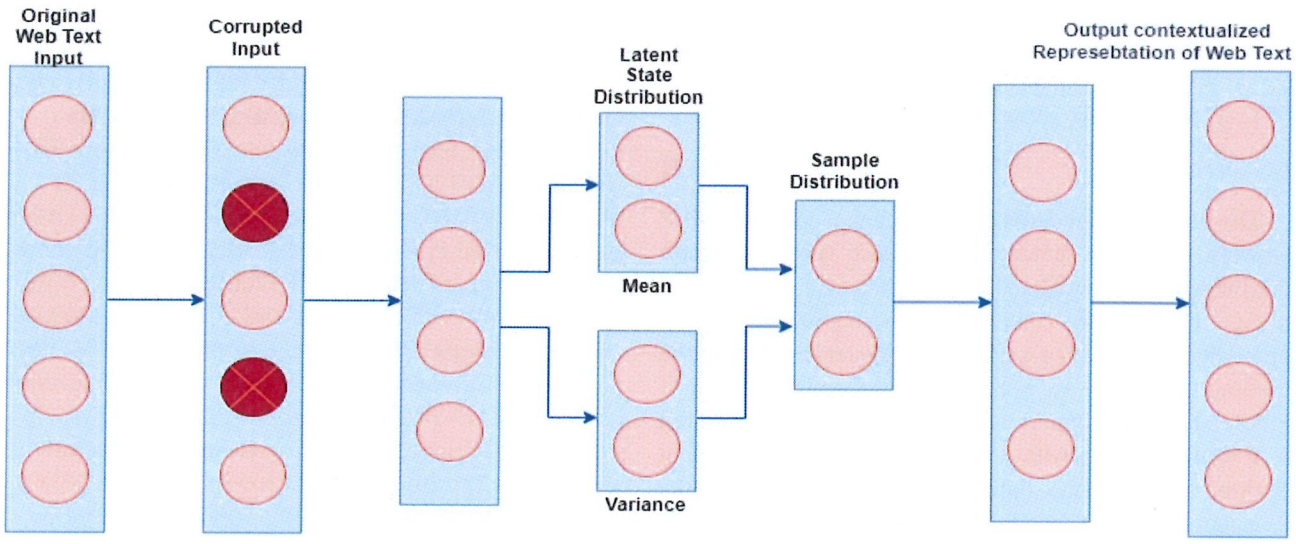


Fig. 3 Word and sentence encoding in SDVAE model

$P_{\varphi}(y|x)$ is the posterior probability of latent variables modeled by SDVAE model with parameter φ. The mean and variance of vectorized words were generated in the latent state distribution to form the contextualized distribution of words. By merging data from both directions for each word, our approach creates word annotations that incorporate contextual information. The forward and backward hidden states are concatenated to create an annotation for a specific word which encapsulates the knowledge of the entire sentence. Then, a bidirectional attentive pointer mechanism is introduced to extract keywords for the sentence's meaning and combine the representations of those informative words to generate a sentence vector. The architecture of the bidirectional attentive pointer mechanism is shown in Fig. 4.

In the attention pointer mechanism, the word annotation is first fed through a one-layer MLP network to get the word prominence vector (v_{jt}) from the hidden representation (H_{jt}) and word context vector (v_w). The word context vector v_w is thought of as an upper demonstration of a query over words, similar to what memory networks utilize. During the training procedure, v_w is initialized randomly and trained jointly. Then, the importance of the word is measured as a similarity of v_{jt} with a normalized prominence weight β_{jt} through the softmax function. The sentence vector (k_j) is then computed as a weighted sum of the word annotations depending on the prominence weights β_{jt} and hidden representation H_{jt}. Bidirectional attentive pointer mechanism for words is explained in Eqs. (2–4),

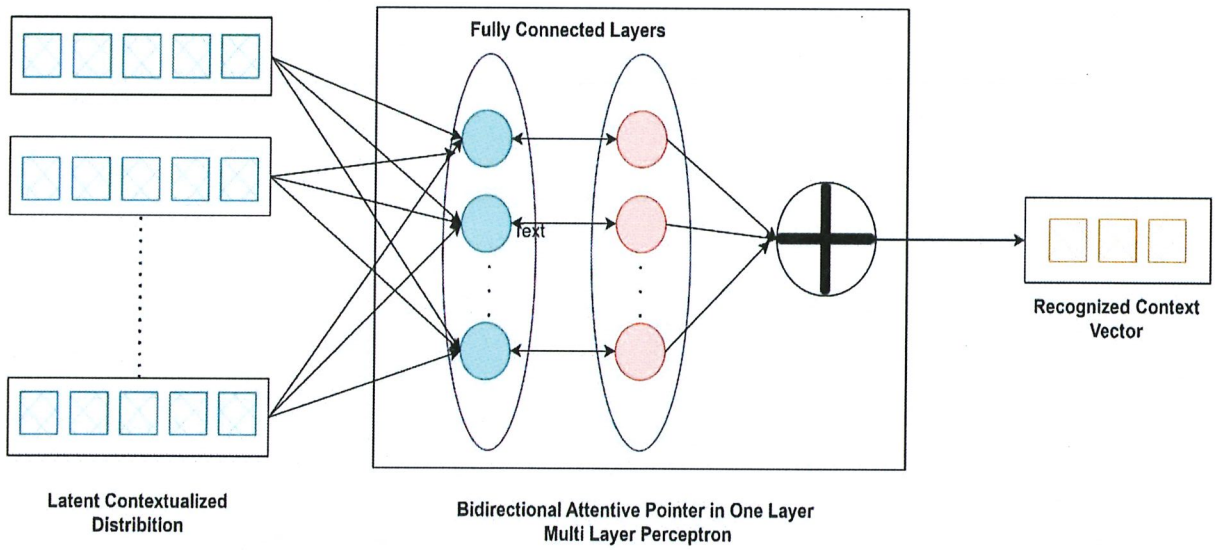


Fig. 4 Architecture of bidirectional attentive pointer mechanism

$$v_{jt} = tanh\left(W_{word}H_{jt} + v_w\right) \tag{2}$$

$$\beta_{jt} = \frac{\exp(v_{jt}^T v_w)}{\sum_t \exp(v_{jt}^T v_w)} \tag{3}$$

$$k_j = vector\ open \tag{4}$$

The sentence annotation is fed through a one-layer MLP network to get the sentence prominence vector (v_j) from the hidden representation (H_j) and sentence context vector(v_s). The sentence context vector v_s is thought of as an upper demonstration of a query over sentences. During the training procedure, v_s is initialized randomly and trained jointly. Then, the importance of the sentence is measured as a similarity of v_j with a normalized prominence weight β_j through the softmax function. The document vector (d) is then computed as a weighted sum of the sentence annotations depending on the prominence weights β_j and hidden representation H_j. Bidirectional attentive pointer mechanism for sentences is explained in Eqs. (5–7),

$$v_j = \tanh(W_{sent}H_j + v_s) \tag{5}$$

$$\beta_j = \frac{\exp(v_j^T v_s)}{\sum_j \exp(v_j^T v_s)} \tag{6}$$

$$d = \sum_j \beta_j H_j \tag{7}$$

The overall process flow of the proposed Hierarchical Attention pointer based Stacked Denoising Variational Autoencoder Model is shown in Fig. 5.

%3CmxGraphModel%3E%3Croot%3E%3CmxCell%20id%3D%220%22%2F%3E%3CmxCell%20id%3D%221%22%20parent%3D%220%22%2F%3E%3CmxCell%20id%3D%222%22%20value%3D%22%26lt%3Bb%26gt%3B%26lt%3Bfont%20style%3D%26quot%3Bfont-size%3A%2018px%3B%26quot%3B%20face%3D%26quot%3BTimes%20New%20Roman%26quot%3B%26gt%3BPre-Processed%26lt%3Bbr%26gt%3Bweb%20Text%26amp%3Bnbsp%3B%26lt%3Bbr%26gt%3Bfrom%20CBoW%26lt%3Bbr%26gt%3BModel%26lt%3B%2Ffont%26gt%3B%26lt%3B%2Fb%26gt%3B%22%20style%3D%22rounded%3D0%3BwhiteSpace%3Dwrap%3Bhtml%3D1%3B%22%20vertex%3D%221%22%20parent%3D%221%22%3E%3CmxGeometry%20x%3D%22-210%22%20y%3D%22,110%22%20width%3D%22,120%22%20height%3D%22,130%22%20as%3D%22geometry%22%2F%3E%3C%2FmxCell%3E%3C%2Froot%3E%3C%2FmxGraphModel%3E.

The two-phase operation of the proposed model is useful in extracting features from contextualized distribution including the words with multiple meanings and semantic characteristics in which the SVDAE model generates smooth latent distribution of contextualized words and sentences with extraction of useful features as well as attention mechanism is adopted to produce context level vector in order to get the information about the contextualized web text with extracting keywords to determine the meaning of sentences and words in web text. Furthermore, after feature extraction, ranking and

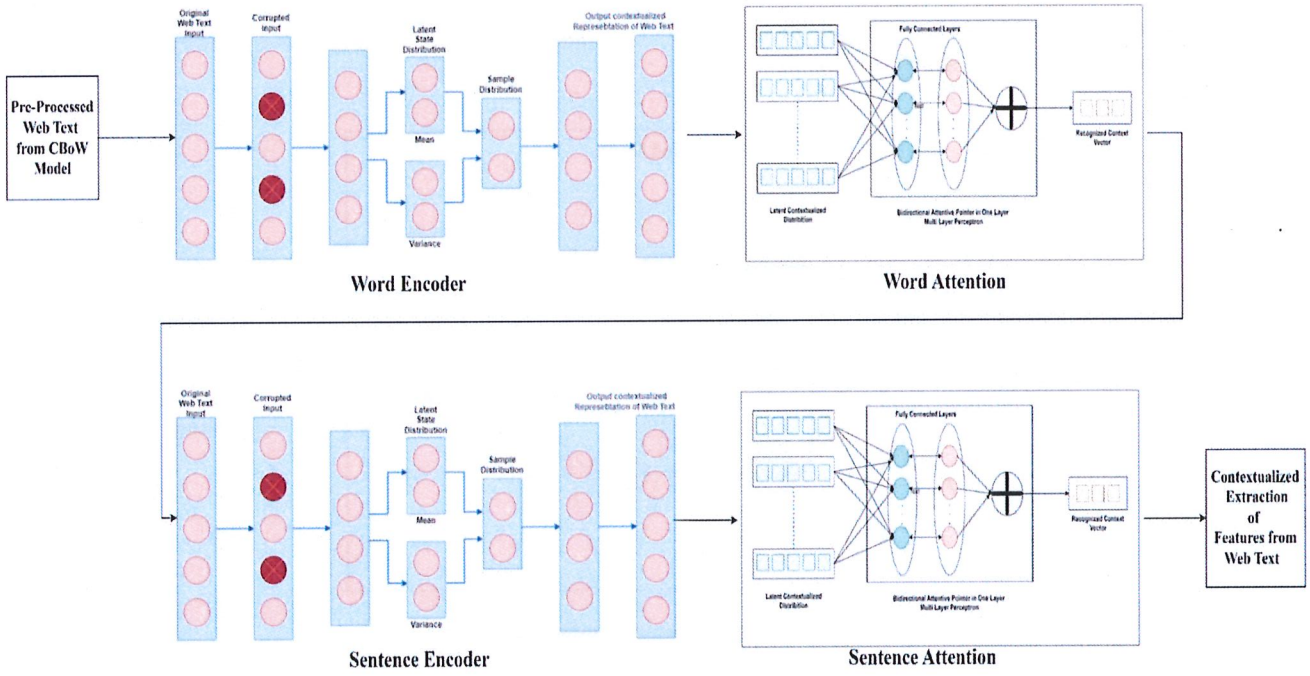


Fig. 5 Architecture of hierarchical attention pointer based stacked denoising variational autoencoder model

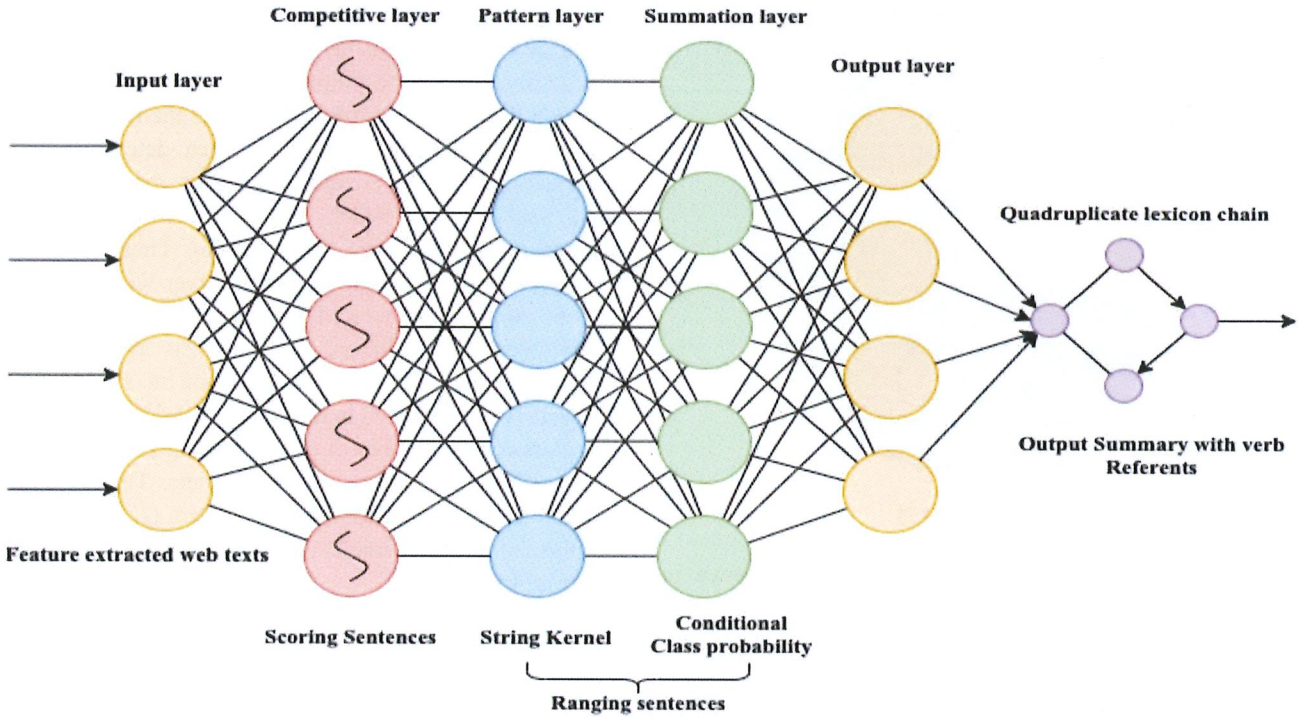


Fig. 6 Ranking and scoring in multilayered competitive probable modular perception model with summary generation using graph-based quadruplicate lexicon chain

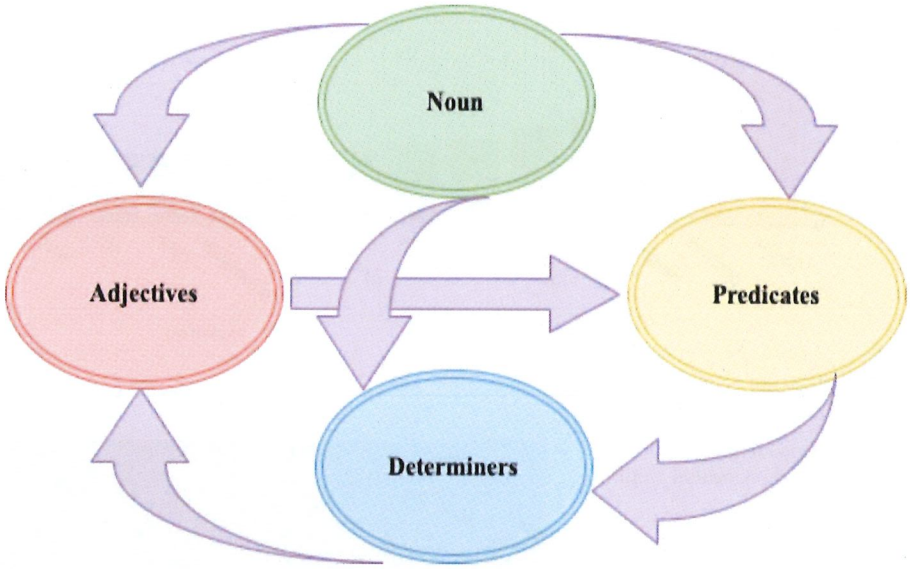


Fig. 7 Graph-based quadruplicate lexicon chains

scoring have to be performed to produce the summary which is explained in the next subsection.

4.3 Multilayered Competitive Probable Modular Perception Model Graph-Based Quadruplicate Lexicon Summarization

The sentence-level context vector formed from the hierarchical model is taken as the input to the Multilayered Competitive Probable Modular Perception Model. This proposed model effectively scores and ranks the sentences based on proximity, singularity, cue words, word frequency, sentence position, and length value. Then, verb referents were checked in these ranked sentences without dangling anaphora using Graph-based Quadruplicate Lexicon Summarization. The process takes place in a multilayered competitive probable modular model and Graph

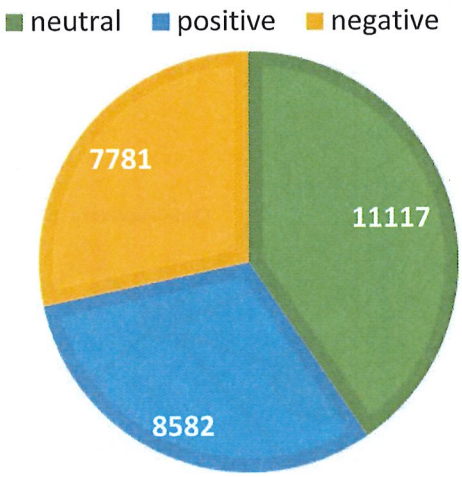


Fig. 8 Polarity of sentiments in the dataset

based Quadruplicate lexicon summarization is shown in Fig. 6.

The input layer processes the feature and keyword-extracted web text and sends the sentence-level context vector to the competitive layer. The competitive probability layer is responsible for scoring the sentences. In this competitive layer, the similarity and proximity measure between the individual input vector and the neuron weight for each competitive neuron has been determined. The neurons in this layer compete with each scoring for each input vector to see which one is the most similar to it thereby effectively score the sentences. Then, the scored sentences were ranked using pattern layer with string kernel and summation layer with conditional probability. The magnitudes of classmates' string kernels are sorted and rated in ascending order. Only a part of the string kernels competes to be included in the class-conditional probability calculation. The number of selected string kernels in class-a is designated as the number of leaders. The ratio μ_a is defined as the ratio of number of leaders (L_a) to the total number of string kernels in each class-$a(C_a)$ which is expressed in Eq. (8):

$$\mu_a = \frac{L_a}{|C_a|} \rightarrow L_a = [\mu_a |C_a|] \tag{8}$$

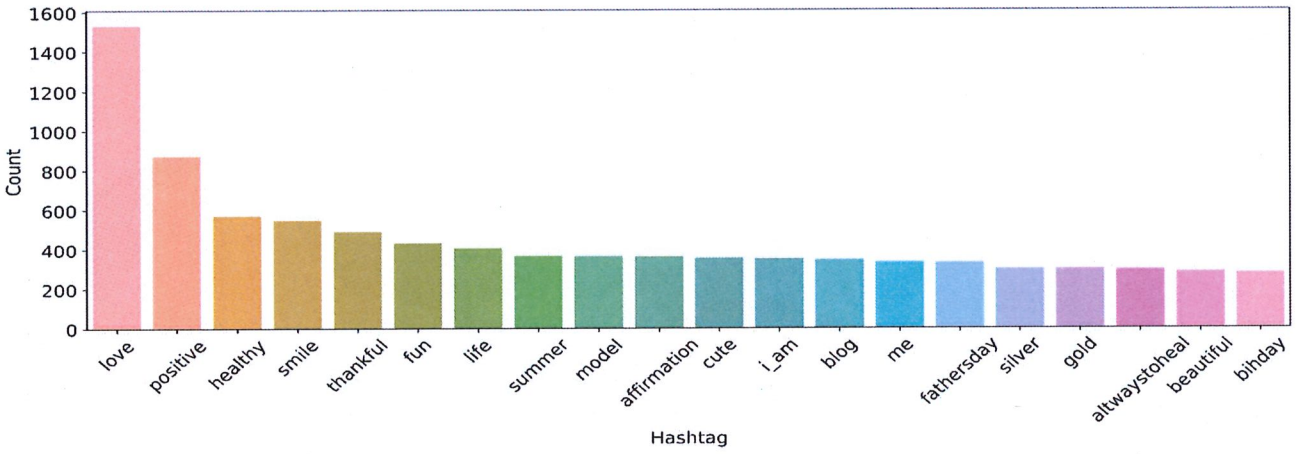


Fig. 9 Keyword count from hashtag

Fig. 10 Smooth latent representation of neutral word features

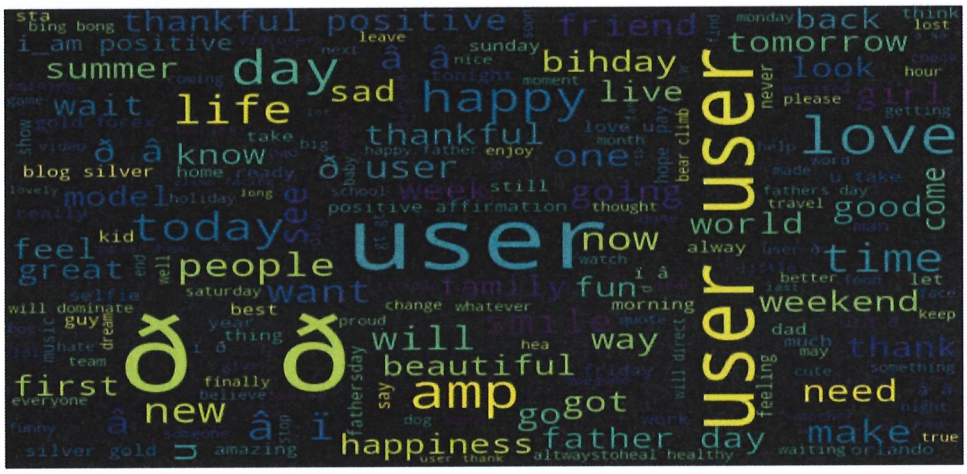




Fig. 11 Extracted keyword features from vocabulary in reviews

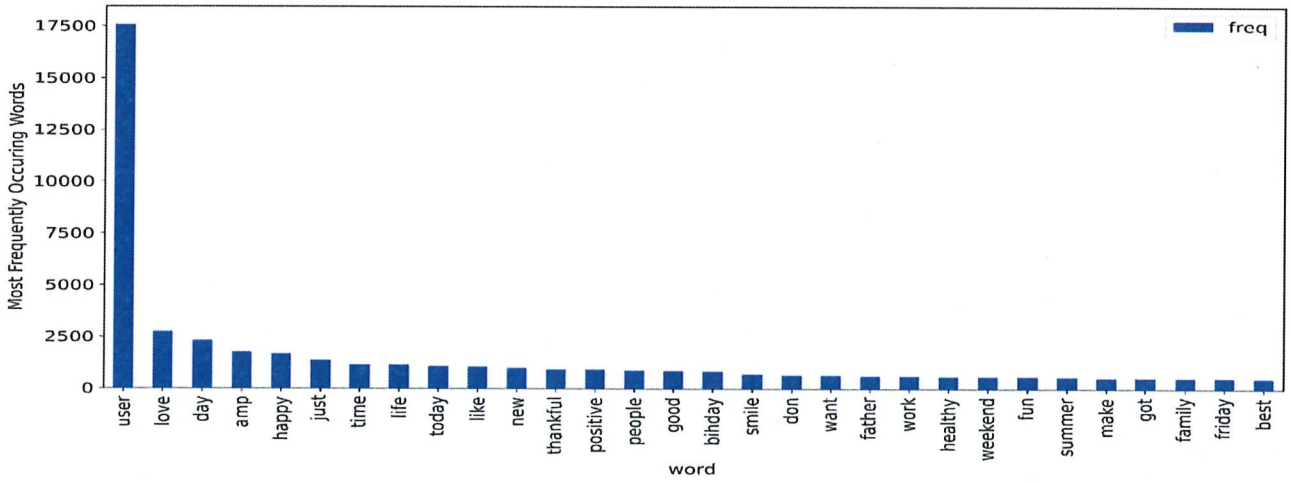


Fig. 12 Scoring and ranking based on similarity and proximity

This ratio μ_a is normalized and minimized which is expressed in Eq. (9):

$$\mu_{amin} = \frac{1}{|C_a|} \leq \mu_a \leq 1 \tag{9}$$

The subscript is a discrete variable with a finite number of possible values within the limit 1. The class-conditional probability P_a of each class is calculated using this Eq. (10):

$$P_a = \frac{1}{L_a} \sum_{k=1}^{L_a} W_{ak} = \frac{1}{[\mu_a|C_a|]} \sum_{k=1}^{[\mu_a|C_a|]} W_{ak} \tag{10}$$

where, W_{ak} is the weight distribution parameter in the classes a and k.

Based on these calculations, the class-conditional probability ranks the scored sentences to form the summary. These scored sentences were given to the graph-based Quadruplicate Lexicon Summarization mechanism in which the lexicon chain is formed with the quadruplicates of nouns, adjectives, determiners, and predictors in a graphical structure which is shown in Fig. 7.

The lexicon chain is formed in a graph with the consideration of noun, predicates, determiners, and adjectives in a sentence to form the summary without dangling anaphora expressions since these quadruplicates form the major grammatical parts of the sentence. Noun candidate words, determiner candidate words, predicate candidate words, and adjective candidate words are the quadruplicate parts of lexical chains' candidate words. All of the words were first steamed. For noun candidate terms, choose all named entities that describe the text's topics. Predicate candidate words are made up of all predicates in each phrase. Each phrase has a single predicate that serves as both the root of the dependency tree and a representation of the attribute that a subject possesses. As adjective (adverb) candidate terms, all noun-qualifying adjectives and adverbs were picked. Determiner words that were used before the nouns are selected as determiner candidate words. The quadruplicate lexicon chain is expressed in Eq. (11) as:

$$QLC = <NChain, DChain, PChain, AChain> \quad (11)$$

Equation (11) includes noun, determiner, predicate, and adverb terms in a quadruplicate lexicon chain format. The graphic structure created in Fig. 7 is an undirected graph, where each word's occurrence is represented by a vertex, and the semantic relationship between two occurrences is represented by an edge. If a word occurs in more than one sense, find and save the most pertinent one. Remove all additional senses and their boundaries. In the traverse semantic graph, words connected by edges create a lexical chain. All building lexical chains contain the final lexical chains. These chains provide sentences with correct meaning and grammatical form without cataphoric and anaphoric expressions. The overall algorithm for web text summarization model has been explained as follows:

Algorithm 1 Algorithm of proposed web text summarization model

Input: *Source web text document*

Output: *Summary of web text*

Initialize *web text summarization model*

do

{

Get *sentences S from web text document D*

/ perform stop word removal, lemmatization, and tokenization as preprocessing steps*/*

for *sentences in D*

stop words = set (stop words. words(S))

tokens= do linguistic analysis (S)

word= word tokenize(S)

for *words in S*

lemmatization= porter stemmer. Stem (W) */*W is words*/*

end

end

Generate *vectorized words (v_w) using the CBoW model*

Construct *contextualized representation of W and S in encoding process*

Determine *the probability of contextualized representation using eqn (1)*

Extract *features and keywords using eqn (2) - (7)*

for *scoring sentences in D*

S[i] = word frequencies[word]+singularity [words]

end

Rank *scored sentences using eqn (8) – (10)*

Generate *extractive summary with correct grammatical form using eqn (11)*

end

Overall, Extractive Web Text Summarization using CBoW Text Vector Model has been presented to generate a summary for web text by preprocessing and vectorizing the web text. Then, features were extracted to understand the meaning of all words and sentences in the web text. Finally, the summary has been generated by efficient ranking with the construction of graph graph-based quadruplicate lexicon chain. The next section explains the result obtained from the proposed model and discusses it in detail.

5 Experiment Result and Discussion

This section includes a thorough analysis of the performance of the proposed system, the implementation results simulated in the Python platform, and a comparison section to make sure the suggested system is appropriate for web text summarizing.

5.1 Dataset Description

Twitter sentiments-text summarization dataset[1] consists of 940 tweets annotated by 22 human annotators with sentiments in positive polarity, negative polarity, and neutral. This dataset is organized with 27,482 rows with 4 columns for describing the text ID, text, selected text, and sentiments. Where the first column the text ID, contains the Twitter ID of the text. The second column Text, presents the original tweet of the text while the third column, selected text, contains only the selected text from the tweets. The fourth column, sentiments, contains positive, neutral, and negative polarity sentiments. Figure 8 depicts the 27,482 rows categorized into three polarity of the sentence 11,117 neutral tweets, 8583 positive tweets, and 7782 negative tweets.

5.2 Experimental Results

Web text data has been preprocessed, vectorized, feature extracted, and summarized using novel techniques, and the results obtained were described in this section. The Dataset was split into 25% for testing a 75% training set.

Figure 9 depicts the important keyword count from the Hashtags in the Twitter sentiments-text summarization dataset. The Twitter text was preprocessed by removing the unwanted stop words, extracting stem words, and tokenization. Then, the CBoW text vectorization model was applied which transforms the text into vectors and forms a distributed combination of words with the most useful keyword counts as shown in Fig. 9.

The smooth latent representation of neutral words in the Twitter reviews were shown in Fig. 10. This contextualized distribution of words was obtained using a Stacked Denoising Variational Autoencoder which extracts all possible meanings of words and performs a denoising mechanism to represent the text without noise.

Figure 11 depicts the keyword features extracted from the vocabulary in reviews. The bidirectional attentive pointer mechanism in the hierarchical model extracts the important keywords and forms sentence vectors in a bidirectional pattern. This mechanism focuses on the keywords from vocabulary in reviews by understanding the meaning of words based on contextualized sentence which is shown in Fig. 11.

The scoring and ranking of words in sentences are depicted in Fig. 12. The words in the sentences were scored based on the similarity and proximity of their occurrence as shown. These scored words in the sentence were ranked by using a pattern layer with string kernel and class-conditional probability mechanism. Then, a quadruplicate lexicon chain is constructed to represent the summary finally.

5.3 Performance Metrics

Precision: The number of positively predicted cases that were truly accurate is known as precision. This formula is used to compute it:

$$Pricision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{12}$$

Recall: It assesses a model's capacity to locate every pertinent instance within the dataset. This formula is used to compute it:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{13}$$

F1-Score: The harmonic mean of recall and precision is the F1 score. It is an effective statistic for striking a balance between precision and recall because it incorporates both into one. This formula is used to compute it:

$$F1_{Score} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{14}$$

A wider measure called accuracy evaluates how accurate a model's predictions are overall. This formula is used to compute it:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{15}$$

[1] https://www.kaggle.com/competitions/tweet-sentiment-extraction/data.

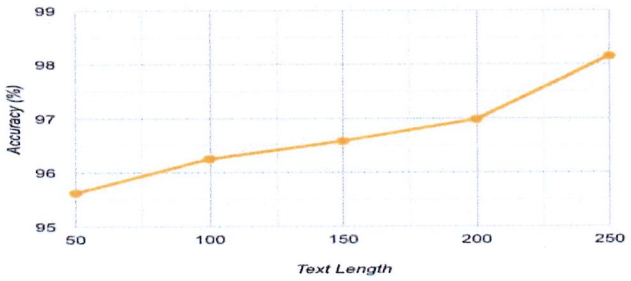


Fig. 13 Accuracy of proposed model based on text lengths

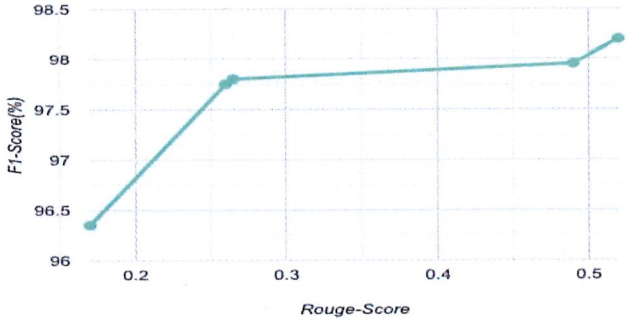


Fig. 15 Rouge score of proposed web text model

ROUGE-N: Calculates how much the generated text and the reference text overlap in terms of n-grams, or word sequences. Unigrams (single words) are referred to as ROUGE-1, bigrams (two-word sequences) as ROUGE-2, and so on. ROUGE-L determines the generated text's longest common subsequence (LCS) from the reference text. It considers how long the longest common subsequence is.

The performance of the proposed approach and the achieved outcome were explained in detail in this section.

The accuracy of the proposed model by varying the text length of the tweet data is demonstrated in Fig. 13. The accuracy attains a maximum value of 98.30% while increasing the text length. The accuracy of the proposed system attains a minimum value of 95.63% while decreasing the length of text. The accuracy of the proposed system is increased by using Multilayered Competitive Probable Modular Perception Model and Graph-based Quadruplicate Lexicon Summarization that effectively rank and generates summary with class conditional probability mechanism and graphical lexicon chains.

Figure 14 depicts the F1-Score of the proposed model and F1-Score was calculated for varying the text length from 50 to 250. The proposed model has a maximum F1-Score of 98.4% and a minimum F1-Score value of 96.2%. The F1-Score of the proposed model increases with the increase in the length of text. The F1-Score of the proposed web text model is improved by a Hierarchical Attention pointer based Stacked Denoising Variational Autoencoder Model which extracts keyword features accurately with a bidirectional attention mechanism. The recall of the proposed web text model by varying the text length of the tweet data is also depicted. The recall attains a maximum value of 98.8% while increasing the text length. The recall of the proposed system attains a minimum value of 97.5% while decreasing the length of text. The recall of the proposed system is increased by using the Multilayered Competitive Probable Modular Perception Model and Graph-based Quadruplicate Lexicon Summarization that accurately generates summary with class conditional probability mechanism and quadruplicate lexicon chains. The overall precision of the proposed model and the precision was calculated by varying the text length from 50 to 250. The proposed model has a maximum precision of 98% and a minimum precision value of 95.3%. The precision of

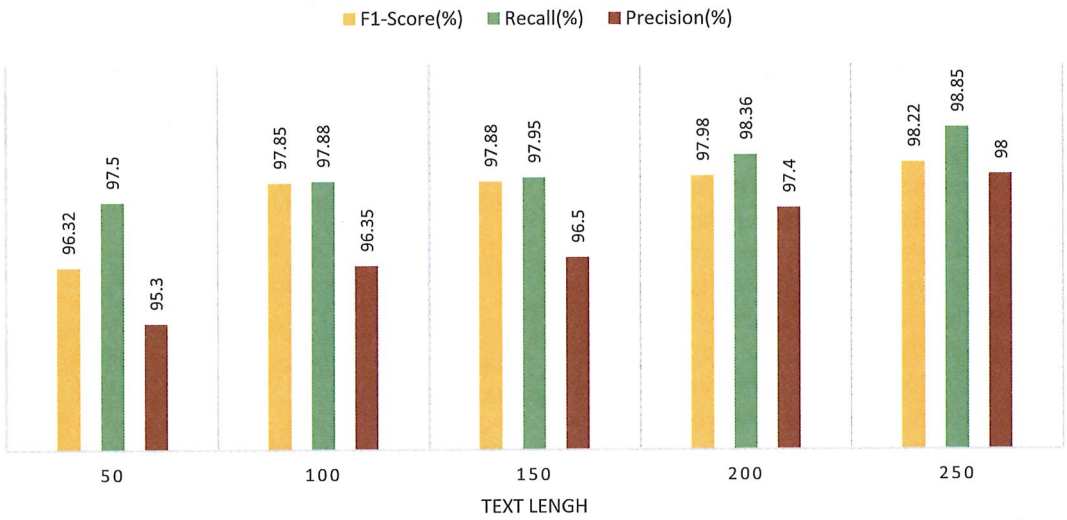


Fig. 14 Recall, F1-Score, precision of over text length of the proposed work

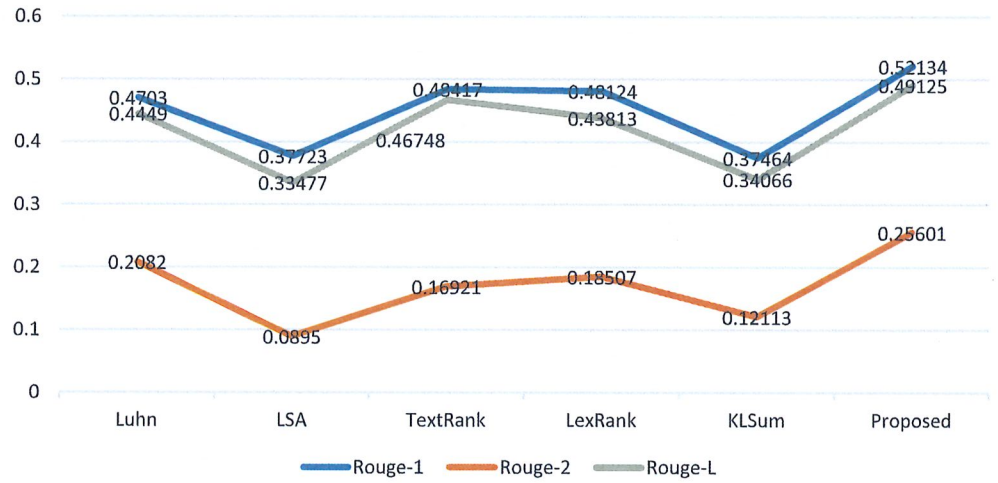


Fig. 16 Comparison of proposed model Rouge-1, Rouge-L and Rouge-2 with the existing approaches in terms of F-measure

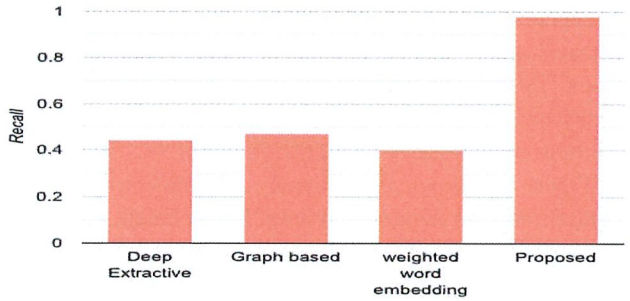


Fig. 17 Comparison of recall

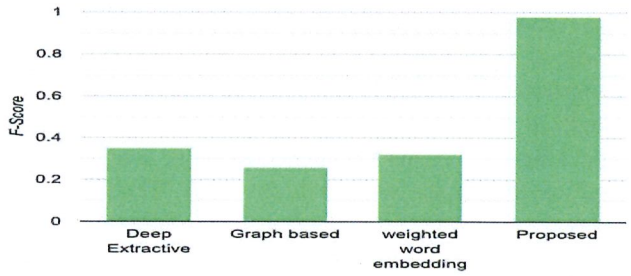


Fig. 19 Comparison of F-score

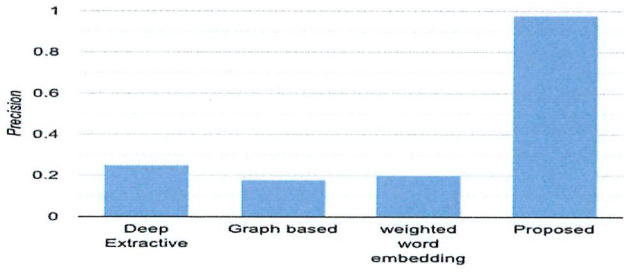


Fig. 18 Comparison of precision

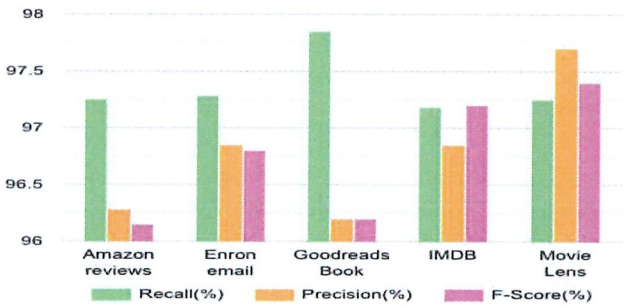


Fig. 20 F1-Score comparison with various datasets

the proposed model increases with the increase in the length of the text. The precision of the proposed model is improved by the Multilayered Competitive Probable Modular Perception Model that precisely scores the sentences based on proximity and singularity using the Competitive probable layer.

The Rouge score of the proposed web text model in terms of F1-Score has been shown in Fig. 15. The Rouge score has the highest value of 0.54 when the F1-score is 98.4% and has the lowest value of 96.2%. The rouge score increases with the increase in the F-score. The rouge-score and F-Score were improved by using the Hierarchical Attention pointer-based Stacked Denoising Variational Autoencoder Model.

5.4 Comparison Results of the Proposed Method

This section highlights the proposed model performance by comparing it to the outcomes of existing approaches and showing their results based on various metrics.

Figure 16 depicts the comparison of rouge-1 that is rouge score of unigram in terms of F-measure. The rouge-1 of the proposed model is compared with existing techniques such as Luhn (Luhn Apr. 1958), LSA (Landauer et al. 1998), TextRank (Mallick, et al. 2019), LexRank (Erkan and Radev 2004), and KLSum (Haghighi and Vanderwende 2009). The rouge-1 of the proposed system has the maximum value of 0.524 whereas the rouge-1 of Luhn, LSA, TextRank, LexRank, and KLSum is 0.472, 0.375, 0.472, 0.426, and 0.375 respectively. The rouge-1 score of the proposed model is high whereas the rouge-1 score of LSA and KLSum is low. Figure 16 depicts the comparison of rouge-2 which is the rouge score of bigram in terms of F-measure. The comparison of rouge-2 of the proposed system with existing techniques such as Luhn, LSA, TextRank, LexRank, and KLSum. The rouge-2 of the proposed model attains the maximum value of 0.3 whereas the rouge-2 of existing techniques such as Luhn, LSA, TextRank, LexRank, and KLSum are 0.22, 0.7, 0.17, 0.18, and 0.13 respectively. Hence, the rouge-2 of the proposed model is high whereas the rouge-2 of LSA is low. The comparison of Rouge-L which is the rouge score of word length in terms of F-measure is also discussed in Fig. 16. The rouge- L of the proposed model is compared with existing techniques such as Luhn, LSA, TextRank, LexRank, and KLSum The rouge- L of the proposed system has the maximum value of 0.484 whereas the rouge L of Luhn, LSA, TextRank, LexRank and KLSum are 0.446, 0.342, 0.470, 0.433 and 0.345 respectively. The rouge- L score of the proposed model is high whereas the rouge- L score of LSA is low.

Figure 17 shows a comparison of the recall of the proposed model with existing techniques such as deep extractive (Bhargava and Sharma 2020), Graph based (Belwal et al. 2021) and weighted word embedding (Rani and Lobiyal 2021). The recall of the proposed system attains a maximum value of 98% whereas the recall of deep extractive, Graph based, and weighted word embedding are 44%, 47%,and 40% respectively. Hence, the proposed web text model has the highest recall, whereas weighted word embedding has the lowest recall.

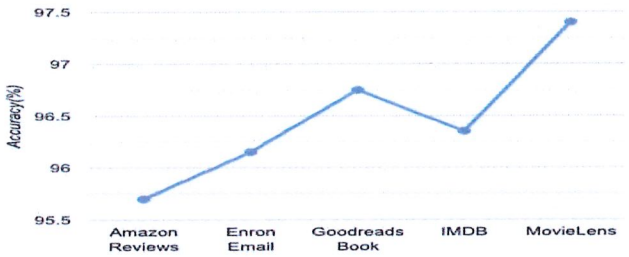


Fig. 21 Accuracy comparison with various datasets

Figure 18 shows a comparison of the precision of the proposed model with existing techniques such as deep extractive, Graph based,and weighted word embedding. The precision of the proposed system attains a maximum value of 98% whereas the precision of deep extractive, Graph based and weighted word embedding are 25%, 18% and 20% respectively. Hence, the proposed web text model has the highest precision, whereas graph based has the lowest precision.

The comparison of the F-measure of the proposed system with existing techniques such as deep extractive, Graph based, and weighted word embedding is shown in Fig. 19. The F-measure of the proposed model attains the maximum value of 98% whereas the F-measure of existing techniques such as deep extractive, Graph based, and weighted word embedding are 35%, 26%, and 32% respectively. Hence, the F-measure of the proposed model is high whereas the F-measure of Graph based is low.

Figure 20 shows the F1-score comparison of various datasets such as Amazon reviews, Enron email, Goodreads Book, IMDB, and MovieLens. It is noticed that the proposed web text model performs well with all datasets with a minimum of 96.15% and a maximum of 97.4% F1-Score.

Figure 21 shows the accuracy comparison of various datasets such as Amazon reviews, Enron email, Goodreads Book, IMDB, and Movie Lens. It is noticed that the proposed web text model performs well with all datasets with a minimum of 96.2% and a maximum of 97.5% accuracy. Figure 20 shows the recall comparison of various datasets such as Amazon reviews, Enron email, Goodreads Book, IMDB, and Movie Lens. It is noticed that the proposed web text model performs well with all dataset with a minimum of 97.18% and maximum of 97.85% recall. Figure 20 shows the precision comparison of various datasets such as Amazon reviews, Enron email, Goodreads Book, IMDB, and Movie Lens. It is noticed that the proposed web text model performs well with all dataset with a minimum of 96.2% and a maximum of 97.7% precision.

Overall, Extractive Web Text Summarization using CBoW Text Vector Model outperforms existing techniques such as Luhn, LSA, TextRank, LexRank, KLSum deep extractive, Graph based, and weighted word embedding by processing text quickly with a continuous bag of words and extracting features accurately with understanding the context using a hierarchical model. Then, the summary is generated using competitive and class-conditional probability mechanisms. Thus, the result achieved has a high rouge score of 0.524, precision of 98%, recall of 98%, and F-score of 98%.

6 Conclusion

In this article, Extractive Web Text Summarization using CBoW Text Vector Model has been introduced to solve the issues such as huge processing time, difficulty in extracting polysemous features, and dangling anaphora in the summary generation of web text data. The vectorized word distribution is obtained using the CBoW text vectorization model and contextualized sentence distribution with consideration of all possible word meanings is extracted using Hierarchical Attention pointer based Stacked Denoising Variational Autoencoder Model and finally, the summary is generated using Multilayered Competitive Probable Modular Perception Model and Graph-based Quadruplicate Lexicon Summarization. The result obtained from Extractive Web Text Summarization using the CBoW Text Vector Model outperforms existing techniques with a high rouge score of 0.524, precision of 98%, recall of 98%, and F-score of 98%. As part of our future work, researchers can take the opportunity to explore these problems for abstractive summarization.

References

Abualigah L, Bashabsheh MQ, Alabool H, Shehab M (2020) Text summarization: a brief review. In: Elaziz MA, Al-Qaness MAA, Ewees AA, Dahou A (eds) Recent advances in NLP: the case of Arabic language. Springer, Cham, pp 1–15

Alami N, Meknassi M, En-nahnahi N (2019) Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. Expert Syst Appl 123:195–211

Al-Maleh M, Desouki S (2020) Arabic text summarization using deep learning approach. J Big Data 7(1):1–17

Belwal RC, Rai S, Gupta A (2021) A new graph-based extractive text summarization using keywords or topic modeling. J Ambient Intell Human Comput 12(10):8975–8990

Bhargava R, Sharma Y (2020) Deep extractive text summarization. Procedia Comput Sci 167:138–146

Chatterjee N, Yadav N (2019) Hybrid latent semantic analysis and random indexing model for text summarization. In: Fong S, Akashe S, Mahalle P (eds) Information and communication technology for competitive strategies. Springer, Singapore, pp 149–156

Doğan E, and Kaya B (2019) Deep learning based sentiment analysis and text summarization in social networks. In: 2019 International artificial intelligence and data processing symposium (IDAP). IEEE

Elbarougy R, Behery G, El Khatib A (2020) Extractive Arabic text summarization using modified PageRank algorithm. Egypt Inform J 21(2):73–81

El-Kassas WS et al (2020) EdgeSumm: graph-based framework for automatic text summarization. Inform Process Manag 57(6):102264

El-Kassas WS et al (2021) Automatic text summarization: a comprehensive survey. Expert Syst Appl 165:113679

Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479

Gangathimmappa M, Subramani N, Sambath V, Ramanujam RAM, Sammeta N, Marimuthu M (2023) Deep learning enabled cross-lingual search with metaheuristic web based query optimization model for multi-document summarization. Concurr Computat: Pract Exp 35(2):e7476

Ghadimi A, Beigy H (2023) SGCSumm: an extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks. Expert Syst Appl 215:119308. https://doi.org/10.1016/j.eswa.2022.119308

Ghodratnama S et al (2020) Extractive document summarization based on dynamic feature space mapping. IEEE Access 8:139084–139095

Haghighi A, and Vanderwende L (2009) Exploring content models for multi-document summarization. In: Proceedings of human language technologies. The 2009 annual conference of the North American chapter of the association for computational linguistics

Hernández-Castañeda Á et al (2020) Extractive automatic text summarization based on lexical-semantic keywords. IEEE Access 8:49896–49907

Hossain Md, Hoque MM (2020) Semantic meaning based Bengali web text categorization using deep convolutional and recurrent neural networks (DCRNNs). In: Misra R, Kesswani N, Rajarajan M, Bharadwaj V, Patel A (eds) International conference on internet of things and connected technologies. Springer, Cham

Joshi A, Fidalgo E, Alegre E, Fernández-Robles L (2023) DeepSumm: exploiting topic models and sequence to sequence networks for extractive text summarization. Expert Syst Appl 211:118442

Kanapala A, Pal S, Pamula R (2019) Text summarization from legal documents: a survey. Artif Intell Rev 51(3):371–402

Kumar Y, Kaur K, Kaur S (2021) Study of automatic text summarization approaches in different languages. Artif Intell Rev 54(8):5897–5929

Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Discourse Process 25(2–3):259–284. https://doi.org/10.1080/01638539809545028

Li S, Pan R, Luo H, Liu X, Zhao G (2021) Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling. Knowl-Based Syst 218:106827

Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Dev 2(2):159–165. https://doi.org/10.1147/rd.22.0159

Luo Z, Xie Q, and Ananiadou S (2023) Chatgpt as a factual inconsistency evaluator for abstractive text summarization. arXiv preprint arXiv:2303.15621

Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, Guo L et al (2023) ImpressionGPT: an iterative optimizing framework for radiology report summarization with chatGPT. arXiv preprint arXiv:2304.08448

Madhuri JN, and Kumar RG (2019) Extractive text summarization using sentence ranking.In: 2019 International conference on data science and communication (IconDSC). IEEE

Magdum PG, Rathi S (2021) A survey on deep learning-based automatic text summarization models. In: Chiplunkar NN, Fukao T (eds) Advances in artificial intelligence and data engineering. Springer, Singapore, pp 377–392

Mallick C et al (2019) Graph-based text summarization using modified TextRank. In: Nayak J, Abraham A, Krishna BM, Sekhar GTC, Das AK (eds) Soft computing in data analytics. Springer, Singapore, pp 137–146

Manjari KU et al. (2020) Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm.In: 2020 4th International conference on trends in electronics and informatics (ICOEI)(48184). IEEE

Mao X et al (2019) Extractive summarization using supervised and unsupervised learning. Expert Syst Appl 133:173–181

Mao, Yuning, et al. (2020) Multi-document summarization with maximal marginal relevance-guided reinforcement learning. arXiv preprint arXiv:2010.00117

Mishra AR, Panchal VK, and Kumar P (2019) Extractive text summarization-an effective approach to extract information from Text. In: 2019 International conference on contemporary computing and informatics (IC3I). IEEE

Patel D, Shah S, Chhinkaniwala H (2019) Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. Expert Syst Appl 134:167–177

Rani R, Lobiyal DK (2021) A weighted word embedding based approach for extractive text summarization. Expert Syst Appl 186:115867

Shini RS, and Kumar VDA (2021) Recurrent neural network based text summarization techniques by word sequence generation. In: 2021 6th International conference on inventive computation technologies (ICICT). IEEE, 2021.

Shirwandkar NS, and Kulkarni S (2018) Extractive text summarization using deep learning. In: 2018 Fourth international conference on computing communication control and automation (ICCUBEA). IEEE

Siautama R, Suhartono D (2021) Extractive hotel review summarization based on TF/IDF and adjective-noun pairing by considering annual sentiment trends. Procedia Comput Sci 179:558–565

Song S, Huang H, Ruan T (2019) Abstractive text summarization using LSTM-CNN based deep learning. Multimed Tools Appl 78(1):857–875

Steinberger J, Poesio M, Kabadjov MA, Ježek K (2007) Two uses of anaphora resolution in summarization. Inf Process Manag 43(6):1663–1680

Suleiman D, and Awajan AA. (2019) Deep learning based extractive text summarization: approaches, datasets and evaluation measures. In: 2019 Sixth international conference on social networks analysis, management and security (SNAMS). IEEE

Uçkan T, Karcı A (2020) Extractive multi-document text summarization based on graph independent sets. Egypt Inform J 21(3):145–157

Wang D et al (2013) Comparative document summarization via discriminative sentence selection. Trans Knowl Discov Data 7(1):2

Zhou Q et al (2020) A joint sentence scoring and selection framework for neural extractive document summarization. IEEE/ACM Trans Audio Speech Lang Process 28:671–681

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 Physics A (2024) 130:748
https://doi.org/10.1007/s00339-024-07902-z

Structural and spectroscopic features of single-phase Dy^{3+} activated $BiYWO_6$ phosphor for luminescent device applications

Indrajeet Maurya[1] · **M. Jayasimhadri**[1]

Received: 1 July 2024 / Accepted: 9 September 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
Dysprosium-activated bismuth yttrium tungstate (Dy^{3+}: $BiYWO_6$) phosphors have been synthesized in an ambient atmosphere using a co-precipitation (CP) method. X-ray diffraction (XRD), field emission scanning electron microscope (FE-SEM) with energy dispersive X-ray (EDX), X-ray photoelectron spectroscopy (XPS), and photoluminescence (PL) measurement techniques were employed to examine the structural, compositional, morphological, elemental composition as well as their oxidation states, and photoluminescent characteristics of the Dy^{3+}: $BiYWO_6$ phosphors, respectively. PL studies of the Dy^{3+} activated $BiYWO_6$ phosphor manifests the CIE color coordinates phosphor falls in near white and yellow regions under 290 and 455 nm wavelengths, respectively. The average decay time (τ_{avg}) for the $^4F_{9/2}$ energy state of Dy^{3+} ions has been found to be in microseconds. The PL spectra measured from room temperature to 448 K ensure good thermal stability of the phosphor. Thus, the synthesized phosphor might be a probable contender for UV/blue excited phosphor-converted white LEDs (pc-WLEDs).

Keywords Rare-earths · X-ray photoelectron spectroscopy · Structural · Luminescence and optical properties

1 Introduction

In the modern era, power saving is essential for reducing the amount of climate change caused by carbon-based gas emissions, produced during the generation of electric power from burning fossil fuels, which can be quantified in terms of both financial worth and environmental cost [1–3]. The majority of power is utilized for both interior and outdoor artificial illumination. The advancements in artificial lighting have been remarkable, commencing with the utilization of oil lanterns and progressing to the development of incandescent and fluorescent lamps, ultimately culminating in the current era of solid-state lighting devices. The increasing adoption of solid state lighting can be attributed to its capacity to provide vibrant illumination, facilitate the emergence of novel lighting paradigms, and contribute to investigating various regions within the electromagnetic spectrum [4]. These days, WLEDs are feasibly the most popular option for interior and exterior illumination. Researchers and scientists have been concentrating on the latest generation of phosphor-based WLEDs to enhance the quality of white light emission, decrease power consumption, and increase financial and environmental benefits. Due to their advantageous chemical, thermal, electrical, and optical abilities, phosphors are particularly suitable to use in pc-WLEDs, solid-state lasers, display devices, medical equipment, lighting devices, as well as other applications. In this regard, rare earth ion- activated inorganic pc-WLEDs have been widely studied. These LEDs offer several advantages such as compact, small size, energy-saving, good material stability, environmental protection, high luminous efficiency, fast switching, and long service life (> 100 000 h) [3, 5, 6]. Though pc-WLEDs have many advantages, they have some drawbacks to restrict their usage in a wide range of applications. There are many ways to make commercial pc-WLEDs, and one of the popular methods to make WLEDs is to combine a blue light emitting InGaN chip with yellow emitting trivalent cerium activated phosphor (YAG: Ce^{3+}) [7–9]. However, owing to the inadequacy of red characteristics, these LEDs have severe drawbacks such as highly correlated colour temperature as well as low colour rendering

✉ M. Jayasimhadri
jayasimha@dtu.ac.in

[1] Luminescent Materials Research Lab (LMRL), Department of Applied Physics, Delhi Technological University, Delhi 110 042, India

Published online: 24 September 2024

index. An alternative way to use UV or blue LED chips in emulsion with multicolour phosphors. Nevertheless, the integration of a UV/blue emitting LED chip through yellow and red glowing phosphors leads to different degradation rates of individual phosphors, inadequate colour temperature stability, phase separation, and significant reabsorption of emitted emission through adjacent phosphors [2, 10]. In order to address these challenges, there has been a notable focus on the advancement of a single-phase UV/blue pumped white illuminating phosphor that offers adjustable emission colour, enhanced luminous output along with low CCT, improved thermal and chemical stability. As a result, single RE ion-based monophase white light emitting phosphors are currently the subject of extensive research.

Selecting an appropriate environmental friendly host material for dispersing Dy^{3+}ions is a critical factor in the process. The crystal structure, morphology influenced by synthesis, and other related factors significantly affect the luminescent material's performance [11–13]. However, high luminous efficiency in commercial WLEDs is accompanied by certain adverse effects, such as the generation of hazardous gases and other toxic by-products during the synthesis process of the phosphors [14]. In this regard, the luminescent, photocatalytic, and optical properties of oxide-based tungstate host materials have gathered substantial interest among researchers due to their non-toxic characteristics, moderate synthesis requirements, cost-effectiveness, environmental friendly nature, exemplary charge stabilization, and elevated thermal, chemical, and hydrolytic stability [15–17]. Additionally, tungstate host material along with bismuth and yttrium can offer rare earth ions a sufficient ligand field environment [18]. Furthermore, the tetrahedron WO_4 exhibits flexibility to provide different coordination environments through modulation of the W–O bond lengths [19].

The prime objective of the present study is to synthesize a series of new Dy^{3+} activated BYW phosphors using the CP route, and their comprehensive investigation was undertaken to examine the crystallinity, structural attributes, particle size as well as morphology analysis. The phase analysis of the synthesized phosphor has been investigated by the diffraction patterns and Rietveld refinement analysis. The morphological and elemental analysis of the host and $BiYWO_6$: xDy^{3+}(x = 3.0 mol%) phosphor using FE-SEM micrograph and EDX measurement. The photoluminescent (PLE and PL), PL decay, chromaticity characteristics as well as temperature dependent emission profiles were thoroughly investigated in detail. Henceforth, all the above-mentioned findings may validate the functionality of Dy^{3+} activated $BiYWO_6$ phosphor to use as an auspicious candidate in the realm of UV/blue excited photonic applications.

2 Experimental section

2.1 Materials and synthesis

Dy^{3+} activated $BiYWO_6$ phosphors and $BiYWO_6$ host material were synthesised via co-precipitation technique. The starting precursors $Y(NO_3)_3.6H_2O$, $Na_2WO_4.2H_2O$, $Bi(NO_3)_3.5H_2O$, and Eu_2O_3 were taken in the stoichiometric ratio. The initial precursors $Y(NO_3)_3.6H_2O$ and $Na_2WO_4.2H_2O$ were dissolved in an appropriate amount of highly pure deionized water while $Bi(NO_3)_3.5H_2O$ and Eu_2O_3 are mixed in nitric acid. The second step involves thoroughly homogenize the precursor solution for a duration of 1 h. The mixed solution has pH ∽ 1, which was further neutralized with ammonia solution and maintained at pH ∽ 7 by continuously stirring for four hours at 60 °C. After a certain time period, the transparent solution has turned into a precipitate. Further, the precipitate was separated via a centrifugation process and dried using hot air oven at 120 °C for 24 h. The sample was ground for 10–15 min and calcined at 900 °C temperature for 2 h to form the final product.

2.2 Characterization techniques

The purity of the phase and crystal structure of the prepared samples BYW: xDy^{3+} (x = 0.0, to 5.0 mol%) were investigated at the scanning rate of 5 °/minute with step size 0.04 using high-resolution X-ray diffractometer (Rigaku, Smartlab) using *Cu-Kα* radiation (3 kilo Watt). The morphology and elemental analysis of the samples BYW and BYW: 3.0 mol% Dy^{3+} samples were observed via FE-SEM (Zeiss, Model-Supra 55). The composition and chemical states of BYW: Dy^{3+} phosphor were analyzed via XPS (PHI, Model-5000 versa probe III). The absorption spectra of Dy^{3+} doped BYW samples were investigated with the spectrophotometer (Jasco, V770). The excitation and emission profiles of the samples were captured using a spectrofluorophotometer (Shimadzu RF-5301PC) at room temperature. An Edinburgh FLS900 was used for the decay measurements with a Xenon flash lamp. Ocean optics spectrophotometer (FLAME-S-XR1-ES) was used to record temperature dependent PL (TDPL) spectra.

3 Results and discussion

3.1 XRD structural analysis

Structural and phase purity of the prepared series of Dy^{3+} (x = 0.0, 0.5, 1.0, 2.0, 3.0, 4.0 and 5.0 mol%) induced $BiYWO_6$ samples were analyzed, and compared with the

standard diffraction pattern (JCPDS card no.: 33–0224) as depicted in Fig. 1(a).

The diffraction pattern corresponding to the host material $BiYWO_6$ was completely matched with standard data without any shifting. Moreover, the diffraction patterns for Dy^{3+} doped $BiYWO_6$ samples were absolutely analogous with the standard data. Hence, the pure crystalline monoclinic structure with space group P2/a of $BiYWO_6$ has been attained without any impurity, which implies Dy^{3+} dopant ions were effectively dispersed in the $BiYWO_6$ host sample. $BiYWO_6$ contains alternating aurivillius-type $[BiYO_2]^{2+}$ layers, which are inserted between $[WO_4]^{2-}$ perovskite layers. According to the described crystal structure, the Y^{3+} ions are dispersed evidently at random positions of Bi^{3+} within the $BiYWO_6$ unit cell [20, 21].

The crystallite size (D) of the as-synthesized BYW: xDy^{3+} (x- 0.0, 0.5, 1.0, 2.0, 3.0, 4.0 and 5.0 mol%) powder samples were calculated using Debye-Scherer's equation, which comprises the location of diffraction peaks and their corresponding full width at half maxima as follows:

$$D = K\lambda / \beta\, Cos\theta \tag{1}$$

where X-ray wavelength ($\lambda = 1.54$ Å, *Cu-Kα* radiation), Bragg's diffraction angle (θ), shape factor ($K = 0.94$), and β is full-width at half maximum are the parameters used to estimate the crystallite size of the samples. With an upsurge in the content of Dy^{3+} ions, the average crystallite size slightly increases and is realized in the range of 30–50 nm.

Rietveld refinement has been performed via Fullprof suite software for the prepared undoped and doped (3.0 mol% Dy^{3+}) $BiYWO_6$ samples. The resultant refinement patterns were depicted in Fig. 1(b, c), which illustrates the experimental and evaluated patterns along with the difference between them. The assessed result discloses that the undoped and doped (3.0 mol% Dy^{3+}) $BiYWO_6$ samples having single monoclinic phase with the lattice parameters, a = 8.1235 Å, b = 3.7535 Å, c = 16.0351 Å and a = 8.1246 Å, b = 3.7538 Å, c = 16.0352 Å, respectively. The goodness of fit was found to be close to the unity denotes good refinement results.

3.2 Morphological studies

The morphology is an important factor influencing both the performance and the luminescence efficiency of the phosphor material. The FE-SEM micrograph of the as-synthesized $BiYWO_6$ host material and $BiYWO_6$: xDy^{3+} (x = 3.0 mol%) phosphor with nano order resolution were depicted in Fig. 2(a & b). FE-SEM images clearly show that the grain sizes are distributed in the range of 100–200 nm diameter for undoped and Dy^{3+} doped samples and no significant changes in the morphology with the doping ions. The micrographs illustrate the size, irregular and non-uniform distribution of closely packed particles of the powder samples, and grains tend to agglomerate. The sub-micron range crystalline particles are appropriate to gain efficient luminescence and are suitable for solid-state lighting applications.

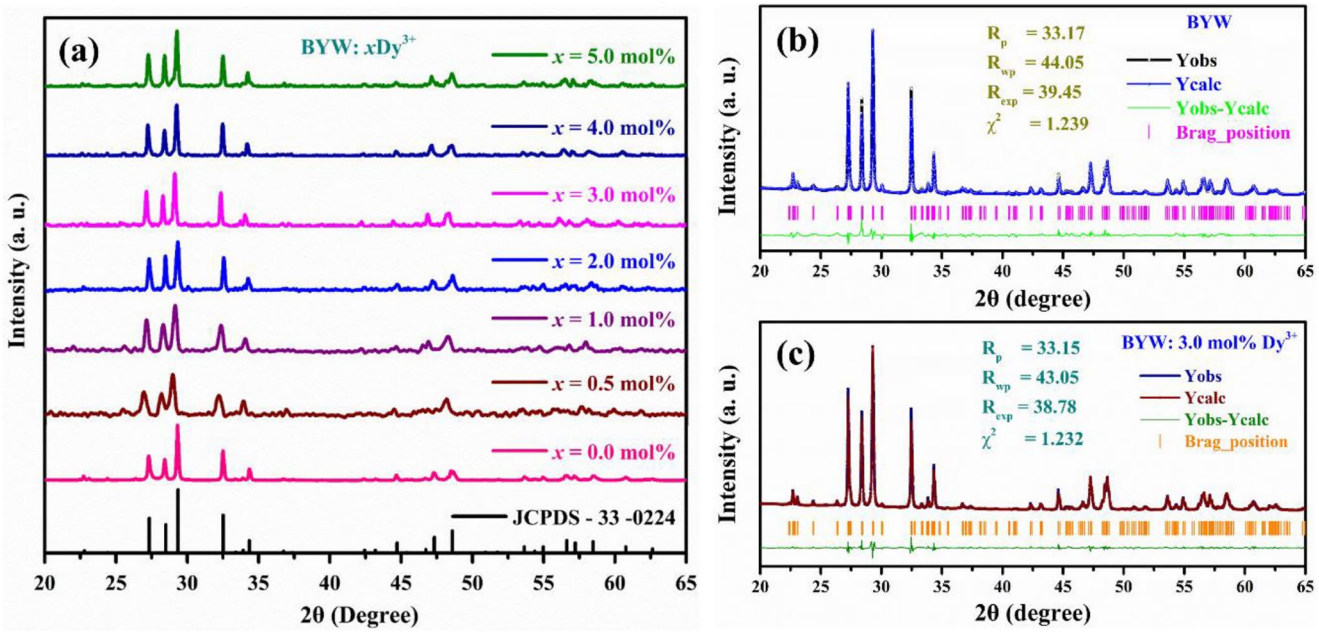
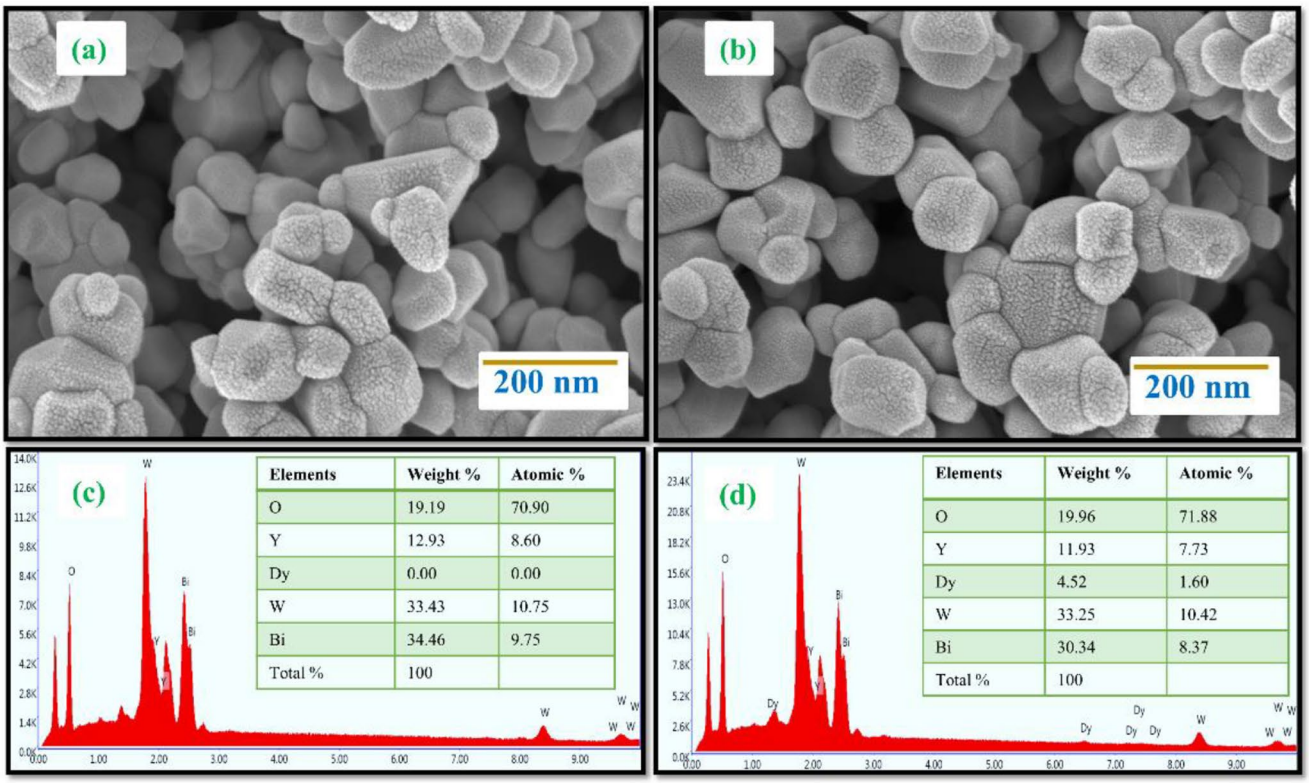


Fig. 1 (**a**) X-ray diffraction patterns of Dy^{3+} doped $BiYWO_6$ samples. The Rietveld refinement profiles of (**b**) $BiYWO_6$, (**c**) 3.0 mol% Dy^{3+} doped $BiYWO_6$ phosphor synthesized via CP route



Elements	Weight %	Atomic %
O	19.19	70.90
Y	12.93	8.60
Dy	0.00	0.00
W	33.43	10.75
Bi	34.46	9.75
Total %	100	

Elements	Weight %	Atomic %
O	19.96	71.88
Y	11.93	7.73
Dy	4.52	1.60
W	33.25	10.42
Bi	30.34	8.37
Total %	100	

Fig. 2 FE-SEM images of (**a**) host $BiYWO_6$ (**b**)3.0 mol% Dy^{3+} doped $BiYWO_6$ phosphor, EDX patterns of (**c**) host $BiYWO_6$ and (**d**) 3.0 mol% Dy^{3+} doped $BiYWO_6$ phosphor (Inset shows the table of elemental composition)

The elemental composition of doped and undoped $BiYWO_6$ was analysed with the help of EDX patterns, as illustrated in Fig. 2 (c & d). The EDX results fully disclose the presence of all compositional elements of the synthesized samples, such as bismuth (Bi), yttrium (Y), tungsten (W), oxygen (O), and dysprosium (Dy). Additionally, the inset table of Fig. 2 (c & d) represents the weight% and atomic% of each compositional element for Dy^{3+} doped $BiYWO_6$ sample, which confirms the presence of all elements along with doping ions (Dy^{3+}).

3.3 XPS

To study the additional clarification related to the composition of elements and their corresponding chemical/oxidation states present in the as-synthesized phosphor material, XPS spectra were demonstrated in Fig. 3. The whole XPS survey of the Dy^{3+}: $BiYWO_6$ phosphor has been displayed in Fig. 3(a). It reveals the existence of the primary elements Bi, Y, W, O, and Dy in Dy^{3+}: $BiYWO_6$ phosphor and does not show any additional impurity peaks. Furthermore, the chemical states of the individual components in the Dy^{3+}: $BiYWO_6$ phosphor are represented by the high-resolution XPS spectra as depicted in Fig. 3 (b-e). The XPS spectrum of Bi 4f & Dy 4d attributed to Bi^{3+} and Dy^{3+}, which is responsible for two peaks at 158.06 and 163.40 eV associated with the $4f_{7/2}$ and $4f_{5/2}$, respectively, as revealed in Fig. 3(b). The dual peaks at the location of 158.24 and 163.60 of the XPS spectrum of Y 3d, which correspond to $3d_{5/2}$ and $3d_{3/2}$, respectively, are indicative of yttrium in the Y^{3+} state as illustrated in Fig. 3(c). Figure 3(d) shows that O 1s peaks at 528.96, 530.14, and 530.96 eV. These peaks are caused by surface adsorbed water molecules (O–H) and metal-oxygen bonds seen in the XPS survey profile of Dy^{3+}: $BiYWO_6$[22, 23]. Further, the W 4f exhibits two peaks at 34.34 and 36.42 eV for $4f_{7/2}$ and $4f_{5/2}$, which represent W^{6+} state of $BiYWO_6$ as shown in Fig. 3(e) [24, 25]. In view of the above, the data confirms that Dy^{3+}: $BiYWO_6$ is present in the as-synthesized phosphor.

3.4 Diffuse reflectance spectral analysis

The diffuse reflectance spectra displayed from 200 to 1000 nm range to provide insight about the optical properties of the samples. Figure 4(a & b) illustrates the diffuse reflectance spectra of BYW: xDy^{3+} (x = 0.0 and x = 3.0 mol%) phosphor. Figure 4(b) signifies that the appearance of new absorption bands beyond 700 nm can be signified by the characteristic peaks of Dy^{3+} ions in the $BiYWO_6$. Both samples demonstrate pronounced absorption bands within the

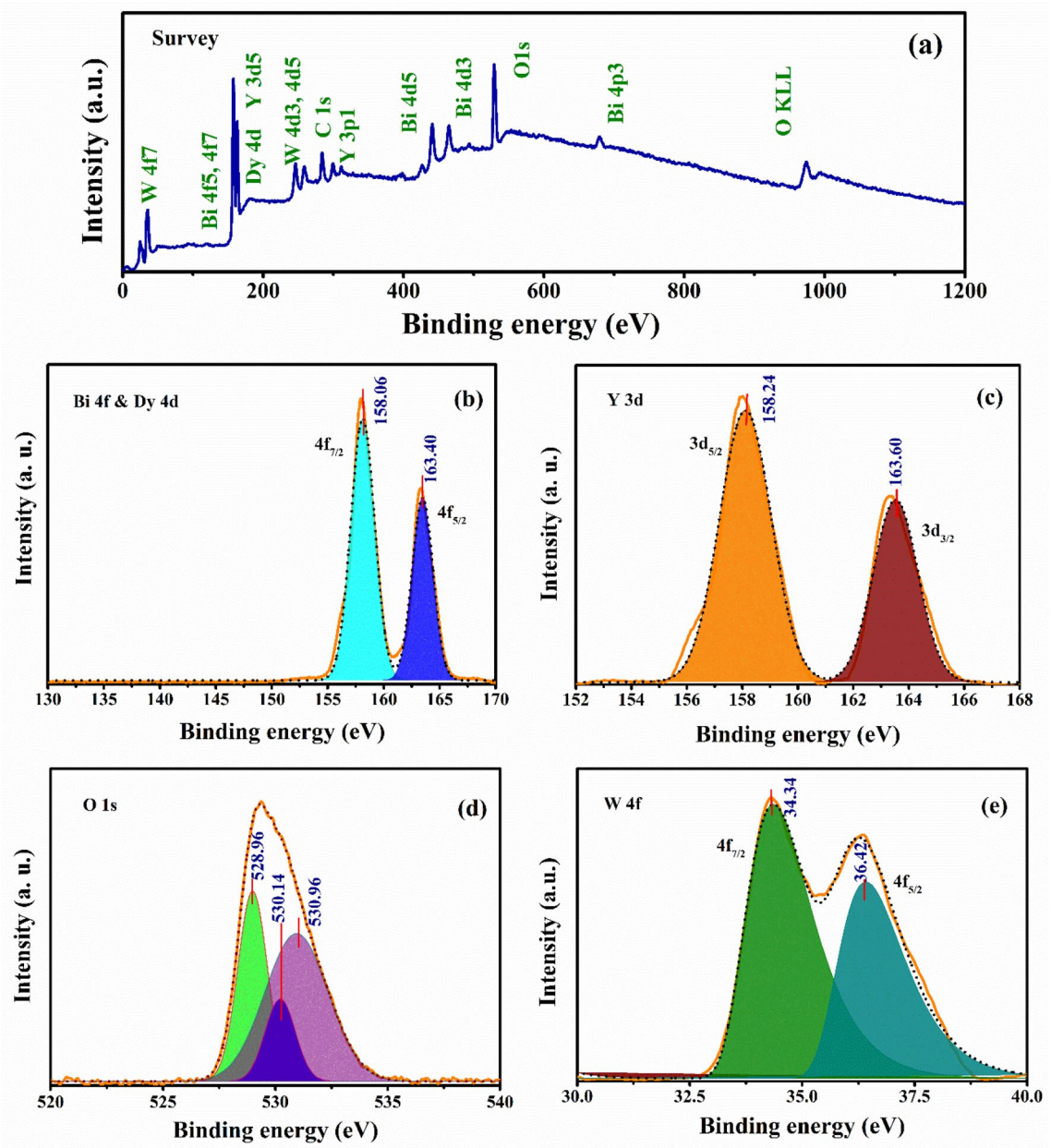


Fig. 3 X-ray photoelectron spectroscopy (XPS) of (**a**) survey, (**b**) Bi, (**c**) Y, (**d**) O, (**e**) W of the Dy^{3+} doped $BiYWO_6$ phosphor

200–400 nm wavelength range, with the absorption being attributed to the transition between O^{2-} - W^{6+} in a charge transfer band [26, 27].

The estimation of the optical energy band gap of the as synthesized materials can be derived from the DR spectra. This estimation is based on the following relationship:

$$F(R) = \frac{(1-R)^2}{2R} = \frac{K}{S} \tag{2}$$

In the aforementioned equation, R (reflection), K (absorption) and S (scattering) constant, respectively. Moreover, it is possible to convert reflectance data into a reflection function ($F(R)$) by utilizing the subsequent mathematical expression:

$$[F(R)h\vartheta]^{1/n} = A(h\vartheta - E_g) \tag{3}$$

where the optical band gap is denoted as E_g, $h\vartheta$ represents the energy of photons, and A is a constant. Additionally, the index n corresponds to the various electronic transitions such as n value equal to ½ and 2 associated with the direct and indirect allowed transitions, while 3/2 and 3 stands for electronic transition corresponding to direct and

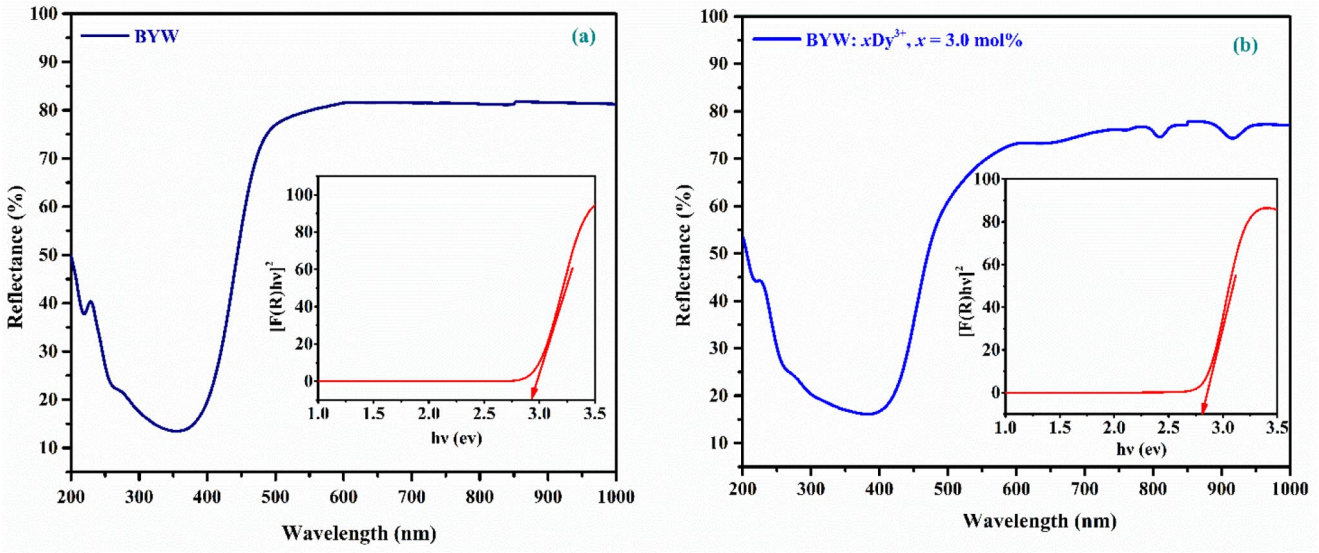


Fig. 4 Absorption spectra of (**a**) host material $BiYWO_6$ and (**b**) 3.0 mol% Dy^{3+} doped $BiYWO_6$ phosphor. The Inset plot reveals the assessment of the band gap by employing the Kubelka-Munk function

indirect forbidden, respectively. As a result, the determination of the optical band gap involves the linear relationship between $[F(R)h\vartheta]^{1/n}$ and $h\vartheta$ by extrapolating the linear graphical region to zero. For the electronic transition n = ½, the representation of fit for the graph depicted in the inset as Fig. 4(a & b) and the estimated band gap values are assessed to be 2.91 and 2.79 eV for the BYW host and Dy^{3+} doped BYW samples, respectively.

3.5 Photoluminescence studies

Figure 5(a) demonstrates the PLE spectrum of BYW: xDy^{3+} (x = 3.0 mol%) phosphor by fixing 575 nm emission wavelength. The excitation spectrum contains absorption peaks in a range of 270–500 nm, which contain characteristics of charge transfer band (CTB) as well as 4f electronic transitions of Dy^{3+} ions. The peaks obtained due to 4f electronic transitions are magnified in a 320–500 nm wavelength range, and revealed in the inset of Fig. 5(a). The excitation peaks are centred at 352, 366, 387, 427, 455, and 475 nm positions and are associated with the transitions from the ground state i.e. $^6H_{15/2}$, to various higher states viz. $^6P_{7/2}$, $^4I_{11/2}$, $^4F_{7/2}$, $^4G_{11/2}$, $^4I_{15/2}$ and $^4F_{9/2,}$ respectively in $BiYWO_6$ lattice [28, 29]. The prominent peaks were observed at 290 and 455 nm as depicted in excitation spectra. The intense broadband situated at 290 nm may be owing to an overlapping between $O^{2-} \rightarrow Dy^{3+}$ charge transfer band and transition of the $(WO_6)^{-6}$ group, indicating the strong covalency between O^{2-} and W^{6+} in the $BiYWO_6$ lattice. The next intense excitation band centered at 455 nm resembles the $^6H_{15/2} \rightarrow {}^4I_{15/2}$ electronic transition of the Dy^{3+} ions. Based on the PLE result, the excitation wavelengths, 290 and 455 nm are selected to record further emission characteristics.

Figure 5(b & c) show the emission spectra of BYW: xDy^{3+} (x = 0.5, 1.0, 2.0, 3.0, 4.0 and 5.0 mol%) phosphors recorded using 290 and 455 nm excitation wavelengths, respectively. Figure 5(b) depicts the emission spectra consisting of a broad emission band, and three emission bands were positioned at 485, 575, and 662 nm wavelengths, which are ascribed to the $^4F_{9/2} \rightarrow {}^6H_{15/2,}$ $^4F_{9/2} \rightarrow {}^6H_{13/2,}$ and $^4F_{9/2} \rightarrow {}^6H_{11/2}$ transitions, under 290 nm excitation wavelength.

In this emission spectra, a broad intense emission band was overlapped with the band positioned at 485 in the 450–500 nm region and a second intense band positioned at 575 in the 560–590 nm region (yellow), with dissimilar emission intensity. The blue emission observed in this study is attributed to the magnetic allowed dipole ($^4F_{9/2} \rightarrow {}^6H_{15/2}$) electronic transition of Dy^{3+} ions. This emission is found to be minimally affected by the strength of the crystal field around the dysprosium ions. On the other hand, the yellow emission is associated with the forced electric dipole ($^4F_{9/2} \rightarrow {}^6H_{13/2}$) transition, which follows the selection rule ΔJ = 2 [30, 31]. In the event that Dy^{3+} ions are positioned at a site of high symmetry featuring an inversion centre, it is expected that the intense transition shall be observed from the $^4F_{9/2}$ to the $^6H_{15/2}$ energy level. Conversely, if ions are located at a site of low symmetry and lacking an inversion centre, the predominant transition observed in the emission spectra will be the hypersensitive transition from $^4F_{9/2}$ to the $^6H_{13/2}$ energy level. According to this principle, Dy^{3+} ions potentially occupy the lower-symmetry sites of Bi^{3+} ions in the $BiYWO_6$ structure without any inversion centre due to their equal oxidation state and smaller ionic radii

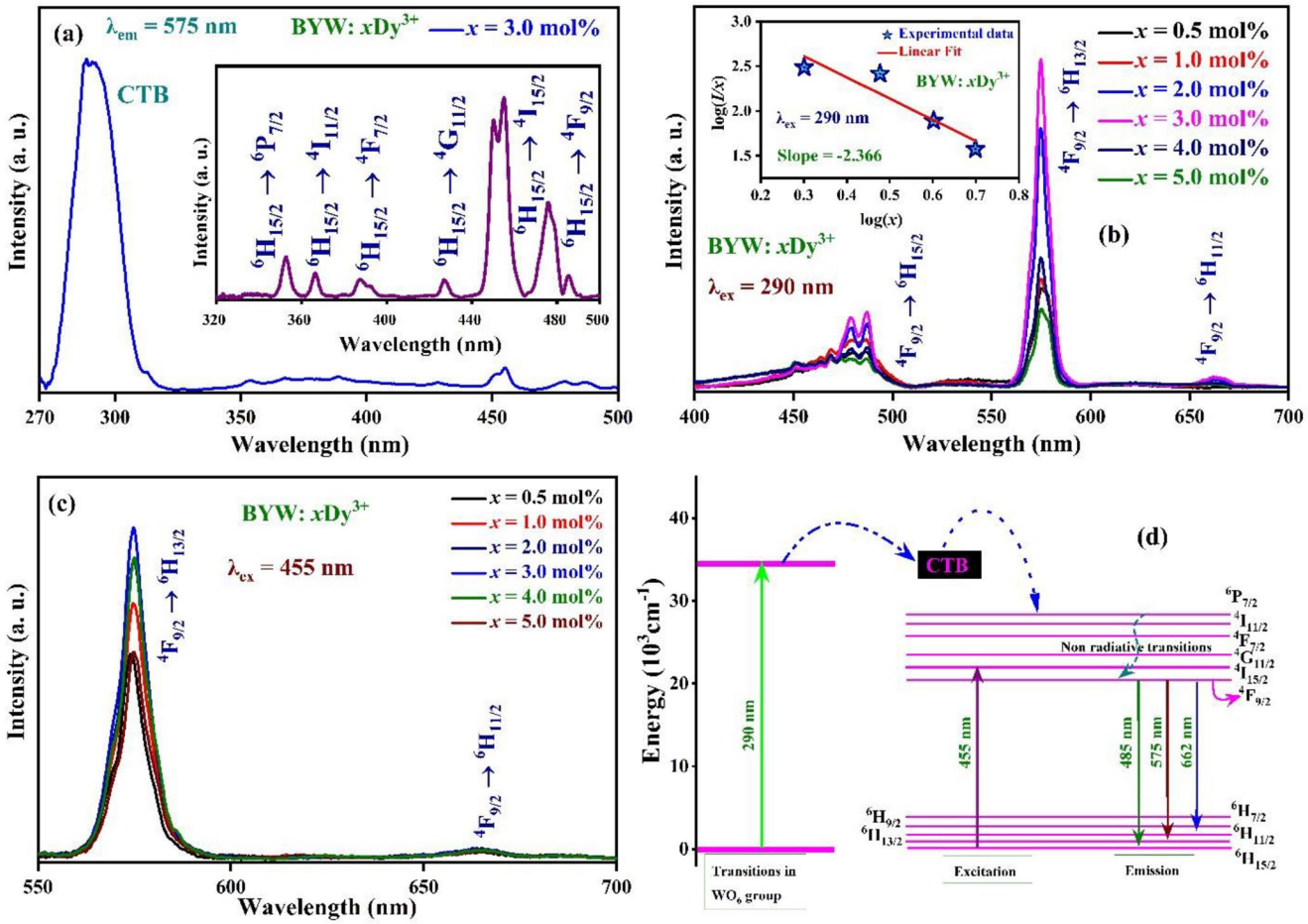


Fig. 5 (**a**) PLE spectrum of 3.0 mol% Dy^{3+} doped $BiYWO_6$ phosphor. (Inset plot shows the magnifying of the PL excitation in 320–500 nm wavelength range.), Emission spectra of Dy^{3+} doped $BiYWO_6$ under excitation wavelength of (**b**) 290 nm (Inset: Plot between log(*I*/*x*) vs. log(*x*) with varying concentrations of Dy^{3+} ions), (**c**) 455 nm, respectively, and (**d**) Partial energy level diagram of Dy^{3+} doped $BiYWO_6$ phosphor

compared to Bi^{3+} (coordination no. (N=6), ionic radii of Bi^{3+} = 1.03 Å and Dy^{3+} = 0.912 Å). Hence, forced electric transition (yellow) emission intensity remained to be stronger than magnetic dipole transition (blue emission intensity) in this BYW: xDy^{3+} phosphor. Whereas the emission spectra recorded using a 455 nm excitation wavelength consist of two emission bands, which are positioned at wavelength 575 and 662 nm corresponding to the $^4F_{9/2} \rightarrow {}^6H_{13/2}$ and $^4F_{9/2} \rightarrow {}^6H_{11/2}$ transitions of Dy^{3+} ions, respectively. In both cases, the emission intensity alters with Dy^{3+} ion content, as displayed in Fig. 5(b & c), and was realized to be highest at 3.0 mol% of Dy^{3+} ions, which is considered as optimized concentration for the $BiYWO_6$ host.

3.5.1 Dexter's theory

In order to examine the impact of dopant ions on the host lattice, the Dy^{3+} ions were systematically varied from x = 0.5 to 5.0 mol% under the excitation wavelength 290 nm, as depicted in Fig. 5(b). The graph illustrates that the emission peak intensity of $^4F_{9/2} \rightarrow {}^6H_{13/2}$ transition initially increases until reaching Dy^{3+} ion content of 3.0 mol%, which subsequently declines owing to concentration quenching. The observed spectacle can be ascribed to the increased proximity of Dy^{3+} - Dy^{3+} ions at higher doping concentrations, resulting in a prevalence of non-radiative energy transfer over radiative emission. Therefore, the non-radiative energy transfer is mostly influenced by the critical distance (R_c) corresponding to the adjacent Dy^{3+} ions within the host lattice. The equation was proposed by Blasse for determining the critical distance (R_c) [32]:

$$R_c \approx 2\left[\frac{3V}{4\pi X_C N}\right]^{1/3} \quad (4)$$

In this equation, V denotes the volume of the unit cell in the host lattice, while N represents the effective number of cations in the crystal lattice. Further, the symbol X_c is used

to denote the optimized concentration corresponding to Dy^{3+} ions dispersed in the host matrix. The critical distance has been found to be 19.63 Å, with the given parameters V=950.87 Å [3], X_c = 0.03 and N=8. Based on the Van Uitert concept, when the assessed value of R_c is significantly beyond the critical distance threshold (5Å), the process of energy transfer takes place via multipolar interaction [33]. In the current scenario, the value of R_c exceeds 5Å, thereby facilitating non-radiative energy transfer through multipolar interaction in agreement with Dexter's theory. This relationship can be expressed as the correlation between luminescence intensity (I), specific type of multipolar interaction (Q), and concentration of dopant ions (x) as follows:

$$\log\left(\frac{I}{x}\right) = A - \frac{Q log\,(x)}{3} \tag{5}$$

The values of Q, specifically 6, 8 and 10, are associated with different types of interactions such as electric dipole-dipole (d-d) interaction, dipole-quadrupole (d-q) and quadrupole-quadrupole (q-q) interactions, respectively [9]. The value of Q can be determined by evaluating the slope of linear regression line between the $\log\left(\frac{I}{x}\right)$ and $log\,(x)$. The inset of Fig. 5(b), illustrates the relation between ($\log\left(\frac{I}{x}\right)$ and $log\,(x)$. This graph is derived from the emission spectra of BYW: xDy^{3+} under the excitation of 290 nm wavelength. The slope of the curve is determined to be -2.366, and the obtained Q value is 7.908 (~8), which provides evidence supporting the hypothesis that the exchange of energy between neighbouring Dy^{3+} ions is primarily attributed to dipole-quadrupole interaction in the concentration quenching mechanism.

3.5.2 Energy level diagram

The process of observed excitation and emission corresponding to Dy^{3+} ions incorporated into BYW host material, represented by the partial energy level diagram is as disclosed in Fig. 5(d). The Dy^{3+} ions are stimulated to fascinating energy levels of CTB or excited state of Dy^{3+} ($^4I_{15/2}$) as shown with upward arrows in Fig. 5(d). When subjected to excitation wavelength 290 nm, the higher excited state of $(WO_6)^{-6}$ groups undergoes a non-radiative relaxation process to the intermediate state. Subsequently, energy is transferred from the intermediate state ($^4F_{9/2}$) to the numerous lower energy levels of Dy^{3+} ions. The energy gap between the $^4F_{9/2}$ energy level of Dy^{3+} ions and the next lower energy state is significantly higher. As a result, only radiative transitions (downward arrows) to lower energy levels of Dy^{3+} ions are possible, leading to emission in the blue, yellow and red regions. A similar mechanism is expected during 455 nm excitation, the energy is non-radiatively discharged to the $^4F_{9/2}$ energy state, followed by a subsequent radiative emission, as described in Sect. 3.5.

3.6 Chromaticity coordinates

To determine the colorimetric performance of the synthesized Dy^{3+} activated BYW phosphor, the chromaticity coordinates have been measured using the emission spectral data under UV/blue excitations, such as 290 and 455 nm, respectively. The colour coordinates (x, y) for the synthesized BYW: xDy^{3+}, x = 3.0 mol% phosphor is found to be (0.46, 0.37) and (0.42, 0.56) under the excitations 290 and 455 nm respectively, as demonstrated in Fig. 6. The estimated (x, y) coordinates for the synthesized BYW: 3.0 mol% Dy^{3+} phosphor falls in the yellow region under 455 nm and near the standard white light region under 290 nm wavelength [34]. Moreover, the correlated colour temperature (CCT) for the synthesized BYW: xDy^{3+}, x = 3.0 mol% phosphor has been computed using the McCamy empirical relation as given below [35]:

$$CCT = -449.0\,n^3 + 3525.0\,n^2 - 6823.0\,n + 5520.3 \tag{6}$$

in the above relation, $n = \frac{x-x_e}{y-y_e}$; $x_e = 0.332$ and $y_e = 0.186$ are the epicenter, respectively. The estimated CCT value for the synthesized BYW: xDy^{3+}, x = 3.0 mol% phosphor was observed to be 2305 and 4047 for the excitation wavelength 290 and 455 nm, respectively.

The observed values of CCT are lower than that of 5000 K, which suggests that the synthesized phosphors emit warm white light under UV/blue excitations. In view of the above, this phosphor material is effectively excited under a UV source with wavelengths of 290 nm as well as a blue source of 455 nm.

3.7 PL decay analysis

The decay curves of BYW: xDy^{3+} (x = 0.5 to 5.0 mol%) phosphors corresponding to the $^4F_{9/2}$ level were measured using 290 nm excitation wavelength by fixing the emission peak at 575 nm, as depicted in Fig. 7**(a)**.

All the BYW samples have exhibited the most accurate alignment decay curves with a bi-exponential function. The equation used to represent the bi-exponential fit for all the decay curves is as follows:

$$I\,(t) = I_0 + A_1 \exp\,(-t/\tau_1) + A_2 \exp(-t/\tau_2) \tag{7}$$

where the variables $I\,(t)$ and I_0 epitomize the luminescence intensities at time t and 0, respectively. The constants A_1 and A_2 are used for fitting purposes. In this exponential

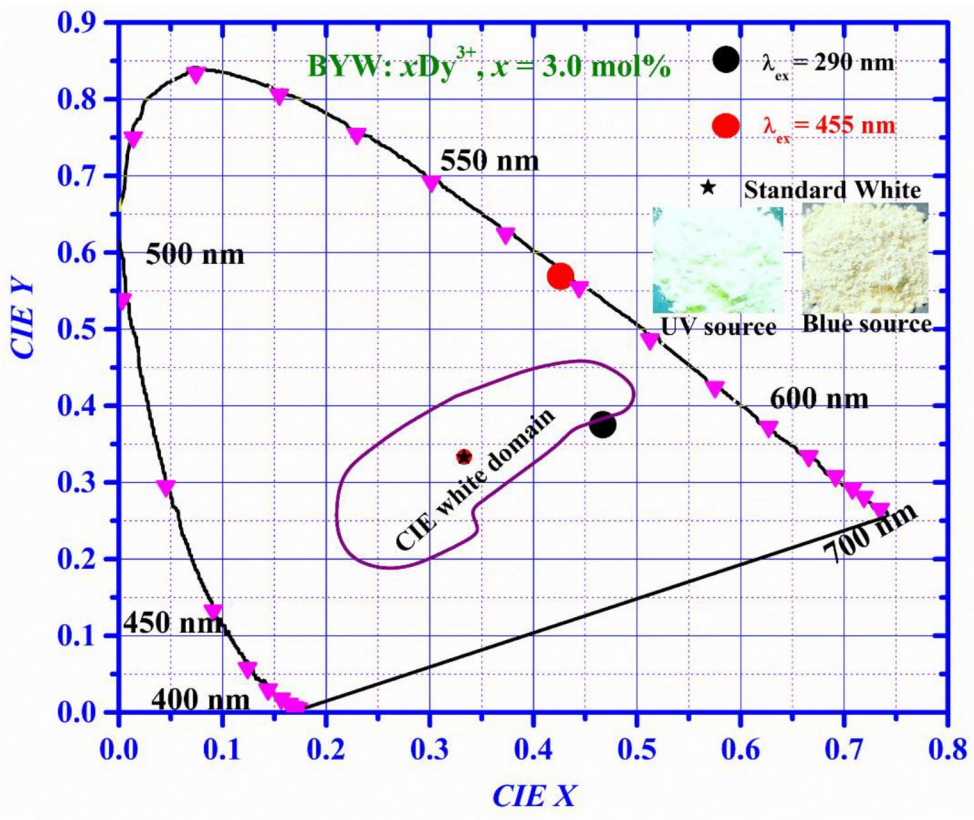


Fig. 6 CIE color coordinates of Dy^{3+} doped $BiYWO_6$ phosphor under excitation wavelength 290 and 455 nm

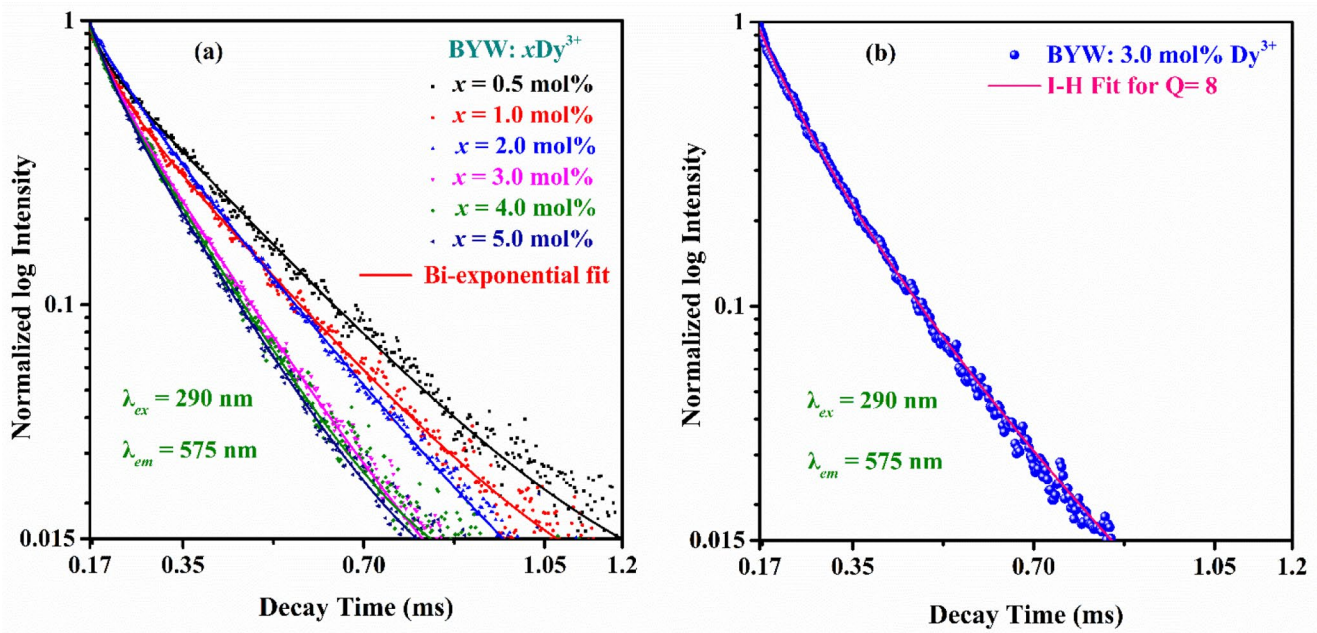


Fig. 7 (**a**) Decay curves for $^4F_{9/2}$ level in the Dy^{3+} doped $BiYWO_6$ phosphor, (**b**) I-H fit for the synthesized 3.0 mol% Dy^{3+} doped $BiYWO_6$ phosphor

mechanism, symbols τ_1 and τ_2 represent the corresponding lifetimes of quick and slow decay. The average value (τ_{avg}) of the decay time for the series of Dy^{3+} activated $BiYWO_6$ phosphor powder samples was estimated from the given formula:

$$\tau_{avg} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \quad (8)$$

The observed average lifetime (τ_{avg}) values for the synthesized BYW: xDy^{3+} (x = 0.5 to 5.0 mol%) phosphor were

found to be 218, 193, 183, 146, 134 and 132 µs, respectively. The average lifetime values of the BYW: xDy^{3+} phosphor decline as the content of Dy^{3+} increases. This decline in the average lifetime values is normally due to a deterioration in the distance between Dy^{3+} ions, which results in energy transfer from donor to acceptor ion of Dy^{3+}. The Inokutti Hirayama (I-H) model provides a clear understanding of the energy transfer mechanism through multipolar interactions between dopant Dy^{3+} ions. The time and decay intensity can be associated with these energy transfers as follows [36]:

$$I_t = I_0 \exp\left[-\frac{t}{t_0} - \eta\left(\frac{t}{t_0}\right)\right]^{\frac{3}{Q}} \quad (9)$$

Where, I_t attributed to the luminesce intensity at time t and I_0 corresponding to the t_0, after the excitation. The lifetime of the donor ions without any acceptor ions is represented as t_0. The energy transfer parameter represented by η. The value of $Q = 6$ due to d-d, $Q = 8$ represents the d-q, and $Q = 10$ attributed to the q-q type of interaction. The decay curve for 3.0 mol% Dy^{3+} doped BYW phosphor is well fitted with the Eq. (9) corresponding to the value of $Q = 8$, as depicted in Fig. 7(b). Therefore, energy transfer mechanism is justified by d-q interaction between Dy^{3+} ions, which was supported by Dexter's theory.

3.8 Temperature dependent PL analysis

In order to achieve comprehensive temperature dependent luminescent characteristics of the BYW: xDy^{3+} (x = 3.0 mol%) phosphor, the photoluminescence spectra were recorded under 455 nm excitation wavelength while exposing the sample within the range from 298 to 448 K temperatures as illustrated in Fig. 8. These spectra provide the information related to the variation in the intensity of the optimised phosphor as a function of temperature. The phosphor material exhibits a decrease in the emission intensity as the temperature rises. The study revealed that the emission intensity corresponding to the $^4F_{9/2} \rightarrow {}^6H_{13/2}$ transition decreased by 55% when the temperature was raised from 298 to 448 K as described in the inset Fig. 8(a).

In addition, the thermal stability of the BYW: xDy^{3+}(x = 3.0 mol%) phosphor can be assessed through the determination of activation energy (ΔE_a). The Arrhenius equation was employed to determine the activation energy (ΔE_a) for thermal quenching. The equation was given by:

$$I_T = \frac{I_0}{1 + C\exp(-\frac{\Delta E_a}{K_B T})} \quad (10)$$

where, C represents an arbitrary constant, I_0 stands for room temperature emission intensity, I_T symbol signifies the emission intensity at a variable temperature $T(K)$ and K_B stands for the Boltzmann constant (8.617×10^{-5} eV/K). The estimation of the activation energy (ΔE_a) can be derived from the linear regression analysis of the graph between $ln[(I_0/I_T)-1]$ and $1/K_BT$ as represented inset Fig. 8(b), and the activation energy (ΔE_a) for BYW: xDy^{3+} (x = 3.0 mol%) phosphor is found to be 0.236 eV. The photoluminescence result, which is dependent on the temperature, indicates that the BYW: xDy^{3+} (x = 3.0 mol%) phosphor consists of good thermal stability.

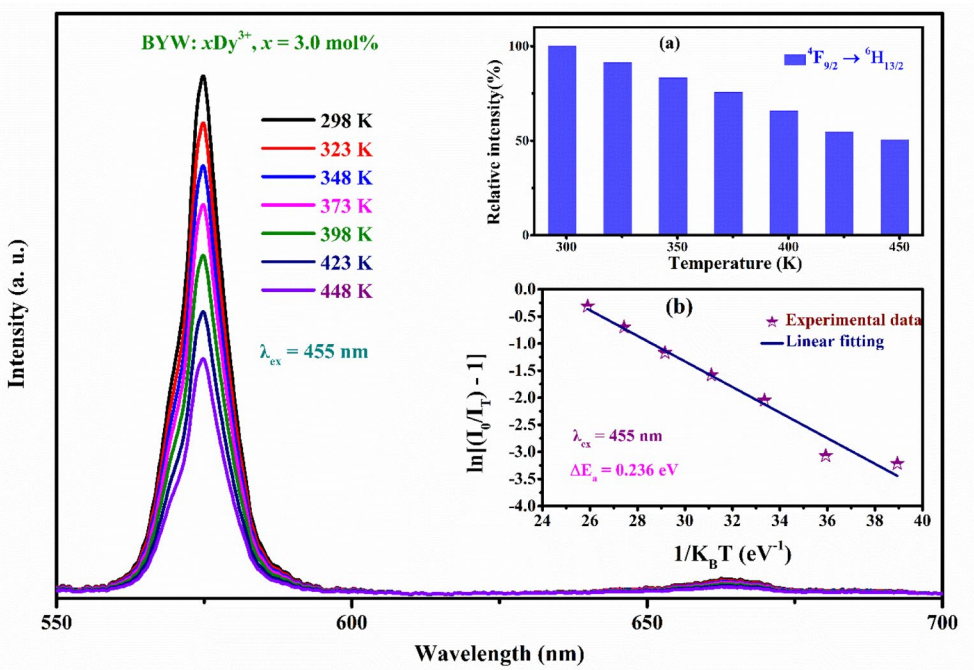


Fig. 8 Temperature dependent PL spectra of Dy^{3+} doped $BiYWO_6$ phosphor, [Inset (**a**) shows the relative emission intensity of $^4F_{9/2} \rightarrow {}^6H_{13/2}$ peak with varying temperature from room temperature (298 K) to 448 K, and (**b**) Plot of $ln[(I_0/I_T)-1]$ Vs $1/K_BT$ of the Dy^{3+} doped $BiYWO_6$ phosphor]

4 Conclusions

Thermally stable multicolour emitting BYW phosphor activated by dysprosium has been successfully synthesized using the CP method. The study of the structural, morphological, oxidation states of the elements and luminescent characteristics of the as-synthesized phosphor have been investigated. The crystalline monoclinic phase purity of Dy^{3+} doped $BiYWO_6$ has been established by the X-ray diffraction and Rietveld refinement. The optical bandgap (E_g) values are found to be 2.91 eV and 2.79 eV for the host (BYW) and BYW: xDy^{3+} (x = 3.0 mol%) phosphor material. The BYW: xDy^{3+} displays the numerous excitation peaks in UV/n-UV and visible regions due to typical transitions of Dy^{3+} ions. Dy^{3+} activated $BiYWO_6$ phosphor manifests excellent emission under excitation wavelengths viz. 290 and 455 nm wavelength, which shows the maximum emission intensity around 575 nm related to the $^4F_{9/2} \rightarrow {}^6H_{13/2}$ transition and gives an intense emission in the yellow region. The emission intensity varies through the content of Dy^{3+} ions and the intense emission was observed at 3.0 mol% among all Dy^{3+} doped $BiYWO_6$ phosphor samples, which has been considered as optimised. The CIE chromaticity coordinates were determined for each concentration, revealing that all samples are located within the near white and yellow regions when excited at wavelengths 290 and 455 nm, respectively. The photoluminescence decay curves of the Dy^{3+} doped BYW phosphors exhibit a bi-exponential nature, characterized by an average decay time in the microsecond range. Further, the energy transfer between the dopant (Dy^{3+}-Dy^{3+}) ions was verified to be a d-q interaction with the help of the I-H model. The luminescent characteristics of the phosphor material display temperature dependence, which serves as evidence for its thermal stability. The CIE coordinates of the Dy^{3+} doped $BiYWO_6$ phosphor demonstrate excitation dependent color-tunable emission, which makes it as a promising candidate for use in various luminescent devices, including pc-white LEDs under UV and blue excitations.

Author contributions Mr. Indrajeet Maurya: Conceptualization, Formal analysis and Validation, Investigation, Methodology, Software, Writing -Original draft. Dr. M. Jayasimhadri: Supervision, Formal analysis, Methodology, Visualization, Validation, Writing - review and editing.

Funding Not applicable.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

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. T. Krishnapriya, R.M. P, A. Jose, J.R. Jose, C. Joseph, P.R. Biju, Synthesis and luminescence characteristics of $Na_6Ca_{(1-X)}P_2O_9$: xDy^{3+} phosphor for outdoor lighting applications. Mater. Chem. Phys. **280**, 125827 (2022). https://doi.org/10.1016/j.matchemphys.2022.125827
2. A.K. Vishwakarma, K. Jha, M. Jayasimhadri, B. Sivaiah, B. Gahtori, D. Haranath, Emerging cool white light emission from Dy^{3+} doped single phase alkaline earth niobate phosphors for indoor lighting applications. Dalton Trans. **44**(39), 17166 (2015). https://doi.org/10.1039/C5DT02436F
3. M.K. Sahu, H. Kaur, B.V. Ratnam, J.S. Kumar, M. Jayasimhadri, Structural and spectroscopic characteristics of thermally stable Eu^{3+} activated barium zinc rthophosphate phosphor for white LEDs. Ceram. Int. **46**(16), 26410 (2020). https://doi.org/10.1016/j.ceramint.2020.07.263
4. K. Naveen Kumar, L. Vijayalakshmi, J. Lim, J. Choi, Non-cytotoxic Dy^{3+} activated $La_{10}W_{22}O_{81}$ nanophosphors for UV based cool white LEDs and anticancer applications. Spectrochimi Acta A: Mol. Biomol. Spectrosc. **278**, 121309 (2022). https://doi.org/10.1016/j.saa.2022.121309
5. A.K. Vishwakarma, K. Jha, M. Jayasimhadri, A.S. Rao, K. Jang, B. Sivaiah, D. Haranath, Red light emitting $BaNb_2O_6$: Eu^{3+} phosphor for solid state lighting applications. J. Alloys Compd. **622**, 97 (2015). https://doi.org/10.1016/j.jallcom.2014.10.016
6. I.V. Baklanova, V.N. Krasil, A.P. Tyutyunnik, Y.V. Baklanova, Journal of Physics and Chemistry of solids Blue- and white-emitting Dy^{3+} -doped aluminum oxide prepared using precursor synthesis. J. Phys. Chem. Solids. **165**, 110683 (2022). https://doi.org/10.1016/j.jpcs.2022.110683
7. Y. Xiang, Y. Zheng, L. Yang, M. Li, Y. Mao, J. Zhu, Exploring a new Dy^{3+}-activated borotellurate phosphor with thermally stable photoluminescence. J. Alloys Compd. **919**, 165837 (2022). https://doi.org/10.1016/j.jallcom.2022.165837
8. K.N. Kumar, L. Vijayalakshmi, J. Lim, J. Choi, Non-cytotoxic Dy^{3+} activated $La_{10}W_{22}O_{81}$ nanophosphors for UV based cool white LEDs and anticancer applications. Spectrochimi Acta Mol. Biomol. Spectrosc. **278**, 121309 (2022). https://doi.org/10.1016/j.saa.2022.121309
9. J. Guo, S. Li, J. Kong, Y. Li, L. Zhou, L. Lou, Q. Lv, R. Tang, L. Zheng, B. Deng, R. Yu, Synthesis and characterization of a new double perovskite phosphor: $NaCaTiTaO_6$: Dy^{3+} with high thermal stability for w-LEDs application. Opt. Laser Technol. **155**, 108347 (2022). https://doi.org/10.1016/j.optlastec.2022.108347
10. A. George, A. Jose, T. Krishnapriya, T. Francis, C. Joseph, P.R. Biju, Tunable luminescence and energy transfer mechanism of $Ca_2Sb_2O_7$: Dy^{3+} host sensitized single phase cool white light emitting phosphor. J. Solid State Chem. **305**, 122625 (2022). https://doi.org/10.1016/j.jssc.2021.122625
11. J. Liao, B. Qiu, H. Wen, W. You, Y. Xiao, Synthesis and optimum luminescence of monodispersed spheres for $BaWO_4$ -based green phosphors with doping of Tb^{3+}. J. Lumin. **130**(5), 762 (2010). https://doi.org/10.1016/j.jlumin.2009.11.028
12. R. Luo, Q. Li, J. Wang, Z. Ning, Y. Zhao, M. Liu, X. Lai, C. Zhong, J. Bi, D. Gao, Bi^{3+} ion doping: an effective strategy to significantly enhance luminescence performances of $NaGdF_4$: Eu^{3+} red-emitting nanocrystals. J. Alloys Compd. **828**, 154375 (2020). https://doi.org/10.1016/j.jallcom.2020.154375
13. Y. Tian, X. Qi, X. Wu, R. Hua, B. Chen, Luminescent properties of $Y_2(MoO_4)_3$: Eu^{3+} red phosphors with flowerlike shape prepared via coprecipitation method. J. Phy Chem. C **113**(24), 10767 (2009). https://doi.org/10.1021/jp901053q
14. T. Gavrilović, J. Periša, J. Papan, K. Vuković, K. Smits, D.J. Jovanović, M.D. Dramićanin, Particle size effects on the structure

and emission of Eu^{3+}:$LaPO_4$ and $EuPO_4$ phosphors. J. Lumin. **195**, 420 (2018). https://doi.org/10.1016/j.jlumin.2017.12.002
15. A.R. Sharits, J.F. Khoury, P.M. Woodward, Evaluating $NaREMgWO_6$ (RE = La, Gd, Y) doubly ordered double perovskites as Eu^{3+} phosphor hosts. Inorg. Chem. **55**(23), 12383 (2016). https://doi.org/10.1021/acs.inorgchem.6b02295
16. N. Xue, X. Fan, Z. Wang, M. Wang, Synthesis process and luminescence properties of Ln^{3+} doped $NaY(WO_4)_2$ nanoparticles. Mater. Lett. **61**(7), 1576 (2007). https://doi.org/10.1016/j.matlet.2006.07.082
17. X. Bai, Y. Cun, Z. Xu, Y. Zi, A.A. Haider, A. Ullah, I. Khan, J. Qiu, Z. Song, Z. Yang, Multiple anti-counterfeiting and optical storage of reversible dual-mode luminescence modification in photochromic $CaWO_4$: Yb^{3+}, Er^{3+}, Bi^{3+} phosphor. J. Chem. Eng. **429**(20), 132333 (2022). https://doi.org/10.1016/j.cej.2021.13233322)
18. P.S. Berdonosov, D.O. Charkin, K.S. Knight, K.E. Johnston, R.J. Goff, V.A. Dolgikh, P. Lightfoot, Phase relations and crystal structures in the systems $(Bi,Ln)_2WO_6$ and $(Bi,Ln)_2MoO_6$ (ln = lanthanide). J. Solid State Chem. **179**(11), 3437 (2006). https://doi.org/10.1016/j.jssc.2006.07.008
19. B. Yan, F. Lei, Molten salt synthesis, characterization and luminescence of $ZnWO_4$:Eu^{3+} nanophosphors. J. Alloys Compd. **507**(2), 460 (2010). https://doi.org/10.1016/j.jallcom.2010.07.203
20. S. Pasternak, D. Levy, Y. Paz, B. Pokroy, Structural analysis and optical properties of the $Bi_{2-x}Y_xWO_6$ system. Cryst. Eng. Comm. **18**(34), 6464 (2016). https://doi.org/10.1039/C6CE01094F
21. H. Liu, J. Yuan, W. Shangguan, Y. Teraoka, Visible-light-responding $BiYWO_6$ solid solution for stoichiometric photocatalytic water splitting. J. Phys. Chem. C **112**(23), 8521 (2008). https://doi.org/10.1021/jp802537u
22. M. Vasilopoulou, A. Soultati, D.G. Georgiadou, T. Stergiopoulos, L.C. Palilis, S. Kennou, N.A. Stathopoulos, D. Davazoglou, P. Argitis, Hydrogenated under-stoichiometric tungsten oxide anode interlayers for efficient and stable organic photovoltaics. J. Mater. Chem. A **2**(6), 1738 (2014). https://doi.org/10.1039/C3TA13975A
23. P. Ranjith, S. Sreevalsa, P. Patra, S. Som, A. Menon, K. Jayanthi, K. Annapurna, N.M.A. Krishnan, A.R. Allu, S. Das, Realizing cool and warm white-LEDs based on color controllable $(Sr,Ba)_2Al_3O_6F$: Eu^{2+}phosphors obtained via a microwave-assisted diffusion method. Phys. Chem. Chem. Phys. **23**(28), 15245 (2021). https://doi.org/10.1039/D1CP01593A
24. T. Bavani, P. Sasikala, S. Arumugam, A. Malathi, P. Praserthdam, J. Madhavan, A novel S-scheme Ws_2/$BiYWO_6$ electrostatic heterostructure for enhanced photocatalytic degradation performance towards the degradation of rhodamine B. Environ. Sci. Pollut Res. **30**(12), 34468 (2023). https://doi.org/10.1007/s11356-022-24614-0
25. Y. Domah, Electrocatalysts for photoelectrochemical water splitting. Meth. Electrocat. Adv. Mater. Allied Appli. p. 353 ISBN 978-3-030-27163-3, 978-3-030-27161-9 (2020). https://doi.org/10.1007/978-3-030-27161-9_14
26. C. Cao, S. Wei, Y. Zhu, T. Liu, A. Xie, H.M. Noh, J.H. Jeong, Synthesis, optical properties, and packaging of Dy^{3+} doped Y_2WO_6, $Y_2W_3O_{12}$, and Y_6WO_{12} phosphors. Mater. Res. Bull. 126, 110846 (2020). 126(February), 110846 (2020). https://doi.org/10.1016/j.materresbull.2020.110846
27. F. Lei, B. Yan, H.H. Chen, J.T. Zhao, Molten salt synthesis, characterization, and luminescence properties of Gd_2MO_6: Eu^{3+} (M = W, Mo) phosphors. J. Ame Ceram. Soc. **92**(6), 1262 (2009). https://doi.org/10.1111/j.1551-2916.2009.03006.x
28. S. Li, J. Guo, W. Shi, X. Hu, S. Chen, J. Luo, Y. Li, J. Kong, J. Che, H. Wang, B. Deng, R. Yu, Synthesis and characterization of wide band excited yellow phosphor $LaNb_2VO_9$: Dy^{3+} and application in potential fingerprint and lip print detection. J. Lumin. **244**, 118681 (2022). https://doi.org/10.1016/j.jlumin.2021.118681
29. R. Yu, D.S. Shin, K. Jang, Y. Guo, H.M. Noh, B.K. Moon, B.C. Choi, J.H. Jeong, S.S. Yi, Photoluminescence properties of novel host-sensitized Y_6WO_{12}: Dy^{3+} phosphors. J. Ame Ceram. Soc. **97**(7), 2170 (2014). https://doi.org/10.1111/jace.12937
30. J. Long, F. Chu, Y. Wang, C. Zhao, W. Dong, X. Yuan, C. Ma, Z. Wen, R. Ma, M. Du, Y. Cao, $M_8MgSc(PO_4)_7$:xDy^{3+} (M = Ca/Sr) single-phase full-color phosphor with high Thermal Emission Stability. Inorg. Chemi. **56**(17), 10381 (2017). https://doi.org/10.1021/acs.inorgchem.7b01268
31. J. Xiao, W. Zhang, T. Wang, J. Zhang, H. Du, Photoluminescence enhancement in a $Na_5Y(MoO_4)_4$:Dy^{3+} white-emitting phosphor by partial replacement of MoO_4^{2-} with WO_4^{2-} or VO_4^{3-}. Ceram. Int. **47**(9), 12028 (2021). https://doi.org/10.1016/j.ceramint.2021.01.046
32. A.K. Vishwakarma, M. Jayasimhadri, Pure orange color emitting Sm^{3+} doped $BaNb_2O_6$ phosphor for solid - state lighting applications. J. Lumin. **176**, 112 (2016). https://doi.org/10.1016/j.jlumin.2016.03.025
33. H. Kaur, M. Jayasimhadri, Color tunable photoluminescence properties in Eu^{3+} doped calcium bismuth vanadate phosphors for luminescent devices. Ceram. Int. **45**(12), 15385 (2019). https://doi.org/10.1016/j.ceramint.2019.05.034
34. M. Vikas, Jayasimhadri, D. Haranath, Optical and luminescent characteristics of thermally stable new Eu^{3+} doped potassium tungstate tellurite glasses for epoxy-free luminescent devices. Curr. Appl. Phys. **58**, 11 (2024). https://doi.org/10.1016/j.cap.2023.11.010
35. M. Vikas, D. Jayasimhadri, Haranath, Colour-tunable and warm white light emitting thermally stable Dy^{3+}/Sm^{3+} co-activated tungstate-tellurite glasses for photonic applications. J. Lumin. **266**, 120276 (2024). https://doi.org/10.1016/j.jlumin.2023.120276
36. M. Jayasimhadri, D. Haranath, Spectroscopic investigations of Dy^{3+} -doped tungstate – tellurite glasses for solid-state lighting applications. Int. J. Appl. Glass Sci. **13**, 645 (2022). https://doi.org/10.1111/ijag.16591

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: 21 February 2024 | Revised: 1 April 2024 | Accepted: 28 August 2024

DOI: 10.1002/mop.34316

RESEARCH ARTICLE

WILEY

Supercontinuum generation in Ga-Sb-S chalcogenide-based PCF using optofluidic approach

Deepak Garg | **Akash Khamaru** | **Ajeet Kumar**

Advanced Photonics Simulation Research Laboratory, Department of Applied Physics, Delhi Technological University, Delhi, India

Correspondence
Ajeet Kumar, Advanced Photonics Simulation Research Laboratory, Department of Applied Physics, Delhi Technological University, Bawana Rd., New Delhi, Delhi 110042, India.
Email: ajeetdph@dtu.ac.in

Abstract

We report the design and theoretical study of a Ga-Sb-S chalcogenide glass-based photonic crystal fiber (PCF) structure for mid-infrared supercontinuum generation. The proposed design is engineered by adjusting the diameter of air holes in the cladding region and pitch, giving us control over the dispersion characteristics of the fiber. The optimized structure design offers the Zero Dispersion Wavelength of 5.2 μm. When pumped with 50 fs secant hyperbolic pulses of peak power 4.9 kW, for a fiber of length 8 mm at pump wavelength of 5 μm, the proposed structure design produces an ultra-broadband supercontinuum spectrum spanning 1.5–14 μm. Proposed PCF design should be helpful in, biomedical imaging, supercontinuum sources, biosesning, and optical coherence tomography.

KEYWORDS

chalcogenide glass, photonic crystal fiber, supercontinuum generation

1 | INTRODUCTION

Supercontinuum is a phenomenon that transforms an ultrashort, high-intensity pulse into a broad spectrum when propagated in a nonlinear medium. Supercontinuum Generation (SCG) was first discovered by Alfano and Shapiro.[1] Supercontinuum is formed when several linear and nonlinear processes such as cross-phase modulation (XPM), four-wave mixing (FWM), self-phase modulation (SPM), and stimulated Raman scattering (SRS) act together on pump beam to contribute to the spectral broadening of pulse. Soliton dynamics at the pump wavelength also plays a significant role in shaping the final broad spectrum. This action of both linear and nonlinear processes transforms a narrow pulse into a broadened spectral pulse.[2] SCG has potential applications in diverse fields that include fluorescence lifetime imaging and optical coherence tomography to microscopy, spectroscopy, and early cancer detection.[3,4] Mid-infrared (M-IR) region holds a particular importance due to the presence of strong fundamental molecular vibration absorption bands. This can help in better understanding of the molecular structure of matter. In the past few years silica has been widely used for SCG but due to its insufficient nonlinearity and high absorption loss after 2.5 μm, it limits the spectral evolution of the input pulse. This reason motivated the researchers to explore different materials like ZBLAN, tellurite, and bismuth. ZBLAN can be used to generate spectrum spanning extending from ultraviolet to mid-IR.[5] Tellurite and Bismuth offer high nonlinear coefficient as compared to silica and offers good transparency in mid-IR region.[6,7]

Among these materials, chalcogenide glasses are excellent candidates due to their high nonlinear refractive index (RI) and excellent IR and mid-IR transmission. Additionally, SCG in mid-IR domain of 2–15 μm conveniently overlaps with the molecular fingerprint spectral region.[2] Hence, chalcogenide glasses are ideal for fabricating mid-IR nonlinear photonic devices.

Many researchers have been exploring the generation of supercontinuum (SC) in chalcogenide fibers and waveguide and the results reported are remarkable. For instance, Shaw et al.[8] demonstrated broadband SC spanning from 2.1 to 3.1 μm in an As_2Se_3 chalcogenide photonic crystal fiber (PCF). Saini et al.[9] theoretically explored an equiangular

 © 2024 Wiley Periodicals LLC.

spiral PCF for SC generation-spanning 1.2–15 μm. Saini et al.[10] also demonstrated the theoretical results of an As_2Se_3 rib waveguide for SCG. Choi et al.[11] characterized and demonstrated the nonlinear behavior of a Ge-Sb-S chalcogenide waveguide for SC generation. Siwicki et al.[12] demonstrated the nonlinear simulation of nanostructured GI chalcogenide-based fiber achieving SC broadening from 3.5 to 8.5 μm when pumped with 6.3 μm wavelength. Jiao et al.[13] experimentally demonstrated SC broadening ranging from 1.6 to 12 μm in a suspended chalcohalide fiber using pump wavelength of 5 μm. Wu et al.[14] also experimentally demonstrated SC generation in mid-IR regime spanning 1.7–11.3 μm using 5 μm femtosecond laser pulses. Furthermore, Chauhan et al.[15] demonstrated broadband SC generation ranging from 1 to 14 μm in the Ga-Sb-S-based chalcogenide PCF pumping with 4 μm central wavelength and 8.19 kW peak power. A ground-breaking development in chalcogenide glass was recently reported by Yang et al., who introduced the highly nonlinear composition $Ga_8Sb_{32}S_{60}$. This material possesses an exceptional nonlinear RI of 12.4×10^{-18} m^2/W, surpassing the capabilities of conventional chalcogenides. Moreover, it exhibits a broad optical transmission window spanning the M-IR range from 0.8 to 14 μm, making it an ideal candidate for diverse M-IR applications.[16] Compared to their As-based counterparts, Ga-Sb-S chalcogenide glasses offer superior stability and lower toxicity, further promoting their potential.[17] Their additional advantage lies in the ease of bonding with heavy metals, which facilitates the generation of multiphoton spectra for laser sources. Due to these advantages offered by Ga-Sb-S chalcogenide glasses, these glasses can help a lot in the next-generation M-IR technologies.

In this paper, a rectangular core $Ga_8Sb_{32}S_{60}$ chalcogenide-based fiber with the implementation of optofluidic approach is studied. Air holes in optical fibers can be filled completely, partially, or selectively with various liquids and gases, allowing for precise control over the transmission of polarized light through the fiber. These liquids include, but are not limited to, ethanol,[18] water,[19] liquid crystals,[20] and polymers.[21] The proposed chalcogenide PCF stands out with its unique combination of flat dispersion, low effective mode area and high nonlinearity. When pumped with 5 μm optical parametric amplifier (OPA) laser source,[13] the proposed PCF structure produces a supercontinuum broadening of 1.5–14.0 μm in only 8 mm length of fiber.

2 | PROPOSED DESIGN

We have designed a novel rectangular lattice pattern PCF based on highly nonlinear $Ga_8Sb_{32}S_{60}$ chalcogenide glass as shown in Figure 1. In cladding region, alternative larger and smaller holes has been arranged in a rectangular lattice

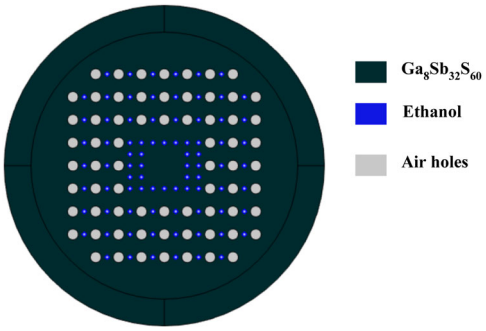


FIGURE 1 Transverse cross-section of the proposed PCF.

pattern. For the better confinement of fundamental mode in the core region, air holes are removed from the center. Larger air holes have diameter d_1 and smaller ethanol-filled holes have diameter d_2. The pitch is kept constant as 3 μm and is represented by Λ. The cladding region is adjoined with a cylindrical layer called perfectly matched layer (PML) to avoid the effect of back reflections into the study area. The thickness of PML is kept 10% of total fiber diameter. Optofluidic approach has been implemented and the smaller holes are filled with ethanol. Experimental techniques for fully and partially filling of air holes in PCF are well developed.[22–24] Due to changes in RI during processing, fabrication of PCF is performed via different techniques.[25]

3 | NUMERICAL ANALYSIS

We employed "COMSOL Multiphysics" software which is based on Finite Element Method (FEM) for modal analysis of the proposed design. The material chosen for cladding is a highly nonlinear $Ga_8Sb_{32}S_{60}$-based chalcogenide glass. For calculating the wavelength-dependence RI of $Ga_8Sb_{32}S_{60}$-based chalcogenide material, Sellmeier equation is used.[15] Sellmeier equation for the chosen material is given as:

$$n(\lambda) = \sqrt{1 + \frac{6.2563\lambda^2}{\lambda^2 - (0.3425)^2} + \frac{2.9444\lambda^2}{\lambda^2 - (34.28)^2}}, \quad (1)$$

where λ symbolizes the pump wavelength in micrometers. For the SCG to be broad and efficient two important factors comes into play namely Group Velocity Dispersion and material nonlinearity. The nonlinearity of material depends on the nonlinear RI and is given as:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}(\lambda)}, \quad (2)$$

where n_2 is the nonlinear RI of material which in our case for $Ga_8Sb_{32}S_{60}$ is 12.4×10^{-18} m^2/W[26] and A_{eff} (λ)

10982760, 2024, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mop.34316 by Delhi Technological University, Wiley Online Library on [23/09/2024]. 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

represents the effective area of the fundamental mode in the core of the fiber which is given as

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

where E is the transverse component of propagating electric field

Dispersion, $D(\lambda)$ is calculated using the wavelength dependence of effective RI (n_{eff}) and is given as:

$$D(\lambda) = \frac{-\lambda}{c}\frac{d^2 Re\left(n_{eff}\right)}{d\lambda^2}, \tag{4}$$

where Re(n_{eff}) represents the real part of effective mode index of fundamental mode.

To obtain SCG a generalized nonlinear Schrodinger wave equation has been employed on our design[26]:

$$\frac{\partial A}{\partial Z} + \frac{\alpha}{2}A - \left(\sum_{n\geq 2}\frac{i^{n+1}}{n!}\beta_n\frac{\partial^n A}{\partial T^n}\right) = i\gamma\left(1 + \frac{i}{\omega_0}\frac{\partial}{\partial T}\right)\left(A(z,t)\int_{-\infty}^{\infty}R(T)\,|A(z,T-T')|^2 dT + i\hbar_R(z,T)\right). \tag{5}$$

The RHS of the above equation represents nonlinear processes and the left-hand side (LHS) represents the linear processes. A(z,T) represents the output pulse envelope and the order of dispersion effect is represented by the symbol "n." We are working under the assumption that $\hbar_R = 0$. In the above equation, α represents the attenuation coefficient and in our calculations, we used the value of the loss coefficient as 0.6 cm^{-1}.[15]

Using Taylor series expansion, the propagation constant (β) can be expanded as[21]:

$$\beta(\omega) = \beta(\omega_0) + \beta_1(\omega-\omega_0) + \frac{1}{2}\beta_2(\omega_0)(\omega-\omega_0)^2 + \frac{1}{3!}\beta^3(\omega-\omega_0)^3..... \tag{6}$$

The nonlinear response, R (t) is expressed as[27]:

$$R(t) = (1-f_R)\delta(t) + f_R h_R(t), \tag{7}$$

where f_R and $h_R(t)$ symbolizes the delayed Raman response and Raman response function respectively. $h_R(t)$ can be expressed as:

$$h_R(t) = \frac{\tau_1^2+\tau_2^2}{\tau_1\tau_2^2}\exp\left[\frac{-t}{\tau_2}\right]\sin\left(\frac{t}{\tau_1}\right), \tag{8}$$

we have chosen $\tau_1 = 23.1$ fs, $\tau_2 = 195$ fs, and $f_R = 0.115$.[15]

For the generation of pump pulses, a tunable hyperbolic secant pulse has been used which can be expressed as[15]:

$$A(z=0,t) = \sqrt{P_0}\,\text{sech}\left(\frac{t}{P_0}\right), \tag{9}$$

where $t_0 = T_{\text{FWHM}}/1.7627$ and the peak power is denoted as P_0.

4 | RESULTS AND DISCUSSION

In the proposed PCF design, dispersion characteristics are controlled by altering the geometric parameters (size of air holes and pitch). We have optimized the structural parameters of our designed PCF to obtain the minimum value of dispersion and effective mode area of the propagating fundamental mode. The effect in dispersion characteristics due to the variation in diameter of air holes is shown in Figure 2A. By keeping $d_2 = 500$ nm and $\Lambda = 3000$ nm, that is, by keeping them constant, and with an increase in diameter (d_1) of the bigger holes, the zero dispersion wavelength (ZDW) shift toward shorter wavelength side. The variation in diameter of bigger air holes ranges from 1200 to 1600 nm to optimize it to get low dispersion at pump wavelength. The value of dispersion changes from 1.1660 to −4.9844 ps/nm/km when d_1 is varied from 1200 to 1600 nm. The optimized value chosen for d_1 is 1400 nm. Similarly, the effect of variation of diameter of smaller ethanol-filled air holes is also studied and is shown in Figure 2B. The optimization of parameter d_2 is more important as with the variation of d_2, the shift in ZDW is seen to be more prominent. On increasing d_2 and keeping other parameters (d_1 and Λ) to be constant as $d_1 = 1400$ nm and $\Lambda = 3000$ nm, the shift in ZDW shifts prominently toward shorter wavelength side. The diameter (d_2) is varied from 400 to 600 nm. The value of dispersion changes from 6.8572 to −2.4546 ps/nm/km when d_2 is varied from 400 to 600 nm at pump wavelength. The optimized value chosen for d_2 is 500 nm. The effect of varying the pitch (Λ) by keeping other parameters d_1 and d_2 fixed is also taken in account and is shown in Figure 2C. With the increase in pitch (Λ), the ZDW shifts toward the larger wavelength but the shift nearly overlaps since the change is much less. The effect of variation of parameter Λ on effective mode area is also negligible.

From the above discussion, it can be concluded that for the chosen optimized parameters $d_1 = 1400$ nm, $d_2 = 500$ nm, and $\Lambda = 3000$ nm the value of dispersion as

10982760, 2024, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mop.34316 by Delhi Technological University, Wiley Online Library on [23/09/2024]. 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

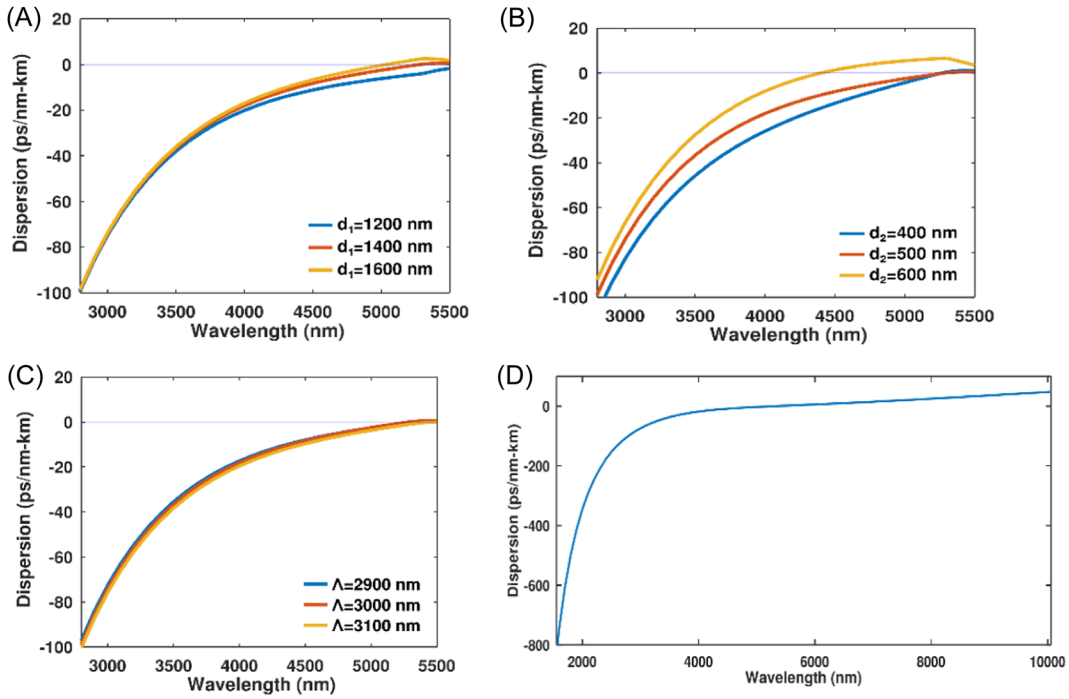


FIGURE 2 Dispersion characteristics with variation in (A) d_1 by keeping d_2 and Λ constant, (B) d_2 by keeping d_1 and Λ constant, (C) Λ by keeping d_1 and d_2 constant, and (D) dispersion characteristics for the optimized parameters.

well as effective mode area at pump wavelength of 5000 nm is minimum. For the optimized parameters (as shown in Table 1), the dispersion characteristics offered by the proposed PCF is shown in Figure 2D. In the wavelength range of 3000–6000 nm the designed PCF exhibits a flat normal dispersion profile. At pump wavelength of 5 μm, the proposed PCF design structure falls under normal dispersion nature. Normal dispersion regime ranges from 1000 to 5200 nm and the anomalous regime covers the range beyond 5200 nm. Kerr Effect is responsible for the broadening in normal dispersion region. A broad and unstructured supercontinuum spectrum can be generated by the designed photonic crystal fiber with optimized parameters while working in the normal dispersion regime.

SCG is studied by solving Equation (5). β_n in the LHS of the equation represents the dispersion parameter where "n" is the order of dispersion coefficients taken into account during calculations. In our simulation, we have considered dispersion coefficients up to 9th order. The effect of higher order coefficients greater than 9th order on the output spectrum is not significant. Due to this reason we have considered the coefficients up to 9th order. SCG spectrum is analyzed and studied with the variations in length of fiber and peak power and are shown in Figures 3 and 4.

The initial broadening of the spectrum is due to the interplay of SPM and XPM. When we input the high-intensity pulse on nonlinear fiber, the intensity dependence RI affects the propagating pulse and helps in broadening of the pulse in a symmetric manner by increasing the spectral content of the pulse. Since, we are working with multiple optical fields and due to the co-propagation of these high-intensity, short pulses, XPM comes into play which helps in the broadening of pulse initially. FWM also comes into play generating new frequencies which helps in spectral broadening. FWM requires the condition of phase matching. Further, interaction of light with molecular lattice vibrations produces another nonlinear effect which is termed as SRS. In the longer wavelength region, SRS is responsible for the broadening of the spectrum.

The effect of fiber length (l) on the supercontinuum spectrum is shown in Figure 3. With an input hyperbolic secant pulse of pulse duration 50 fs and peak power of 4.9 kW, length of fiber is varied from 0 to 8 mm in steps of 4 mm. The broadening is

TABLE 1 Optimized parameters of proposed design.

Parameters	Value (nm)
d_1	1600
d_2	500
Λ	3000

10982760, 2024, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mop.34316 by Delhi Technological University, Wiley Online Library on [23/09/2024]. 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

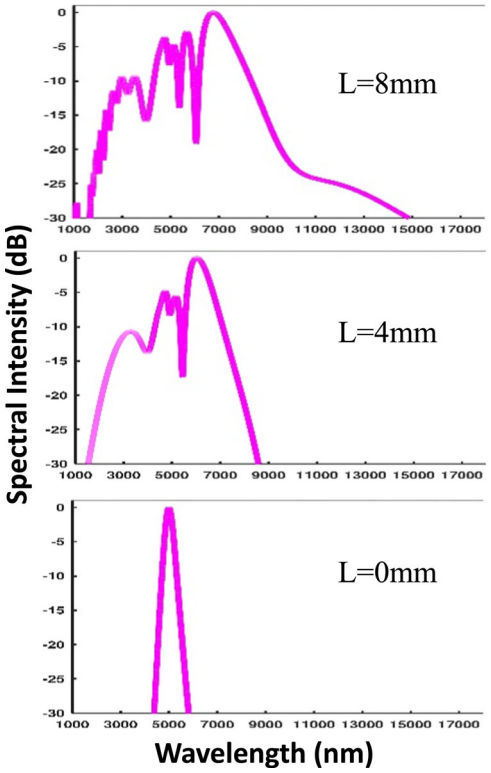


FIGURE 3 SC generation with variation in length of fiber.

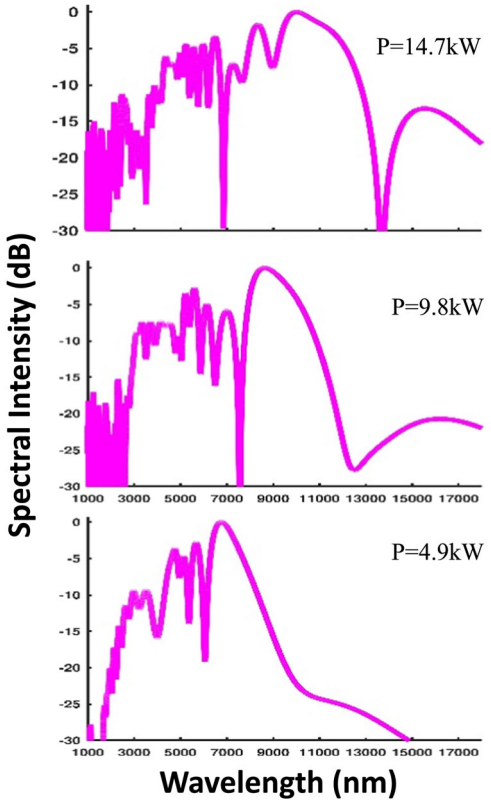


FIGURE 4 SC generation with variation in Peak Power.

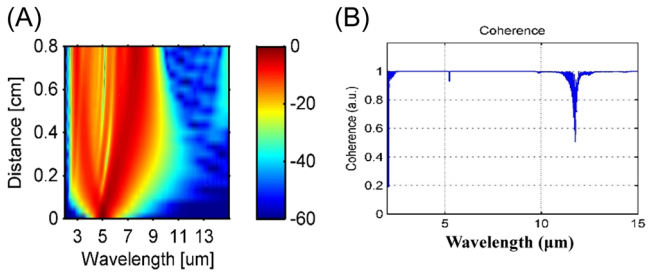


FIGURE 5 Spectral evolution and Coherence of SC at 4.9 kW.

due to XPM and SPM followed by FWM and SRS. The maximum broadening is achieved at 8 mm length of fiber with pulse width and peak power taken as 50 fs and 4.9 kW respectively. The achieved broadening in 8 mm length of fiber is from 1.5 to 14 μm.

The effect of variation in peak power (P) on the SC spectrum is also studied and is shown in Figure 4. In the long wavelength region SRS plays part in the spectral broadening. Due to less intensity at low peak power the contribution to spectral broadening is less but with the increase in peak power of the input pulse the intensity also increases resulting in greater broadening in long wavelength region. The Spectral broadening continues up-to a certain power limit after that there is no effect of increasing power on spectral broadening. Peak power is varied in equal intervals of 4.9 kW ranging from 4.9 to 14.7 kW with a pulse width of 50 fs and in 8 mm length of fiber. The broadening of SC spectrum increases with the increase in peak power up to 4.9 kW. The achieved SC broadening when simulated with 4.9 kW peak power, 50 fs pulse width and 8 mm fiber length ranges from 1.5 to 14 μm which spans the mid-IR regime. The spectral evolution and coherence of the SC spectrum is shown in Figure 5A,B, respectively, at 8 mm length of fiber, 50 fs pulse width and 4.9 kW peak power.

The study of SC spectrum in PCF using $Ga_8Sb_{32}S_{60}$ chalcogenide-based material has been done previously. Medjouri et al. demonstrated the numerical investigation of SCG in a hexagonal geometry PCF spanning 1.65–9.24 μm at 20 dB level using peak power of 20 kW. Chauhan et al. also demonstrated SC broadening from 1 to 14 μm in a 9 mm long PCF using 4 μm pump source at 8.19 kW. In our study, we used $Ga_8Sb_{32}S_{60}$ chalcogenide-based PCF incorporated with optofluidic approach generating SC spectrum ranging from 1.5 to 14 μm when pumped with 5 μm OPA laser source and at relatively low peak power of 4.9 kW.

10982760, 2024, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mop.34316 by Delhi Technological University, Wiley Online Library on [23/09/2024]. 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 | CONCLUSIONS

A new design of highly nonlinear rectangular core $Ga_8Sb_{32}S_{60}$ chalcogenide-based ethanol-filled PCF is being proposed for ultra-broadband SCG in mid-IR and far-IR region. The broadband SC spectra from 1.5 to 14 μm is produced when the input pulse is pumped by a 5 μm OPA laser with 50 fs pulse width having peak power 4.9 kW in only 8 mm fiber length of proposed photonic crystal fiber. The proposed PCF structure possess high nonlinearity of 670 W^{-1} km^{-1} at pump wavelength of 5 μm. The proposed photonic crystal fiber structure design is able to generate a broad supercontinuum spectra which can be helpful in various diverse application including biosesning, chemical sensing, food quality control, military security, and optical coherence tomography.

ACKNOWLEDGMENTS

This work was supported by the Delhi Technological University, Bawana Road, New Delhi, Delhi 110042, India.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Ajeet Kumar http://orcid.org/0000-0003-4632-2841

REFERENCES

1. Alfano RR, Shapiro SL. Emission in the region of 4000 to 7000Å via four-photon coupling in glass.*Phys Rev Lett*. 1970;24:584. doi:10.1103/PhysRevLett.24.584
2. Chaitanya AGN, Saini TS, Kumar A, Sinha RK. Ultra broadband mid-IR supercontinuum generation in Ge_11.5As_24Se_64.5 based chalcogenide graded-index photonic crystal fiber: design and analysis. *Appl Opt*. 2016;55:10138. doi:10.1364/AO.55.010138
3. Dudley JM, Genty G, Coen S. Supercontinuum generation in photonic crystal fiber. *Rev Mod Phys*. 2006;78:1135. doi:10.1103/RevModPhys.78.1135
4. Kaminski CF, Watt RS, Elder AD, Frank JH, Hult J. Supercontinuum radiation for applications in chemical sensing and microscopy. *Appl Phys B Lasers Opt*. 2008;92:367.
5. Jiang X, Joly NY, Finger MA, et al. Deep-ultraviolet to mid-infrared supercontinuum generated in solid-core ZBLAN photonic crystal fibre. *Nat Photonics*. 2015;9:133-139. doi:10.1038/NPHOTON.2014.320
6. Domachuk P, Wolchover NA, Cronin-Golomb M. et al. Over 4000 nm bandwidth of mid-IR supercontinuum generation in sub-centimeter segments of highly nonlinear tellurite PCFs. *Opt. Express*. 2008;16:7161. doi:110.1364/OE.16007161
7. Buczynski R, Bookey HT, Pysz D, et al. Supercontinuum generation up to 2.5 μm in photonic crystal fiber made of lead-bismuth-galate glass. *Laser Phys Lett*. 2010;7:666-672. doi:10.1002/lapl.201010039
8. Shaw LB, Thielen PA, Kung FH, Nguyen VQ, Sanghera JS, Aggarwal ID. IR Supercontinuum Generation in As-Se Photonic Crystal Fiber. In: *Advance Solid State Photonics*. Technical Digest (Optica Publishing Group, 2005), paper TuC5; 2005. doi:10.1364/ASSP.2005.TuC5
9. Saini TS, Baili A, Kumar A, Cherif R, Zghal M, Sinha RK. Design and analysis of equiangular spiral photonic crystal fiber for mid-infrared supercontinuum generation. *J Mod Opt*. 2015;62:1560. doi:10.1080/09500340.2015.1051600
10. Saini TS, Kumar A, Sinha RK. Design and modelling of dispersion-engineered rib waveguide for ultra broadband mid-infrared supercontinuum generation. *J Mod Opt*. 2017;64:143-149. doi:10.1080/09500340.2016.1216190
11. Choi JW, Han Z, Sohn BU, et al. Nonlinear characterization of GeSbS chalcogenide glass waveguides. *Sci Rep*. 2016;6:39234. doi:10.1038/srep39234
12. Siwicki B, Filipkowski A, Kasztelanic R, Klimczak M, Buczyński R. Nanostructured graded-index core chalcogenide fiber with all-normal dispersion–design and nonlinear simulations. *Opt Express*. 2017;25:12984. doi:10.1364/OE.25.012984
13. Jiao K, Wang XG, Liang X, et al. Single-mode suspended large-core chalcohalide fiber with a low zero-dispersion wavelength for supercontinuum generation. *Opt Express*. 2022;30:641. doi:10.1364/OE.447571
14. Wu B, Zhao Z, Wang X, et al. Mid-infrared supercontinuum generation in a suspended-core tellurium-based chalcogenide fiber.*Opt Mater Express*. 2018;8:1341. doi:10.1364/OME.8.001341
15. Chauhan P, Kumar A, Kalra Y. Mid-infrared broadband supercontinuum generation in a highly nonlinear rectangular core chalcogenide photonic crystal fiber. *Opt Fiber Technol*. 2018;46:174-178. doi:10.1016/j.yofte.2018.10.004
16. Zhang B, Yu Y, Zhai C, et al. High brightness 2.2–12 μm mid-infrared supercontinuum generation in a nontoxic chalcogenide step-index fiber. *J Am Ceram Soc*. 2016;99:2565-2568. doi:10.1111/jace.14025
17. Saini TS, Sinha RK. Mid-infrared supercontinuum generation in soft-glass specialty optical fibers: a review. *Prog Quantum Electron*. 2021;78:100342. doi:10.1016/j.pquantelec.2021.100342
18. Yiou S, Delaye P, Rouvie A, et al. Stimulated Raman scattering in an ethanol core microstructured optical fiber.*Opt Express*. 2000;13:4786. doi:10.1364/OPEX.13.004786
19. Martelli C, Canning J, Lyytikainen K, Groothoff N. Water-core fresnel fiber. *Opt Express*. 2005;13:3890. doi:10.1364/OPEX.13.003890
20. Alkeskjold TT, Alkeskjold TT, Lægsgaard J, et al. Highly tunable large-core single-mode liquid-crystal photonic bandgap fiber. *App. Optics*. 2006;45:2261. doi:10.1364/AO.45.002261
21. Cox FM, Argyros A, Large MCJ. Liquid-filled hollow core microstructured polymer optical fiber. *Opt Express*. 2006;14:4135. doi:10.1364/OE.15.013675
22. Nielsen K, Noordegraaf D, Sørensen T, Bjarklev A, Hansen TP. Selective filling of photonic crystal fibres. *J Opt A: Pure Appl Opt*. 2005;7:L13. doi:10.1088/1464-4258/7/8/L02

10982760, 2024, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mop.34316 by Delhi Technological University, Wiley Online Library on [23/09/2024]. 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

23. Wang Fei, Yuan W, Hansen O, Bang O. Selective filling of photonic crystal fibers using focused ion beam milled micro-channels. *Opt Express*. 2011;19:17585. doi:10.1364/OE.19.017585
24. Kumar P, Fiaboe KF, Roy JS. Design of nonlinear photonic crystal fibers with ultra-flattened zero dispersion for supercontinuum generation. *ETRI J*. 2020;42:282-291. doi:10.4218/etrij.2019-0024
25. Jian F, Hou Z, Zhou G, et al. Experimental study on all Yb-doped photonic crystal fiber laser. in: *Fiber Lasers XIV: Technology and Systems*. 2017:459-466. https://doi.org/10.1117/12.2264091
26. Saini TS, Tiwari UK, Sinha RK. Rib waveguide in Ga-Sb-S chalcogenide glass for on-chip mid-IR supercontinuum sources: design and analysis. *J Appl Phys*. 2017;122:053104. doi:10.1063/1.4997541
27. Agrawal GP. Nonlinear Fiber Optics 5th ed. (Academic, San Diego, California. 2013.

How to cite this article: Garg D, Khamaru A, Kumar A. Supercontinuum generation in Ga-Sb-S chalcogenide-based PCF using optofluidic approach. *Microw Opt Technol Lett*. 2024;66:e34316. doi:10.1002/mop.34316

10982760, 2024, 9, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mop.34316 by Delhi Technological University, Wiley Online Library on [23/09/2024]. 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

Research Article
doi.org/10.1002/slct.202401639
Chemistry Europe
European Chemical Societies Publishing

www.chemistryselect.org

Synthesis and Assessment of Acacia Gum-Based Hydrogel as a Promising Novel Biopolymeric Matrix for Delivery of Ciprofloxacin

Anshul,[a] Garima Tomar,[a] Priyanka Meena,*[a] and Sudhir G. Warkar*[a]

In this current study, hydrogels based on acacia gum (AG), polyacrylamide (PAM), and carboxymethyl tamarind kernel gum (CMTKG) were successfully developed and ciprofloxacin (drug) incorporated into it. To optimize the hydrogel, the concentrations of crosslinker and initiator were varied, and their effects on the swelling were examined. The hydrogel was subjected to characterization techniques such as Powder X-ray Diffraction (PXRD), Scanning Electron Microscopy (SEM), Attenuated Total Reflection-Fourier Transform Infrared spectroscopy (ATR-FTIR), and Thermogravimetric analysis (TGA). The hydrogel swelling and in vitro drug release were assessed in pH 1.2 and 7.4 buffer solutions. The findings revealed that an alkaline pH yielded superior results for the swelling and in vitro drug release experiments compared to an acidic pH. The drug release kinetic of the ciprofloxacin-loaded AG/PAM/CMTKG hydrogel followed the Korsmeyer–Peppas model, demonstrating non-fickian diffusion at pH 7.4 and fickian mechanism at pH 1.2. The pH-dependent behavior of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel displayed its potential implications for site-specific ciprofloxacin release.

1. Introduction

Hydrogel is a three-dimensional network of fluid-absorbing hydrophilic polymers with chemical or physical cross-links that allow it to swell without dissolution. These polymers contain various hydrophilic groups, such as $-NH_2$, $-OH$, $-COOH$ $-CONH_2$, $-SO_3H$, etc.[1] Hydrogels possess properties like super absorbency, hydrophilicity, and softness, making them valuable in biomedical and pharmaceutical applications.[2] Their bio-adhesive features enhance tissue permeability and drug residence time, making hydrogels excellent for controlled drug delivery.[1]

In recent years, hydrogels have gained popularity as drug delivery vehicles, particularly for oral drug delivery, due to their biocompatibility and non-toxicity.[3] Additionally, hydrogels that respond by altering their volume when subjected to specific environmental changes are described as intelligent or smart hydrogels. These are sensitive to physical stimuli like light, pressure, temperature, electric field, magnetic field, ultrasound, or chemical stimuli like pH, ionic strength, CO_2, glucose or biological stimuli like enzymes, antigens, glutathione, and DNA.[1] In all, pH-sensitive hydrogels have been extensively researched, as they play a critical role in the development of improved smart drug delivery systems.[3] The pH-sensitive hydrogels are composed of a polymer with weak acidic or basic groups that readily ionize at either higher or lower pH values, respective to their pKa values.[4]

[a] *Anshul, G. Tomar, P. Meena, S. G. Warkar*
Department of Applied Chemistry, Delhi Technological University, Delhi, India
E-mail: priyankameena.du2016@gmail.com
sudhirwarkar@gmail.com

Polyacrylamide is a synthetic pH-sensitive polymer obtained through the crosslinking of acrylamide monomers.[5,6] The PAM hydrogels exhibit significant volume transitions in response to various physical and chemical stimuli.[7] PAM hydrogels are acknowledged for their mechanical strength and biocompatible characteristics, exhibiting no toxic effects on human fibroblasts. Moreover, these hydrogels are non-carcinogenic and non-allergenic, thereby minimizing the likelihood of mutagenic effects while demonstrating resistance to specific body enzymes. Consequently, these hydrogels underscore their versatility in various biomedical applications.[6]

Recently, there has been significant interest in hybrid hydrogels, offering a combination of desirable attributes from both natural and synthetic polymers. A noteworthy example is carboxymethyl tamarind kernel gum (CMTKG), a derivative of Tamarind kernel gum (TKG). CMTKG is a biopolymer consisting of 3:2:1 molar ratio of glucose, galactose, and xylose. This chemical modification of TKG enhances features such as shell integrity, drug loading efficiency, swelling, bio adhesiveness, pH-sensitivity and hydrophilicity.[8]

Acacia Gum is a polysaccharide, having a high branch structure, obtained as a dried exudate extracted from the stems and branches of Acacia Senegal.[9,10] It has 4-O-methyl-ß-D-glucuronopyranosyl, α-L-arabinofuranosyl, ß-D-glucuronopyranosyl and α–L -rhamnopyranosyl units as its main structural constituents.[9] It is commonly found in regions extending from West Africa to the Indian Peninsula.[11] It is soluble in both hot and cold water but insoluble in organic solvent.[12]

There are numerous antibacterial materials reported in the literatures.[13] Among these, ciprofloxacin, an antibiotic drug, has been relatively less explored. Ciprofloxacin is classified within the fluoroquinolone family and is identified as 1-cyclopropyl-6-fluoro-4-oxo-7-(piperazin-1-yl)-quinoline-3-carboxylic acid. Its structure features a quinolone core with a fluorine

© 2024 Wiley-VCH GmbH

atom located at position 6 and a piperazine ring situated at position 7.[14–16] It is commonly employed to treat infections, including urinary tract, skin and gastrointestinal infections.[16]

This work deals with the preparation of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel tailored for pH-dependent release of ciprofloxacin, employing varying amounts of initiator and crosslinker. The characterization of the developed hydrogels was conducted through ATR–FTIR, SEM, PXRD, and TGA analysis. The swelling and drug release analysis were investigated in simulated physiological solutions of pH 1.2 and 7.4, respectively. Various kinetic models, including First order, Hixson Crowell, Higuchi, Zero-order, Korsmeyer–Peppas were employed to explore the drug release mechanisms.

2. Material and Methods

CMTKG (Hindustan Gum and Chemicals Ltd., Bhiwani, Haryana), *N*, *N*′-methylene bisacrylamide (MBA, Merck, Germany), ciprofloxacin (Unicure Pvt. Ltd.), acrylamide (AM, CDH, New Delhi), acacia gum (AG, Thermo Fisher Scientific India Pvt. Ltd. Mumbai, India), and potassium persulphate (KPS, Fischer Scientific, Bombay, India) were utilized as provided.

2.1. Synthesis of AG/PAM/CMTKG Hydrogels and Ciprofloxacin–Loaded AG/PAM/CMTKG Hydrogel

The various formulations of AG/PAM/CMTKG hydrogels with different concentrations of MBA (B1-B5) and KPS (B6-B9) were prepared, as detailed in Table 1. To formulate the AG/PAM/CMTKG hydrogels, specific quantities of AG, AM, and CMTKG were dissolved in distilled water and stirred up for an hour. Afterward, KPS and MBA were added while maintaining constant stirring for 1 hour. The solution was then transferred to the test tube and kept in a water bath at 60 °C for an hour. After extraction from the test tube, the hydrogel was cut into pellets and dried for 48 hr at 60 °C in an oven.[7]

The same process, with slight modifications, was employed for fabrication of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel. Ciprofloxacin was introduced during the addition of MBA and KPS, while the remaining steps followed the same procedures as depicted in Figure 1.[17]

2.2. Swelling Analysis

The swelling of the hydrogels were measured by using the gravimetric technique.[18] Hydrogel discs, initially dried, were immersed in HCl—KCl buffer (pH 1.2) and phosphate buffer saline (pH 7.4).[19] The hydrogel was taken out of the buffer at 1 hour intervals. The filter paper was used to blot the extra liquid from the hydrogel disc, and the discs were weighed.[20]

Table 1. Amount of materials involved in the AG/PAM/CMTKG hydrogels synthesis, with their swelling %.

Sample code	KPS (mg)	MBA (mg)	CMTKG (g)	PAM (g)	AG (g)	Swelling (%)	
						pH 1.2	pH 7.4
B-1	36	12	0.25	0.5	0.2	837	968
B-2	36	17	0.25	0.5	0.2	746	935
B-3	36	22	0.25	0.5	0.2	662	863
B-4	36	27	0.25	0.5	0.2	685	737
B-5	36	32	0.25	0.5	0.2	615	723
B-6	18	12	0.25	0.5	0.2	702	857
B-7	24	12	0.25	0.5	0.2	756	907
B-8	30	12	0.25	0.5	0.2	801	968
B-9	42	12	0.25	0.5	0.2	709	815

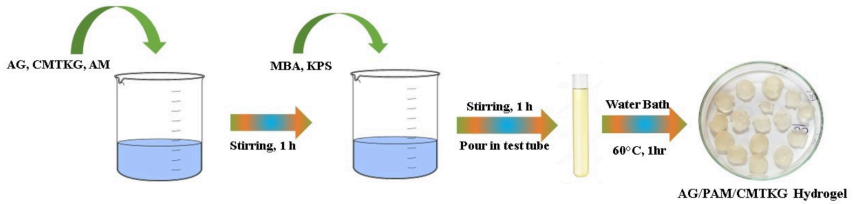


Figure 1. Synthesis of AG/PAM/CMTKG hydrogels.

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Yet again, the discs were immersed in the same solution, and the equilibrium swelling was assessed using the given equation.[21]

$$\text{Equilibrium Swelling } (\%) = \frac{W_f - W_i}{W_i} \times 100 \quad (1)$$

where W_f and W_i are the weights of the swollen and dried hydrogel samples, respectively.[22]

2.3. Drug Loading and Drug Encapsulation Efficiency

Drug loading and drug encapsulation efficiency were evaluated for ciprofloxacin-loaded AG/PAM/CMTKG hydrogel. A loaded-hydrogel pellet was submerged in a 100 ml pH 7.4 buffer solution and left to swell for 24 h.[23] The absorbance was measured using a UV-visible spectrophotometer, and the amount of ciprofloxacin was determined through a calibration curve.[24] The given equation was used to calculate the drug loading (DL%) and drug encapsulation efficiency (DEE%).[23,25]

$$DL\% = \frac{\text{Amount of drug loaded in hydrogel pellet}}{\text{Weight of hydrogel pellet}} \times 100 \quad (2)$$

$$DEE\% = \frac{\text{Amount of drug loaded in hydrogel pellet}}{\text{Theoretical amount of drug in a hydrogel pellet}} \times 100 \quad (3)$$

2.4. In Vitro Ciprofloxacin Release

The *in vitro* release study for ciprofloxacin-loaded AG/PAM/CMTKG hydrogel was conducted in an orbital incubator shaker under pH 1.2 and 7.4 buffer solutions. A hydrogel pellet was immersed in 100 ml of buffer solution at 37 °C. The 3 ml of the buffer solution was taken off after 1 hour, and the equivalent volume of fresh buffer solution was added. The experiment was performed in triplicate, and the absorbance of released ciprofloxacin was evaluated by a UV-visible spectrophotometer at λ_{max} 277 nm.[26] The calibration curve drawn for ciprofloxacin was used to determine drug content in the sample and given equation was used to calculate the drug release %.[27]

$$\text{Cumulative drug release } \left(\frac{M_t}{M_\infty}\right) = \frac{V_t C_t + V_e \sum_1^{n-1} C_{t-1}}{M_\infty} \quad (4)$$

Here, V_t is the total volume of buffer solution, V_e is the volume of buffer solution extracted, C_t and C_{t-1} are the concentration of drug release at time t and t-1.

2.5. Kinetic Modelling

The kinetic modelling was carried out to analyse the release of ciprofloxacin from a ciprofloxacin-loaded AG/PAM/CMTKG hydrogel using Hixson-Crowell, Zero-order, Higuchi, Korsmeyer–Peppas, and First-order models. Each model's coefficient of determination (R^2) values were compared, and the best-fit model is the one whose R^2 value is closest to unity.[28]

2.6. Characterization

PXRD was carried out using CuKα radiation at 2Θ range of 10–70° with a Bruker D8 diffractometer. The ATR-FTIR spectrophotometer (Model: PerkinElmer spectrum 2) was used to record ATR-FTIR spectra in the wavenumber range of 450–4000 cm^{-1}. The thermogravimetric analysis was performed using PerkinElmer TGA 4000 in an N_2 atmosphere at a heating rate of 10 °C/min from 30–600 °C. The SEM images were captured using SEM (Model: JEOL Japan Mode: JSM 6610LV).

3. Results and Discussion

3.1. Proposed Mechanism of AG/PAM/CMTKG and Ciprofloxacin-Loaded AG/PAM/CMTKG Hydrogel Formation

The hydrogels were prepared using a free radical mechanism with variable quantities of KPS and MBA, as presented in Table 1. The breakdown of KPS at 60 °C produced sulfate radicals, which react with AM to produce radicals on it through the cleavage of its vinyl bond. These AM radicals then initiated the propagation by attacking other acrylamide molecules, ultimately resulting in the formation of PAM. Moreover, the sulfate radical also produces the radical on CMTKG and AG by cleavage of the –OH bond. Additionally, the MBA forms the covalent crosslinked bond by joining these active radicals, forming a three-dimensional (3-D) AG/PAM/CMTKG hydrogel network. The hydrogen bonding interaction between ciprofloxacin and PAM contributes to the synthesis of a 3-D network of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel, as illustrated in Scheme 1.[29]

3.2. Swelling Analysis

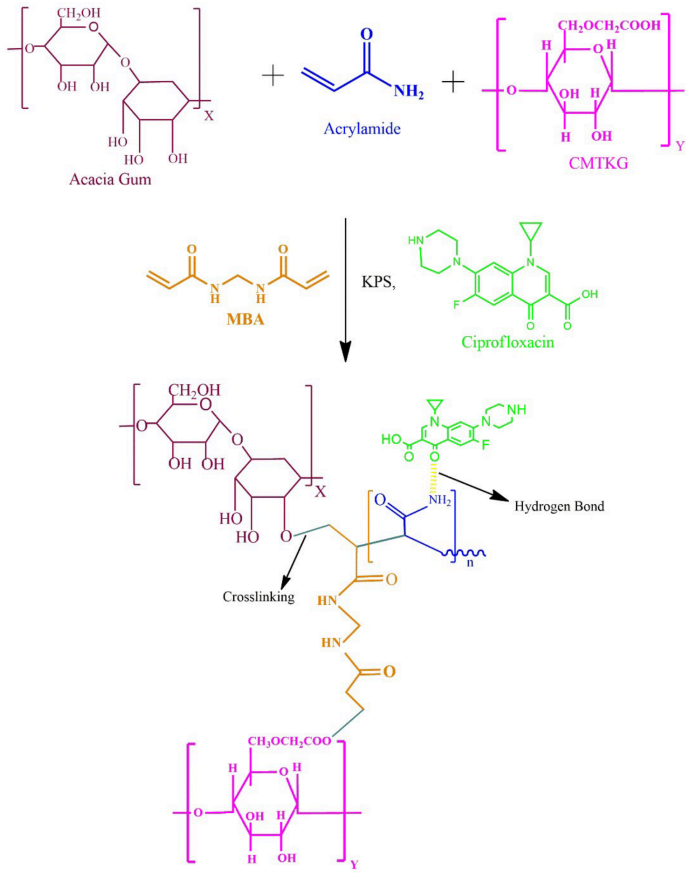
The swelling of the prepared hydrogels is reported at pH 1.2 and 7.4, as shown in Table 1. The impact of the initiator and cross-linker amount on the swelling of the hydrogels is displayed in Figure 2.

3.2.1 Impact of Cross-Linker

A notable decrease in the swelling capacity with an increase in MBA amount was observed in Figure 2(a). This observed decline in swelling can be attributed to the increased crosslink density

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Scheme 1. Mechanism of synthesis of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.

resulting from the rise in concentration of the crosslinker, which limits its ability to absorb the fluid effectively. At concentrations lower than 12 mg, the hydrogel does not exhibit significant crosslinking, resulting in no hydrogel network formation.[30]

3.2.2. Impact of Initiator

The MBA concentration held constant at 12 mg, and swelling effects were studied for variations in KPS content. The decrease in swelling is observed below 36 mg of KPS content, it is attributed to the presence of unreacted monomers due to a lesser number of initiator free radicals. Furthermore, at initiator concentrations exceeding 36 mg, swelling diminishes as oligomers form upon the rapid collision of monomer radicals (Figure 2(b)). Consequently, the solubility of the oligomeric components took place during the swelling experiment and hence, the overall swelling ratio of hydrogel decreases.[17]

Among the various hydrogels, formulation B-1 demonstrates the highest degree of swelling, as illustrated in Figure 3. Notably, all hydrogels show increased swelling at pH 7.4 ascribed to deprotonation of the carboxylic acid of CMTKG, resulting in the formation of COO^- ions. The repulsion between COO^- ions causes the polymer framework to relax, leading to extra free space for fluid to enter, hence an increased swelling observed. Conversely, the polymer network contracts due to the hydrogen bonds that form between the polymeric chains at pH 1.2 and thus swelling decreases.[29]

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

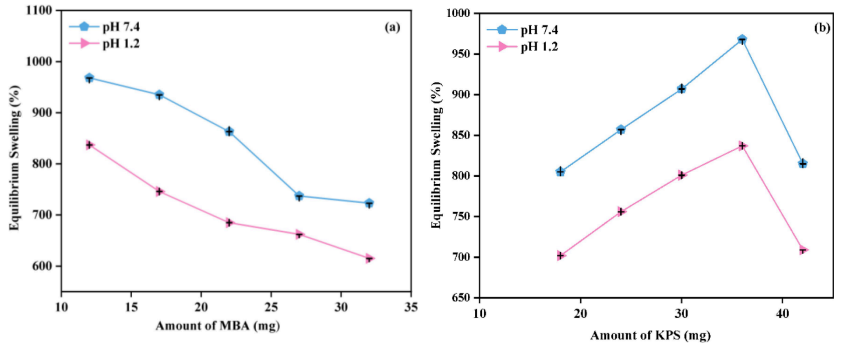


Figure 2. Impact of (**a**) MBA and (**b**) KPS on the equilibrium swelling % of synthesized hydrogels.

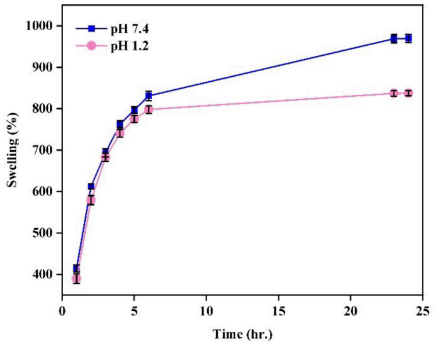


Figure 3. Swelling curve of AG/PAM/CMTKG hydrogel (B-1) in pH 1.2 and 7.4.

3.3. Characterization

PXRD analysis was done for ciprofloxacin, AG/PAM/CMTKG hydrogel and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel as shown in Figure 4. The ciprofloxacin sharp peak observed at 2Θ values of 11.5, 24.72, 19.39, 26.4, revealing its crystalline nature.[31] The PXRD of both ciprofloxacin-loaded AG/PAM/CMTKG and AG/PAM/CMTKG hydrogel demonstrates a broad-band, indicating their amorphous nature. Moreover, the ciprofloxacin-loaded AG/PAM/CMTKG-loaded hydrogel does not exhibit the peaks of ciprofloxacin, indicating that the ciprofloxacin has been thoroughly distributed throughout the amorphous crosslinked network.[32]

The ATR-FTIR of the AG/PAM/CMTKG hydrogel (B-1), ciprofloxacin-loaded AG/PAM/CMTKG, and MBA are shown in Figure 5. In the spectra of AG/PAM/CMTKG hydrogel and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel, the peaks at 3438 cm^{-1} and 3437 cm^{-1} are attributed to –OH and –NH overlapping stretching. The COO^- symmetric peak is evident at 1316 cm^{-1} and 1319 cm^{-1}, while the asymmetric peak appears at 1657 cm^{-1} and 1659 cm^{-1} for AG/CMTKG/PAM hydrogel and ciprofloxacin-loaded AG/CMTKG/PAM hydrogel. The symmetric C–H stretching peaks of AG/CMTKG/PAM hydrogel and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel are observed at 2925 cm^{-1} & 2854 cm^{-1}, and 2923 cm^{-1} & 2853 cm^{-1}, respectively.[17] Additionally, the C–O–C peak was observed at 1101 cm^{-1} and 1097 cm^{-1} for AG/PAM/CMTKG hydrogel and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel, respectively.[33] In the ATR-FTIR spectrum of MBA, the peaks are at 1304 cm^{-1}, referring to the C–N stretch, moved to 1316 cm^{-1} and 1319 cm^{-1} in the AG/PAM/CMTKG and ciprofloxacin-loaded AG/PAM/CMTKG, indicating decreased conjugation ascribed to cross-linking. Consequently, the AG/PAM/CMTKG and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel have been successfully developed. When comparing ciprofloxacin-loaded AG/PAM/CMTKG hydrogel to the unloaded hydrogel, no additional peak was observed, indicating that the drug and polymer chain only interact weakly or physically, primarily via hydrogen bonding.[34]

TGA analysis of AG/CMTKG/PAM and ciprofloxacin-loaded AG/CMTKG/PAM hydrogel revealed three distinct decomposition stages, as depicted in Figure 6. The initial degradation, occurring between 30–238 °C, is attributed to the loss of bound water and the initial breakdown of the polymeric chain in AG/CMTKG/PAM and ciprofloxacin-loaded AG/CMTKG/PAM hydrogel, respectively, resulting in weight losses of 20.17% and 21.61%.[35] The second decomposition stage, observed at 238–375 °C is associated with the removal of carboxymethyl and hydroxyl groups from the polymeric chain, with weight losses of 47.8% and 30.7% in AG/CMTKG/PAM and ciprofloxacin-loaded AG/CMTKG/PAM hydrogel. In the third stage, the breakdown of crosslinking in the polymeric network is observed, leading to weight losses of and 28.15% and 41.35% from 375–600 °C. The ciprofloxacin-loaded AG/CMTKG/PAM hydrogel exhibited increased thermal stability compared to the AG/CMTKG/PAM hydrogel alone, attributed to the hydrogen bonding (physical interactions) between the ciprofloxacin and the polymer.[35]

SEM images of the AG/PAM/CMTKG hydrogel and ciprofloxacin-loaded AG/CMTKG/PAM hydrogel are presented in Fig-

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

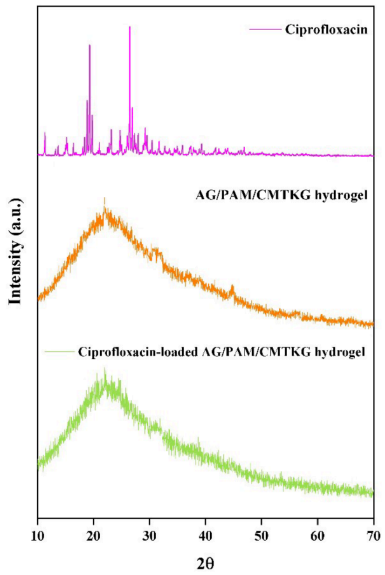


Figure 4. PXRD pattern of ciprofloxacin, AG/PAM/CMTKG hydrogel (B-1), and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.

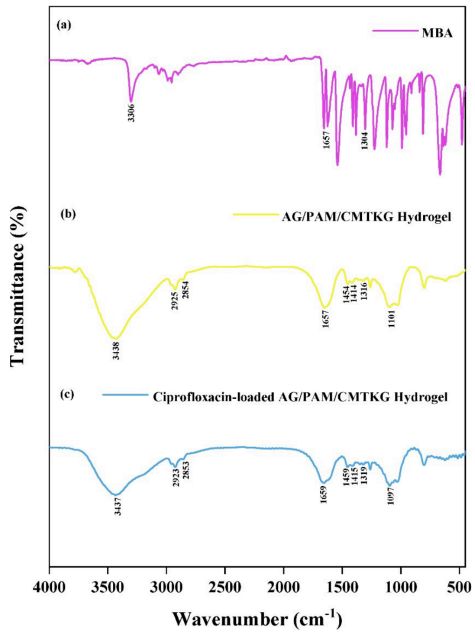


Figure 5. ATR-FTIR spectra of MBA, AG/PAM/CMTKG hydrogel and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.

ure 7. The AG/PAM/CMTKG hydrogel exhibits a porous structure and uneven surface morphology, suggesting strong fluid absorption and drug loading efficiency. On the other hand, SEM micrographs of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel reveal a smoother and less porous surface than those of unloaded hydrogel as the drug occupies the porous hydrogel matrix.[36]

3.4. Drug Loading and Drug Encapsulation Efficiency

The ciprofloxacin-loaded AG/PAM/CMTKG hydrogel was assessed for drug encapsulation efficiency and drug loading. The drug loading (%) is noticed as 18.22%, and the drug encapsulation efficiency (%) is observed as 58.17%.

3.5. In Vitro Ciprofloxacin Release

The *in vitro* release study is conducted for the ciprofloxacin-loaded AG/PAM/CMTKG hydrogel, as depicted in Figure 8. Notably, the highest release of ciprofloxacin from this hydrogel was recorded as 16.6% at pH 1.2 and 56.3% at pH 7.4. A notable greater drug release is observed at pH 7.4 compared to pH 1.2, indicating the deprotonation of the carboxylic acid group of CMTKG. These deprotonated carboxylate ions produce an electrostatic repulsion, expanding the hydrogel network, leading to increased swelling, which in turn increases the

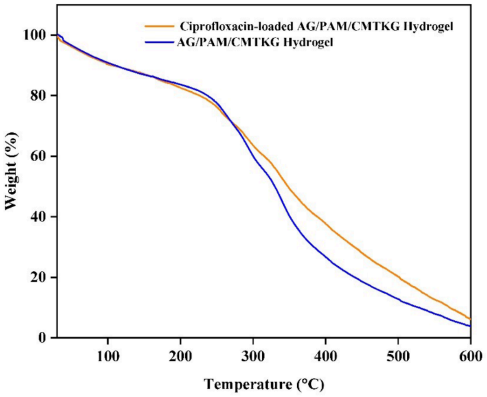


Figure 6. TGA curve of AG/PAM/CMTKG hydrogel and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.

diffusion of ciprofloxacin from the hydrogel. Conversely, at pH 1.2, the polymeric network contracts due to the hydrogen bonding among polymers. Consequently, the swelling decreases, causing a lower ciprofloxacin release. The ciprofloxacin

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

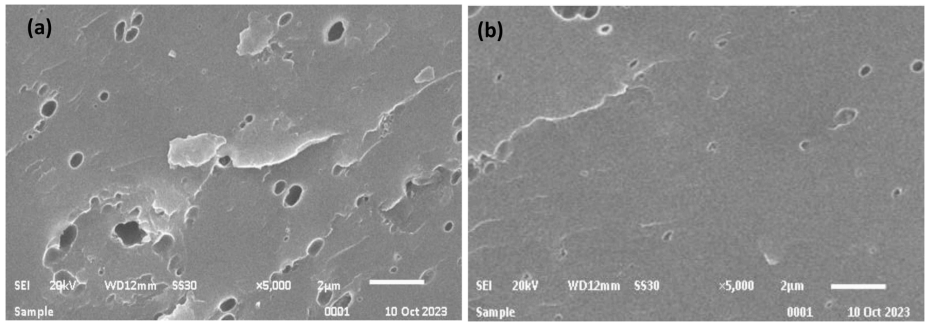


Figure 7. SEM image of AG/PAM/CMTKG and ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.

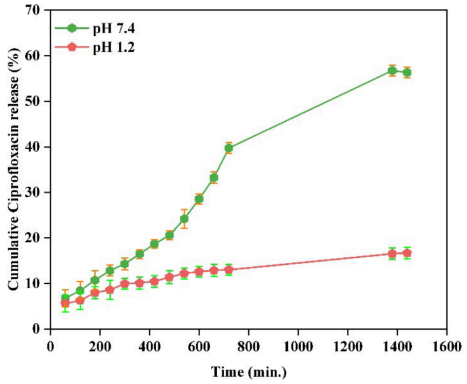


Figure 8. Ciprofloxacin release from ciprofloxacin-loaded AG/PAM/CMTKG hydrogel in pH 7.4 and 1.2.

was released without a burst effect, indicating a gradual release from the hydrogels. This slow release ensures an optimal and sustained drug level over extended periods, reducing the need for frequent dosage. The findings suggest that the pH-dependent release of ciprofloxacin can be achieved using the ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.[30]

3.6. Kinetic Modeling

Several models were employed to study the ciprofloxacin release kinetics of the ciprofloxacin-loaded AG/PAM/CMTKG hydrogel. Among these models, the Korsmeyer–Peppas model shows the highest R^2 value, as observed in Table 2. At pH 7.4, the R^2-value and n-value for the Korsmeyer–Peppas model were determined to be 0.9885 and 0.622, respectively. At pH 1.2, the R^2 value was obtained as 0.9947, and the n value as 0.3149 (Figure 9). These n values suggest that the release of ciprofloxacin at pH 1.2 follows the Fickian diffusion mechanism, signifying a diffusion process followed for drug release. However, non-Fickian diffusion is observed at pH 7.4, suggest-

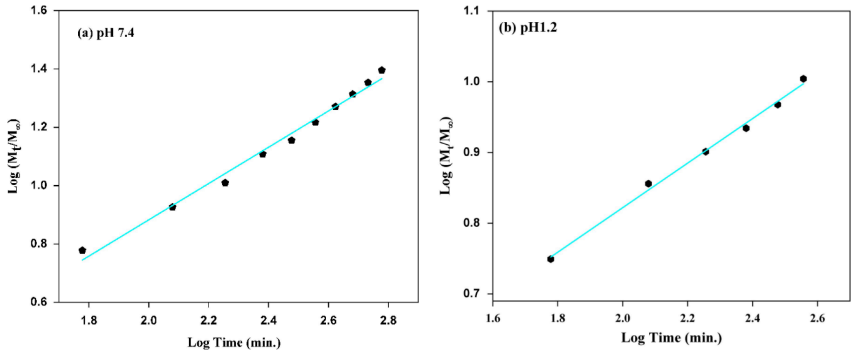


Figure 9. Release kinetics plot of ciprofloxacin corresponding to the Korsmeyer–Peppas model in (a) pH 7.4 and (b) pH 1.2.

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. Kinetic modelling data of ciprofloxacin-loaded AG/PAM/CMTKG hydrogel.

Model	Equation	pH 7.4		pH 1.2		Ref.
		n	R^2	n	R^2	
Zero Order	$M_t = M_\infty + k_0\, t$	–	0.9759	–	0.8586	[38]
Korsmeyer–Peppas	$M_t/M_\infty = kt^n$ k = kinetic constant n = diffusion exponent	0.622	0.9885	0.3149	0.9947	[39]
First Order	$\text{Log } M_t = \text{Log } M_\infty + \frac{kt}{2.303}$ k = rate constant	–	0.7659	–	0.9741	[40,41]
Higuchi	$M_t/M_\infty = k_H t^{1/2}$ k_H = kinetic constant	–	0.9537	–	0.9897	[42]
Hixson-Crowell	$(M_t)^{1/3} - (M_\infty)^{1/3} = k_{HC}.t$ k_{HC}=Hixson Crowell constant	–	0.8248	–	0.8575	[39]

ing the rate of diffusion and rate of polymer chain relaxation were similar.[37]

3.7. Conclusions

A pH-sensitive ciprofloxacin-loaded AG/PAM/CMTKG hydrogel was fabricated with KPS (initiator) and MBA (crosslinker). The synthesized hydrogel was successfully characterized by ATR-FTIR, PXRD, SEM and TGA techniques. Furthermore, the impact of the KPS and MBA concentrations on the swelling of developed hydrogel was examined. The findings imply that swelling decreases with a rise in MBA and increases with rise in KPS content, but a further increase in KPS beyond a particular concentration resulted in a decrease in swelling capacity. The swelling and in vitro drug release study were accomplished at pH 1.2 and 7.4 and perceived to be greater at pH 7.4. The release kinetics of ciprofloxacin revealed a good fit with the Korsmeyer–Peppas model at both pH 1.2 and 7.4, demonstrating a Fickian diffusion mechanism at pH 1.2 and non–Fickian diffusion at pH 7.4. Consequently, the fabricated AG/PAM/CMTKG hydrogel shows potential for pH-dependent delivery of ciprofloxacin.

Ethical Approval

Not applicable.

AcknowledgementsFunding

The authors are grateful to the Delhi Technological University, New Delhi, India and the Council of Scientific & Industrial Research (CSIR), India.

AcknowledgementsFunding

No funding was received for conducting this study.

Conflict of Interests

The authors have no relevant financial or non-financial interests to disclose.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Keywords: Acacia gum · Biopolymer · Drug delivery · Hydrogel · Kinetic modelling

[1] M. Rizwan, R. Yahya, A. Hassan, M. Yar, A. D. Azzahari, V. Selvanathan, F. Sonsudin, C. N. Abouloula, *Polymer* **2017**, *9*, 137https://doi.org/10.3390/polym9040137.
[2] Z. Yang, H. Peng, W. Wang, T. Liu, *J. Appl. Polym. Sci.* **2010**, *116*, 2658–2667.
[3] K. Wang, Q. Fu, X. Chen, Y. Gao, K. Dong, *RSC Adv.* **2012**, *2*, 7772–7780.
[4] K. Sohail, I. U. Khan, Y. Shahzad, T. Hussain, N. M. Ranjha, *Brazil. J. Pharm. Sci.* **2014**, *50*, 173–184.
[5] U. D. Shivhare, K. B. Jain, V. B. Mathur, K. P. Bhusari, A. A. Roy, *Digest J. Nanomat. Biostruct.* **2009**, *4*, 285–290.
[6] G. Sennakesavan, M. Mostakhdemin, L. K. Dkhar, A. Seyfoddin, S. J. Fatihhi, *Polym. Degrad. Stab.* **2020**, *180*, 109308.
[7] Q. B. Wei, F. Fu, Y. Q. Zhang, Q. Wang, Y. X. Ren, *J. Polym. Res.* **2014**, *21*, 453 https://doi.org/10.1007/s10965-014-0453-0.
[8] P. Meena, P. Singh, S. G. Warkar, *Colloid Polym. Sci.* **2023**, *302*, 377–391 https://doi.org/10.1007/s00396-023-05198-8.
[9] H. M. Aldawsari, S. Singh, N. A. Alhakamy, R. B. Bakhaidar, A. A. Halwani, S. M. Badr-Eldin, *Pharmaceutica* **2021**, *13*, 1554, https://doi.org/10.3390/pharmaceutics13101554.
[10] B. Singh, S. Sharma, A. Dhiman, *Carbohydr. Polym.* **2017**, *165*, 294–303.
[11] D. Balodhi, S. Khushbu, G. Warkar, A. Rani, R. K. Gupta, *Indian J. Chem. Technol.* **2021**, *28*, 262–273.
[12] C. Sanchez, M. Nigen, V. Mejia Tamayo, T. Doco, P. Williams, C. Amine, D. Renard, *Food Hydrocolloids* **2018**, *78*, 140–160.
[13] L. Zhang, H. Shi, X. Tan, Z. Jiang, P. Wang, J. Qin, *Front. Chem.* **2022**, *10*, 1–11.
[14] K. M. Hosny, *AAPS PharmSciTech* **2010**, *11*, 241–246.
[15] J. Du, I. M. El-Sherbiny, H. D. Smyth, *Ageing Int.* **2014**, *15*, 1535–1544.
[16] D. Das, S. Pal, *Int. J. Biol. Macromol.* **2015**, *72*, 171–178.
[17] P. Meena, P. Singh, S. G. Warkar, *Eur. Polym. J.* **2023**, *197*, 112340.
[18] C. Chang, B. Duan, L. Zhang, *Polymer* **2009**, *50*, 5467–5473.
[19] H. Khan, R. N. Shukla, A. K. Bajpai, *Mater. Sci. Eng. C* **2016**, *61*, 457–465.

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[20] L. S. Tan, H. L. Tan, K. Deekonda, Y. Y. Wong, S. Muniyandy, K. Hashim, J. Pushpamalar, *Carbohydr. Polym. Technol. Appl.* **2021**, *2*, 100084.
[21] A. R. Karimi, B. Rostaminejad, L. Rahimi, A. Khodadadi, H. Khanmohammadi, A. Shahriari, *Int. J. Biol. Macromol.* **2018**, *118*, 1863–1870.
[22] C. García-Astrain, A. Gandini, C. Peña, I. Algar, A. Eceiza, M. Corcuera, N. Gabilondo, *RSC Adv.* **2014**, *4*, 35578–35587.
[23] B. Niu, J. Jia, H. Wang, S. Chen, W. Cao, J. Yan, X. Gong, X. Lian, W. Li, Y. Y. Fan, *Int. J. Biol. Macromol.* **2019**, *141*, 1191–1198.
[24] J. X. Fu, H. J. Wang, Y. Q. Zhou, J. Y. Wang, *Mater. Sci. Eng. C* **2009**, *29*, 1161–1166.
[25] N. Dilbaghi, H. Kaur, M. Ahuja, P. Arora, S. Kumar, *J. Exp. Nanosci.* **2014**, *9*, 1015–1025.
[26] D. H. Hanna, G. R. Saad, *Bioorg. Chem.* **2019**, *84*, 115–124.
[27] S. Alinavaz, G. R. Mahdavinia, H. Jafari, M. Hazrati, A. Akbari, *J. Mol. Struct.* **2021**, *1225*, 129095.
[28] A. Ashames, K. Ullah, M. Al-Tabakha, S. A. Khan, N. Hassan, A. Mannan, M. Ikram, M. Buabeid, G. Murtaza, *PLoS One* **2022**, *17*, e0271623, https://doi.org/10.1371/journal.pone.0271623.
[29] F. Sabbagh, I. I. Muhamad, *J. Taiwan Inst. Chem. Eng.* **2017**, *72*, 182–193.
[30] S. A. Ghumman, S. Noreen, H. Hameed, M. A. Elsherif, R. Shabbir, M. Rana, K. Junaid, S. N. A. Bukhari, *Gels* **2022**, *8*, 291, https://doi.org/10.3390/gels8050291.
[31] B. Mallikarjuna, M. S. Rao, C. V. Prasad, K. Chowdoji Rao, K. S. V. Krishana Rao, M. C. S. Subha, *J. Appl. Pharmacol.* **2011**, *1*, 171–177.
[32] M. Suhail, A. Khan, J. M. Rosenholm, M. U. Minhas, P. C. Wu, *Gels* **2021**, *7*, 1–16.
[33] K. K. Mali, S. C. Dhawale, R. J. Dias, *Int. J. Biol. Macromol.* **2017**, *105*, 463–470.
[34] H. Roy, C. Brahma, S. Nandi, K. Parida, *Int. J. Appl. Basic Med. Res.* **2013**, *3*, 55.
[35] A. G. Rusu, M. I. Popa, G. Lisa, L. Vereştiuc, *Thermochim. Acta* **2015**, *613*, 28–40.
[36] H. Shoukat, F. Pervaiz, S. Noreen, M. Nawaz, R. Qaiser, M. Anwar, *Polym. Bull.* **2020**, *77*, 1869–1891.
[37] M. Suhail, Y. H. Hsieh, A. Khan, M. U. Minhas, P. C. Wu, *Gels* **2021**, *7*, 68, https://doi.org/10.3390/gels7020068.
[38] N. Goonoo, A. Bhaw-Luximon, R. Ujoodha, A. Jhugroo, G. K. Hulse, D. Jhurry, *J. Control. Rel.* **2014**, *183*, 154–166.
[39] N. Rafique, M. Ahmad, M. U. Minhas, S. F. Badshah, N. S. Malik, K. U. Khan, *Polym. Bull.* **2022**, *79*, 4535–4561.
[40] D. Das, R. Das, J. Mandal, A. Ghosh, S. Pal, *J. Appl. Polym. Sci.* **2014**, *131*, 1–12.
[41] Tushar, Y. Saraswat, P. Meena, S.G. Warkar, *Colloid Polym. Sci.* **2023**, https://doi.org/10.1007/s00396-023-05135-9.
[42] Kajal, R. Kumar, P. Meena, S. G. Warkar, *Colloid Polym. Sci.* **2023**, https://doi.org/10.1007/s00396-023-05152-8.

Manuscript received: April 6, 2024

© 2024 Wiley-VCH GmbH

23656549, 2024, 34, Downloaded from https://chemistry-europe-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/10.1002/slct.202401639. By Harvard University, Wiley Online Library on [26/09/2024]. See the Terms and Conditions (https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

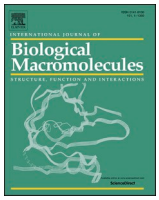
International Journal of Biological Macromolecules 280 (2024) 136029



Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac



Research Paper



Tailoring pH-sensitive carboxymethyl tamarind kernel gum-based hydrogel for an efficient delivery of hydrophobic drug indomethacin

Priyanka Meena, Poonam Singh *, Sudhir G. Warkar *

Department of Applied Chemistry, Delhi Technological University, Delhi, India

ARTICLE INFO

Keywords:
Biopolymer
Drug release
Hydrogel

ABSTRACT

Polyacrylamide hydrogels have gained attention in the drug delivery field for their pH-dependent nature. Nevertheless, their limited degradability and lower entrapment efficiency for hydrophobic drugs hinder their utility in controlled drug release. This research aims to design a degradable pH-sensitive hydrogel for delivering the hydrophobic drug indomethacin to the colon. This work developed and optimized the hydrogels based on β-cyclodextrin, carboxymethyl tamarind kernel gum, and polyacrylamide with varying amounts of polyethylene glycol diacrylate. The optimized hydrogel exhibits 76.52 % gel fraction, 89.21 % porosity, 1000.27 % swelling, and 90.0 % equilibrium water content. The hydrogel was characterized using Attenuated Total Reflection-Fourier Transform Infrared spectroscopy, confirming the successful crosslinking of the synthesized hydrogel. Powder X-ray Diffraction revealed their amorphous nature while Scanning Electron Microscopy showed a porous surface morphology of the hydrogel. Moreover, rheology confirmed the hydrogel's elastic nature. Notably, the hydrogel demonstrated a drug release of 60.26 % at pH 7.4. Kinetic modelling of indomethacin release data indicated a Fickian diffusion release mechanism. Cytotoxicity tests on HCT-116 cells showed 79 % viability, and the hydrogel fully degraded within 10 days. These results confirmed the potential of synthesized β-CD/PAM/CMTKG hydrogel for controlled indomethacin delivery to the colon.

1. Introduction

The emergence of new-generation drugs has sparked increased interest in innovative controlled drug delivery systems [1]. Unlike conventional drug delivery systems, where the concentration of the drug in the blood rapidly rises after intake, followed by a decline over time, controlled drug delivery systems prioritize maintaining a consistent delivery rate. Hydrogels play a crucial role as a controlled drug delivery system due to their time-dependent swelling behavior and pH-dependent nature. Hydrogels are three-dimensional polymeric networks with hydrophilic functional groups such as -OH, $-SO_3H$, $-CONH_2$, etc., facilitating them to absorb bio-fluids [2].

Moreover, developing environment-sensitive delivery systems responsive to changes in external factors like pH, temperature, light, ionic strength, etc., is crucial for targeted and controlled drug delivery [3]. Among these, pH-sensitive hydrogels have been extensively investigated for their ability to respond to pH variations that occur in the gastrointestinal tract. The pH-sensitive hydrogels can shield drugs from enzymatic hydrolysis and stomach acid, making them ideal for drug-delivery systems. Nowadays, natural biopolymer-based pH-sensitive hydrogels, particularly those derived from polysaccharides such as sodium alginate, chitosan, carboxymethyl tamarind kernel gum, gelatin, etc., have gained interest due to their low cost, good biodegradability, and biocompatibility [1,4]. For example, Akhlaq et al. synthesized a pH-sensitive methotrexate-loaded gelatin and polyvinyl alcohol hydrogel, which exhibited 94.3 % drug release at pH 1.2 and 83.21 % at pH 7.4 [5]. Patel et al. developed hydrogel using carboxyethyl xanthan gum and carboxymethyl xanthan gum for repaglinide drug delivery, showing approximately 80 % drug release in gastric fluids [6]. Additionally, Nawaz et al. reported gelatin/polyvinyl pyrrolidone hydrogels for oral captopril delivery, showing faster drug release at pH 1.2 compared to pH 7.4 (colon pH) [7]. However, the authors disclosed controlled oral drug delivery potential, but they did not focus on colon targeting, even the research on carboxymethyl tamarind kernel gum is limited. Moreover, they did not study with polyacrylamide combination.

Tamarind kernel gum (TKG), extracted from the seeds of the *Tamarindus indica* tree, is a natural, cost-effective, and biocompatible polysaccharide used in drug delivery [8,9]. For example, Pal et al. developed

* Corresponding authors.
E-mail addresses: poonam@dtu.ac.in (P. Singh), sgwarkar@dtu.ac.in (S.G. Warkar).

https://doi.org/10.1016/j.ijbiomac.2024.136029
Received 24 April 2024; Received in revised form 17 September 2024; Accepted 24 September 2024
Available online 25 September 2024
0141-8130/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

tamarind kernel gum/alginate hydrogel to deliver gliclazide [10]. Similarly, Jana et al. synthesized chitosan and tamarind kernel gum polymeric network for aceclofenac release [11]. However, its limited swelling capacity restricted its broader utility in biomedical applications. Hence, the carboxymethyl derivative of TKG, termed carboxymethyl tamarind kernel gum (CMTKG), has attracted significant attention. [8]. It is a water-soluble biopolymer composed of glucose, xylose, and galactose in a molar ratio of 3:2:1. It possesses high viscosity, extended shelf life, exhibits anionic characteristics, and regulates high swelling and drug release properties [8,9]. For example, Manya et al. synthesized a sodium alginate/poly(sodium acrylate)/CMTKG hydrogel, which exhibited a maximum swelling of 2500 % after 24 h, resulting in a higher drug release (80 %) at pH 7.4 [12].

Moreover, biopolymer-based hydrogels often lack mechanical strength, whereas synthetic polymers have good mechanical properties. Therefore, a blend of natural polymers with biocompatible synthetic ones such as polyacrylamide (PAM) can result in hydrogels with enhanced characteristics. Polyacrylamide (PAM) is a pH-sensitive, water-soluble polymer known for its non-toxicity and excellent biocompatibility, making it an ideal candidate for drug delivery [13].

PAM, being hydrophilic, interacts mainly with hydrophilic drugs, resulting in high loading efficiency for hydrophilic drugs but low efficiency for hydrophobic drugs [14,15]. For instance, Qiao et al. synthesized gelatin/PAM hydrogel for the controlled release of the hydrophilic drug lidocaine HCl, which exhibited a drug loading of 75 % [14]. Pandey et al. reported the synthesis of pectin/pam hydrogel for the delivery of the hydrophobic drug budesonide. The hydrogel exhibited a drug loading capacity of 8.89 % [15]. Moreover, exploration in the realm of hydrophobic drugs has been relatively limited. Furthermore, existing studies on hydrogel-based hydrophobic drug delivery still face challenges, including difficulty in achieving targeted delivery. For example, Sarfraz et al. developed β-cyclodextrin-g-poly(2-acrylamido-2-methylpropane sulfonic acid) hydrogel to deliver the hydrophobic drug indomethacin. Although the finding revealed that 27–36 % indomethacin was released at pH 1.2, this non-targeted high release in the stomach can lead to gastric irritation [16]. This study seeks to address this research gap by developing a pH-responsive hydrogel matrix and investigating the impact of beta-cyclodextrin (β-CD) on the controlled release of the hydrophobic drug indomethacin in the colon.

The β-CD has a truncated cone-like structure with seven glucose units, having a hydrophobic inner cavity and a hydrophilic outer surface, demonstrating significant potential in forming inclusion host-guest complexes with numerous hydrophobic guest (drug molecules), such as aspirin, indomethacin, etc. [17,18]. This property enables β-CD to enhance the loading of hydrophobic drugs such as indomethacin, thereby extending their therapeutic efficacy.

Indomethacin, classified as a non-steroidal anti-inflammatory drug (NSAID), finds extensive use in addressing chronic pain and inflammatory conditions, including rheumatoid and osteoarthritis [19]. Although indomethacin was approved in 1984, there is very limited literature on its delivery using hydrogel-based drug delivery systems [20]. Hydrogels, being hydrophilic, are primarily used for hydrophilic drug delivery, and there is limited research on their use for hydrophobic drugs like indomethacin. Additionally, its shorter half-life necessitates repeated dosing, which can lead to serious health issues, including diarrhoea, irritation, gastric bleeding, and nausea [21].

Therefore, in this study, efforts are directed toward the synthesis of a hydrogel based on β-CD, CMTKG, and PAM via the free radical mechanism for controlled indomethacin release. The fabricated hydrogels are characterized using ATR-FTIR, PXRD, and SEM techniques. Sol-gel fraction, swelling (%), equilibrium water content, porosity, pore volume, pore diameter, and network parameters are investigated for all the synthesized hydrogels. Furthermore, the drug release mechanism is explored using various mathematical models, including Hixson-Crowell, Korsmeyer-Peppas, First-Order, Zero-Order, and Higuchi. Moreover, degradability and cytotoxicity studies have been conducted for the synthesized hydrogel.

2. Materials and methods

2.1. Materials

β-Cyclodextrin (β-CD, MW = 1134.98 g/mol, CDH, New Delhi, India), Acrylamide (AM, MW = 71.08 g/mol, Fischer Scientific, Mumbai, India), Poly(ethylene glycol diacrylate) (PEGDA, M_n = 2000, Sigma Aldrich, India), potassium persulfate (KPS, MW = 270.30 g/mol, Fischer Scientific, Mumbai, India), and indomethacin (Unicure India Pvt. Ltd., Noida, India) were used as obtained. CMTKG (0.20° of substitution, MW = 9.14×10^5 g/mol) was courteously gifted by Hindustan Gum Pvt. Ltd., Haryana, India.

2.2. Synthesis of PAM/CMTKG and β-CD/PAM/CMTKG hydrogels

The series of β-CD/PAM/CMTKG hydrogels (W-1 to W-4) with varied quantities of the PEGDA (1.5–3 ml) were fabricated using a free radical mechanism, as displayed in Table 1. To fabricate β-CD/PAM/CMTKG hydrogels, a specific quantity of CMTKG, β-CD, and AM was dissolved in distilled water and stirred for 30 min. To the above solution, the fixed amount of KPS and PEGDA were added and stirred for 30 min. Then, the solution was poured into the test tubes and placed inside the water bath, set at 60 °C for 1 h. Finally, the developed hydrogel was extracted from the test tube, cut into slices, and oven-dried at 55 °C for 48 h [22].

The procedure for the fabrication of control (PAM/CMTKG hydrogel) was the same as the one described above. A fixed quantity of CMTKG and AM was dissolved in distilled water and stirred for 30 min. To this solution, the fixed amount of KPS and PEGDA were added and stirred for 30 min. Then, the mixture was poured into the test tubes and placed inside the water bath, set at 60 °C for 1 h. The developed hydrogel was extracted from the test tube, cut into slices, and oven-dried at 55 °C for 48 h (Fig. 1).

2.3. Characterization

PXRD measurements were performed via a Bruker D8 Discover Diffractometer utilizing Cu Kα radiation (λ = 1.5418 Å), within 2θ range of 5–70°. The ATR-FTIR spectra were captured with a Nicolet iS50 FTIR Tri-detector spectrometer within the wavenumber range from 450 to 4000 cm^{-1}. DSC analysis was carried out using PerkinElmer DSC8000 within a temperature range of 0–250 °C at a heating rate of 10 °C/min. The micrographic images were captured using SEM (Model: JEOL Japan Mode: JSM 6610LV).

2.4. Swelling analysis

A swelling assessment was carried out for all synthesized hydrogels in both HCl-KCl buffer (pH 1.2) and phosphate buffer saline (pH 7.4) for 24 h, using gravimetric analysis. In the study, a specific amount (M_{DH}) of dried hydrogels was placed into the buffer solution. At time interval (1 h), the hydrogels were removed from the solutions, carefully dried using a filter paper, and reweighed (M_{SH}). The given equation was applied for the calculation of swelling [23].

$$\text{Swelling } (\%) = \frac{M_{SH} - M_{DH}}{M_{DH}} \times 100 \quad (1)$$

Moreover, when hydrogel reaches equilibrium, its swelling is termed equilibrium swelling %, calculated using the same formula mentioned above.

2.5. Equilibrium water content (EWC)

The equilibrium water content of hydrogel discs was measured in HCl-KCl buffer (pH 1.2) and phosphate buffer saline (pH 7.4). The

Table 1
Compositions of CMTKG-based hydrogels and their equilibrium swelling % ($p < 0.05$, $n = 3$).

Sample code	CMTKG (g)	β-CD (g)	AM (g)	PEGDA (ml)	Equilibrium swelling (%)	
					pH 1.2	pH 7.4
W-1	0.1	0.1	2	1.5	750.21 ± 0.32	1000.27 ± 0.42
W-2	0.1	0.1	2	2.0	675.15 ± 0.98	925.18 ± 0.85
W-3	0.1	0.1	2	2.5	574.89 ± 0.84	822.12 ± 0.91
W-4	0.1	0.1	2	3.0	457.17 ± 0.75	700.28 ± 0.93

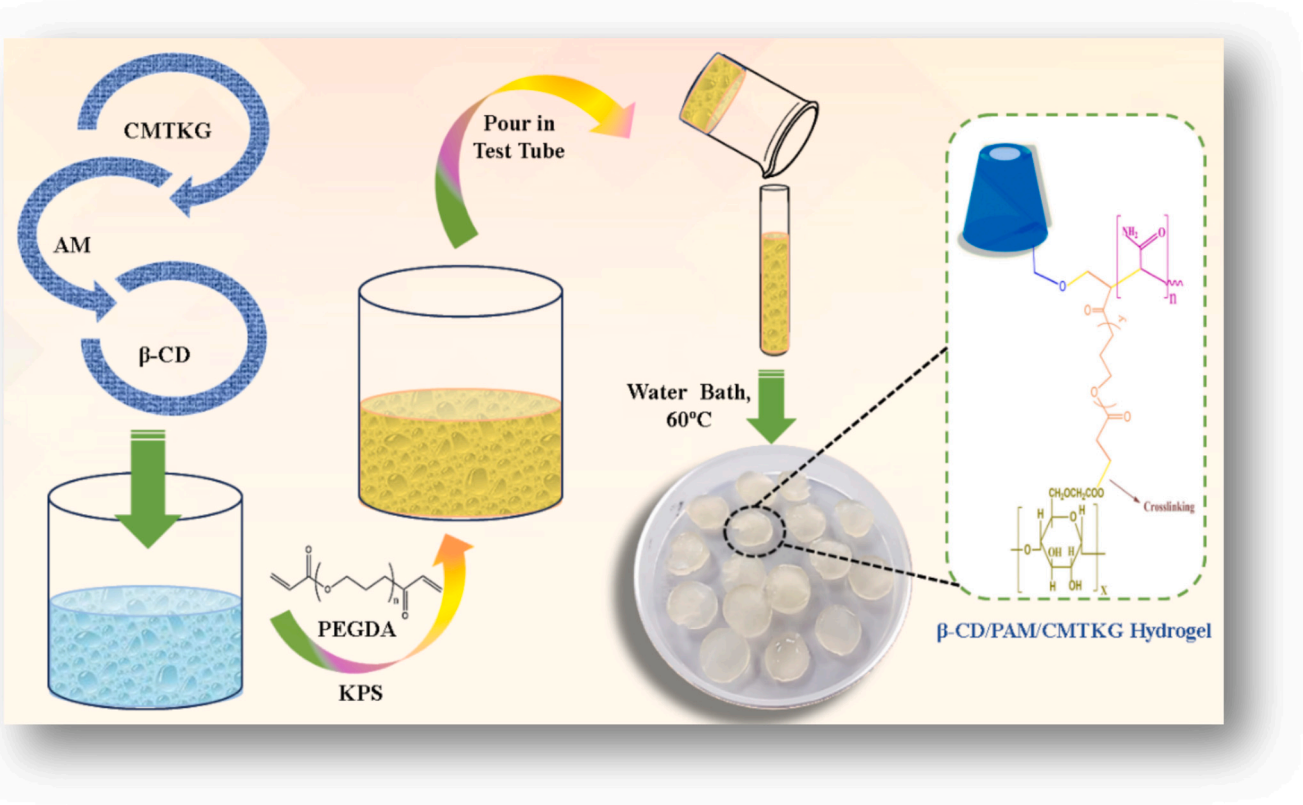


Fig. 1. Schematic illustration of the β-CD/PAM/CMTKG hydrogel fabrication.

specific amount (M_{DR}) of hydrogel discs was allowed to swell for 24 h to achieve equilibrium. After wiping away the excess surface fluid, the swollen hydrogel discs were weighed (M_{EH}). The equilibrium water content percent (EWC %) was calculated using Eq. (2) [24].

$$\text{EWC}\ (\%)\frac{M_{EH} - M_{DR}}{M_{EH}} \times 100 \tag{2}$$

2.6. Determination of hydrogel network parameters

To determine all the network parameters for the synthesized hydrogel, the Flory and Rehner swelling theory was employed. The calculations were performed based on swelling data obtained from hydrogel swelling in pH 7.4 and 1.2 over 24 h.

2.6.1. Volume fraction

The proportion of the total volume of a hydrogel occupied by the polymer component represents the polymer volume fraction (ϕ), calculated using given Eq. (3) [2].

$$\phi = \frac{V_{DH}}{V_{SH}} \tag{3}$$

where V_{SH} and V_{DH} denote the volume of swollen and dried hydrogels, respectively, determined using given Eq. (4).

$$V = \pi r^2 h \tag{4}$$

Here, r represents the radius and h is the height of the cylindrical hydrogel discs, calculated using a Vernier calliper [25].

2.6.2. Polymer-solvent interaction parameter (χ)

The assessment of polymer compatibility with the surrounding fluid was conducted through the computation of polymer-solvent interaction parameters (χ), utilizing the given Eq. (5) [2].

$$\chi = \frac{1}{2} + \frac{\phi}{3} \tag{5}$$

2.6.3. Average molecular weight between cross-links ($\overline{M}_c$)

This parameter represents the average molecular weight of the polymer chains between two crosslinking points in a polymer network and is computed using Eq. (6).

$$\overline{M}_c = -\frac{d_{SH} V_w \left(\phi^{\frac{1}{3}} - \frac{1}{2}\phi\right)}{\left(\ln(1-\phi) + \phi + \chi\phi^2\right)} \tag{6}$$

where V_W is the volume of the solvent, i.e., water [26]. The d_{SH} is the

density of swollen hydrogel, assessed using Eq. (7).

$$d_{SH} = \frac{W_{SH}}{V_{SH}} \tag{7}$$

where W_{SH}denotes the weight of the swollen hydrogel.

2.6.4. Crosslinking density

Crosslinking density (ρ) is a measure of how densely packed the polymer chains are in a three-dimensional network, computed using given Eq. (8) [25,27].

$$\rho = \frac{d_{SH}}{\overline{M}_c} \tag{8}$$

2.7. Sol-gel analysis

To analyze the non-crosslinked (sol) and crosslinked (gel) components of the developed hydrogels, a sol-gel analysis was performed. The dried hydrogel weight was initially measured (M_i) and then immersed in distilled water for 24 h. Then, the hydrogel was subjected to drying in the oven at 50 °C until a constant weight (M_e) was reached. The sol and gel fraction (%) was assessed using Eqs. (9) and (10) [2].

$$\text{Sol Fraction (\%)} = \frac{M_i - M_e}{M_e} \times 100 \tag{9}$$

$$\text{Gel fraction (\%)} = 100 - \text{Sol fraction} \tag{10}$$

2.8. Porosity

The porosity evaluation for all the synthesized hydrogels was carried out using the solvent displacement technique. Initially, the weighed dried hydrogel disc (W_i) was immersed in 5 ml of hexane for 1 h. Then, the hydrogel discs were extracted from the hexane, the disc was wiped with filter paper, and weight (W_f) was noted [28]. The porosity of the hydrogel discs was determined using the given Eq. (11).

$$\text{Porosity (\%)} = \frac{W_f - W_i}{\rho V} \times 100 \tag{11}$$

where, ρ represents hexane density (0.659 g/ml), while V corresponds to the volume of the cylindrical hydrogel disk, calculated using the formula $V = \pi r^2 h$, where 'r' is the radius and 'h' is the height of the disk, quantified using a Vernier calliper [29].

2.9. BET analysis

The specific surface area, pore volume, and pore diameter evaluation for all the synthesized hydrogels were carried out using the BET-BJH (Barrett-Joyner-Halenda) nitrogen adsorption-desorption technique with Quanta chrome Nova-1000 using liquid nitrogen temperature at 77 k.

2.10. Rheology

Rheological analysis was conducted using a stress-controlled Anton Paar Modular Compact Rheometer 302 (MCR) with parallel plates with a 25 mm diameter and a 1 mm gap. Measurements were taken across a frequency range from 0.01 to 100 rad/s, maintaining a constant temperature of 25 °C. The storage modulus (G'), loss modulus (G"), and complex viscosity were assessed as frequency-dependent parameters.

2.11. Indomethacin loading and entrapment efficiency

Out of all the fabricated formulations, the hydrogel (W-1) was selected for the drug loading due to its highest swelling capacity. The drug loading and entrapment efficiency of β-CD/PAM/CMTKG hydrogel (with β-CD) and control PAM/CMTKG hydrogel (without β-CD) was examined using a weight method. The specified amount of hydrogel disc (M_{UH}) was immersed into a 100 ml buffer solution with a pH of 7.4 containing a specific amount (M_D) of the indomethacin for 24 h [30]. Afterward, a hydrogel disc was taken out, and oven-dried at 60 °C, and weight (M_{LH}) was noted. The DL% and DEE% were determined using the following equations [30].

$$\text{DL (\%)} = \frac{M_{LH-}\, M_{UH}}{M_{UH}} \times 100 \tag{12}$$

$$\text{DEE (\%)} = \frac{M_{LH-}\, M_{UH}}{M_D} \times 100 \tag{13}$$

2.12. In vitro indomethacin release study

To investigate the release of indomethacin from the hydrogel, a precise amount of indomethacin-loaded hydrogel disc was immersed in 100 ml pH 1.2 buffer solutions for 2 h, followed by pH 6.8 for next 2 h, and lastly in pH 7.4 till equilibrium achieved [31]. The solutions were agitated at 200 rpm and 37 °C in an orbital incubator shaker. To keep the volume constant, 3 ml of released medium was withdrawn and replaced with fresh buffer every hour for up to 24 h. The withdrawn solutions were diluted with ethanol, and the absorbance of the released drug in the medium was quantified using a UV–visible spectrophotometer (Cary 300 UV–Vis model) at the λ_{max} 318 nm for pH 1.2, λ_{max} 319 nm for pH 6.8 and λ_{max} 320 nm for pH 7.4 solution [29]. The drug content was determined by utilizing the standard calibration curve of indomethacin drawn in a mixture of ethanol & pH 7.4 buffer solution (1:9 v/v %), ethanol & pH 6.8 buffer solution (1:9 v/v %) and ethanol & pH 1.2 buffer solution (1:9 v/v %).

$$\text{Cumulative drug release} \left(\frac{M_t}{M_\infty}\right) = \frac{V_t C_t + V_W \sum_1^{n-1} C_{t-1}}{M_\infty} \tag{14}$$

Here, M_t is the amount of drug released at time t, M_∞ is the amount of drug loaded, V_t is the total volume of release medium, V_W is the volume of solution withdrawn, C_t and C_{t-1} are the concentration of solution at time t and t-1 [32].

2.13. Kinetic modelling

Different mathematical models, such as the Hixson-Crowell, Korsmeyer-Peppas, First-Order, Zero-Order, and Higuchi, were employed to analyze the drug release mechanism. The evaluation of regression coefficient (R^2) values was conducted for each model. The model exhibiting an R^2 value closest to 1 was identified as the most suitable model for the description of the drug release mechanism [33,34].

2.14. Cytotoxicity

The impact of indomethacin-loaded β-CD/PAM/CMTKG hydrogel on the HCT 116 human colon cell line, obtained from NCCS Pune, was examined through the MTT Assay. Cells were initially seeded at a density of 10,000 cells per well in a 96-well plate in DMEM medium, supplemented with 10 % FBS and 1 % antibiotic solution, at 37 °C with 5 % CO_2 for 24 h. Following this, the cells were subjected to various concentrations (ranging from 0.78 to 6.25 μg/ml) of the hydrogel, while wells lacking hydrogel solution were designated as control. After 24-h incubation, MTT Solution was introduced and allowed to incubate for 2 h. Subsequently, the supernatant was removed, and the cell layer was dissolved in 100 μl of Dimethyl Sulfoxide. To evaluate cell viability, absorbance readings were taken at 540 nm and 660 nm using an Elisa plate reader (iMark, Biorad, USA). Furthermore, images were captured under an inverted microscope (Olympus ek2) equipped with a 10 MP Aptima CMOS digital camera (AmScope).

2.15. Degradation

Degradation experiments were executed in a pH 7.4 buffer, employing an orbital incubator shaker set at 37 °C. The hydrogel disc was immersed in a buffer until it achieved complete swelling and was then weighed(W_{IH}). Subsequently, the hydrogel discs were taken out every hour, wiped with filter paper, weighed (W_{FH}), and then dipped back into the buffer solution. The degradation (%) was computed using the following equation [35,36].

$$\text{Degradation}\ (\%) = \frac{W_{IH} - W_{FH}}{W_{IH}} \times 100 \quad (15)$$

2.16. Statistical analysis

Each experiment was carried out three times ($n = 3$), and the results are presented as mean ± standard deviation. The analysis was performed using ANOVA (One-way analysis of variance), with significance set at $p < 0.05$ for all experiments.

3. Result and discussion

3.1. Mechanism of β-CD/PAM/CMTKG and indomethacin-loaded β-CD/PAM/CMTKG hydrogel synthesis

Several β-CD/PAM/CMTKG hydrogels (W-1 to W4) were synthesized through the free radical mechanism, employing KPS as an initiator and PEGDA as a crosslinker. Initially, the thermal decomposition of KPS at 60 °C produces a sulfate radical, triggering the polymerization of acrylamide (AM) to form polyacrylamide and simultaneously generating radicals on CMTKG & β-CD. PEGDA acts as a crosslinker, fostering the formation of a crosslinked polymeric network by establishing covalent bonds between CMTKG, PAM, and β-CD (hydrophilic exterior), ultimately leading to the fabrication of a three-dimensional β-CD/PAM/CMTKG hydrogel [18,37].

In indomethacin-loaded β-CD/PAM/CMTKG hydrogels, β-CD (guest) hydrophobic interior forms a host-guest complex with indomethacin (guest) entrapped in its cavity, enhancing drug loading and enabling a slow & controlled release over time. Scheme 1 outlines the mechanism involved in the synthesis of the indomethacin-loaded β-CD/PAM/CMTKG hydrogel [38].

3.2. Characterization

PXRD was conducted to investigate the crystalline or amorphous nature of the material. The PXRD pattern of indomethacin displayed sharp peaks at 2θ values of 12.9°, 18.2°, 19.3°, 28°, and 39.8°, revealing its high crystallinity [19]. The broad peak observed in the β-CD/PAM/CMTKG and indomethacin-loaded β-CD/PAM/CMTKG hydrogel indicates their amorphous structure (Fig. 2). Additionally, in the indomethacin-loaded β-CD/PAM/CMTKG hydrogel, no extra peaks attributable to indomethacin (drug) were detected, implying a homogeneous dispersion of indomethacin within the hydrogel matrix [39]. A relatable article published by Tushar et al. reported that the synthesized metformin HCl-loaded CMTKG/poly(sodium acrylate)/xanthan gum hydrogel did not exhibit a distinct sharp peak for the drug metformin HCl [22].

The ATR-FTIR spectra of the PEGDA, indomethacin, β-CD/PAM/CMTKG hydrogel (W-1), and indomethacin-loaded β-CD/PAM/CMTKG hydrogel, depicted in Fig. 3 with corresponding data presented in Table 2. The C═C stretch at 1646 cm^{-1} in PEGDA shifted to lower wavenumbers (1602 cm^{-1} and 1603 cm^{-1}) in β-CD/PAM/CMTKG hydrogel, and indomethacin-loaded β-CD/PAM/CMTKG hydrogel respectively, indicating crosslinking. Furthermore, the C═O peak shifted to higher wavenumbers (1668 cm^{-1} and 1665 cm^{-1}) in β-CD/PAM/CMTKG hydrogel and indomethacin-loaded β-CD/PAM/CMTKG hydrogel respectively, indicates a reduction in conjugation and hence, confirming polymerization of AM [31].

Additionally, in indomethacin-loaded β-CD/PAM/CMTKG hydrogel, no new peaks emerged other than drug compared to β-CD/PAM/CMTKG hydrogel, suggesting a weak physical interaction (hydrogen bonding) between indomethacin (drug) and the hydrogel [31,37]. Meena et al. also reported that the introduction of ibuprofen drug in xanthan gum/PAM/polyvinyl pyrrolidone hydrogel did not produce any new peaks in the FTIR spectra of the drug-loaded hydrogel as compared to unloaded hydrogel [31].

The DSC curve of indomethacin, β-CD/PAM/CMTKG hydrogel and indomethacin-loaded β-CD/PAM/CMTKG hydrogel is shown in Fig. 4. The DSC curve of Indomethacin shows endothermic peaks at 160 °C corresponding to melting temperature (T_m) of drug, indicating the crystalline nature of the indomethacin [42]. Moreover, the endothermic peak at 75 °C and 81 °C observed in the DSC curve of both β-CD/PAM/CMTKG hydrogel and indomethacin-loaded β-CD/PAM/CMTKG hydrogel correspond to the moisture loss from the polymeric network. It is noticed that the endothermic peak of indomethacin disappears in indomethacin-loaded β-CD/PAM/CMTKG hydrogel, which signifies that the drug is uniformly dispersed into the amorphous hydrogel matrix [2,43,44]. Mali et al. synthesized a chitosan and CMTKG hydrogel for aceclofenac delivery and observed that the drug-loaded hydrogel did not exhibit a characteristic endothermic melting peak of the drug [8].

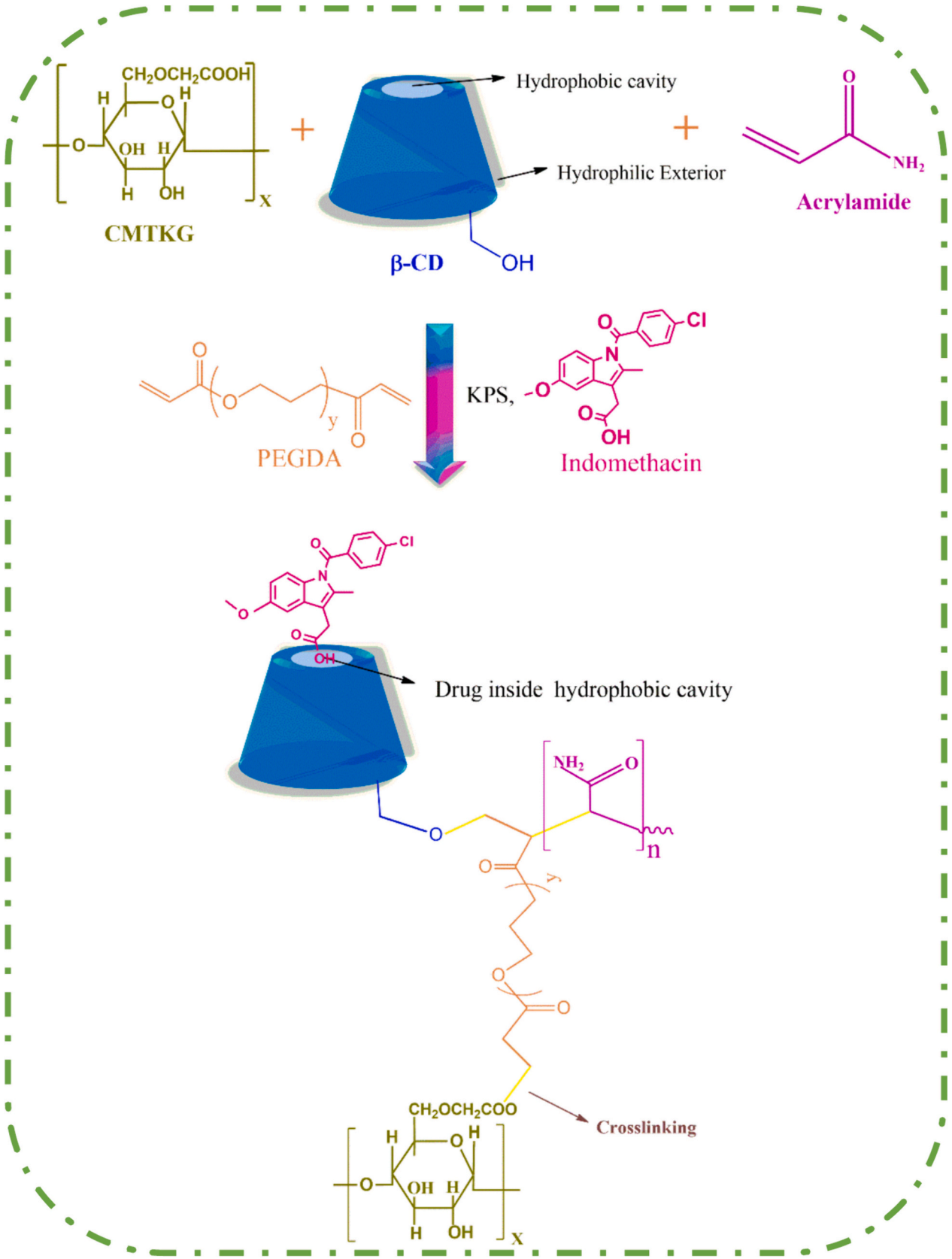
SEM micrographs of β-CD/PAM/CMTKG hydrogel (W-1) and indomethacin-loaded β-CD/PAM/CMTKG hydrogel were presented in Fig. 5. The β-CD/PAM/CMTKG hydrogel exhibited a highly porous surface, emphasizing its capacity for swelling and drug loading. Additionally, the surface morphology analysis of the indomethacin-loaded β-CD/PAM/CMTKG hydrogel displayed a very low porous surface. Thus, it suggests that indomethacin has been successfully encapsulated within the polymer network, resulting in a decrease in the free void in hydrogel [31]. Kiran et al. synthesized a β-CD-grafted acrylamide-co–N-(hydroxymethyl)-acrylamide hydrogel for controlled diclofenac sodium release. The unloaded hydrogel displayed a porous surface, which became smooth and less porous upon the addition of diclofenac sodium drug [39].

3.3. Swelling studies

The swelling behavior of hydrogels is affected by different PEGDA concentrations (W-1 to W-4), as shown in Fig. 6(a). It was observed that with an increase in PEGDA concentration (1.5–3 ml), a decrease in equilibrium swelling occurs. The increased PEGDA content contributes to an increased crosslinking density, thus reducing the available free space for fluid absorption, resulting in diminished swelling [31]. The formulation W-1, containing 1.5 ml of PEGDA, exhibits higher swelling maximum swelling of 1000.27 % ± 0.42 at pH 7.4, with a recorded swelling of 750.21 % ± 0.32 at pH 1.2. A notable difference in swelling was observed in pH 7.4 and 1.2 buffer solutions, as shown in Fig. 6(b). This phenomenon is attributed to the deprotonation of the carboxylic acid (COOH) within CMTKG at pH 7.4. Under this condition, electrostatic repulsion occurs between the deprotonated COO^- ions, leading to network expansion and, consequently, more available space for fluid absorption, leading to increased swelling ($p < 0.05$). While, at pH 1.2, the carboxyl groups may remain in their un-ionized state, so no network expansion, resulting in a lower swelling compared to that observed at pH 7.4 [9]. Liu et al. Rehman et al. synthesized tamarind gum/β-CD *graft*-poly(methacrylic acid) hydrogel for capecitabine delivery, which showed higher swelling in pH 7.4 compared to pH 1.2 [44].

3.4. Equilibrium water content

Equilibrium water content (EWC) refers to the amount of water that a hydrogel can absorb and retain when it reaches a state of equilibrium, and its corresponding values are presented in Table 3. Fig. 7.



Scheme 1. Mechanism for indomethacin-loaded β-CD/PAM/CMTKG hydrogels synthesis.

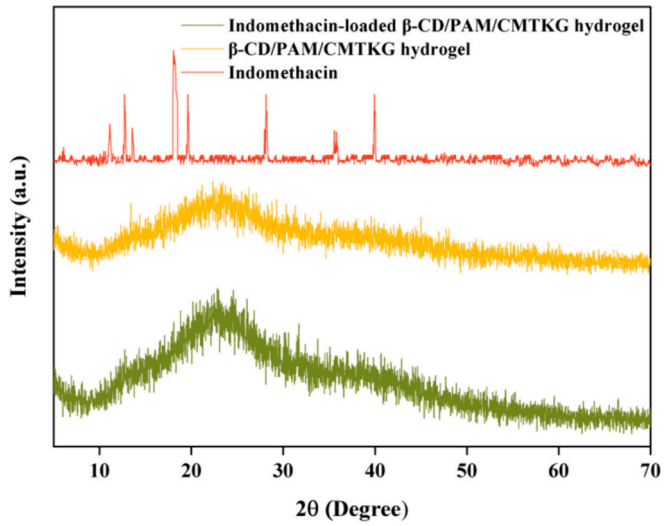


Fig. 2. PXRD of Indomethacin, β-CD/PAM/CMTKG hydrogel (W-1), and indomethacin-loaded β-CD/PAM/CMTKG hydrogel.

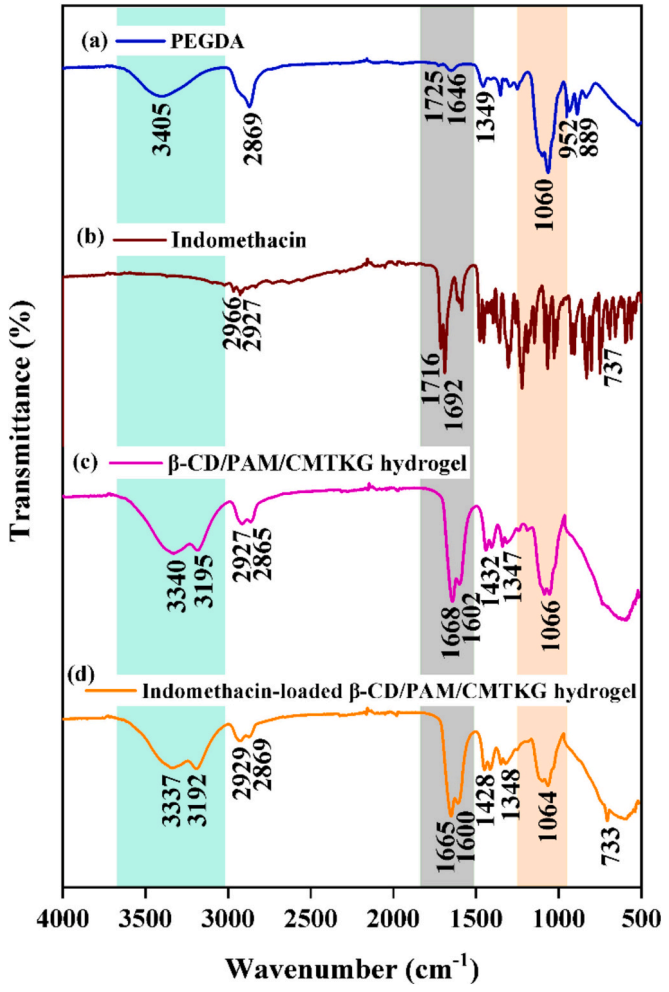


Fig. 3. ATR-FTIR spectra of (a) PEGDA, (b) Indomethacin, (c) β-CD/PAM/CMTKG hydrogel, and (d) indomethacin-loaded β-CD/PAM/CMTKG hydrogel.

demonstrates that increased crosslinker concentration leads to a decrease in the hydrogel's EWC ($p < 0.05$). This is because a higher concentration of crosslinkers results in a more densely crosslinked network, which reduces the available space for water absorption and retention within the hydrogel structure. Consequently, the W-1 hydrogel with a higher EWC can absorb and retain more water, facilitating efficient and controlled drug release [45]. Anirudhan et al. synthesized a hydrogel using β-CD, gelatin, and oxidized dextrin for 5-FU delivery, showing that the equilibrium water content (EWC) decreased as the crosslinker amount increased [46].

3.5. Network parameter

The hydrogels network parameter is influenced by varying PEGDA concentrations (W-1 to W-4), as detailed in Table 4. Higher PEGDA concentrations lead to a decrease in $\overline{M}_c$ and an increase in χ, ρ, and ϕ values. Increased PEGDA content results in greater crosslinking and an increased crosslinking density (ρ). Moreover, an inverse relationship exists between crosslinking density (ρ) and $\overline{M}_c$. Consequently, with an increase in ρ, a decrease in $\overline{M}_c$ was noticed [47]. Shukla et al. observed similar results with their synthesis of a poly(vinyl alcohol)/poly(acrylic acid–co–Acrylamide) hydrogel, showing that an increase in crosslinker amount increases the value of ρ and decreases $\overline{M}_c$ [48]. Additionally, it was already found that an increase in PEGDA content leads to a decrease in swelling, thereby increasing the polymer volume fraction in the swollen state (ϕ). Furthermore, the higher χ values represent the weak interaction between the polymer and solvent. Thus, increasing the crosslinker boosts crosslinking, strengthening the interaction between the polymer-polymer and weakening the polymer-solvent interaction (χ) [25]. Moreover, χ values are lower in pH 7.4 compared to pH 1.2 buffer, suggesting stronger polymer-solvent interaction at pH 7.4, characterized by greater swelling capacity compared to pH 1.2 buffer [27].

3.6. Sol-gel analysis

The sol-gel analysis for the fabricated hydrogels is illustrated in Fig. 8, and the corresponding gel fraction data is outlined in Table 3. As the quantities of PEGDA (W-1 to W-4) increase, a significant increase in cross-linking density is observed ($p < 0.05$). Consequently, more cross-linking within the hydrogel leads to a more compact structure and an increase in the insoluble gel part [49]. Similar outcomes were observed with the tamarind gum/β-CD *graft*-poly (methacrylate) hydrogel developed by Shafiq et al. for the delivery of acyclovir. This hydrogel showed increased gel fraction as the crosslinker concentration was raised [50].

3.7. Porosity

The porosity data, as detailed in Table 3, reveals an increase in the quantity of the crosslinker PEGDA (W-1 to W-4) leads to a corresponding decrease in porosity ($p < 0.05$), as presented in Fig. 9. This phenomenon arises from an increase in the entanglement and crosslinking between the polymeric chains of CMTKG, β-CD, and PAM, creating a denser structure and consequently, reduced porosity [43]. Naeem et al. synthesized a β-CD/chitosan/poly(vinyl alcohol-co-acrylic acid) composite crosslinked with N, N′-Methylene-bis-acrylamide. The study found that as the amount of crosslinker increased, the porosity decreased [51].

3.8. BET analysis

The specific surface area, pore volume, and pore diameter data, as detailed in Table 5, reveals that an increase in the quantity of the crosslinker PEGDA (W-1 to W-4) leads to a corresponding decrease in specific surface area, pore volume, and pore diameter. As the PEGDA quantity increases, the polymer chains in the hydrogel become more tightly interconnected, thereby decreasing the overall specific surface area of the hydrogel. Moreover, these polymer chains are more densely packed and entangled, leading to smaller pore diameters. Furthermore, the denser network with fewer and smaller pores leads to a decrease in

Table 2
ATR-FTIR data of PEGDA, Indomethacin, β-CD/PAM/CMTKG hydrogel, and indomethacin-loaded β-CD/PAM/CMTKG hydrogel.

Functionality	Wavenumber (cm^{-1})				Ref.
	PEGDA	Indomethacin	β-CD/PAM/CMTKG hydrogel	Indomethacin-loaded β -CD/PAM/CMTKG hydrogel	
C=C	1646	1692	1602	1600	[31]
OH stretching	3405	–	3340	3337	[31]
COO^- stretching	1349	–	1347	1348	[9,40]
C-H stretching	2869	2966 & 2927	2927 & 2865	2928 & 2869	[18]
C-O-C	1060	–	1066	1064	[18,40]
C=O	1725	1716	1668	1665	[31]
C-Cl	–	737	–	733	[41]

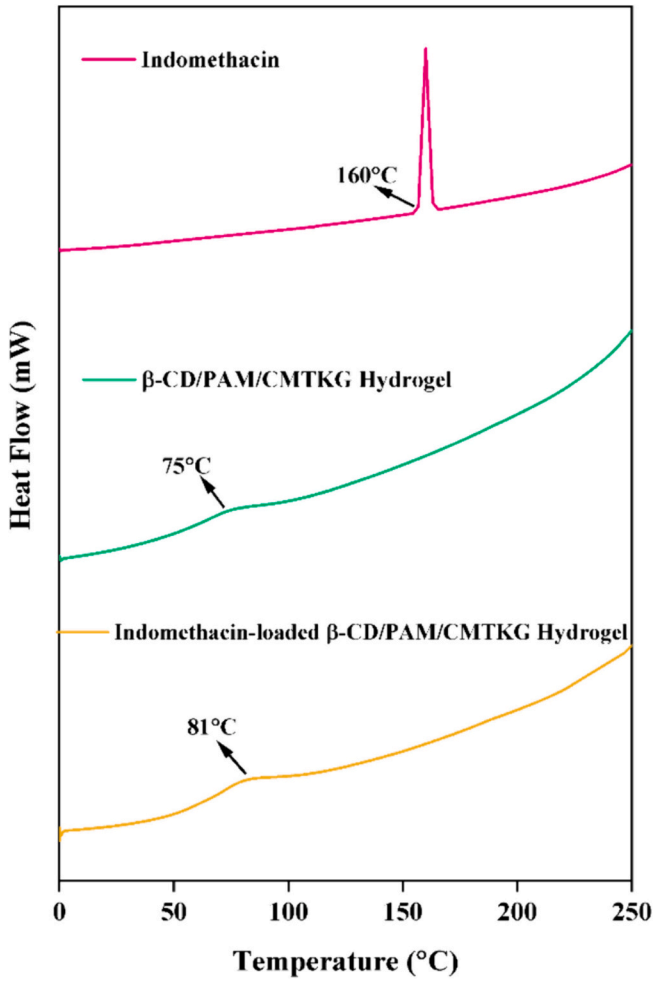


Fig. 4. DSC curve of indomethacin, β-CD/PAM/CMTKG hydrogel, and indomethacin-loaded β-CD/PAM/ CMTKG hydrogel.

the total pore volume of the hydrogel [52]. This data correlated well with the above porosity data, indicating a decrease in porosity and, consequently, lower pore diameter and pore volumes. Machado et al. developed a chitosan/silica composite crosslinked with glutaraldehyde and observed that increasing the amount of the crosslinker resulted in a reduction in surface area [53].

3.9. Rheology

The rheological properties of PAM/CMTKG and β-CD/PAM/CMTKG hydrogel (W-1) were investigated by analyzing changes in storage modulus (G′), loss modulus (G″), and complex viscosity with varying angular frequency. Fig. 10 (a) illustrates that as angular frequency increases from 0.1 to 100 rad/s, the complex viscosity decreases; hence, the material is pseudoplastic [54]. The β-CD/PAM/CMTKG hydrogel (W-1) exhibits a larger complex viscosity value than the control (PAM/CMTKG hydrogel), suggesting its potential as a preferred choice for controlled drug release compared to PAM/CMTKG hydrogels. Moreover, the G' > G" reveals the elastic nature and strong mechanical strength of the synthesized hydrogels (Fig. 10 (b)). Furthermore, the addition of β-CD improves both G' and G" values, with the β-CD/PAM/CMTKG hydrogel showing larger G' values due to increased cross-linking. This heightened crosslinking also enhances resistance to deformation, leading to higher energy dissipation and a higher loss modulus than the PAM/CMTKG hydrogel [36]. Similar findings were reported by Pinelli et al., who synthesized a polyvinyl alcohol/agarose/β-CD hydrogel for ibuprofen delivery. This hydrogel demonstrated an elastic nature, with $G' > G''$ [55].

3.10. Indomethacin loading and entrapment efficiency

The β-CD/PAM/CMTKG and PAM/CMTKG hydrogels were evaluated for drug loading and entrapment efficiency, and the corresponding values are detailed in Table 6. The β-CD/PAM/CMTKG hydrogel shows a significant increase in drug loading and entrapment efficiency ($p < 0.05$) compared to the PAM/CMTKG hydrogel. This is attributed to β-CD, which entrapped the hydrophobic indomethacin drug within its hydrophobic cavity, thus enhancing the loading and entrapment efficiency of β-CD/PAM/CMTKG hydrogel [38]. A comparable result was observed with the gelatin/β-CD hydrogel developed by Rehman et al. for methotrexate delivery. The drug loading was found to improve with the introduction of β-cyclodextrin [38].

3.11. In vitro indomethacin release analysis

In vitro release analysis was conducted for indomethacin-loaded PAM/CMTKG hydrogel and indomethacin-loaded β-CD/PAM/CMTKG hydrogel. The indomethacin-loaded β-CD/PAM/CMTKG hydrogel exhibited lesser drug release in comparison to the indomethacin-loaded PAM/CMTKG hydrogel, as depicted in Fig. 11. This is attributed to the hydrophobic cavity of β-CD, which is responsible for entrapping indomethacin inside the cavity and lowering down its release from the hydrogel [38]. At pH 7.4, the maximum release of indomethacin from PAM/CMTKG and β-CD/PAM/CMTKG hydrogels was noted as 72.24 ± 0.70 % and 60.26 ± 0.65 %, respectively. Moreover, at pH 6.8, the maximum indomethacin release (%) for PAM/CMTKG and β-CD/PAM/CMTKG hydrogels were 26.32 ± 0.73 % and 24.44 ± 0.62 %, respectively. However, at pH 1.2, the maximum indomethacin release (%) for PAM/CMTKG and β-CD/PAM/CMTKG hydrogels were 11.94 ± 0.49 % and 10.71 ± 0.61 %, respectively. Indomethacin release was significantly higher ($p < 0.05$) in both hydrogels at pH 7.4 compared to pH 6.8 and pH 1.2. The increased release at pH 7.4 is attributed to higher swelling (as discussed in Section 2.3), facilitating greater diffusion of indomethacin from the hydrogel [17]. This characteristic makes β-CD/PAM/CMTKG suitable for targeted drug delivery to the colon. A related result was reported by Nayak et al., who developed a hydrogel using

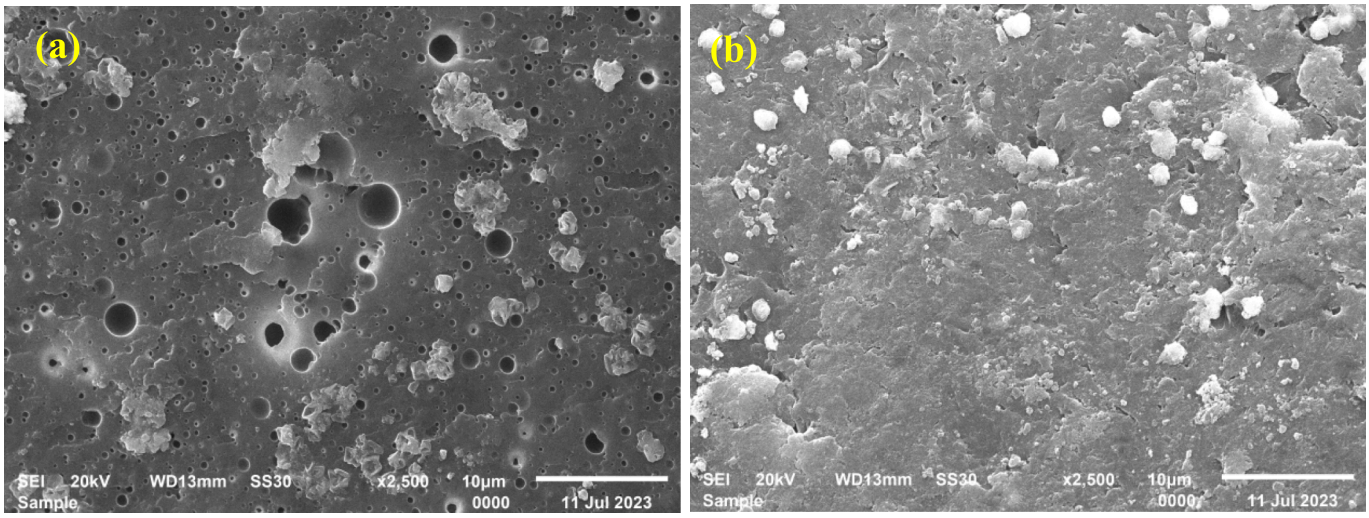


Fig. 5. SEM micrographs of (a) β-CD/PAM/CMTKG, and (b) indomethacin-loaded β-CD/PAM/CMTKG hydrogels.

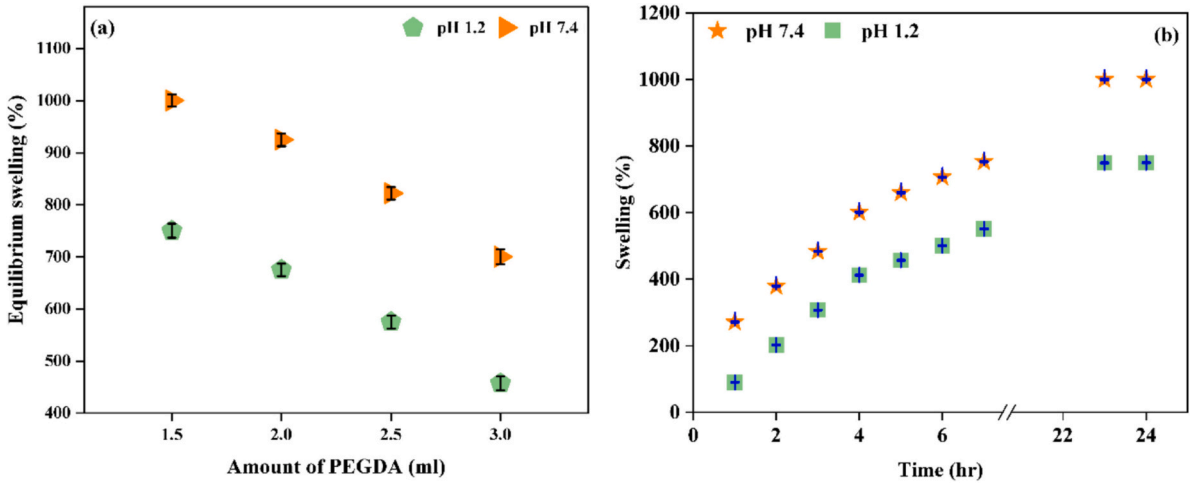


Fig. 6. **(a)** Effect of the PEGDA on the equilibrium swelling (%) and **(b)** Rate of swelling of W-1 hydrogel vs time.

Table 3
Gel fraction, porosity, and equilibrium water content data of synthesized β-CD/PAM/CMTKG hydrogels ($p < 0.05$, $n = 3$).

Formulation	Porosity %	Gel fraction (%)	Equilibrium water content (%)	
			pH 7.4	pH 1.2
W-1	89.21 ± 0.68	76.52 ± 0.67	90.9 ± 0.92	88.23 ± 0.87
W-2	87.50 ± 0.74	77 ± 0.83	90.24 ± 0.73	87.09 ± 0.76
W-3	84.35 ± 0.71	78.34 ± 0.82	89.15 ± 0.84	85.18 ± 0.72
W-4	82.56 ± 0.83	80.25 ± 0.92	87.5 ± 0.75	82.05 ± 0.82

tamarind seed and alginate for the release of diclofenac sodium, where drug release was significantly higher at pH 7.4 (96.07 %) compared to pH 1.2 (7 %) [56]. Moreover, Nayak et al. developed tamarind seed/gellan gum hydrogel, showing 79 % release of metformin HCl at pH 7.4 in 10 h [57]. However, in our study, the introduction of β-CD resulted in a prolonged drug release for 24 h, thus minimizing the repeated dosage condition. This is attributed to the ability of β-CD to encapsulate the drug within its cavity, leading to prolonged release. Similar results were reported by Liu et al., where gelatin/ β-CD hydrogel showed prolonged release of methotrexate as compared to gelatin/dextran hydrogel [38].

Additionally, Shaw et al. synthesized a gelatin/CMTKG hydrogel, which exhibited 45 % swelling and 32 % release of fluoroquinolone at pH 7.4 [58]. However, in our study, the introduction of PAM resulted in significantly higher drug release. This is primarily due to the hydrophilic amide group, which increases the swelling ability (1000.27 %) and, consequently, facilitates a greater release of the drug (60.26 %) in the colon [59]. Hence, the synthesized β-CD/PAM/CMTKG hydrogel shows the potential to facilitate the better-controlled release of indomethacin in the colon.

3.12. Kinetic modelling

The correlation coefficients (R^2) of various models were compared to assess the release mechanism of indomethacin from indomethacin-loaded PAM/CMTKG hydrogel and indomethacin-loaded β-CD/PAM/CMTKG hydrogel (Table 8). For indomethacin-loaded β-CD/PAM/

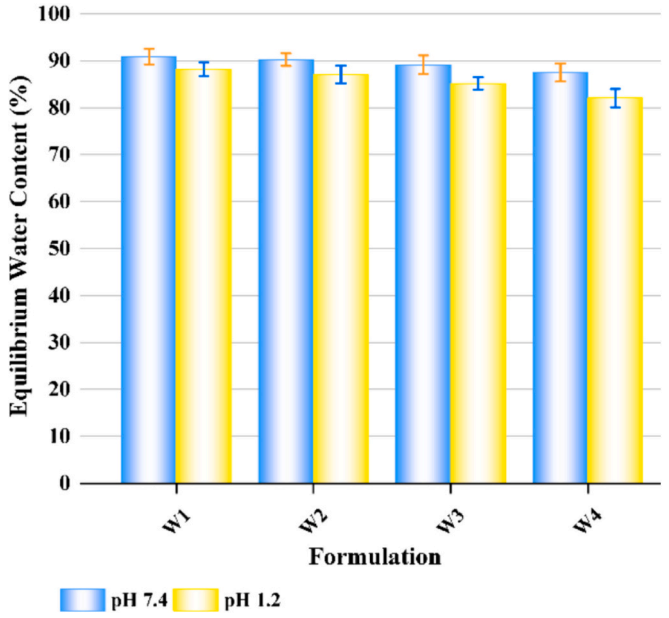


Fig. 7. Equilibrium water content plot for all synthesized β-CD/PAM/CMTKG hydrogels.

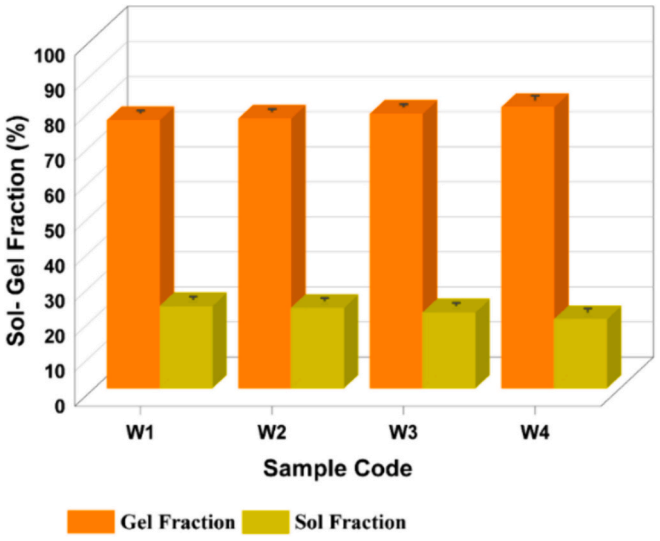


Fig. 8. Sol-gel analysis plot for all fabricated β-CD/PAM/CMTKG hydrogels.

CMTKG hydrogel, the Higuchi model demonstrated the highest R^2 value, indicating the best fit to describe the drug release mechanism (Fig. 12). According to the Higuchi model, drug release predominantly occurs through Fickian diffusion. Fickian diffusion occurs when drug molecules move through the hydrogel due to a concentration gradient. This movement is governed by the diffusion principle, where drug molecules move from regions of higher concentration (inside the hydrogel) to regions of lower concentration (surrounding buffer solution) [49]. Moreover, indomethacin-loaded PAM/CMTKG hydrogel exhibited the Korsmeyer-Peppas model as the most fitted one, with a release exponent (n) value as 0.4902 ($n < 0.5$), also indicating Fickian diffusion drug release behavior [43]. Similarly, Ahmed et al. developed a hydrogel using tamarind gum, β-CD, and poly(methacrylic acid), with drug release data following Fickian diffusion [60].

3.13. Cytotoxicity

Evaluating cytotoxicity is important for ensuring the biocompatibility of fabricated hydrogel for biomedical applications. The analysis was performed for the different concentrations (0.78, 1.56, 3.125, 6.25 μg/ml) of synthesized indomethacin-loaded β-CD/PAM/CMTKG hydrogel and a control (without hydrogel), against the HCT-116 human colon cell line (Fig. 13). The fabricated indomethacin-loaded β-CD/PAM/CMTKG hydrogel shows 79 % cell viability at the highest concentration (6.25 μg/ml), highlighting the non-toxic behavior of synthesized matrix. This result is observed due to the inherent biocompatibility of CMTKG and β-CD present in the hydrogel [18,31,62]. Additionally, microscopic examination of HCT-116 cells revealed polygonal-shaped cells and similar cell density in control and hydrogel-treated cell lines, signifying no change in cell shape and very

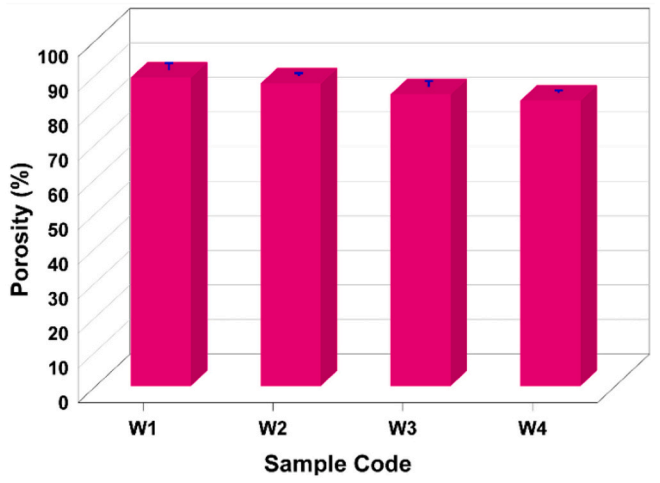


Fig. 9. Porosity plot for all formulated β-CD/PAM/CMTKG hydrogels.

Table 5
Surface area, pore volume, and porosity diameter of synthesized β-CD/PAM/CMTKG hydrogels.

Formulation	Specific surface area (m^2/g)	Pore diameter (nm)	Pore volume (cc/g)
W-1	22.34	25.86	0.147
W-2	19.43	21.64	0.112
W-3	17.86	19.02	0.094
W-4	14.95	16.39	0.076

Table 4
Network parameter data for all synthesized β-CD/PAM/CMTKG hydrogels.

Sample code	χ		ϕ		$\overline{M}_c \times 10^6$ (g/mol)		ρ × 10^{-6} (mol/ml)	
	pH 7.4	pH 1.2	pH 7.4	pH 1.2	pH 7.4	pH 1.2	pH 7.4	pH 1.2
W-1	0.5173	0.5363	0.0520	0.1090	2.204	1.6572	0.2276	0.3018
W-2	0.5178	0.5383	0.0535	0.1151	2.043	1.4914	0.2455	0.3354
W-3	0.5180	0.5433	0.0542	0.1302	1.8160	1.2699	0.2763	0.4077
W-4	0.5183	0.5486	0.0550	0.1462	1.5469	1.0098	0.3243	0.4954

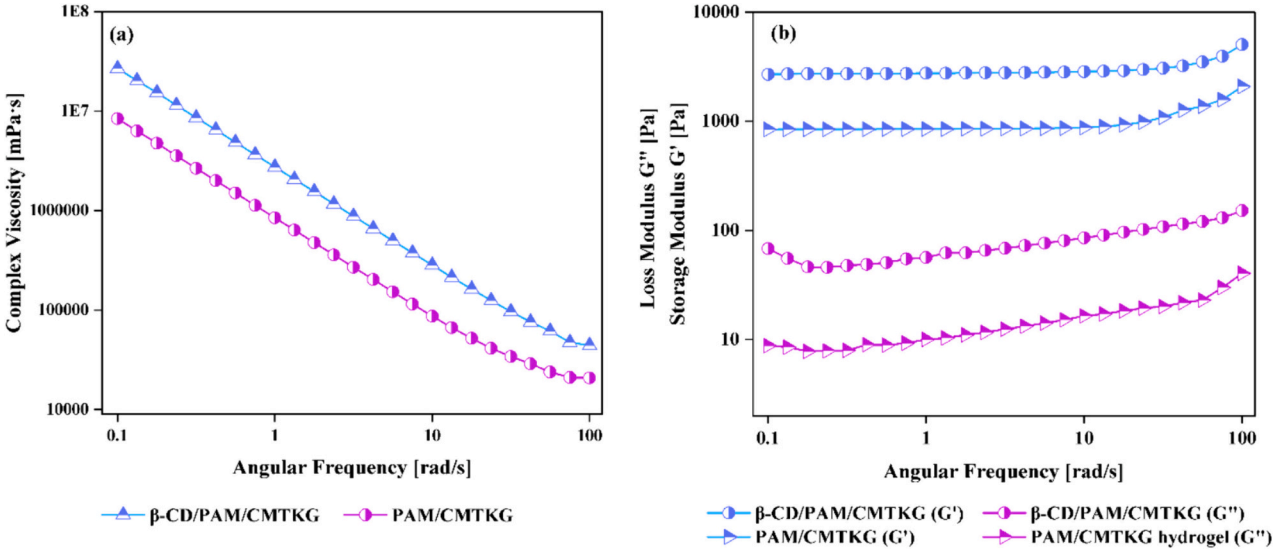


Fig. 10. Plot of angular frequency versus (a) complex viscosity, and (b) G' & G" for PAM/CMTKG and β-CD/PAM/CMTKG hydrogels.

Table 6
Indomethacin loading and entrapment efficiency values for β-CD/PAM/CMTKG and PAM/CMTKG hydrogel ($p < 0.05$, $n = 3$).

Hydrogel	Drug loading	Drug entrapment efficiency
PAM/CMTKG	20.36 ± 0.49	72.36 ± 0.29
β-CD/PAM/CMTKG	25.84 ± 0.52	79.89 ± 0.37

little impact on the cell viability induced by the synthesized hydrogel. Anirudhan et al. synthesized hydrogel based on β-CD, gelatin, and oxidized dextrin for 5-fluorouracil delivery, which demonstrated 78 % cell viability, even at higher concentrations (25 mg/ml) against HT-29 colon cell lines [46].

3.14. Degradation

The biodegradability of materials plays a pivotal role in facilitating their effectiveness and versatility in various biomedical applications. The degradation profile of the synthesized indomethacin-loaded β-CD/PAM/CMTKG hydrogel, depicted in Fig. 14, reveals a degradation attributed to the breakdown of glycosidic (C-O-C) linkages, which led to complete degradation (100 %) within 10 days. Similarly, Jitendra et al., synthesized carboxymethyl chitosan and *Moringa oleifera* for ciprofloxacin HCl delivery, which showed gradual degradation (18 %) in pH 7.4 buffer solution [63]. Thus, the remarkable degradability highlights the potential of this fabricated indomethacin-loaded β-CD/PAM/CMTKG hydrogel for safe and effective drug delivery applications.

3.15. Conclusion

This research successfully synthesized CMTKG, β-CD, and PAM-based hydrogels for controlled indomethacin release. The concentration of PEGDA was observed to impact the porosity (%), gel fraction (%), equilibrium water content (%), and swelling behavior of the hydrogel. The rheological analysis confirmed the elastic nature of the synthesized β-CD/PAM/CMTKG hydrogel. Moreover, the network parameter assessment revealed an increase in crosslinking density with an increase in concentration of PEGDA. The addition of β-CD revealed an increase in drug loading and entrapment efficiency, with the drug release data fitting a Higuchi model. Remarkably, hydrogel exhibited higher release (60.26 %) at pH 7.4, making it ideal for targeted drug delivery to the colon. Notably, the synthesized hydrogel degrades completely (100 %) within 10 days and exhibits no cytotoxicity against HCT-116 human colon cell lines. Consequently, the synthesized indomethacin-loaded β-CD/PAM/CMTKG hydrogel demonstrates non-toxicity and exhibits pH-responsive properties, emphasizing its potential for the controlled release of the hydrophobic drug indomethacin. Furthermore, the limitations of the current study include that the results cannot confirm the

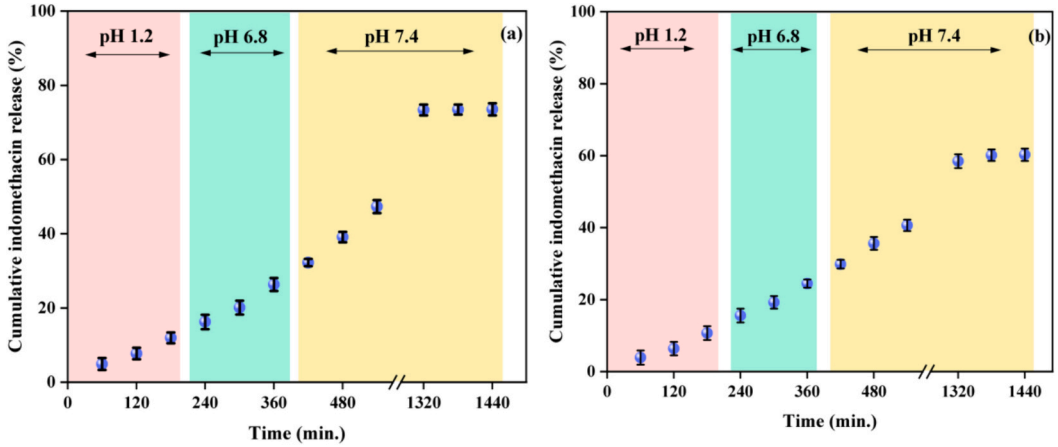


Fig. 11. Indomethacin release profiles for indomethacin-loaded a) PAM/CMTKG hydrogel and b) β-CD/PAM/CMTKG hydrogel at pH 1.2, 6.8 and 7.4.

Table 8
Regression coefficient (R^2) data for indomethacin-loaded PAM/CMTKG hydrogel, and indomethacin- loaded β-CD/PAM/CMTKG hydrogel.

Model	Equation	Indomethacin-loaded PAM/CMTKG hydrogel	Indomethacin-loaded β-CD/PAM/CMTKG hydrogel	Ref.
		R^2	R^2	
Hixson-Crowell	$(M_t)^{1/3}$-$(M_\infty)^{1/3}$ = $k_{HX}.t$ k_{HX} = Hixson Crowell constant	0.9689	0.9537	[61]
First Order	Log M_t = Log M_∞+ $\frac{kt}{2.303}$ k = First order rate constant	0.9632	0.9685	[61]
Korsmeyer–Peppas	$M_t/M_\infty = kt^n$ k = kinetic constant n = diffusion exponent	0.9847	0.9786	[31]
Zero Order	$M_t = M_\infty + k_0 t$	0.9798	0.9769	[43]
Higuchi	$M_t/M_\infty = k_{HG}.t^{½}$ k_{HG} = kinetic constant	0.9801	0.9802	[31]

therapeutic efficacy of synthesized indomethacin-loaded β-CD/PAM/CMTKG hydrogel. Moreover, varying the amounts of initiator and biopolymer was necessary to optimize the hydrogel's properties. These insights point to the opportunities for further development of efficient hydrogel-based drug delivery systems.

CRediT authorship contribution statement

Priyanka Meena: Writing – original draft, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Poonam Singh:** Writing – review & editing, Supervision. **Sudhir G. Warkar:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Conceptualization.

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.

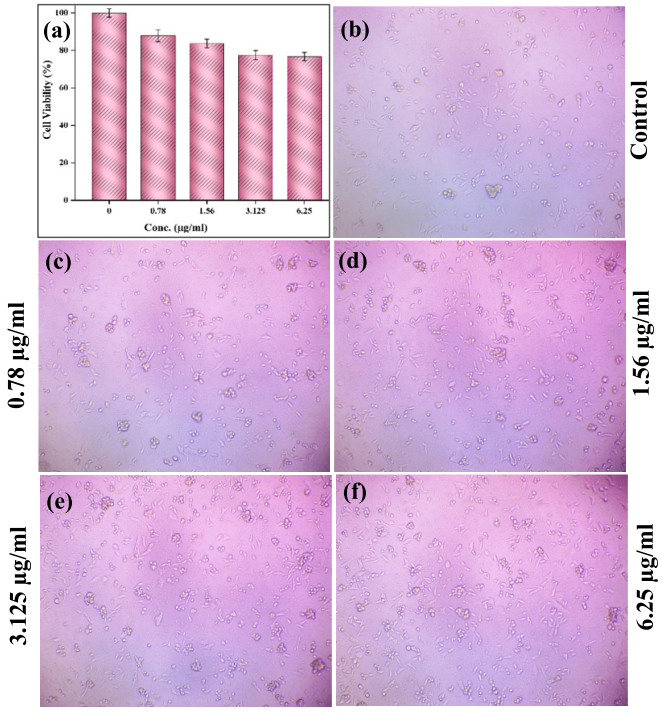


Fig. 13. (a) Cytotoxicity plot for indomethacin-loaded β-CD/PAM/CMTKG hydrogel, and Optical microscopic images of HCT-116 cells treated with **(b)** control **(c)** 0.78 μg/ml, **(d)** 1.56 μg/ml **(e)** 3.125 μg/ml and **(f)** 6.25 μg/ml of indomethacin-loaded β-CD/PAM/CMTKG hydrogel.

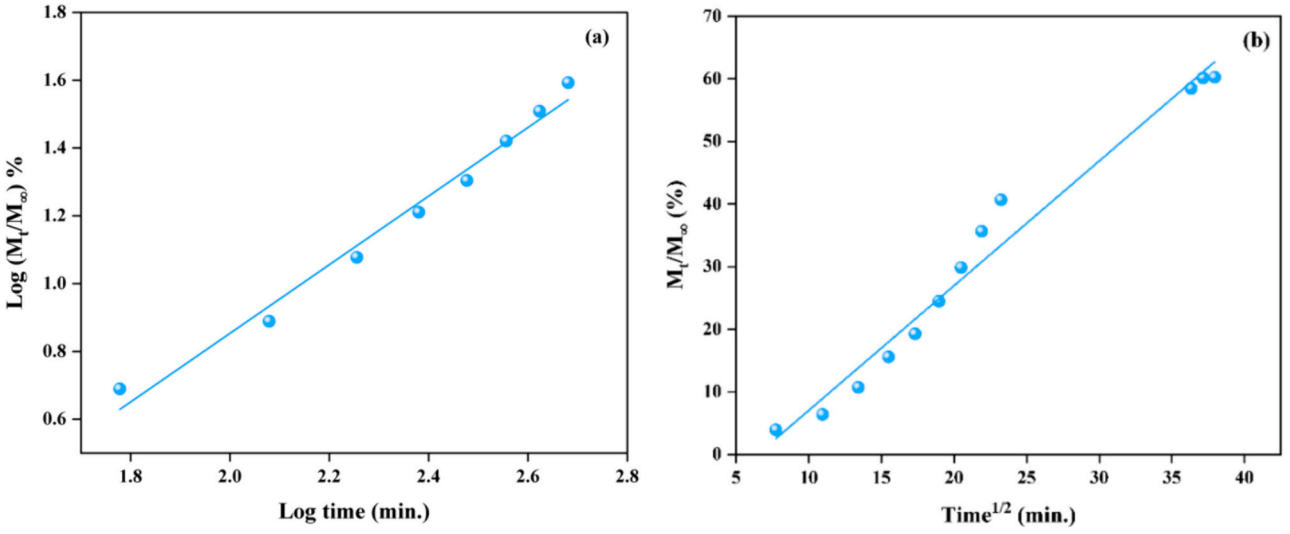


Fig. 12. Release Kinetic plot as per (a) Korsmeyer–Peppas model for indomethacin-loaded PAM/CMTKG hydrogel, and (b) Higuchi model for indomethacin-loaded β-CD/PAM/CMTKG hydrogel.

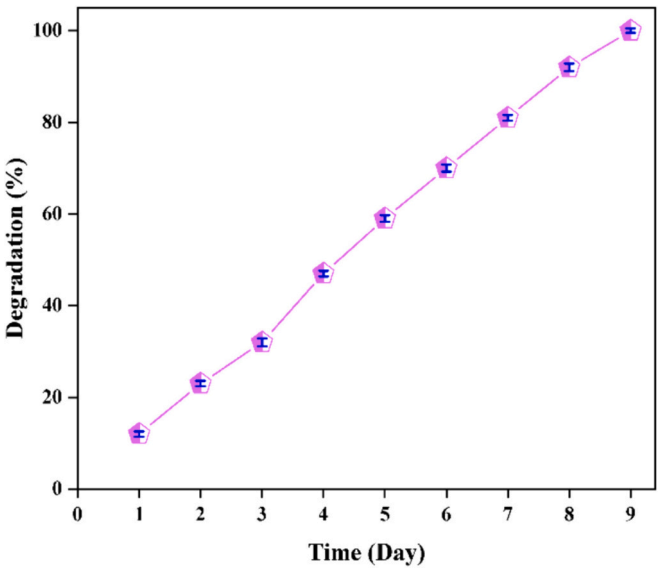


Fig. 14. Degradation curve for indomethacin-loaded β-CD/PAM/CMTKG hydrogel.

Data availability

Data will be made available on request.

Acknowledgements

The authors Priyanka Meena, Poonam Singh, and Sudhir G. Warkar thank the Delhi Technological University, New Delhi, India, for its infrastructural facilities, and Priyanka Meena also thanks the Council of Scientific & Industrial Research (CSIR) for providing senior research fellowship.

Data and code availability

Not applicable.

Funding

No funding was received for conducting this study.

References

[1] W.Q. Kong, C.D. Gao, S.F. Hu, J.L. Ren, L.H. Zhao, R.C. Sun, Xylan-modified-based hydrogels with temperature/pH dual sensitivity and controllable drug delivery behavior, Materials. 10 (2017). doi:https://doi.org/10.3390/ma10030304.
[2] F. Naeem, S. Khan, A. Jalil, NM. Ranjha, A. Riaz, MS. Haider, S. Sarwar, F. Saher, S. Afzal, pH Responsive cross-linked polymeric matrices based on natural polymers: Effect of process variables on swelling characterization and drug delivery properties, BioImpacts. 7 (2017) 177–192. doi:10.15171/bi.2017.21.
[3] K. Thirupathi, T.T.V. Phan, M. Santhamoorthy, V. Ramkumar, S.-C. Kim, pH and Thermoresponsive PNIPAm-co-polyacrylamide hydrogel for dual stimuli-responsive controlled drug delivery, Polymers 15 (2022) 167, https://doi.org/10.3390/polym15010167.
[4] M.A. Raza, S.H. Park, Irradiated Ch/GG/PVP-based stimuli-responsive hydrogels for controlled drug release, Journal of Applied Polymer Science. 137 (2020). doi: https://doi.org/10.1002/app.49041.
[5] M. Akhlaq, A.K. Azad, I. Ullah, A. Nawaz, M. Safdar, T. Bhattacharya, A.B.M. H. Uddin, S.A. Abbas, A. Mathews, Methotrexate-loaded gelatin and polyvinyl alcohol (gel/PVA) hydrogel as a pH-sensitive matrix, Polymers 13 (2021) 2300, https://doi.org/10.3390/polym13142300.
[6] J. Patel, S. Maiti, N.S.H.N. Moorthy, Repaglinide-laden hydrogel particles of xanthan gum derivatives for the management of diabetes, Carbohydr. Polym. 287 (2022) 119354, https://doi.org/10.1016/j.carbpol.2022.119354.
[7] S. Nawaz, S. Khan, U. Farooq, M.S. Haider, N.M. Ranjha, A. Rasul, A. Nawaz, N. Arshad, R. Hameed, Biocompatible hydrogels for the controlled delivery of anti-hypertensive agent: development, characterization and in vitro evaluation, Des. Monomers Polym. 21 (2018) 18–32, https://doi.org/10.1080/15685551.2018.1445416.
[8] K.K. Mali, S.C. Dhawale, R.J. Dias, V.D. Havaldar, P.R. Kavitake, Interpenetrating networks of carboxymethyl tamarind gum and chitosan for sustained delivery of aceclofenac, Marmara Pharmaceutical Journal. 21 (2017) 771–782. doi:10.12991/mpj.2017.20.
[9] S. Jana, R. Sharma, S. Maiti, K.K. Sen, Interpenetrating hydrogels of O-carboxymethyl tamarind gum and alginate for monitoring delivery of acyclovir, Int. J. Biol. Macromol. 92 (2016) 1034–1039, https://doi.org/10.1016/j.ijbiomac.2016.08.017.
[10] D. Pal, A.K. Nayak, Novel tamarind seed polysaccharide-alginate mucoadhesive microspheres for oral gliclazide delivery: In vitro-in vivo evaluation, Drug Delivery. 19 (2012) 123–131. doi:10.3 109/10717544.2012.657717.
[11] S. Jana, A. Saha, A.K. Nayak, K.K. Sen, S.K. Basu, Aceclofenac-loaded chitosan-tamarind seed polysaccharide interpenetrating polymeric network microparticles, Colloids Surf. B: Biointerfaces 105 (2013) 303–309, https://doi.org/10.1016/j.colsurfb.2013.01.013.
[12] M. Shrila, Ananya, P. Meena, S.G. Warkar, Preparation and characterization of sodium alginate-based hydrogel for delivery of hydrophilic drug metformin hydrochloride, ChemistrySelect 9 (2024) e202401773, https://doi.org/10.1002/slct.202401773.
[13] P. Meena, P. Singh, S.G. Warkar, Development and assessment of carboxymethyl tamarind kernel gum-based pH-responsive hydrogel for release of diclofenac sodium, Eur. Polym. J. 197 (2023) 112340, https://doi.org/10.1016/j.eurpolymj.2023.112340.
[14] Z. Qiao, L. Tran, J. Parks, Y. Zhao, N. Hai, Y. Zhong, H. Ji, Highly stretchable gelatin-polyacrylamide hydrogel for potential transdermal drug release, Nano Select. 2 (2021) 107–115. doi:10 .1002/nano.202000087.
[15] M. Pandey, H. Choudhury, S.K. Segar Singh, N.C. Annan, S.K. Bhattamisra, B. Gorain, M.C.I.M. Amin, Budesonide-loaded pectin/polyacrylamide hydrogel for sustained delivery: fabrication, characterization and in vitro release kinetics, Molecules 26 (2021) 1–13, https://doi.org/10.3390/molecules26092704.
[16] R.M. Sarfraz, M. Ahmad, A. Mahmood, M.R. Akram, A. Abrar, Development of β-cyclodextrin-based hydrogel microparticles for solubility enhancement of rosuvastatin: An in vitro and in vivo evaluation, Drug Design, Development and Therapy. 11 (2017) 3083–3096. doi:10.2 147/DDDT.S143712.
[17] N.S. Malik, M. Ahmad, M.U. Minhas, Cross-linked β-cyclodextrin and carboxymethyl cellulose hydrogels for controlled drug delivery of acyclovir, PLoS ONE. 12 (2017) 1–17. doi:10.1 371/journal.pone.0172727.
[18] A.F. Zakaria, S. Kamaruzaman, N. Abdul Rahman, N. Yahaya, Sodium Alginate/β-Cyclodextrin Reinforced Carbon Nanotubes Hydrogel as Alternative Adsorbent for Nickel(II) Metal Ion Removal, Polymers. 14 (2022). doi:https://doi.org/10.3390/polym14245524.
[19] M. Yuan, J. Niu, Q. Xiao, H. Ya, Y. Zhang, Y. Fan, L. Li, X. Li, Hyaluronan-modified transfersomes based hydrogel for enhanced transdermal delivery of indomethacin, Drug Deliv. 29 (2022) 1232–1242, https://doi.org/10.1080/10717544.2022.2053761.
[20] B.J. Elmunzer, I. Hernandez, W.F. Gellad, The skyrocketing cost of rectal indomethacin, JAMA Intern. Med. 180 (2020) 631–632, https://doi.org/10.1001/jamainternmed.2020.0099.
[21] K. Dewangan, K. Nagori, M. Sharma, A. Singh, S. Chandrakar, V.D. Sahu, G. Sharma, S. Gupta, H. Solanki, M. Majumdar, D.K. Tripathi, A. Alexander, Ajazuddin, formulation, evaluation and release studies of indomethacin suppositories, Res. J. Pharm. Technol. 9 (2016) 942–944, https://doi.org/10.5958/0974-360X.2016.00181.5.
[22] Tushar, Y. Saraswat, P. Meena, S.G. Warkar, Synthesis and characterization of novel xanthan gum - based pH - sensitive hydrogel for metformin hydrochloride release, Colloid Polym. Sci. 301 (2023) 1147–1158, https://doi.org/10.1007/s00396-023-05135-9.
[23] I. Gholamali, M. Asnaashariisfahani, E. Alipour, A.A. Sepahi, In-situ synthesized carboxymethyl chitosan/poly(vinyl alcohol) bio-nanocomposite hydrogels containing nanoparticles with drug-delivery properties, Bulletin of Materials Science. 43 (2020). doi:https://doi.org/10.1007/s12034-020-02231-2.
[24] W. Zhao, Y. Zhu, J. Zhang, T. Xu, Q. Li, H. Guo, J. Zhang, C. Lin, L. Zhang, A comprehensive study and comparison of four types of zwitterionic hydrogels, J. Mater. Sci. 53 (2018) 13813–13825, https://doi.org/10.1007/s10853-018-2535-6.
[25] E. Larrañeta, M. Imízcoz, J.X. Toh, N.J. Irwin, A. Ripolin, A. Perminova, J. Domínguez-Robles, A. Rodríguez, R.F. Donnelly, Synthesis and characterization of lignin hydrogels for potential applications as drug eluting antimicrobial coatings for medical materials, ACS Sustain. Chem. Eng. 6 (2018) 9037–9046, https://doi.org/10.1021/acssuschemeng.8b01371.
[26] S.M.H. Bukhari, S. Khan, M. Rehanullah, N.M. Ranjha, Synthesis and characterization of chemically cross-linked acrylic acid/gelatin hydrogels: effect of pH and composition on swelling and drug release, Int. J. Polym. Sci. 2015 (2015), https://doi.org/10.1155/2015/187961.
[27] B. Singh, S. Sharma, A. Dhiman, Acacia gum polysaccharide based hydrogel wound dressings: synthesis, characterization, drug delivery and biomedical properties, Carbohydr. Polym. 165 (2017) 294–303, https://doi.org/10.1016/j.carbpol.2017.02.039.
[28] K. Kalantari, E. Mostafavi, B. Saleh, P. Soltantabar, T.J. Webster, Chitosan/PVA hydrogels incorporated with green synthesized cerium oxide nanoparticles for wound healing applications, Eur. Polym. J. 134 (2020) 109853, https://doi.org/10.1016/j.eurpolymj.2020.109853.
[29] M. Suhail, I.H. Chiu, M.C. Hung, Q.L. Vu, I.L. Lin, P.C. Wu, In vitro evaluation of smart and pH-sensitive chondroitin sulfate/sodium polystyrene sulfonate hydrogels for controlled drug delivery, Gels 8 (2022) 1–18, https://doi.org/10.3390/gels8070406.

[30] B. Niu, J. Jia, H. Wang, S. Chen, W. Cao, J. Yan, X. Gong, X. Lian, W. Li, Y.Y. Fan, In vitro and in vivo release of diclofenac sodium-loaded sodium alginate/carboxymethyl chitosan-ZnO hydrogel beads, Int. J. Biol. Macromol. 141 (2019) 1191–1198, https://doi.org/10.1016/j.ijbiomac.2019.09.059.
[31] P. Meena, P. Singh, S.G. Warkar, Fabrication and evaluation of stimuli-sensitive xanthan gum-based hydrogel as a potential carrier for a hydrophobic drug ibuprofen, Colloid Polym. Sci. 302 (2024) 377–391, https://doi.org/10.1007/s00396-023-05198-8.
[32] Y. Hu, X. Dong, L. Ke, S. Zhang, D. Zhao, H. Chen, X. Xiao, Polysaccharides/mesoporous silica nanoparticles hybrid composite hydrogel beads for sustained drug delivery, J. Mater. Sci. 52 (2017) 3095–3109, https://doi.org/10.1007/s10853-016-0597-x.
[33] M. Suhail, P.C. Wu, M.U. Minhas, Development and characterization of pH-sensitive chondroitin sulfate-co-poly(acrylic acid) hydrogels for controlled release of diclofenac sodium, J. Saudi Chem. Soc. 25 (2021) 101212, https://doi.org/10.1016/j.jscs.2021.101212.
[34] N.Thombare Khushbu, S.G. Warkar, Controlled release and release kinetics studies of boron through the functional formulation of carboxymethyl tamarind kernel gum-based superabsorbent hydrogel, Polym. Bull. 79 (2022) 2287–2303, https://doi.org/10.1007/s00289-021-03634-9.
[35] M.A. Raza, S.H. Park, Irradiated Ch/GG/PVP-based stimuli-responsive hydrogels for controlled drug release, J. Appl. Polym. Sci. 137 (2020) 49041, https://doi.org/10.1002/app.49041.
[36] Z. Jing, A. Xu, Y.Q. Liang, Z. Zhang, C. Yu, P. Hong, Y. Li, Biodegradable poly (acrylic acid-co-acrylamide)/ poly(vinyl alcohol) double network hydrogels with tunable mechanics and high self-healing performance, Polymers. 11 (2019). doi: https://doi.org/10.3390/polym11060952.
[37] I. Rani, S.G. Warkar, A. Kumar, Nano ZnO embedded poly (ethylene glycol) diacrylate cross-linked carboxymethyl tamarind kernel gum (CMTKG)/poly (sodium acrylate) composite hydrogels for oral delivery of ciprofloxacin drug and their antibacterial properties, Mater. Today Commun. 35 (2023) 105635, https://doi.org/10.1016/j.mtcomm.2023.105635.
[38] C. Liu, Z. Zhang, X. Liu, X. Ni, J. Li, Gelatin-based hydrogels with β-cyclodextrin as a dual functional component for enhanced drug loading and controlled release, RSC Adv. 3 (2013) 25041–25049, https://doi.org/10.1039/c3ra42532k.
[39] R.Tiwari Kiran, V.K. Singh, M.K. Singh, S. Krishnamoorthi, K. Kumar, Synthesis, characterization of β-CD based novel hydrogels with dual objectives of drug release and dye removal, Iran, Polym. J. 29 (2020) 615–623, https://doi.org/10.1007/s13726-020-00826-4.
[40] A.K. Nayak, D. Pal, K. Santra, Swelling and drug release behavior of metformin HCl-loaded tamarind seed polysaccharide-alginate beads, Int. J. Biol. Macromol. 82 (2016) 1023–1027, https://doi.org/10.1016/j.ijbiomac.2015.10.027.
[41] B. Soltani, H. Nabipour, N. Ahmadi Nasab, Fabrication, controlled release, and kinetic studies of indomethacin—layered zinc hydroxide nanohybrid and its effect on the viability of HFFF2, Journal of Dispersion Science and Technology. 39 (2018) 1200–1207. doi:10.1080 /01932691.2017.1388178.
[42] G. Germini, L. Peltonen, 3D printing of drug nanocrystals for film formulations, Molecules 26 (2021) 1–12, https://doi.org/10.3390/molecules26133941.
[43] S.A. Ghumman, S. Noreen, H. Hameed, M.A. Elsherif, R. Shabbir, M. Rana, K. Junaid, S.N.A. Bukhari, Synthesis of pH-Sensitive Cross-Linked Basil Seed Gum/Acrylic Acid Hydrogels by Free Radical Copolymerization Technique for Sustained Delivery of Captopril, Gels. 8 (2022). doi:https://doi.org/10.3390/gels8050291.
[44] U. Rehman, R.M. Sarfraz, A. Mahmood, T. Mahmood, N. Batool, B. Haroon, Y. Benguerba, Tamarind/β-CD-g-poly (MAA) pH responsive hydrogels for controlled delivery of Capecitabine: fabrication, characterization, toxicological and pharmacokinetic evaluation, J. Polym. Res. 30 (2023) 1–20, https://doi.org/10.1007/s10965-022-03422-7.
[45] B.B. Mandal, S. Kapoor, S.C. Kundu, Silk fibroin/polyacrylamide semi-interpenetrating network hydrogels for controlled drug release, Biomaterials 30 (2009) 2826–2836, https://doi.org/10.1016/j.biomaterials.2009.01.040.
[46] T.S. Anirudhan, A.M. Mohan, Novel pH switchable gelatin based hydrogel for the controlled delivery of the anti cancer drug 5-fluorouracil, RSC Adv. 4 (2014) 12109–12118. doi:10 .1039/c3ra47991a.
[47] K. Sohail, I.U. Khan, Y. Shahzad, T. Hussain, N.M. Ranjha, Ph-sensitive polyvinylpyrrolidone-acrylic acid hydrogels: impact of material parameters on swelling and drug release, Braz. J. Pharm. Sci. 50 (2014) 173–184, https://doi.org/10.1590/S1984-82502011000100018.
[48] S. Shukla, A.K. Bajpai, Preparation and characterization of highly swelling smart grafted polymer networks of poly(vinyl alcohol) and poly(acrylic acid-co-acrylamide), J. Appl. Polym. Sci. 102 (2006) 84–95, https://doi.org/10.1002/app.23088.
[49] N. Rafique, M. Ahmad, M.U. Minhas, S.F. Badshah, N.S. Malik, K.U. Khan, Designing gelatin-based swellable hydrogels system for controlled delivery of salbutamol sulphate: characterization and toxicity evaluation, Polym. Bull. 79 (2022) 4535–4561, https://doi.org/10.1007/s00289-021-03629-6.
[50] K. Shafiq, A. Mahmood, M.M. Salem-Bekhit, R.M. Sarfraz, A.S. Algarni, E.I. Taha, A.A. Mansour, S. Al Zahrani, Y. Benguerba, Development and Optimization of Tamarind Gum-β-Cyclodextrin-g-Poly(Methacrylate) pH-Responsive Hydrogels for Sustained Delivery of Acyclovir, Pharmaceuticals. 15 (2022). doi:https://doi.org/10.3390/ph15121527.
[51] A. Naeem, Y. Chengqun, Hetonghui, Z. Zhenzhong, Z. Weifeng, G. Yongmei, β-Cyclodextrin/chitosan-based (polyvinyl alcohol-co-acrylic acid) interpenetrating hydrogels for oral drug delivery, International Journal of Biological Macromolecules. 242 (2023) 125149. doi:10. 1016/j.ijbiomac.2023.124506.
[52] H. Kaco, S. Zakaria, C.H. Chia, M.S. Sajab, A.S.M. Saidi, Characterization of Aldehyde Crosslinked Kenaf Regenerated Cellulose Film, BioResources. 10 (2015) 6705–6719. doi:10.15376/biores.10.4.6705-6719.
[53] T.S. Machado, L. Crestani, G. Marchezi, F. Melara, J.R. de Mello, G.L. Dotto, J.S. Piccin, Synthesis of glutaraldehyde-modified silica/chitosan composites for the removal of water-soluble diclofenac sodium, Carbohydrate Polymers. 277 (2022). doi:https://doi.org/10.1016/j.carbpol.2021.118868.
[54] F. Cuomo, M. Cofelice, F. Lopez, Rheological characterization of hydrogels from alginate-based nanodispersion, Polymers. 11 (2019). doi:https://doi.org/10.3390/polym11020259.
[55] F. Pinelli, M. Ponti, S. Delleani, F. Pizzetti, V. Vanoli, F.B. Vangosa, F. Castiglione, H. Haugen, L.P. Nogueira, A. Rossetti, F. Rossi, A. Sacchetti, β-Cyclodextrin functionalized agarose-based hydrogels for multiple controlled drug delivery of ibuprofen, International Journal of Biological Macromolecules. 252 (2023) 126284. doi:https://doi.org/10.1016/j.ijbiomac.2023.126284.
[56] A.K. Nayak, D. Pal, Development of pH-sensitive tamarind seed polysaccharide-alginate composite beads for controlled diclofenac sodium delivery using response surface methodology, International Journal of Biological Macromolecules. 49 (2011) 784–793. doi:10.10 16/j.ijbiomac.2011.07.013.
[57] A.K. Nayak, D. Pal, K. Santra, Tamarind seed polysaccharide-gellan mucoadhesive beads for controlled release of metformin HCl, Carbohydrate Polymers. 103 (2014) 154–163. doi:10.101 6/j.carbpol.2013.12.031.
[58] G.S. Shaw, K. Uvanesh, S.N. Gautham, V. Singh, K. Pramanik, I. Banerjee, N. Kumar, K. Pal, Development and characterization of gelatin-tamarind gum/carboxymethyl tamarind gum based phase-separated hydrogels: a comparative study, Des. Monomers Polym. 18 (2015) 434–450, https://doi.org/10.1080/15685551.2015.1041075.
[59] B. Singh, M. Mohan, B. Singh, Synthesis and characterization of the azadirachta indica gum–polyacrylamide interpenetrating network for biomedical applications, Carbohydr. Polym. Technol. Appl. 1 (2020) 100017, https://doi.org/10.1016/j.carpta.2020.100017.
[60] I. Ahmed, A. Mahmood, O.S. Qureshi, R.M. Sarfraz, H. Ijaz, M. Zaman, M.R. Akram, S. Usmani, Zulcaif, M. Malook, Development of tamarind gum/β-CD-co-poly (MAA) hydrogels for pH-driven controlled delivery of capecitabine, Springer Berlin Heidelberg, 2024. doi:https://doi.org/10.1007/s00289-023-04999-9.
[61] D. Das, R. Das, J. Mandal, A. Ghosh, S. Pal, Dextrin crosslinked with poly(lactic acid): a novel hydrogel for controlled drug release application, J. Appl. Polym. Sci. 131 (2014) 1–12, https://doi.org/10.1002/app.40039.
[62] Z. Fan, P. Cheng, M. Liu, S. Prakash, J. Han, Z. Ding, Y. Zhao, Z. Wang, Dynamic crosslinked and injectable biohydrogels as extracellular matrix mimics for the delivery of antibiotics and 3D cell culture, RSC Adv. 10 (2020) 19587–19599, https://doi.org/10.1039/d0ra02218g.
[63] J. Kumar, R. Purwar, Self-healing biocompatible injectable hydrogel based on multialdehyde Moringa oleifera, ChemistrySelect 9 (2024) e202400309, https://doi.org/10.1002/slct.202400309.

Tex-ViT: A Generalizable, Robust, Texture-based dual-branch cross-attention deepfake detector

Deepak Dagar[1], Dinesh Kumar Vishwakarma[2,*]
Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Bawana Road, Delhi-110042, India
deepakdagargate@gmail.com1, dvishwakarma@gmail.com[2,*]

Abstract
Deepfakes, which employ Generative Adversarial Networks (GANs) to produce highly realistic facial modification, are widely regarded as the prevailing method. Traditional Convolutional Neural Networks (CNNs) have been able to identify bogus media, but they struggle to perform well on different datasets and are vulnerable to adversarial attacks due to their lack of robustness. Vision transformers have demonstrated potential in the realm of image classification problems, but they require enough training data. Motivated by these limitations, this publication introduces Tex-ViT (Texture-Vision Transformer), which enhances CNN features by combining ResNet (Residual Networks) with a vision transformer. The model combines traditional ResNet features with a texture module that operates in parallel on sections of ResNet before each down-sampling operation. The texture module then serves as an input to the dual branch of the cross-attention vision transformer. It specifically focuses on improving the global texture module, which extracts feature map correlation. Empirical analysis reveals that fake images exhibit smooth textures that do not remain consistent over long distances in manipulations. Experiments were performed on different categories of FaceForensics++ (FF++), such as Deepfakes (DF), Face2Face (f2f), Faceswap (FS), and Neural Texture (NT), together with other types of GAN datasets in cross-domain scenarios. Furthermore, experiments also conducted on FF++, DFDCPreview, and Celeb-DF dataset underwent several post-processing situations, such as blurring, compression, and noise. The model surpassed the most advanced models in terms of generalization, achieving a 98% accuracy in cross-domain scenarios. This demonstrates its ability to learn the shared distinguishing textural characteristics in the manipulated samples. These experiments provide evidence that the proposed model is capable of being applied to various situations and is resistant to many post-processing procedures.

Keywords: Deepfake detector, Texture, Gram matrices, Generalization, Robustness.

1 Introduction

With the advancements in technology, especially GANs, it is possible to generate highly realistic content that can easily deceive the naked eye. Deepfake is a current state of the art of visual and audio manipulation. Deepfake is a technology where highly astonishing, realistic, and believable content is created using deep learning technology (Figure 1). Visual deepfakes can be classified into five categories: lip sync, attribute manipulation, full-image synthesis, body re-enactment, and face ap [1]. The application of the deepfake has benefitted the education and entertainment industry in various ways. Still, its malicious aspects have threatened the individual's reputation, creating chaos in society and threatening the nation's peace [1].

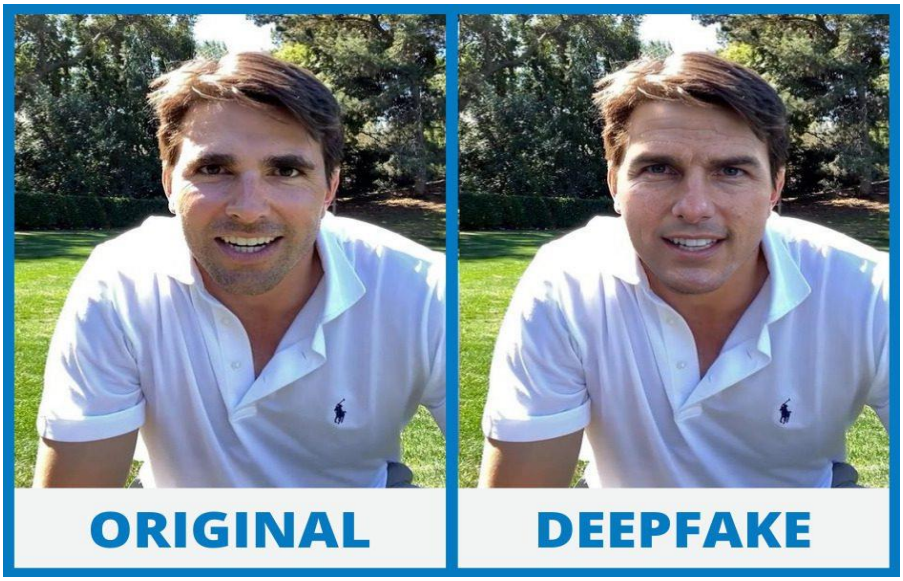


Figure 1: Example of Deepfake [2]

Various approaches have been devised to counter the anti-social elements' malicious intent. These approaches use various clues like GAN-based artifacts [3] [4], visual artifacts [5] [6], biological signals [7] [8], temporal inconsistency between frames [9] [10], audio-visual inconsistency [11] [12], while other let the Deep Neural Networks models [13] [14] learn the intrinsic features of the distribution. The majority of these models exhibit strong performance when used inside the same domain as the training and testing data, but they struggle when applied to different domains. Also, these methods result in a decline in performance in real-world scenarios where visual data can undergo compression, experience the addition of noise, and encounter unanticipated data blur [15]. Nevertheless, certain approaches have partially addressed these issues [16] [17] [18]. More work needs to be done to bridge this gap. Furthermore, CNN algorithms with narrower receptive fields [25] may fail to capture the long-term correlation of features. Long-range feature modeling necessitates the incorporation of an extra module within the CNN model to account for global texture at various semantic levels.

In the cross-domain context, a model needs to identify features that are consistent across the false data sets. Texture is an example of such a feature. To address the above limitations, a texture-based multi-scale cross-attention transformer has been developed. The model posits that counterfeit images exhibit smooth surfaces, and the overall texture remains consistent across the synthetic data samples. The model combines traditional ResNet features with a texture module that operates in parallel on sections of ResNet before each down-sampling operation. The texture module then serves as an input to the dual branch of the cross-attention vision transformer. The motivation for integrating CNN with vision transformer stems from the limitations of CNNs in capturing global properties that exhibit variations over greater distances, due to their inherent inductive biases such as equivariance, translation, and locality of reference. Nevertheless, vision transformers, equipped with inherent attention mechanisms, are capable of precisely capturing such global aspects. The paper's contribution can be summarized as follows:

- An empirical analysis from human and CNN's perspective states that texture statistics play a vital role in the discrimination of fake media. Also, the analysis found that longer-range texture correlation is not preserved in the simulated images.

- Experimentation was done on the FF++ and GAN images dataset categories to show the generalizable abilities of the cross-dataset distribution. The model easily outperforms the other state-of-the-art models and confirms that texture statistics play a vital role in detecting fake manipulated samples.
- Experimentation was done on the Celeb-DF, FF++, and DFDCPreview with different post-processing operations like blurring; the addition of noise and compression shows that the model can beat the scores of others easily, showing robustness in various scenarios.
- Ablation studies were performed on each component of the model to demonstrate the component's effectiveness and the efficacy of their combination and prove that texture with a cross-attention mechanism is a powerful feature.
- Complexity computational analysis is performed against various state-of-the-art computer vision and deepfake detection models to exhibit that the proposed model is light-weight with 43 million parameters and performs better than the former models in diverse scenarios.

The remaining sections of this manuscript are organized as follows: section 2 describes the related work done in this field, including deepfake generation and detection methods. Section 3 discusses the empirical analysis to demonstrate the effectiveness of texture as a feature for classification. Section 4 presents the proposed model and a description of its various components. Section 5 explains the experimentation setup, data pre-processing and data augmentation techniques, experimentation on cross-domain settings on the FF++ and various GAN images, and the in-domain setting for various post-processing operations. Section 7 presents the ablation studies of different components to discuss their importance in feature extraction. Section 8 presents the visualization results for each dataset image.

2 Related Work

The significant research contributions for deepfake detection are discussed in this section.

2.1 Deepfake Image Synthesis

The advancement of GAN technology has taken significant strides in generating realistic images since its inception by Goodfellow [19]. Deepfake Image synthesis can be classified as identity Swap, Attribute Manipulation, and Entire face synthesis [1]. **An identity swap transfers** identity from the source to the target. Face-swap [20] and Fake-App [21] use auto-encoder and decoder for identity swap. Their standard methodology for FS relies on a dual auto-encoder and decoder framework [1]. The initial and widely recognized instance of an identity swap is FaceSwap [20], which was utilized in the creation of the FaceForensics++ dataset [22]. The recent face-swap techniques use GANs owing to their ability to generate highly realistic faces[1]. In recent years the face swap technology has become more sophisticated, owing to the sophistication of the deep learning approaches [2], [3], [4], [5], [6]. StyleSwap[7] The style-based GAN employs an architecture that effectively incorporates the needed information from both branches. Additionally, a layer masking branch is included to enhance the blending information further. FSGANv2[8] employs the technique of reenacting delaunay triangulation and barycentric coordinates, followed by building a face-blending network to achieve a smoothing effect. Model[9] also uses a 3D aware face swapping technique for swapping faces in 3D, which produces high-quality and visually consistent swapped faces from single-view source and target images. It exploits the robust geometric and textural characteristics of 3D human faces, whereby the 2D faces are mapped onto the underlying space

of a 3D generative model by separating the identification and attribute characteristics in the latent space. Other models also incorporate the 3D-aware techniques with a mask diffusion process; Zhao et al.[10] uses a mid-point estimation method that consists of a two-point segmentation method, which allows for the incorporation of identity constraints to produce high-quality faces. Faceshifter [23] is another method that generates highly realistic images using an encoder and generator in a self-supervised manner. **Attribute Manipulation** is another image manipulation category where various facial attributes like expressions, eyes, age, and mustache are manipulated in an image. Such methods generally employ ED or a combination of ED with GANs with a conditioned attribute[11], [12], [13], [14], [15]. StarGAN [24] and STGAN [25] are typical examples of it over the years. Yang et al. [14] provide a solution based on GAN utilizing dilated convolutions to adjust the receptive fields of shallow layers in StyleGAN. This approach enables the expansion of fixed-size tiny features at shallow layers into larger ones that can adapt to varying resolutions, hence enhancing their ability to describe unaligned faces accurately. Facial attributes generally use the disentanglement of attributes in the latent space, which allows manipulation to be carried out in the desired way. Liu et al. [13] alter the underlying code in the 'style space' and provide 'Style Intervention,' a streamlined optimization-based technique, to enhance the visual accuracy of manipulation outcomes. Huang et al.[11] introduce a new method called AdaTrans, which is an adaptive non-linear latent transformation for disentangled and conditional face editing. The approach breaks down the manipulation process into multiple smaller steps, wherein at each step, the direction and size are determined based on both the facial features and the hidden codes. Current works frequently fail to retain the domain-invariant component of the image, such as the identity of human faces. Liu et al.[12] introduce an implicit style function (ISF) as a method to efficiently do image-to-image translation across multiple modes and domains using pre-trained unconditional generators. The ISF modifies the latent representation of an underlying code so that the resulting image falls inside the intended visual range.

Entire Image synthesis, where the whole image is synthesized using powerful GANs. ProGAN[16] and StyleGAN[17] are the popular examples that have been proposed over the years. The authors have enhanced the architecture of StyleGAN2[18] by incorporating the normalization technique employed in the generator. In addition, they modified the progressive GAN method by imposing constraints on the mapping of the generator, ensuring a more controlled transformation from the latent code to images. In StyleGAN3[19], the authors observe the aliasing effect by considering all signals in the network as continuous and may deduce simple architectural modifications that are universally applicable and ensure that undesired information cannot infiltrate the hierarchical synthesis process. Many techniques struggle to strike a balance between the accuracy and diversity of produced samples. Diffusion-based models, such as ConPreDiff[20], have been employed for diffusion-based synthesis. This significant method focuses on preserving semantic connections between nearby pixels, resulting in the development of high-quality image synthesis. Another approach uses Latent diffusion models to generate high-resolution images. Podell et al. [20] created the SDXL, a system that integrates multiple novel conditioning techniques and is trained on different aspect ratios. In addition, they provide a refinement model that improves the visual quality of samples generated by SDXL using a post-hoc image-to-image technique. Various famous datasets have been proposed: Faceforensics, Faceforensics++, Celeb-DF, DFDC, and Deeperforensics. These datasets use face manipulation techniques or a combination of several face manipulation techniques.

2.2 Deepfake detection

Since the advent of this technology, there have been very famous approaches that look for various clues/features for fake detection; after all these technological advancements in the respective fields, these methods do not generalize well on unseen samples. In other words, these methods fail to perform well on the cross-dataset, and they are not robust enough against various pre-processing operations like blurring, compression, and addition of noise. Different detection approaches have solved it to a certain extent. Khalid et al. [13] approach-based class detection approaches use a Variational Encoder, which is trained on real human face images, and other classes are considered anomalies, such as deepfakes. However, the method relies on the RMSE function to compute the reconstruction scores of the images. Their performance is unclear to the GAN-generated images; hence, the model is not generalizable. Li et al. [5] use face X-rays for a more general face forgery detection. The method assumes the existence of blending boundaries and decomposes the input image into integrating two or more sources. However, the method may fail for entirely synthetic images, and an adversary can develop an adversarial example that can bypass the detection mechanism and, hence, is not generalizable to different types of manipulations and robust against the various detection scenarios. Wang et al. [26] proposed a novel approach that is based on monitoring the neuron's behavior layer by layer, and such interaction patterns produce subtle artifacts for detection. Still, the algorithm's performance is far from satisfactory, and adversaries can develop adversarial examples that can evade the detection mechanism and, hence, are not robust enough. Other methods([27] [28]) use co-occurrence matrices as input on the image pixels to compute the disparities along various spatial and spectral band channels and then pass through CNNs for fake face detection. However, the information included in the raw data is lost due to handicraft features.

Generalization for the deepfake detector has been a critical issue, and various researchers have addressed it [29] [30] [31] [32] [33] [34]. A generalizable detector can perform on the cross-datasets, enabling it to perform well on unseen datasets and the dataset that undergoes various adversarial operations like noise, blur, and compression of samples. Nadimpalli et al. [35] use a combination of reinforcement learning and supervised learning. The approach enables the choice of top-k augmentation for the test samples, which are averaged together for final classification. The model gives good results for generalizability, but their evaluation is limited to the Celeb-Df and FF++ datasets. The generalization model needs to uncover various common features among various kinds of manipulation. Researchers worked in that direction [36] [37]. Yu et al. [37] use a module's independently trained specific forgery feature extractor and U-Net architecture to examine the module's validity. However, the method assumes the presence of similar forgery traces, which could be hidden or concealed by the adversary. Atama et al. [38] use image noise residual domain to examine the training impact of spatial and spatial-temporal features to improve generalization, but their methodology is based on high-frequency information, which is insufficient for the data compression scenarios. Li et al. [39] designed a code detection tool that uses a codebook to capture the actual space distribution of authentic and forged images. The model performs well with various cross-dataset images and compressed images, but the model performance remains unclear to various other and wild datasets. The generalizability of the detector has been a critical challenge that needs to be explored for a detector to work in various adversarial scenarios

Texture-based detection: Texture of CNNs has grabbed much attention in the last few years. Various authors ([40] [41] [18] [16]) worked on the textures, as CNN models are intrinsically biased towards textures rather than shapes. Yang et al. [41] described a novel Multi-scale Texture Difference Model that provides a stationary description of texture difference by merging the information from pixels' gradient and intensity. The model performs well on

various distortions, but the performance is significantly reduced for unseen manipulation. Bonomi et al. [17] use Local Derivative Patterns on Three Orthogonal Planes (LDP-TOP) to explore the textural dynamics of the spatial and temporal dimensions. Though the model performs well on various manipulations of FF++, their performance remains questionable on various kinds of GAN images and heavy compression video samples. Yadav et al. [42] have a unique end-to-end architecture for facial manipulation detection that extracts discriminative "textural" (T) and "manipulation residuals" (MR) features from two branches in unequal proportions. The performance of the model remains questionable to unseen scenarios. Gao et al. [43] explore the similarities between various forged content in terms of texture characteristics and the presence of artifacts, based on that designed an ensemble which results in the model's generalization abilities. Method performance is not suitable for highly compressed scenarios. Yang et al. [40] approach discover the texture difference between fake and real images in image saliency, further enhanced by the guided filter with a saliency map. The method requires much data and must be trained again for unseen images. Liu et al. [16] leverage global image texture representations using gram matrices for robust fake image detection. Their performance remains unclear to other deepfake manipulations. These detection methods either lack generalization capabilities to unseen manipulation or are not powerful enough against various detection scenarios like the addition of noise, compression, and blurring. This paper proposed a texture-based approach that uses CNNs and multi-scale vision transformers to improve these limitations. The texture is computed using gram matrices on top of ResNet layers, which serve as input to ViT and conventional CNN features. The discriminative model shows its effectiveness in improving generalizability and robustness towards the manipulated artifacts.

3 Empirical Analysis

The question confronting the research community is finding the common discriminative features for different fake images generated. Various researchers experimented in this regard [16] [18]; one of the critical findings of such experiments was that Texture statistics plays a vital role for various CNNs-based detectors like ResNet and successfully capture such texture differences for discrimination. This section presents the importance of texture as a feature for classification from different perspectives.

3.1 Texture as a clue

Texture refers to the appearance of the surface characterized by the shape, size, density, and proportionate arrangement of its elementary parts. In computer vision terminology, it is the repeated occurrence of the grey pixel level in the space [41]. An empirical analysis of the fake and real images is done to reveal the differences between textural characteristics. Texturized images are generated for accurate and fake images using a texturized generating algorithm, and it can be seen from Figure 2 that fake images lack texturized details compared to authentic images. This could be because forming fake tampered data samples usually involves three steps, i.e., pre-processing, face generation, and post-processing operation. Post-processing operations are generally done to hide such texture defects; as a result, counterfeit images tend to have a smoother surface and fewer texture characteristics [44]. Also, looking more closely at Figure 3, one finds that manipulated images come from smoother surfaces. Hence, the lack of a texturized surface would be a potential clue for fake image detection.



Figure 2: Real and fake Images are shown with their texturized images. Texturized images are generated from the images using the texture-based algorithm.



Figure 3: Fake images on a closer look showing that fake images tend to have smoother surfaces.

Human and Deep neural networks discern the images differently for their identification as real or fake. Humans usually identify artificial objects based on the shapes and colors of artifacts such as 'asymmetrical eyes,' 'uneven teeth,' 'irregular skin tone,' 'fairly smooth surfaces,' and more. [18]. Most of their identification is based on the shapes rather than the texture of the surface [45]. Reliance on these noticeable or high-level artifacts for identification misleads them easily, especially for entire-GAN image synthesis.

3.2 CNN Result Analysis

Many studies [46] [47] [48] have been done to capture the significance of various layers of CNNs and what each layer processes and learns. The widely accepted notion is that earlier layers of CNNs learn the low-level features like edges, lines, corners, etc., while the deeper ones learn the complex representations like faces, trees, etc. One of the more profound findings by the authors [1] is that CNN models are biased or inclined towards texture rather than shapes. In other words, CNN's models identify objects based on their surface rather than their shapes. Also, one of the exciting findings is that CNN's models cannot incorporate long-range information due to their small receptive fields [49]. Hence, CNNs need an additional module to integrate texture information into the module over much longer distances.

4 Proposed Model

The model comprises two components(Figure 4): texture Architecture and dual-branches cross-attention vision transformer. Texture architecture consists of two branches which serve as inputs to the parallel branches of the cross-attention vision transformer.

4.1 Texture Architecture:

Texture architecture constitutes resnet-18 architecture as a backbone, and the texture block is computed at the input and before every down-sampling operation incorporating global texture at various levels. The texture block consists of convolutional layers and gram matrices. Gram matrices are used to extract texture correlation, while convolution layers is then applied to enhance the representation, and pooling layers are used to align the computed features with the ResNet backbone features for the next level. The global Texture module is computed at multiple semantic levels before to each ResNet downsampling operation to model long-range texture features [16]. The main backbone of ResNet-18 learns the conventional features representation of the input images at various levels with a skip connection. It improves the gradient flow between the layers of the multi-scale features while the Texture block learns the global textures' semantic information at various scales.

Gram matrices as Texture features: Within a model, the texture is represented by the correlations among the features map responses in various model layers [50]. The Gram matrix quantifies the correlation between different feature map responses over different layers. These correlations, which are determined up to a constant factor, are represented by the gram matrices. Gram matrices $G^l \epsilon\, R^{N_l \times N_l}$ computes linear dependence between the layers:

$$G_{ij}^{l} = \sum_{k} F_{ik}^{l} F_{jk}^{l} \qquad \text{(i)}$$

Equation (i) represents the gram matrix. G_{ij}^{l} which is the inner product between the ith and jth feature of layer l, where F^l represents the lth feature map vectorized representation and F_{ik}^{l} represents the kth activation of the ith filter at position k in layer l. A texturized model, as defined, does not consider spatial information and is distinguished by the correlations among the feature maps. The texture is generated by computing gram matrices, which are calculated in the model prior to each downsampling operation of the ResNet. These matrices are then concatenated and used as input to the cross-attention mechanism of the vision transformer.

4.2 Dual branch Cross-Attention Vision Transformer

The vision transformer, free of inductive biases, is known for capturing long-range, global relationships between the pixels, courtesy of their self-attention mechanism and capacity for holding semantic information. The proposed architecture uses two parallel branches, and patches input into these branches are comparable in scale. The model takes the texture architecture's input, and then positional embedding is added into each patch, including the CLS token, to embed positional information into the model. Then, these tokens are passed through the stacked transformer encoder. Each transformer encoder contains a dual branch and is composed of Multi-headed self-attention(MSA) followed by the feed-forward Network(FFN) [51]. FFN includes two layers of the multi-layer perceptron, and the GELU non-linear layer is applied at the end of the first layers. Layer-Norm (LN) is used at the end of every block, with residual skip-connection applied after every block. The input to the $\mathbf{x}_0$ ViT and lth processing of the transformer encoder can be written as:

$$\mathbf{x}_0 = [\mathbf{x}_{\text{class}} || \mathbf{x}_{\text{patch}} \mathrm{E}] + \mathrm{E}_{\text{pos}} \qquad E \in R^{P^2.C \; x \; D}, E_{pos} \in R^{(N+1) \; x \; D} \qquad \text{(ii)}$$

$$\dot{z}_l = \dot{z}_{l-1} + MSA\big(LN(\dot{x}_{l-1})\big), \qquad l = 1. \ldots . L \tag{iii}$$

$$\dot{x}_l = \dot{z}_l + FFN\big(LN(\dot{z}_l)\big), \qquad l = 1 \ldots \ldots L \tag{iv}$$

where is $\dot{x}_{clsemb} \in R^{1\times C}$, $\dot{x}_{patchemb} \in R^{N\times C}$ and $\dot{x}_{posemb} \in R^{(N+1)\times C}$ are the cls, patch, and positional embedding tokens, respectively(C and N are the embedding's dimension and the number of the tokens). Afterward, the CLS token of one branch, which has learned the abstract information, acts as a token query to interact with the patch tokens of the other branch through an attention mechanism resulting in multi-scale features. Similarly, the CLS token interacts with the patch tokens of the other branch. The cross-attention mechanism is represented in the subsequent equations where $\dot{x}$ is the input to MSA(Multi-headed self-attention module:

$$\dot{y}^1 = [\dot{x}^1_{cls} || \dot{x}^2_{patch}] \quad \dot{x}^1 \in token\ I^{st} branch, \quad \dot{x}^2 \in tokens\ II^{nd}\ branch \tag{v}$$

$$q = \dot{y}^1_{cls} W_q, \qquad [k, v] = \dot{y}^1 W_{kv} \qquad W_q, W_{kv} \in R^{D\ x\ 3D_h} \tag{vi}$$

$$A = softmax\left(\frac{qk^T}{\sqrt{D_h}}\right) \qquad A \in R^{NxN} \tag{vii}$$

$$SA(\dot{y}^1) = Av \tag{viii}$$

$$MSA(\dot{y}^1) = [SA_1(\dot{y}^1); SA_2(\dot{y}^1); \ldots . ; SA_k(\dot{y}^1)] W_{msa} \qquad W_{msa} \in R^{k.D_h xD} \tag{ix}$$

$$\acute{y}^1_{cls} = \dot{x}^1_{cls} + MSA(LN([\dot{x}^1_{cls} || \dot{x}^2_{patch}])) \tag{x}$$

Where q, k, and v are the query, key, and value, respectively, n+1 is the number of patches, d is the model dimension, k is the number of heads, and Dh(d/k) is the head dimension. W_q, W_{kv}, and W_{msa} are the learnable parameters for the query, key, value, and MSA, respectively. Following fusion with other branch tokens, the CLS token at the next transformer encoder engages with its patch tokens once more. Here, it imparts knowledge from the other branch to its patch tokens, enhancing each patch token's representation. Then, these tokes are passed through the Layer Norm to MLP(Multi-Layer Perceptron) for parameter learning:

$$\ddot{y}^1 = \acute{y}^1_{cls} + \dot{x}^1_{patch} \tag{xi}$$

$$\check{z} = MLP(LN(\ddot{y}^1)) \tag{xii}$$

Finally, these classification tokens are concatenated for the final predictions.

Algorithms 1: Tex-ViT for Deepfake classification

Parameter Initialisation:

- Input: I = {I_1, I_2,…..I_n} be the set of images, and L = {0, 1} be the set of labels, 0 being the real and 1 being the deepfake image
- n is the size of the dataset
- Split I into three subsets for 70% training, 15% validation, and 15% testing.

1: **For 1 to 100 epochs, do**
2: Input image I into ResNet for feature extraction.
3: Compute the texture using texture block before every down sampling operation in ResNet and keep concatenating them.
4: ResNet computed Fed conventional CNNs, and texture features were calculated at step 1 and step 2 into the dual branch of the vision transformer.
5: Split the features into patches (fixed sizes) and flatten them at each branch.

6: With these image patches flattened, create linear embeddings in lower dimensions.
7: Include positional embeddings with CLS token.
8: Feed the sequence into the transformer encoder at each branch.
9: Create tokens by querying the CLS token of the I^{st} branch with patch tokens of another branch and vice-versa.
10: Concatenate the tokens of both branches for classification.
11: Train the model end-to-end and update weights using the Adam optimizer.
12: Evaluate the validation set and save the weights of the model that performs well.
13: **end for**
14: Load the weights of the model saved at step 11.
15: Evaluate the performance on the test set

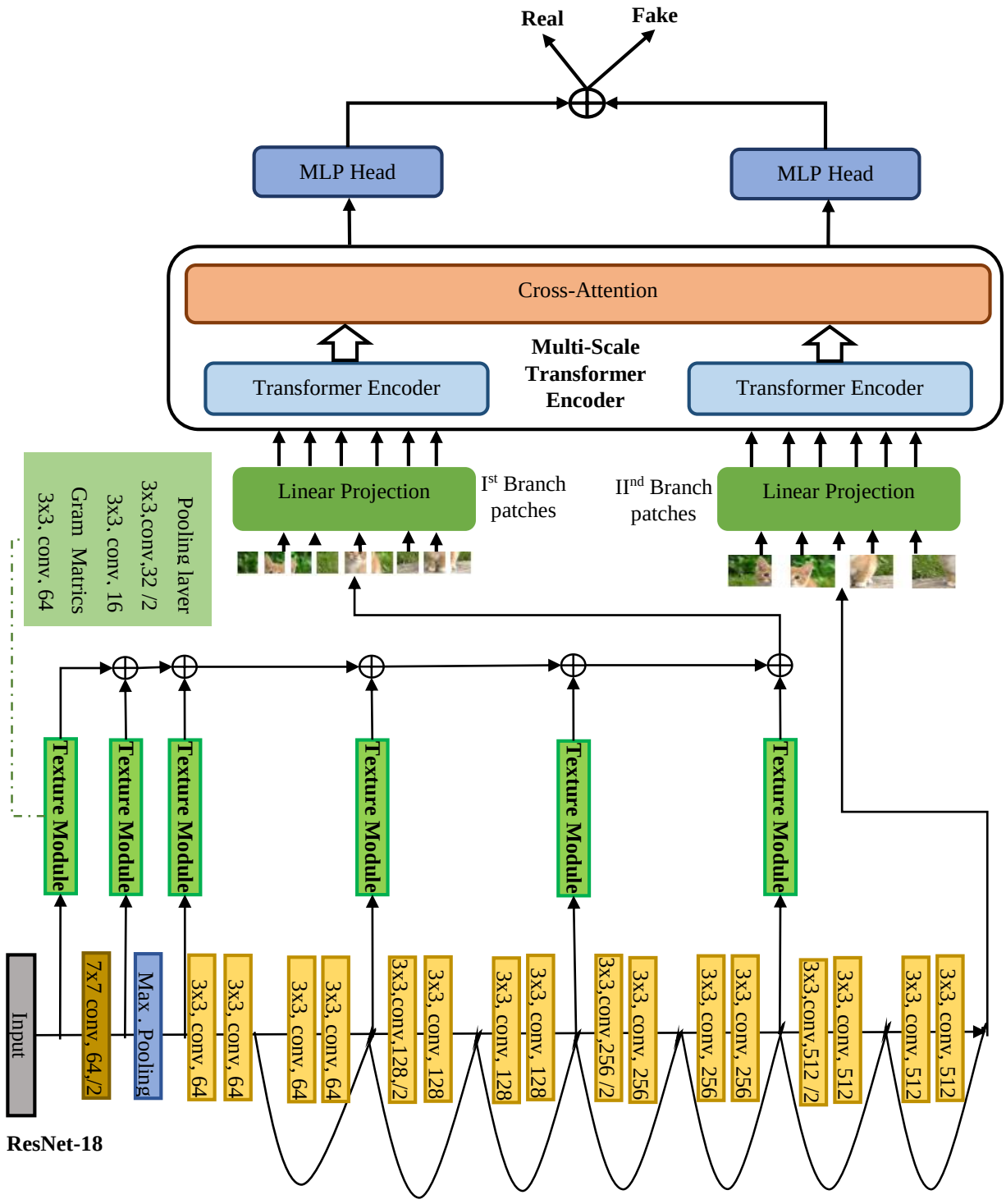


Figure 4: Proposed model consisting of texture module and ResNet serving as an input to dual-branch vision transformer with cross-attention mechanism

5 Experiments

This section will detail the choice of training hyperparameters, different datasets, the choice of face extractor, and the different experiment scenarios conducted.

5.1 Experimental Settings

The starting learning rate is set to 0.01. Adam optimizer is used to update the model's parameters, as it is being widely used and took less time for parameter updation in my case.

The batch size is taken as 64, due to the memory constraints at my computer systems. Each experiment is run for 100 epochs, as the model performance hits the saturation point after this specific number of epochs. The experiments are run for 24GB NVIDIA TITAN RTX GPUs.

5.2 Dataset Pre-Processing

Three deepfake datasets are used to evaluate the models: Celeb-DF, DFDC-Preview, and Faceforensics++, as these are being widely used and also the state-of-the-art dataset in the current scenarios. These datasets consist of short facial videos from which the frames are extracted utilizing the RetinaFaceResNet50 face extractor, as the RetinaFaceResNet50 face extractor has a lesser failure rate than MTCNN. One hundred frames are extracted from each video. For a DFDC and Celeb-DF dataset, most faces have aspect ratios of [1, 1.5] and heights between [151, 200] pixels. Lastly, an FF++ has a size of [151, 200] and an aspect ratio of [1,1.5](Table 1). Different GAN images are also used to evaluate the model. Fake images: ProGAN and StyleGAN images and real image datasets: CelebA-HQ, CelebA, and FFHQ are downloaded from their respective repositories. StarGAN and STGAN images are generated by executing the code from their GitHub repositories.

Table 1: Details for the training, validation, and testing dataset with their resolutions

Dataset	**Training Set**	**Validation set**	**Testing set**	**Image Resolution**
FF++(DeepFakes)	8000 authtentic, 8000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	128x128
FF++(face2face)	8000 authtentic, 8000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	128x128
FF++(Faceswap)	8000 authtentic, 8000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	128x128
FF++(Neural Texture)	8000 authtentic, 8000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	2000 authtentic, 2000 manipulated image frames.	128x128
DFDCPreview	10,000 authtentic, 8000 manipulated image frames.	1500 real, 1500 fake image frames.	1500 real, 1500 fake image frames.	128x128
Celeb-DF	10,000 authtentic, 10,000 manipulated image frames.	8000 authtentic, 8000 manipulated image frames.	8000 authtentic, 8000 manipulated image frames.	128x128
CelebA-HQ & ProGAN	10,000(CelebA-HQ) & 10,000(ProGAN)	1500(CelebA-HQ) & 1500(ProGAN)	1500(CelebA-HQ) & 1500(ProGAN)	1024x1024
CelebA-HQ& StyleGAN	10,000(CelebA-HQ) & 10,000(StyleGAN)	1500(CelebA-HQ) & 1500(StyleGAN)	1500(CelebA-HQ) & 1500(StyleGAN)	1024x1024
FFHQ and StyleGAN	10,000(FFHQ) & 10,000(StyleGAN)	1500(FFHQ) & 1500(StyleGAN)	1500(FFHQ) & 1500(StyleGAN)	1024x1024
CelebA & StarGAN	10,000(CelebA) & 10,000(StarGAN)	1500(CelebA) & 1500(StarGAN)	1500(CelebA) & 1500(StarGAN)	128x128
CelebA & STGAN	10,000(CelebA) & 10,000(StarGAN)	1500(CelebA) & 1500(StarGAN)	1500(CelebA) & 1500(StarGAN)	128x128

5.3 Data Augmentation and Scaling

Usually, a vision transformer needs data for training to perform at par with the CNN model, as shown by the original ViT model [52]. However, with the rich set of careful data-augmentation techniques, DeiT [53] has shown promising results with fewer data and comparable

performance with the CNN model. For the proposed model, various data-augmentation techniques have been used, which include rand augmentation [54], cut mix [55], and mixup [56], along with the random-erasing [57] and drop path regularisation model techniques to improve the overall results of the classifications.

5.4 Experiments on the cross-domain settings for the Faceforensics++ dataset

Experimentation has been performed on the various categories of Faceforensics++ [22]. The model is trained on one class of FF++ and tested on the same as on other varieties of FF++. Weights of the models that perform well on the validation are saved for evaluation on the test dataset. Even though a lot of state-of-the-art models have been introduced recently, it is still difficult to compare them fairly. This is partially because there is a dearth of publicly available codes for the models and training procedures that are unavailable to the research community. Consequently, we advocate for the community to embrace open-source software and for the generation strategies of large-scale datasets to be evaluated independently of the model's success. These kinds of actions are essential to maintaining equity and encouraging further developments in this area. For the comparison, four models are taken into consideration:

a) MesoInception-4 model [58].
b) Capsule Network [59].
c) Combining Vision Transformers and Efficient Net(E-ViT) [60].
d) UCF [36].
e) IID [61].
f) SIA[21]
g) UIA[22]

Code for these models has been taken from their GitHub repository and customized according to the dataset, and more evaluation metrics have been added for comprehensive evaluation. These models have been trained on one category of manipulation and tested on other categories of manipulation of FF++. These experiments are necessary to test the performance of the models against various manipulations, which is necessary to validate the detector's generalization abilities.

Table 2 represents the score of the various models when trained on the DeepFakes category and tested on the various categories of the FF++. Various models score very well for the same type of manipulation, and few even score perfectly, or it can be said that they are overfitting. These overfitted models' performance degraded heavily when asked to classify other categories of FF++. MesoNet and CapsuleNet, which are considered the most advanced models for detecting forgery, get an accuracy score of approximately 50% when it comes to identifying manipulated images in the FF++ dataset. This is due to their limited ability to learn just traditional features from convolutional neural networks (CNN), which is insufficient for effectively detecting cross-manipulation scenarios. Among the latest techniques, the Uia and Ucf approaches, which are renowned for identifying shared characteristics among different types of manipulation, do not meet the necessary criteria for an effective deepfake detector that can be applied universally. Our model outperforms the other models, with an accuracy of 72%, specifically for the DF category of the dataset. The majority of models struggle to accurately categorize the face-swap category in FF++, and a small number of models performed below 50% accuracy. Our model's performance demonstrates that texture is a consistent characteristic that remains present across different types of facial alterations.

Table 3 represents the scores for the models trained on the face2face categories and tested on the other categories of the FF++ dataset. Once again, the different models excessively suit the face2face category, and their effectiveness significantly declines when applied to other

categories of manipulation. The CapsuleNet approach exhibits the poorest performance among all methods for cross-manipulation. When trained on the face2face category, all models perform significantly better than in the previous scenarios in cross-manipulation settings. This is likely because face2face involves facial re-enactment techniques that use video frames to create a highly detailed reconstruction of the face, taking into account different lighting conditions and facial emotions. These inherent characteristics enable the models to learn the implicit features of the manipulation, which in turn aids in cross-manipulation scenarios. Uia is a highly effective model with an accuracy score of approximately 70%. This model utilizes an unsupervised technique and incorporates an inconsistency-aware module to detect discrepancies among the patch-level data. The Sia approach has poor performance due to its heavy reliance on the attention mechanism, which can occasionally miss tiny abnormalities in the manipulations. Furthermore, several models exhibited subpar performance without any notable improvements.

Conversely, Tex-ViT maintains the general characteristics, avoids overfitting when the training and testing data are from the same distribution, and generalizes well to alternative distribution categories. The manipulation achieved scores of 73% and 71% for the DF and NT categories, respectively. The FS category score for the manipulation has increased compared to the score in the preceding table.

Table 4 represents the performance of the model when trained in the FS category and tested on additional variations of FF++. All of the models exhibit overfitting for the same category and demonstrate inadequate performance for the other manipulation categories. This is mostly attributed to the Faceswap construction technique, which utilizes facial landmark points to generate a 3D template. This template is then projected onto the target shape in order to minimize the disparity between the projected shape and the landmark points. The meticulous process, which involves precise shape mixing and color correction, poses a greater challenge for the detector to identify accurately. The Ucf and Uia model exhibits poor performance, with an accuracy score of approximately 50% when trained on the face swap manipulation. This highlights the fragility of the model under different circumstances. Sia has poor performance, although IID demonstrates slightly higher performance compared to previous models. This improvement can be attributed to the model's capability to learn both implicit and explicit characteristics. However, the overall performance still falls well short of the desired score. MesoNet and CapsuleNet exhibit substandard performance; however, the transformer-based model E-ViT demonstrates significantly superior performance owing to its hybrid model structure, which combines CNN and ViT to capture long-range dependencies. Our model encounters difficulty in accurately identifying this category of manipulation, with an accuracy score ranging from 62-67%. This is because the face swap-generating mechanism incorporates intricate generation techniques that our model finds challenging to detect.

Table 5 represents the final category of manipulation, wherein the model is trained using the NT category of manipulation and subsequently tested on the remaining categories. The models trained in this specific category of manipulation exhibit a commendable level of performance when compared to other categories of manipulation. The NT generation mechanism utilizes the texturing technique to understand the inherent characteristics of the data sample, enabling it to excel in the categories of manipulation. All the models exhibit comparable performance when trained on the NT category of manipulation, indicating that texture is one of the invariant properties that assist in their ability to perform in diverse types of manipulation. The majority of the models achieved a score above 70% when tested on the DF category of manipulation. With the exception of Ucf and IID, all models in the face2face

category perform well and even outperform the former category. However, the manipulation category (FS) consistently shows poor performance. Furthermore, the issue of overfitting persists in the NT category. Once again, our model outperforms the other models. The state-of-the-art (SoTA) model achieves an accuracy score of 77% in face-to-face manipulation and 76% in Deepfake manipulation categories, demonstrating higher performance compared to other models.

So, other models score almost perfectly when training and testing come from the same distribution but fail to generalize well for other distributions. In contrast, based on the texture module and cross-attention mechanism, our model performs well in almost various manipulations, effectively proving that texture is a potential feature that persists among different manipulations. However, every performance suffers from the FS category of manipulation of FF++. Figure 5 represents the ROC curves of various models for images trained on face2face and testing on different types.

5.5 Experiments on the cross-domain settings for GAN dataset images

Extensive testing has been done on various GAN-generated images. Complete image synthesis, such as StyleGAN, ProGAN, and Attribute manipulation images of StarGAN and STGAN, has been considered for analysis. High-resolution authentic images are taken from the FFHQ, CelebA-HQ, while low-resolution images are taken from the CelebA. Five real and fake image datasets have been designed for the fair and comprehensive evaluation: CelebA-HQ ProGAN, CelebA-HQ StyleGAN, FFHQ StyleGAN, CelebA StarGAN, and CelebA STGAN. To compare the results of these datasets, four state-of-the-art models have been considered:

a) Xception with depth-wise separable convolution(Xception-Net) [62].
b) CNN's generated images are easy to spot now(CNN-Net) [63].
c) Efficient-Net [64].
d) UCF [36].
e) IID [61].
f) SIA[21]
g) UIA[22]

Once again, code has been extracted from the GitHub source and tailored to suit the dataset used for these models. Additionally, additional evaluation criteria have been incorporated to facilitate comparison evaluation. The model exhibits exceptional performance and surpasses the scores achieved by several state-of-the-art methodologies. Similarly, when it comes to FF++ models designed for cross-forgery, their current level of performance is still rather distant from the optimal score required for their practical implementation in real-world situations. Table 6 represents the score of these models on various datasets. It is evident that the scores of different models vary between datasets; they excel in one dataset but do poorly in another. It is evident that when the testing dataset contains both fake and actual images, models can promptly recognize them due to their training on these types of images, but they face difficulty in identifying the other class. For instance, in the initial row, a model that was trained on CelebA-HQ ProGAN and tested on CelebA-HQ StyleGAN, most models can accurately recognize the CelebA-HQ category, as indicated by the precision value. However, they struggle to correctly classify the other category, as indicated by their recall value. ProGAN and StyleGAN employ distinct modification techniques, leading to disparate feature spaces.

Consequently, models trained in one category exhibit suboptimal performance on the other. There is a noticeable similarity between the CelebA-HQ and FFHQ real datasets in terms of their characteristics, as a model trained on one dataset can achieve good performance on the other. For instance, a model that has been trained using the CelebAHQ StyleGAN has strong

performance when combined with the FFHQ StyleGAN. Among these scenarios, the IID model has performed exceptionally well. This is because the datasets used are completely synthetic, and the IID model is particularly adept at identifying implicit inconsistencies within the feature space. Tex-ViT consistently beats other models across a range of settings. The evidence indicates that the model acquires the typical distinguishing characteristics, including the overall texture that remains consistent despite different types of modification, whether it is generating a full image or altering various qualities. The recall metric score indicates that all the models struggle to identify the ProGAN images accurately. Remarkably, nearly all the models awarded a flawless rating to the photos that were trained using StarGAN and assessed using STGAN, and vice versa. This could potentially elucidate the rationale behind the utilization of comparable counterfeiting techniques in their production. . Figure 6 represents the ROC curves of various models for images trained and tested on different GAN image datasets.

5.6 Experiments on the in-domain settings for various post-processing operations of FF++, Celeb-DF, and DFDCPreview dataset

One of the limitations of the different models is that they are not robust enough for various post-processing operations like blurring, compression, the addition of noise, scaling, translation, etc. [1]. To demonstrate the model's robustness, the primary post-processing operations on the test dataset are blurring, compression, and addition of noise. For blurring the images, Gaussian blur PyTorch transformation has been used with a kernel size of 7x7 and sigma 25; for the addition of noise, zero mean and standard deviation of 0.2 have been designed, and finally, for the compression, quality of the images has been degraded by 3x times(Figure 7). Models have been trained on the regular images but tested on the images undergoing various post-processing operations. Three primary deepfake datasets have been considered for evaluation. Again, four models have been used for the comparative assessment:

a) MesoInception-4 model [58].
b) Capsule Network [59].
c) CNN's generated images are easy to spot now(CNN-Net) [63].
d) UCF [36].
e) IID [61].
f) SIA[21]
g) UIA[22]

Testing dataset does not undergo any post-processing operations. The first row of the table represents the results when the image has not undergone any processing operations; in that case, all the models have performed perfectly(Table 7). Without undergoing post-processing operations, every model in the FF++ dataset overfits and ultimately performs poorly for various post-processing activities. When comparing FF++ to DFDCPreview and Celeb-DF, FF++ models perform better. However, the score for the Celeb-DF dataset is lower, possibly because this dataset contains high-quality images that closely resemble genuine images, with constant lighting and texture, making it slightly more challenging. MesoNet has gotten the lowest amount compared to other competitive approaches, mostly due to its poor capacity to capture the traditional CNN features. The IID approach outperforms the other model, achieving a near-perfect score for the DFDCPreview dataset. The model's capacity to concentrate on the inconsistent characteristics connected to identity makes it particularly useful for face-swapping manipulations.

Testing dataset undergoes blurring operations: Here, the second row for each dataset represents the score for blurring operations. The performance has not degraded significantly

for the blurring operation, showing that the blurry images retain the manipulated artifacts of the non-blurry images. However, the images have been blurred to a significant extent. The performance of the models in the case of FF++ decreased by at least 12% when exposed to blur operations. For the DFDCPreview and Celeb-Df datasets, the performance showed a slight decline, around 3-6%, with a few exceptions for certain models. MesoNet exhibits a more pronounced decrease in performance for Celeb-DF and DFDCPreview because of their heavy reliance on traditional CNN capabilities. The IID model has seen a decline in performance, specifically for the FF++ dataset, but there is a slight reduction in performance for the other dataset. Sia and Uia exhibit superior resilience compared to other models when exposed to blurring operations. Other models are significantly impacted. Our model has improved its ability to withstand and recover from challenges, as indicated by a mere 1-2% decrease in performance for the DFDCPreview and Celeb-Df datasets. Moreover, almost 12% of the data samples continue to display discriminatory artifacts even after undergoing substantial blurring.
Testing dataset undergoes compression: The quality of the photographs was reduced by treble as a consequence of the compression. Compressing samples for the DFDCPreview and Celeb-DF datasets has a minimal effect on the models' performance, suggesting that the altered artifacts are not significantly affected by the reduction in size. The performance of models such as UCF, CNN-Net, Sia, and IID in the FF++ dataset has been significantly impacted, indicating that these models are not specifically designed for compression circumstances. Additionally, the dataset contains a variety of manipulations, which exacerbates the model's complication in comprehending the extensive distribution of features. The models' efficacy was minimally affected by the Celeb-DF dataset, while the DFDCPreview dataset had a slightly greater impact. Compression typically entails a reduction in the resolution of data sampling, which affects the smaller details and leads to distortions such as ringing, banding, blocking, and halo. The gradients in the smooth portions are significantly impaired by these distortions. Models that concentrate on specific artifacts encounter difficulties when confronted with a diverse array of intricate properties. Models that concentrate on a variety of artifacts at differing levels of detail are more likely to accurately classify intricate feature patterns. Our model prioritizes intricate attributes, commencing with the integration of conventional CNN features at multiple levels and texturing. It employs a cross-attention method to comprehend both global and local details by utilizing the capabilities of transformers. This enables the model to acquire nuanced, intricate characteristics at multiple levels with greater efficacy.
Addition of Noise to the testing dataset: The PyTorch transformation modifies the data samples with noise, which has a mean of zero and a standard deviation of 0.2. The results revealed a substantial decrease in the scores of all models, with IID, MesoNet CNN-Net, Sia, and Uia scoring as low as 50%. This underscores the susceptibility of these detection methods to the presence of noise. The IID model, which is intended to detect face-swap, has been significantly impacted by the introduction of noise in the FF++ and DFDC Preview datasets. This has resulted in erroneous classification and has rendered the model susceptible to adversarial perturbations. MesoNet is susceptible to a variety of adversarial strategies as a result of a significant decrease in its classification score. The quality of images is reduced by the presence of noise, which masks anomalies or inconsistencies and introduces random variations and patterns. Consequently, accurate or complete feature extraction is impeded. In order to achieve optimal performance on a noisy dataset, a model must either employ the attention mechanism to leverage multi-scale features that can effectively capture both local and global features that are resilient to noise, or employ sophisticated data augmentation techniques that introduce noise to aid in the model's classification learning. By utilizing the latter approach, our model

has been able to acquire intricate and resilient features that can withstand a variety of adversarial techniques. When the accuracy of the other model in the DFDCPreview dataset is less than 80%, the Tex_ViT model consistently outperforms it with an accuracy score of 98%. Nevertheless, the model remains susceptible to the introduction of disturbance to a certain extent.

Table 2: Models trained on deepfake dataset of FF++ and tested on its different categories. Here, bold values represent the highest score among competitive methods.

Test	Ucf					IID					MesoNet					CapsuleNet					E-ViT					Sia					Uia					Tex-ViT(ours)				
	Pr	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**
Df	0.9989	0.9985	0.9987	0.9987	0.9987	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	1.0	0.999	0.9994	1.0	0.9995	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.999	1.0	0.9995	1.0	0.9995	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9829	0.982	0.9824	0.9985	0.9825
F2F	0.791	0.263	0.263	0.697	0.6023	0.7790	0.2715	0.4026	0.7218	0.5972	0.6667	0.1455	0.2389	0.6345	0.5365	0.6239	0.151	0.2431	0.6159	0.53	0.755	0.165	0.2708	0.6479	0.5557	0.6464	0.3455	0.4503	0.6345	0.5783	0.7286	0.192	0.3039	0.6508	0.5602	**0.7242**	**0.6126**	**0.6637**	**0.7048**	**0.6948**
FS	0.224	0.018	0.033	0.711	0.5396	0.4637	0.016	0.0309	0.5621	0.4987	0.2168	0.0155	0.0289	0.2962	0.4797	0.3764	0.0335	0.0615	0.4207	0.489	0.392	0.265	0.4964	0.4700	0.4927	0.3245	0.0865	0.1365	0.3999	0.4532	0.4652	0.0435	0.0709	0.5062	0.49625	**0.6952**	**0.040**	**0.07564**	**0.6547**	**0.6291**
NT	0.8164	0.238	0.368	0.6693	0.5923	0.8765	0.245	0.3829	0.7263	0.6025	0.7703	0.208	0.3275	0.6972	0.573	0.7132	0.25	0.3702	0.6639	0.5747	0.789	0.257	0.3877	0.6829	0.5942	0.6616	0.5025	0.5711	0.6707	0.6227	0.7794	0.304	0.4374	0.7097	0.609	**0.7248**	**0.6989**	**0.6541**	**0.7958**	**0.7028**

Table 3: Models trained on the face2face dataset of FF++ and tested on its different categories. Here, bold values represent the highest score among competitive methods.

Test	Ucf					IID					MesoNet					CapsuleNet					E-ViT					Sia					Uia					Tex-ViT(ours)				
	Pr	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**
Df	0.8531	0.3165	0.4617	0.7533	0.631	0.8638	0.387	0.5345	0.8117	0.663	0.7459	0.367	0.4919	0.7085	0.621	0.8161	0.253	0.3862	0.6304	0.598	0.6751	0.2775	0.3933	0.6733	0.572	0.6217	0.544	0.5802	0.6560	0.6065	0.7305	0.6305	0.6768	0.777	0.6989	**0.7365**	**0.664**	**0.6983**	**0.7958**	**0.7133**
F2F	0.9975	0.9985	0.9980	0.9986	0.998	0.9995	1.0	0.9997	1.0	0.9997	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9960	0.9995	0.9977	0.9999	0.9977	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9714	0.9875	0.9794	0.9986	0.9793
FS	0.644	0.124	0.2079	0.5282	0.5277	0.5780	0.1315	0.2143	0.6605	0.5177	0.5871	0.266	0.3662	0.6064	0.5395	0.6371	0.151	0.2441	0.566	0.5325	0.6399	0.2390	0.3480	0.5922	0.552	0.5676	0.5015	0.5325	0.6079	0.5597	0.7120	0.4315	0.5373	0.6473	0.6284	**0.6952**	**0.5982**	**0.6433**	**0.6987**	**0.6593**
NT	0.8092	0.314	0.4524	0.6827	0.62	0.7891	0.378	0.5111	0.7877	0.6384	0.6937	0.3375	0.4541	0.6843	0.5942	0.8122	0.2855	0.4224	0.6666	0.6097	0.7209	0.4855	0.5802	0.7106	0.649	0.6741	0.694	0.6839	0.7217	0.6792	0.7523	0.638	0.6904	0.7949	0.7139	**0.7604**	**0.689**	**0.7229**	**0.8152**	**0.7360**

Table 4: Models trained on the face swap dataset of FF++ and tested on its different categories. Here, bold values represent the highest score among competitive methods.

Test	Ucf					IID					MesoNet					CapsuleNet					E-ViT					Sia					Uia					Tex-ViT(ours)				
	Pr	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**	**Pr**	**Re**	**F1**	**AUC**	**Acc**
Df	0.371	0.0305	0.0564	0.4577	0.4894	0.7414	0.3885	0.5098	0.7262	0.6226	0.5892	0.522	0.5535	0.6321	0.579	0.5375	0.0895	0.1534	0.5415	0.5065	0.6872	0.311	0.4282	0.6712	0.5847	0.6432	0.274	0.3843	0.6093	0.5610	0.5309	0.253	0.3427	0.5789	0.5147	**0.7746**	**0.2486**	**0.3764**	**0.6847**	**0.6248**
F2F	0.7403	0.1625	0.2665	0.5369	0.5527	**0.6943**	**0.4315**	**0.5322**	**0.6964**	**0.6207**	0.6216	0.5455	0.5811	0.6354	0.6067	0.6383	0.1985	0.3028	0.5568	0.5430	0.6700	0.3005	0.4149	0.573	0.5762	0.7391	0.35	0.4751	0.6554	0.6132	0.6722	0.4605	0.5465	0.6516	0.618	0.7546	0.2283	0.3505	0.6446	0.6078
FS	0.9969	0.9949	0.9949	0.9949	0.995	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9853	0.9755	0.9804	0.9982	0.9805
NT	0.6137	0.058	0.105	0.4929	0.5107	0.6113	0.2855	0.3892	0.5926	0.552	0.5903	0.539	0.5635	0.6205	0.5825	0.5342	0.1365	0.2174	0.5350	0.5087	0.6246	0.2355	0.3420	0.5574	0.547	0.5770	0.219	0.3175	0.5467	0.5292	0.5275	0.2585	0.3469	0.5181	0.5135	**0.7589**	**0.6378**	**0.6931**	**0.7088**	**0.6728**

Table 5: Models trained on NTdataset of FF++ and tested on its different categories. Here, bold values represent the highest score among competitive methods.

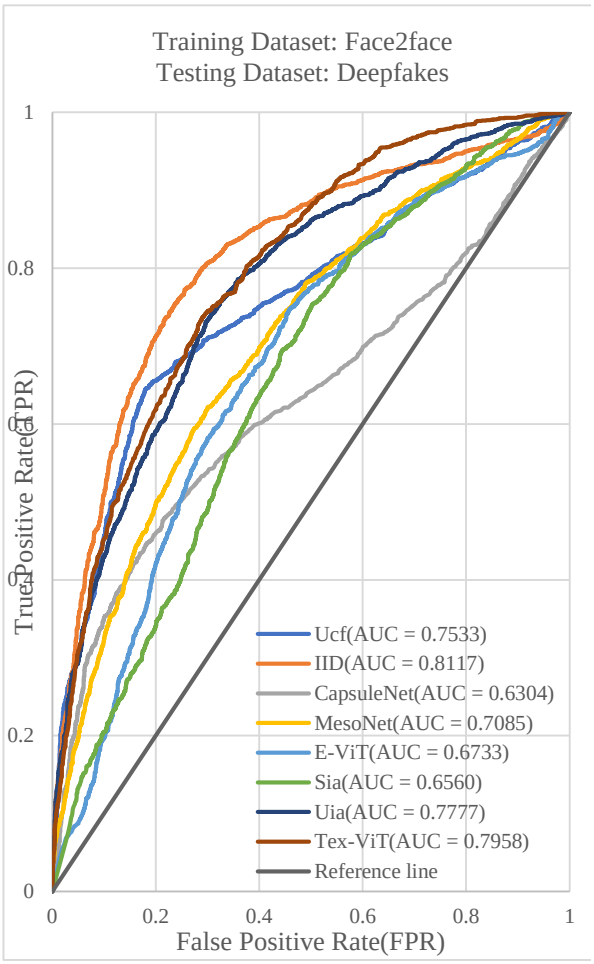
Test	Ucf					IID					MesoNet					CapsuleNet					E-ViT					Sia					Uia					Tex-ViT(Ours)				
	Pr.	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re**	**F1**	**AUC**	**Acc**	**Pr.**	**Re**	**F1**	**AUC**	**Acc**	**Pr.**	**Re**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**
Df	0.8814	0.613	0.7231	0.7860	0.7652	**0.7941**	**0.729**	**0.7601**	**0.8542**	**0.77**	0.7347	0.8145	0.7725	0.8287	0.7602	0.7619	0.616	0.6812	0.7801	0.7117	0.7894	0.6505	0.7133	0.8325	0.7385	0.7236	0.6335	0.6755	0.7699	0.6957	0.6966	0.6395	0.6668	0.7383	0.6805	0.7405	0.8045	0.7714	0.8394	0.7612
F2F	0.7989	0.379	0.5146	0.6384	0.642	0.7169	0.6245	0.6675	0.7843	0.6899	0.7276	0.8615	0.7889	0.8342	0.7695	0.7100	0.5045	0.5898	0.7265	0.6492	0.7583	0.739	0.7485	0.8374	0.7517	0.7369	0.7565	0.7466	0.8053	0.7432	0.7532	0.748	0.7506	0.8183	07514	**0.7672**	**0.7995**	**0.7830**	**0.8558**	**0.7785**
FS	0.2403	0.0465	0.078	0.4005	0.4497	0.3039	0.1295	0.1816	0.4045	0.4156	0.5013	0.3745	0.4287	0.5296	0.501	0.4557	0.1675	0.2449	0.4886	0.4837	0.4966	0.2235	0.3082	0.5057	0.4985	0.4798	0.304	0.3722	0.5189	0.4872	0.5543	0.429	0.4836	0.5606	0.542	**0.7072**	**0.5844**	**0.6399**	**0.8148**	**0.6408**
NT	0.9954	0.993	0.9945	0.9972	0.9945	0.9494	0.9955	0.9719	0.9999	0.9712	0.9985	1.0	0.9992	1.0	0.9992	1.0	0.9925	0.9959	0.9978	0.996	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9645	0.9795	0.9719	0.9971	0.9718

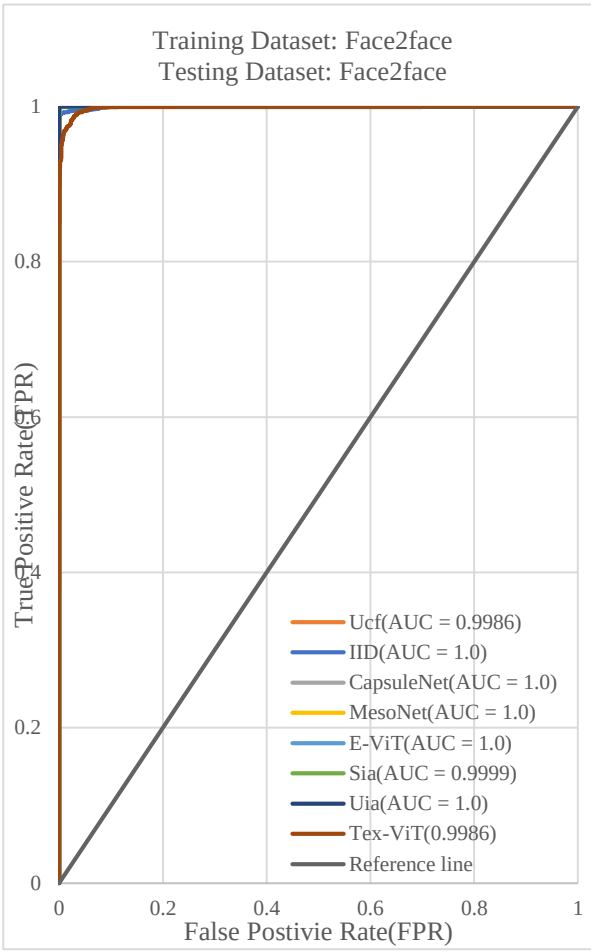
Table 6: Models trained and tested on the GAN datasets. Here, bold values represent the highest score among competitive methods.

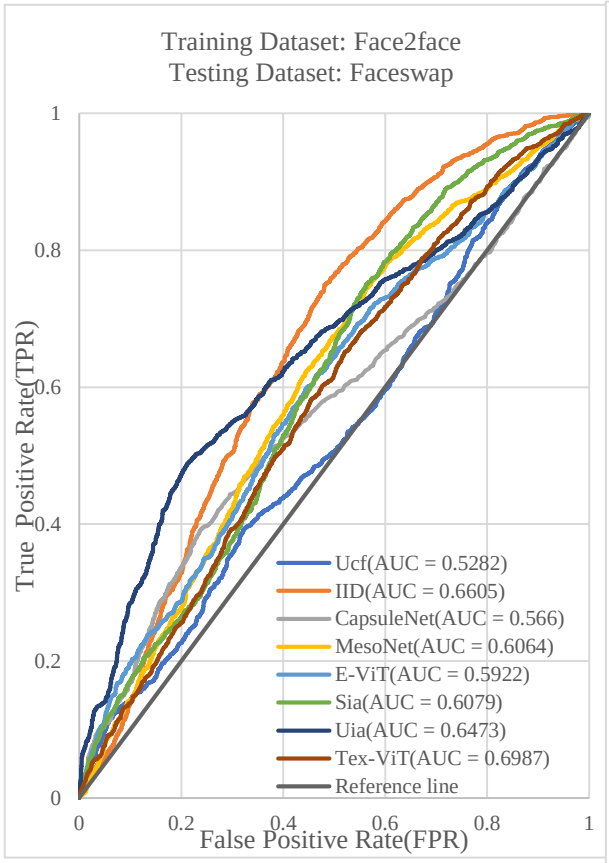
Train	Test	Xception					Ucf					IID					CNN-Net					Efficient-Net					Sia					Uia					Tex-ViT(ours)				
		Pr.	**Re.**	**F1**	**AUC**	**Acc.**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**
CelebA-HQ_ProGAN	CelebAHQ_StyleGAN	1.0	0.06	0.113	0.789	0.53	0.9	0.006	0.011	.4645	0.502	1.0	0.003	0.006	0.684	0.501	0.971	0.191	0.321	0.970	0.595	1.0	0.008	0.015	0.846	0.504	1.0	0.165	0.283	0.937	0.582	0.8	0.005	0.010	0.865	0.502	0.880	**0.64**	**0.741**	**0.892**	**0.776**
	FFHQ_StyleGAN	0.904	0.050	0.095	0.754	0.522	0.652	0.997	0.788	0.777	0.716	0.8	0.002	0.005	0.604	0.501	0.609	0.346	0.446	0.653	0.570	0.721	0.038	0.072	0.720	0.511	0.778	0.154	0.258	0.725	0.555	0.966	0.019	0.037	0.728	0.509	**0.769**	**0.638**	**0.698**	**0.809**	**0.723**
CelebA-HQ_StyleGAN	CelebAHQ_ProGAN	1.0	0.024	0.046	0.687	51.2	0.471	0.005	0.010	0.416	0.496	0.5	0.5	0.5	0.546	0.5	0.788	0.003	0.006	0.836	0.501	1.0	0.006	0.001	0.809	0.503	1.0	0.005	0.010	0.845	0.502	0.5	0.5	0.5	0.470	0.5	**0.713**	**0.544**	**0.617**	**0.681**	**0.606**
	FFHQ_StyleGAN	0.862	1.0	0.926	0.988	0.920	0.732	0.999	0.845	0.773	0.816	0.863	**1.0**	**0.926**	**0.999**	**0.920**	0.993	0.998	0.861	0.994	0.84	0.742	0.999	0.851	0.997	0.826	0.610	1.0	0.758	0.978	0.681	0.702	1.0	0.825	0.996	0.788	0.840	0.948	0.891	0.945	0.884
FFHQ_StyleGAN	CelebAHQ_ProGAN	0.980	0.236	0.381	0.811	0.616	0.5	0.5	0.5	0.5	0.5	0.666	0.001	0.002	0.602	0.503	0.772	0.012	0.023	0.779	0.506	0.903	0.018	0.036	0.795	0.508	0.636	0.004	0.009	0.577	0.501	0.964	0.036	0.069	0.786	0.517 33	0.842	**0.362**	**0.506**	**0.745**	**0.647**
	CelebAHQ_StyleGAN	0.994	0.999	0.997	0.999	0.997	0.996	0.99	0.993	0.997	0.993	1.0	**0.999**	**0.999**	**0.999**	**0.999**	0.999	0.996	0.996	0.999	0.996	0.996	0.999	0.999	0.999	0.999	0.998	0.997	0.997	0.999	0.998	0.998	0.998	0.998	0.999	0.998	0.981	0.979	0.980	0.996	0.980
CelebA StarGAN	CelebA STGAN	0.999	1.0	0.999	1.0	0.999	0.999	1.0	0.999	1.0	0.999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.977	0.927	0.951	0.988	0.953
CelebA StarGAN	CelebA STGAN	0.998	1.0	0.999	0.999	0.999	1	1	1	1	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	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.995	0.949	0.972	0.996	0.972

Table 7: Models trained on various datasets and tested under various conditions. Here, bold values represent the highest score among competitive methods.

Training Dataset	Testing Dataset	Ucf					IID					MesoNet					CapsuleNet					CNN-Net					Sia					Uia					Tex-ViT(ours)				
		Pr.	**Re.**	**F1**	**AUC**	**Acc.**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc.**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**	**Pr.**	**Re.**	**F1**	**AUC**	**Acc**
FF++	FF++	0.988	0.9915	0.9898	0.9968	0.9897	0.9897	0.9951	0.9923	0.9996	0.9923	0.8322	0.9998	0.9084	0.9968	0.8991	**0.9925**	**0.9993**	**0.9993**	**0.9997**	**0.9993**	0.9998	0.9996	0.9968	0.9999	0.9968	0.9973	0.9985	0.9979	0.9999	0.9979	0.9982	0.9992	0.9986	0.9999	0.9987	0.9368	0.9424	0.9396	0.9882	0.9395
FF++	FF++ Blurry	0.673	0.9282	0.7803	0.8297	0.7386	0.575	0.9975	0.7299	0.8699	0.6304	0.5305	1.0	0.6933	0.9819	0.5575	0.628	0.9975	0.7715	0.8788	0.7039	0.9807	0.9982	0.8329	0.9831	0.7998	0.7540	0.958	0.8438	0.9336	0.8227	0.7151	1.0	0.8338	0.9854	0.8007	**0.7720**	**0.9894**	**0.8673**	**0.9644**	**0.8486**
FF++	FF++ Noisy	0.642	0.6445	0.6433	0.6479	0.6415	0.5	0.5	0.5	0.5903	0.5	0.5235	0.9928	0.6854	0.7079	0.5445	0.6981	0.046	0.0092	0.5612	0.5013	0.5	0.5	0.5	0.5	0.5	0.5	0.0095	0.0186	0.4979	0.50	0.7489	0.1103	0.1923	0.6546	0.5366	**0.6425**	**0.7839**	**0.7061**	**0.7409**	**0.673**
FF++	FF++ Compression	1.0	0.0025	0.0044	0.5784	0.5011	1.0	0.0094	0.0186	0.912	0.5047	**0.9262**	**0.9868**	**0.9555**	**0.9940**	**0.9541**	1.0	0.0032	0.0064	0.6939	0.5016	0.9376	0.0764	0.1419	0.9300	0.5381	0.9973	0.231	0.3751	0.9002	0.6151	0.9992	0.6947	0.8196	0.9914	0.8471	0.9713	0.8651	0.9159	0.9839	0.9206
DFDCPreview	DFDCPreview	1.0	0.9996	0.9997	0.9997	0.9998	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.9900	0.9993	0.9947	0.9999	0.9946	1.0	0.9993	0.9996	0.9999	0.9996	0.9999	0.9986	0.9989	0.9999	0.999	1.0	0.9986	0.9993	1.0	0.9993	1.0	0.9993	0.9996	0.9999	0.9996	0.9926	0.9893	0.9909	0.9995	0.9910
DFDCPreview	DFDCPreviewBlurry	0.998	0.7926	0.8838	0.9287	0.8956	1.0	0.8493	0.9185	0.9940	0.9246	0.9985	0.8913	0.9418	0.9981	0.945	1.0	0.748	0.8558	0.9881	0.874	0.9967	0.72	0.8372	0.9965	0.86	1.0	0.8273	0.9055	0.9957	0.9136	1.0	0.9433	0.9708	0.9999	0.9716	**0.9933**	**0.9893**	**0.9913**	**0.9992**	**0.9913**
DFDCPreview	DFDCPreview Noisy	0.5	1.0	0.6666	0.5447	0.5	0.5	1.0	0.6666	0.50	0.5	0.5	1.0	0.6666	0.5	0.5	0.4999	0.9986	0.6662	0.7179	0.4996	0.5598	1.0	0.6666	0.6015	0.5	0.2727	0.01	0.019	0.4388	0.4916	0.4978	0.99	0.6625	0.6841	0.4956	**0.9788**	**0.986**	**0.9824**	**0.9985**	**0.982**
DFDCPreview	DFDCPreview Compression	0.757	0.9999	0.8619	0.7940	0.84	0.9375	1.0	0.9677	0.9999	0.9666	0.9695	0.9986	0.9839	0.9994	0.9836	0.5468	1.0	0.7070	0.9304	0.5856	0.9943	1.0	0.7587	0.9938	0.9939	0.9247	0.9986	0.9602	0.9989	0.9586	0.9621	0.9993	0.9803	0.9998	0.98	**0.9980**	**0.9986**	**0.9983**	**0.9999**	**0.9998**
Celeb-DF	Celeb-DF	0.954	0.9543	0.9544	0.9605	0.9544	0.9227	0.9451	0.9338	0.9773	0.9330	0.8222	0.9075	0.8628	0.9259	0.8556	0.9418	0.9462	0.9441	0.9741	0.9439	0.9848	0.9376	0.9406	0.9814	0.9408	0.8483	0.9008	0.8738	0.9431	0.8698	0.9235	0.9298	0.9266	0.9823	0.9264	**0.976**	**0.9442**	**0.9603**	**0.9928**	**0.9610**
Celeb-Df	Celeb-DF Blurry	0.830	0.9511	0.8865	0.8937	0.8782	0.7483	0.9397	0.8332	0.9174	0.8118	0.6291	0.9757	0.7649	0.8822	0.7002	0.8793	0.756	0.8130	0.8979	0.8261	0.9600	0.9421	0.8912	0.9574	0.885	0.8191	0.8792	0.8481	0.9189	0.8425	0.8717	0.9345	0.9020	0.9631	0.8985	**0.9502**	**0.9258**	**0.9379**	**0.9812**	**0.9387**
Celeb-Df	Celeb-DF Noisy	0.5	1.0	0.6666	0.5083	0.50	0.6241	0.5	0.5	0.5427	0.50	0.5	0.5	0.5	0.5267	0.50	0.5	0.5	0.5	0.5696	0.5	0.5	0.5	0.5	0.513	0.50	0.4766	0.09312	0.15558	0.4947	0.4954	0.6924	0.1513	0.2484	0.6218	0.5421	**0.6649**	**0.7581**	**0.7084**	**0.7365**	**0.6721**
Celeb-Df	Celeb-DF Compression	0.973	0.5248	0.6819	0.9124	0.7551	0.9048	0.7524	0.8216	0.9321	0.8366	0.9468	0.334	0.4938	0.9193	0.6576	0.9278	0.5146	0.6620	0.8951	0.7373	0.9344	0.534	0.6881	0.9328	0.758	0.7512	0.9086	0.8228	0.9093	0.8044	0.9192	0.778	0.8427	0.9481	0.8548	**0.9203**	**0.9252**	**0.9227**	**0.9793**	**0.9226**







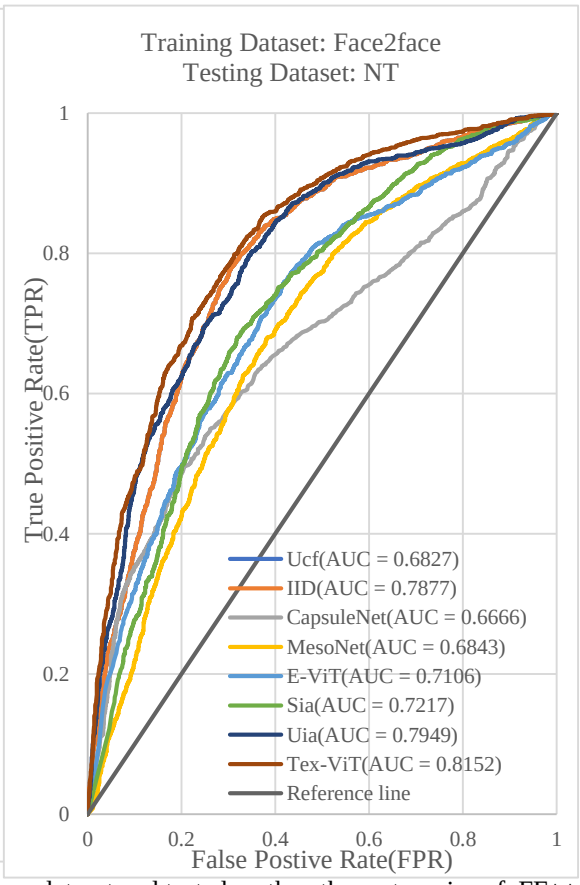


Figure 5: ROC curve for the model trained on Face2face dataset and tested on the other categories of FF++

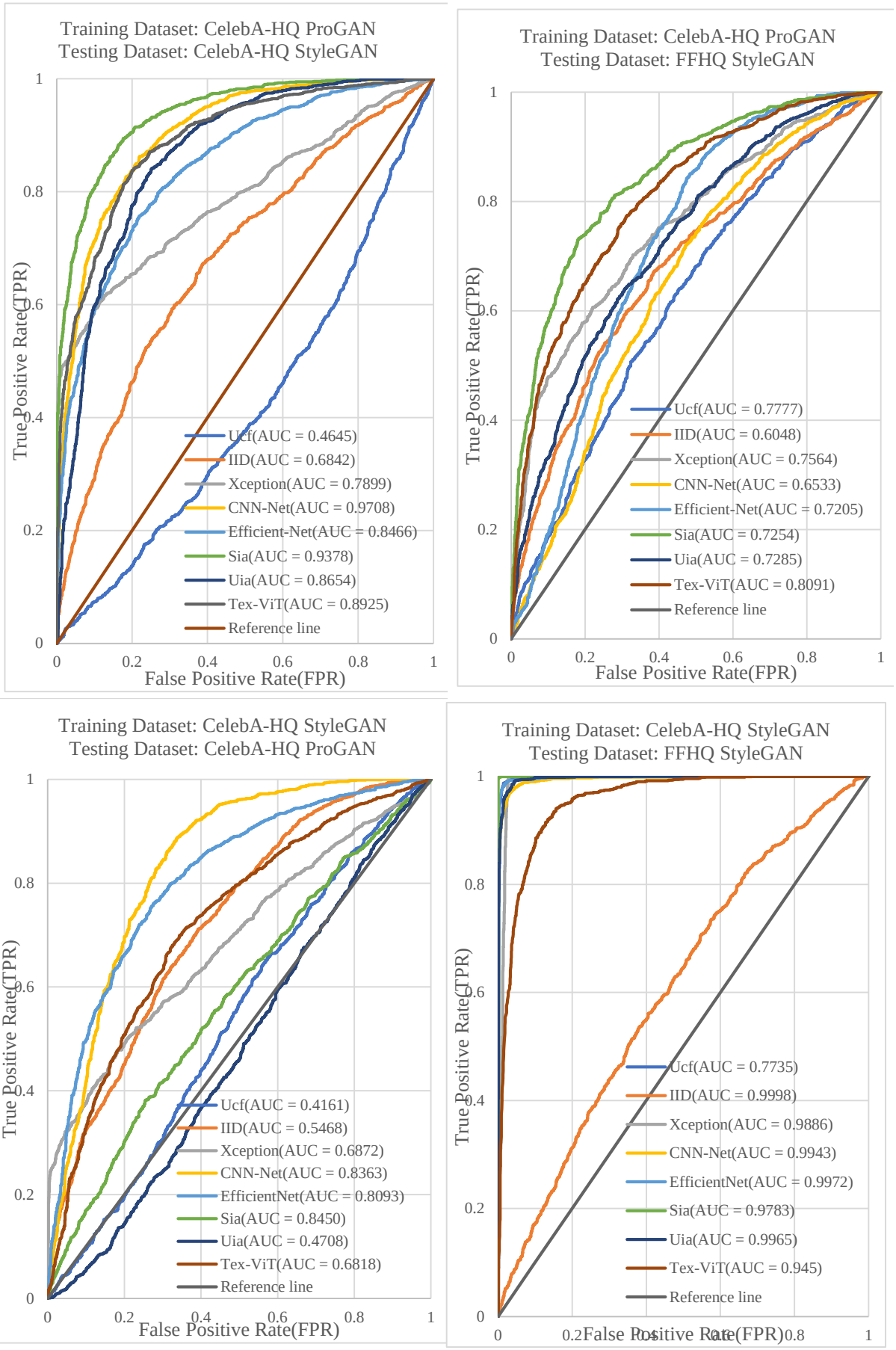


Figure 6: ROC curve for the model trained and tested on the different types of GAN images



Original Image **Blurred Image** **Noisy_Image** **Compressed Image**

Figure 7: Image undergoes different post-processing operations

6 Ablation Studies

The ablation study is performed to confirm the validity of each Tex-ViTcomponent, which is presented in this section. The experiments are performed on different categories of the Faceforensics++ dataset. The components are trained on three categories of FF++ and tested on the fourth category, and the process is done recursively for each category.

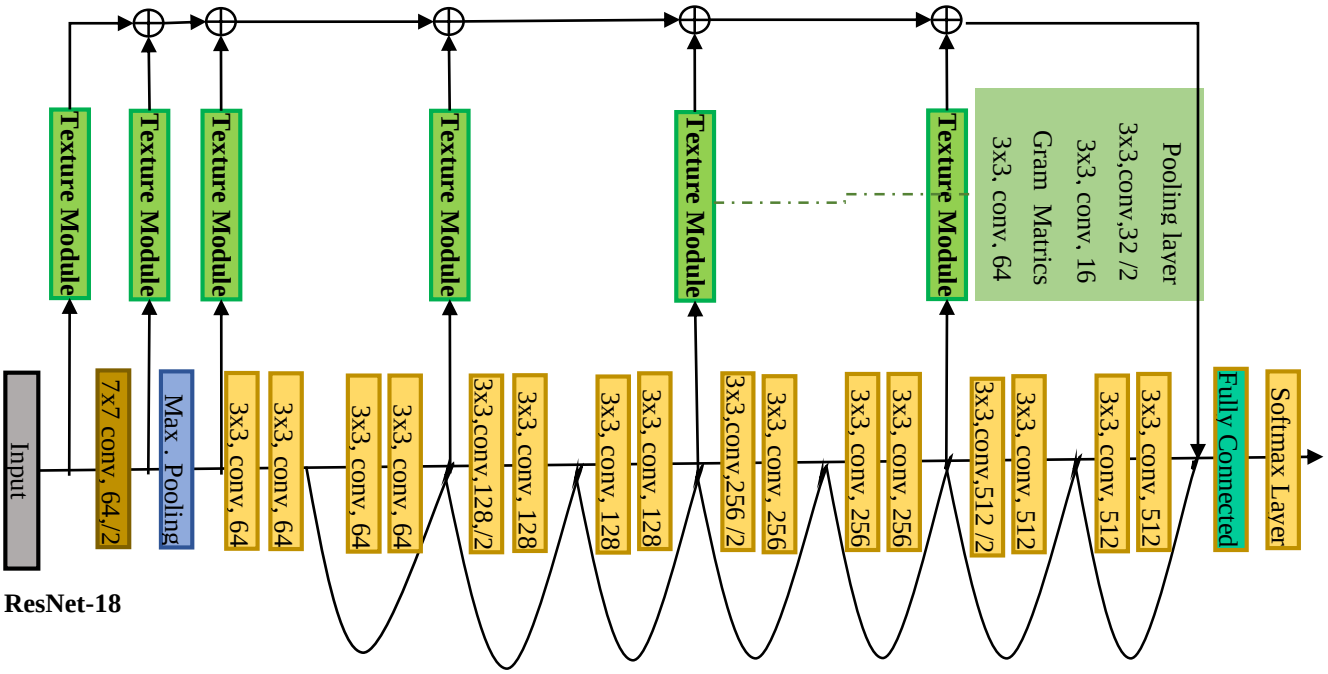


Figure 8: The ResNet18 model design includes a Texture block, in which the characteristics from both branches are combined using concatenation to make the final prediction.

6.1 ResNet with Texture module

This section features ResNet18 architecture as the backbone, and the texture block is computed before every downsampling operation. Furthermore, such texture features keep concatenating further and finally merging with the main branch. Finally, these concatenated features are passed fully connected and sigmoid layer for final prediction. Figure 8 shows the illustration of the component. Table 8 shows the classification results. The average accuracy score varies from 70 to 78%, showing that the texture, even without the attention mechanism, is a powerful feature representing the innate, inherent classification. Again, the component is struggling big-time for the face-swap category during classification, which can be understood as face-swap boundary inconsistencies that have a short range that may not be captured. The model can

effectively capture the discriminative features for other categories, showing that texture is impacted by the diverse manipulation method.

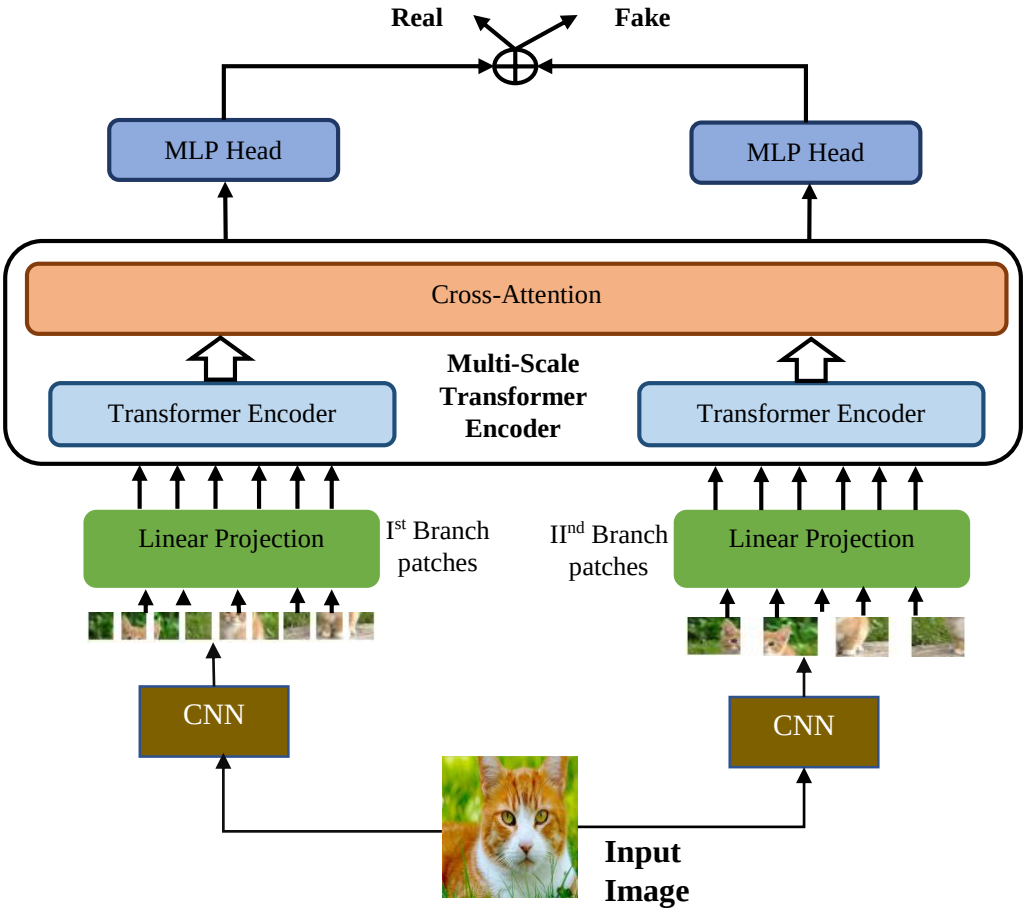


Figure 9: The Dual-branch cross-attention Vision transformer employs a model diagram in which the input image undergoes convolutional neural networks (CNNs) to generate feature maps of varying sizes.

6.2 Dual-branch cross-attention vision transformer

An illustration of this component is exhibited in Figure 9. Here, the input is passed through two CNNs at different branches, which produce feature maps of different sizes, which are subsequently passed to the transformer encoder for a cross-attention mechanism. Then, these features are passed to a multi-layer perceptron, and results are concatenated for final prediction. Table 8 shows the classification result of this component. Accuracy results range from 62 to 70%, which are below par compared to the former component, showing that conventional CNN features with cross-attention are insufficient to classify diverse manipulated images. For face swap samples, the score is more than the previous component, indicating that the attention mechanism has a significant role in places where texture features are not discriminative. The bottom line for this component is that a transformer with cross-attention is not enough to capture the significant latent inherent features that persist among various diverse manipulated samples.

6.3 Tex-ViT(ResNet+Texture Block + Cross-Attention Transformer)

Model, which is the combination of ResNet with texture block along with cross-attention transformer, has been proposed for this paper. As seen from the above component analysis,

texture could be a potential clue for deepfake detection that persists among various kinds of deepfake manipulation. At the same time, a cross-attention vision transformer, free of inductive biases, focuses on the specifics of the manipulation. The texture is computed and concatenated with conventional CNN features, and such features are further amplified, focussed, and augmented by the cross-attention mechanism, leading to a powerful feature representation and a better performance. The table clearly demonstrates that this potent combination results in an accuracy of eighty percent for Deepfake and Face2face samples, more than seventy for NT, and over sixty percent for face-swap. This confirms that the model is capable of acquiring distinguishing characteristics that are consistent across different types of samples.

Table 8 shows the classification results of the different components. Components are trained on different categories of FF++ and tested on the fourth category recursively, where Case A = F2F+FS+NT, Case B = DF+FS+NT, Case C = DF+F2F+NT, Case D= DF+F2F+FS

Train	Test	Precision	Recall	F1 Score	AUC	Acc
ResNet with Texture block						
Case A	DF	0.83	0.7045	0.7624	0.8784	0.78
Case B	F2F	0.882	0.7625	0.8179	0.9207	0.83
Case C	FS	0.4646	0.151	0.2279	0.5686	0.48
Case D	NT	0.7958	0.569	0.6635	0.8091	0.71
Dual-branch transformer augmented with Cross-Attention Mechanism						
Case A	DF	0.6578	0.7748	0.7115	0.7759	0.7001
Case B	F2F	0.6784	0.6956	0.6869	0.7489	0.6845
Case C	FS	0.6512	0.5512	0.5974	0.6748	0.6306
Case D	NT	0.7002	0.6785	0.68917	0.7458	0.6978
Tex-ViT (Our final model)						
Case A	DF	0.7901	0.949	0.8622	0.9275	0.8485
Case B	F2F	0.8364	0.8565	0.8463	0.9169	0.8463
Case C	FS	0.5773	0.543	0.5596	0.5900	0.57275
Case D	NT	0.7739	0.777	0.754	0.8600	0.775

Table 9 Tex-ViT's computational complexity is compared to well-known computer vision models. The column labeled "parameter" represents the number of trainable parameters. The accuracy score, number of trainable parameters, and GPU-CPU times taken for each input batch size are also included in the table.

Model	Accuracy	Parameter (millions)	CPU time(sec)	GPU time(sec)
CNN-Net	52.47	23.51	1.99s	0.379s
ResNext	79.55	81.41	4.61s	0.338s
CapsuleNet	60.56	1.5	2.49s	0.154
EfficientNetV2	80.5	118.51	3.93s	0.40s
XceptionNet	71.6	20.81	2.45s	0.66s
MesoInception	62.3	0.028	2.1s	0.72s
Swin Transformer	75.4	59.96	4.92s	0.41s
ResNet152	79.3	58	3.03s	0.309s
ConvNext	83.39	88.57	3.62	0.37s
Efficient_b7	75.49	66.34	4.07s	0.35s
Tex-ViT	**84.85**	**43**	**3.06s**	**1.02s**

7 Complexity Analysis of Tex-ViT

This section examines the computational complexity of the Tex-ViT architecture in comparison to established computer vision models. The computational factors to be taken into account are the quantity of trainable parameters, the level of precision achieved on the DF (FF++) dataset, and the inference times on both CPU and GPU.

Table 9 presents the tabular view of the complexity analysis of Tex-ViT against the various standard model of computer vision models. It can be easily seen that Tex-ViT is a reasonably light-weight model compared to various computer vision models, including EfficientNetv1, and has also achieved more excellent performance in diverse scenarios. Figure 10 represents the accuracy score on the y-axis and the number of trainable parameters on the x-axis. Regarding the number of trainable parameters, Efficientv2 and ConvNext are on the higher sides and have an accuracy score of around 80%; owning a higher number of parameters tends

to learn more discriminative information about the manipulation. Deepfake Detection models like MesoInceptionNet, Capsule Net, and CNN-Net are relatively light-weight, but their performance suffers in diverse scenarios. The execution time is measured in seconds for a batch size of 32 during inference time. Heavy models take more execution time than light-weight models due to their number of parameters. Xception is a light-weight model with 20 million parameters and an accuracy score of 70%.

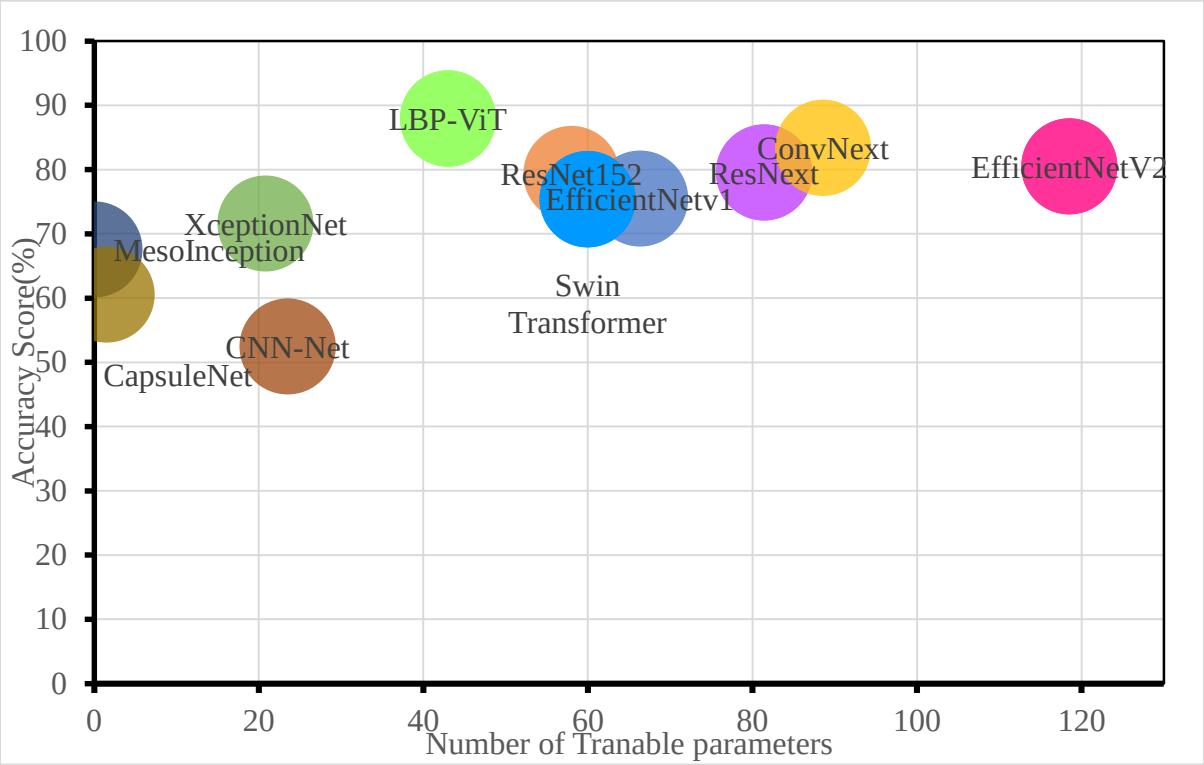


Figure 10 Tex-ViT's complexity analysis in comparison to well-known computer vision models.

8 Visualization results of the Tex-ViT's predictions

This section graphically shows the region of focus for the Tex-ViT prediction. Images are considered from every dataset, the model is trained, and then during prediction, GradCAM class activation maps are used to produce the region of focus for the classifier. Figure 11 shows the region of focus for the model's prediction. The model focuses on different regions of texture to produce the visualization results.



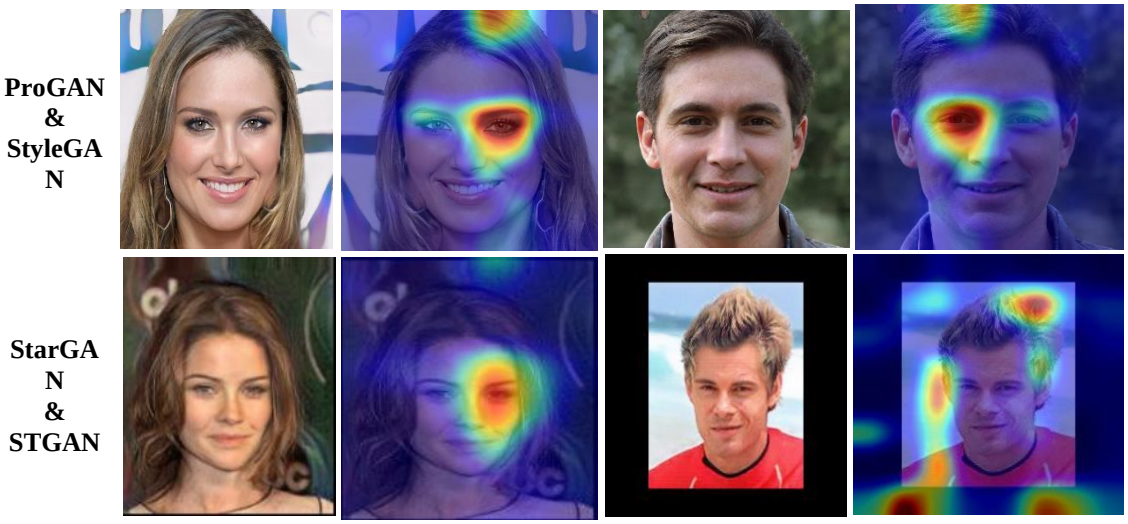


Figure 11: Tex-ViT's model's region of focus of feature for various datasets

9 Conclusion

In this paper, empirical analysis has been done on the human visual and CNN results to demonstrate that human vision is based on the shapes and CNN layers that use texture to identify objects. This finding indicates that texture is one crucial indicator for deepfake detection. Furthermore, it has been seen that texture correlation is not preserved in the case of fake images and that images tend to have smoother surfaces. Inspired by that, the proposed Tex-ViT uses the conventional CNN features using ResNet and texture modules using the features of ResNet. Then, the output of these two parallel branches serves as an input to the dual-branch vision transformers operating on patches with the cross-attention mechanism. Experimental results show that the model performs well in the cross-domain setting of FF++ datasets and different types of GAN image datasets and outperforms the other state-of-the-art models. Experimentation is done on various post-processing image scenarios, and it was found that the model is robust enough for different adversarial attacks. These experiments show that the model learns the common discriminative features that persist along several fake images. One limitation of the method is that it does not perform well in cross-domain settings when testing images from the faceswap category, and this could be future work, improving the model's accuracy on the Faceswap category of FF++. The evaluation shows that texture is an invariant feature that persists among various manipulation methods, and learning such a feature would eventually result in a good performance for the model. However, the model has a low score in the FS category of FF++. Experimentation is also done on the different types of GAN image datasets and outperforms the other state-of-the-art models. It again shows the model's superior learning abilities for different feature spaces. Experimentation is done on various post-processing image scenarios, and it was found that the model is robust enough for different adversarial operations. However, the model needs to improve its score for compressed scenarios, but its score is still better than the other SoTA models. These experiments show that the model learns the common discriminative features that persist along several fake images.

Future work would involve improving the model’s accuracy for the FS manipulation category of FF++, which could incorporate additional modules to enhance the learning of features. Also, improving the model's robustness against the compressed data samples would be one of the futuristic works.

10 Author's contributions, compliance with Ethical standards, declaration of competing interests, availability and informed consent of data.

Author's contribution

Deepak Dagar: Software, Validation, Investigation, Data Curation, Writing – Original Draft, Visualization.

Dinesh Kumar Vishwakarma: Conceptualization, Methodology, Formal Analysis, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Compliance with Ethical Standards

The authors affirm that the manuscript has adhered to the ethical standards specified by the Journal of Information Security and Applications.

Competing interests

The authors affirm that they do not possess any identifiable conflicting financial or non-financial interests or personal ties that could have potentially influenced the findings presented in the research.

Research Data Policy and Data Availability Statements

Given that the dataset and its corresponding code are open-source and accessible to the public. Therefore, it is unnecessary to acquire informed consent from the pertinent stakeholders before use the dataset.

11 References

[1] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: generation, detection, and applications," *International Journal of Multimedia Information Retrieval,* vol. 11, pp. 219-289, 2022.

[2] "All you need to know about deepfakes and why they are dangerous," [Online]. Available: https://www.zee5.com/articles/all-you-need-to-know-about-deepfakes-and-why-they-are-dangerous.

[3] S. McCloskey and . M. Albright, "Detecting GAN-generated Imagery using Color Cues," in *arXiv:1812.08247v1*, 2018.

[4] N. Yu, L. Davis and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 2019.

[5] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen and B. Guo, "Face X-ray for More General Face Forgery Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, 2020.

[6] F. Matern, C. Riess and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa, 2019.

[7] X. Yang, Y. Li and S. Lyu, “Exposing Deep Fakes Using Inconsistent Head Poses,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, 2019.

[8] Y. Li, M.-C. Chang and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” in *IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, 2018.

[9] D. Güera and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks,” in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, 2018.

[10] I. Amerini, L. Galteri, R. Caldelli and A. D. Bimbo, “Deepfake Video Detection through Optical Flow based CNN,” in *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, 2019.

[11] S. Agarwal, H. Farid, O. Fried and M. Agrawala, “Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, 2020.

[12] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera and D. Manocha, “Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues,” in *ACM International Conference on Multimedia*, New York, 2020.

[13] H. Khalid and S. S. Woo, “OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, 2020.

[14] X. Xuan, B. Peng, W. Wang and J. Dong, “On the Generalization of GAN Image Forensics,” in *Chinese Conference on Biometric Recognition*, Zhuzhou, 2019.

[15] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma and Y. Liu, “Countering Malicious DeepFakes: Survey, Battleground, and Horizon,” in *arXiv:2103.00218v1*, 2021.

[16] Z. Liu, X. Qi and H. S. P. Torr, “Global Texture Enhancement for Fake Face Detection In the Wild,” in *Conference on Computer Vision and Pattern Recognition(CVPR)*, Virtual, 2020.

[17] M. Bonomi, C. Pasquini and G. Boato, “Dynamic texture analysis for detecting fake faces in video sequences,” *Journal of Visual Communication and Image Representation*, vol. 79, 2021.

[18] J. Yang, S. Xiao, A. Li, . W. Lu, X. Gao and Y. Li, “MSTA-Net: Forgery Detection by Generating Manipulation Trace Based on Multi-scale Self-texture Attention,” *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, vol. 32, no. 7, pp. 4854 - 4866, 2021.

[19] I. 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*, Montreal Canada, 2014.

[20] "FaceSwap," [Online]. Available: https://faceswap.dev/. [Accessed August 2024].

[21] "FakeApp," [Online]. Available: https://www.malavida.com/en/soft/fakeapp/. [Accessed August 20245].

[22] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

[23] L. Li, J. Bao, H. Yang, D. Chen and F. Wen, "Advancing High Fidelity Identity Swapping for Forgery Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, 2020.

[24] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[25] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo and S. Wen, "STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[26] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang and Y. Liu, "FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[27] M. Barni, K. Kallas, E. Nowroozi and B. Tondi, "CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, New York, 2020.

[28] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury and B. S. Manjunath, "Detecting GAN generated Fake Images using Co-occurrence Matrices," 2019.

[29] A. Jevnisek and S. Avidan, "Aggregating Layers for Deepfake Detection," in *International Conference on Pattern Recognition (ICPR)*, Montreal, QC, Canada, 2022.

[30] Y. Lin, H. Chen, B. Li and J. Wu, "Towards Generalizable DEEPFAKE Face Forgery Detection with Semi-Supervised Learning and Knowledge Distillation," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022.

[31] H. Chen, Y. Lin, B. Li and S. Tan, "Learning Features of Intra-Consistency and Inter-Diversity: Keys Toward Generalizable Deepfake Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 11468-1480, 2023.

[32] Z. Hou, Z. Hua, K. Zhang and Y. Zhang, “CDNet: Cluster Decision for Deepfake Detection Generalization,” in *IEEE International Conference on Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 2023.

[33] W. Guan, W. Wang, J. Dong and B. Peng, “Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization,” *IEEE Transactions on Information Forensics and Security,* vol. 19, 2024.

[34] Z. Guo, L. Wang, W. Yang, G. Yang and K. Li, “LDFnet: Lightweight Dynamic Fusion Network for Face Forgery Detection by Integrating Local Artifacts and Global Texture Information,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 34, no. 2, pp. 1255 - 1265, 2024.

[35] A. V. Nadimpalli and A. Rattani, “On Improving Cross-dataset Generalization of Deepfake Detectors,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 2022.

[36] Z. Yan, Y. Zhang, Y. Fan and B. Wu, “UCF: Uncovering Common Features for Generalizable Deepfake Detection,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023.

[37] P. Yu, J. Fei, X. Zhihua, Z. Zhili and J. Weng, “Improving Generalization by Commonality Learning in Face Forgery Detection,” *IEEE Transactions on Information Forensics and Security,* vol. 17, pp. 547-558, 2022.

[38] M. Atamna, I. Tkachenko and S. Miguet, “Improving Generalization in Facial Manipulation Detection Using Image Noise Residuals and Temporal Features,” in *IEEE International Conference on Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 2023.

[39] J. Li, Z. Yu, G. Luo and Y. Zhu, “CodeDetector: Revealing Forgery Traces with Codebook for Generalized Deepfake Detection,” in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 2024.

[40] J. Yang, S. Xiao, A. Li , G. Lan and H. Wang, “Detecting fake images by identifying potential texture difference,” *Future Generation Computer Systems,* vol. 125, pp. 127-135, 2021.

[41] J. Yang , A. Li, S. Xiao, W. Lu and X. Gao, “MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference,” *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY,* vol. 16, pp. 4234 - 4245, 2021.

[42] A. Yadav and D. K. Vishwakarma, “MRT-Net: Auto-adaptive weighting of manipulation residuals and texture clues for face manipulation detection,” *Expert Systems with Applications,* vol. 232, 2023.

[43] J. Gao, M. Micheletto , G. Orrù, S. Concas , X. Feng, G. L. Marcialis and F. Roli, “Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection,” *Engineering Application of Artificial Intelligence,* vol. 133, 2024.

[44] J. Yang, S. Xiao, A. Li, . W. Lu, X. Gao and Y. Li, "MSTA-Net: Forgery Detection by Generating Manipulation Trace Based on Multi-scale Self-texture Attention," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY,* vol. 32, no. 7, pp. 4854 - 4866, 2022.

[45] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *International Conference on Learning Representations (ICLR),* New Orleans, 2019.

[46] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision*, Zurich, 2014.

[47] H. Lee, R. Grosse, R. Ranganath and A. Y. Ng , "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, 2009.

[48] D. Dai, Y. Li, Y. Wang, H. Bao and G. Wang, "Rethinking the image feature biases exhibited by deep convolutional neural network models in image recognition," *CAAI Transactions on Intelligence Technology,* vol. 7, no. 4, pp. 721-731, 2022.

[49] W. Luo, Y. Li, R. Urtasun and R. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, 2016.

[50] L. A. Gatys, A. S. Ecker and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Quebec, Canada, 2015.

[51] C.-F. (. Chen, Q. Fan and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," in *Computer Vision and Pattern Recognition*, Nashville, Tennessee, 2021.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, . X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *2021*, In International Conference on Learning Representations.

[53] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[54] E. D. Cubuk, B. Zoph, J. Shlens and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, 2020.

[55] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," in *IEEE/CVF International Conference on Computer Vision (ICCV),* Seoul, Korea, 2019.

[56] H. Zhang, . M. Cisse, Y. N. Dauphin and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in *In International Conference on Learning Representations*, 2018.

[57] Z. Zhong, L. Zheng, G. Kang, S. Li and Y. Yang, “Random Erasing Data Augmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence,*, 2020.

[58] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” in *IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong, China, 2019.

[59] H. H. Nguyen, J. Yamagishi and I. Echizen, “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[60] D. Coccomini, . N. Messina, . C. Gennaro and F. Falchi, “Combining EfficientNet and Vision Transformers for Video Deepfake Detection,” in *International Conference on Image Analysis and Processing*, 2022.

[61] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang and D. Ye, “Implicit Identity Driven Deepfake Face Swapping Detection,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023.

[62] F. Chollet, “Xception: Deep Learning with Depth-wise Separable Convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[63] S.-Y. Wang, O. Wang, R. Zhang, A. Owens and A. A. Efros, “CNN-Generated Images Are Surprisingly Easy to Spot… for Now,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[64] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *International Conference on Machine Learning*, 2019.

2024 Asia Pacific Conference on Innovation in Technology (APCIT)
Mysuru, India. Jul 26-27, 2024

The Effectiveness of Advance Deep Learning Architectures for Classification of Stress using Raw EEG Data

1st Ayush Tibrewal
Department of Applied Physics
Delhi Technological University
Delhi, India
ayushtibrewal2004@gmail.com

2nd Shikha
Department of Computer
Science and Engineering
Delhi Technological University
Delhi, India
shikha.9126@gmail.com

3rd Divyashikha Sethia
Department of Software Engineering
Delhi Technological University
Delhi, India
divyashikha@dtu.ac.in

2024 Asia Pacific Conference on Innovation in Technology (APCIT) | 979-8-3503-6153-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/APCIT62007.2024.10673521

***Abstract*—Understanding and detecting stress in today's fast-paced world remains challenging. Previous research has explored various physiological signals, including EEG (Electroencephalography), to quantify stress. However, accurately assessing stress poses a significant challenge, particularly in preprocessing and feature extraction of EEG signals. Hence, this paper employs advanced deep learning algorithms, including ResNet18-1D, DenseNet-1D, and VGG16-1D, on raw EEG signals for automated stress classification. The study utilizes the publicly available EEG-based SAM40 dataset for stress classification and further validates the results on the EEG-based MAT dataset. The results demonstrate the efficacy of the ResNet18-1D model by achieving the highest of 98.73% and 97.24% on the SAM40 and MAT datasets, respectively. The findings pave the way for further exploration and utilization of raw EEG signals across diverse research domains beyond stress classification.**

***Index Terms*—Raw EEG, DeepLearning, Stress Classification, VGG16, ResNet18, DenseNet**

I. Introduction

Stress has become a common aspect of our daily lives, affecting individuals across various ages. Stress is not merely a physiological response to any challenging situation but is linked to how individuals perceive situations cognitively [1]. Kemeny et al. [2] discuss the impact of stressful life experiences on various physiological systems, such as the autonomic nervous system and the immune system. It introduces how a person sees and feels about stressful situations, which can lead to different emotional and physical responses. Stress measurement commonly utilizes Electroencephalography (EEG) signals, which record the brain's electrical activity. This technique involves placing electrodes on the scalp to measure the electrical signals generated by the brain [3]. The signals obtained are then analyzed to determine an individual's stress levels. This approach provides a direct method of assessing psychological states, making EEG a valuable tool.

Researchers preprocess EEG data to extract relevant features in various studies before applying machine learning models [4] [5]. These preprocessing steps often involve feature extraction algorithms along with filtering and artifact removal according to the specific characteristics of EEG data. While this traditional approach has produced promising results, it typically requires domain expertise and manual intervention to design and select appropriate features, which can be time-consuming and may introduce subjectivity.

With the advancement in deep learning [6], researchers expect classifiers to autonomously learn required features rather than relying on hand-engineered features. Nisar et al. [7] introduce a deep learning model by modifying the VGG and ResNet models for emotion classification. They find that restructuring ResNet18 to optimize EEG signal processing significantly improves accuracy in emotion classification and reduces model parameters. By investigating the sequence of convolutional dimension arrangement, they identify the importance of temporal-dimension convolution before spatial-dimension convolution for achieving optimal performance. This finding motivates further exploration, particularly in stress classification.

This paper explores deep learning models like ResNet18, VGG16, and DenseNet, and the goal of this work is to induce architectural changes in the original image-oriented models and transform them into ResNet18-1D, DenseNet-1D, and VGG16-1D, which have improved raw EEG signal performance significantly for stress classification. The approach in the following study involves feeding the raw EEG data directly into advanced deep-learning models. Using this method, the models autonomously recognize meaningful patterns for stress classification without relying on manually crafted features. This study aims to show how deep learning models optimize raw EEG data, opening the door to a more straightforward and practical stress classification approach.

The study utilizes two publicly available EEG-based raw datasets, SAM40 [8] and MAT [9]. Notably, the results show that the given models, when fed with raw EEG signals, demonstrate a remarkable ability to identify moments of stress accurately, which makes the stress-spotting process more efficient and suggests exciting possibilities for advancing the understanding of stress and its impact on one's well-being.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:50 UTC from IEEE Xplore. Restrictions apply.

A. Contributions

This paper's contributions are:

1) Advanced deep learning architectures ResNet18-1D, VGG16-1D, and DenseNet-1D classify raw EEG data without requiring feature extraction methods.
2) To compare advanced deep learning models using the SAM40 [8] dataset and then verify the results on the MAT [9] dataset for stress classification.

The remaining paper reads as follows: Section II covers previous studies in this area, Section III explains datasets and the development of models Additionally, Section IV discusses the results of the entire study. The paper concludes with the conclusion in Section V.

II. Related Work

This section discusses the relevant literature that forms the basis for the research on stress classification. Exploration of these studies aims to understand the methods and discoveries influencing stress classification, as highlighted in Table I.

Marthinsen et al. [10] discovered that using machine learning classifiers K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) on electroencephalography (EEG) data can successfully identify mental stress. The authors employ Hjorth parameters and a two-times ICA-filtered data set that results in an accuracy of 96.50%. These findings imply that time-domain features, Hjorth parameters, and filtered data all help the classifiers perform better. The study mentioned that by reducing the channels to 8, the Accuracy only decreased by 6.22%.

Ahammed et al. [11] used 36 participants in cognitive tasks to test their approach. The findings demonstrated notable differences in EEG signal complexity between task and resting states and between brain areas and activities. This result implies that developing a stress detection system using this technology could be done practically. They also found a link between higher stress levels and improved task performance. They also achieved good results with 80% sensitivity, 100% specificity, and 90% accuracy in classifying between resting and task states using SVM.

In this study by Nirde et al. [12], 36 patients' EEG signals were recorded using a mathematical arithmetic methodology in three distinct states: no task, normal task, and difficult task. They gathered signals from 21 brain channels representing different brain areas. Further, they employed three machine learning classifiers for the EEG dataset: CNN, SVM, and Decision Tree. The features included in the dataset were kurtosis, standard deviation, skewness, and variance. The authors also use models like ANN and LSTM to analyze the unprocessed EEG data. The deep learning models outperformed the single classifier (SVM) in performance. While ANN outperformed LSTM and machine learning classifiers, achieving the maximum Accuracy of 96.80%, SVM received the best Accuracy of 96.70% and an average of 91.72%.

Ajra et al. [13] presented a shallow neural network that analyzes EEG data using convolutional neural network (CNN) layers. The spectral-temporal features were extracted from EEG signals using the event-related spectral power (ERSP) approach, obtaining an accuracy of 90.68%, a considerable improvement over earlier techniques. This model demonstrated robustness with a standard deviation accuracy of only 2.3%, indicating consistent performance across different trials. The study used freely accessible EEG data from subjects meditating and completing mental math exercises. The classification relied on extracting spectral-temporal properties from EEG data.

The previous papers primarily focus on traditional machine learning classifiers and mainly rely on manually crafted features from the EEG data, potentially overlooking the benefits of deep learning architectures like ResNet18-1D, DenseNet-1D, and VGG16-1D. This study aims to show that advanced deep learning architectures are capable of automatically learning relevant features from the raw EEG data, thereby reducing the need for manual preprocessing and potentially improving efficiency and accuracy. Furthermore, some previous studies utilize smaller or less diverse datasets, limiting the generalizability of their findings. In contrast, this paper addresses this issue by leveraging publicly available EEG datasets like SAM40 [8] and MAT [9], offering more extensive and diverse data for training and validation.

III. Methodology

This section discusses the datasets, the deep learning architectures, and their training for stress classification.

A. Dataset

1) SAM40 Dataset [8]: : It comprises EEG data collected from 40 subjects, with 14 females and 26 males, aged between 18 and 25, with a mean age of 21.5 years. The Emotiv Epoc Flex gel kit has been used to capture 32 channels of EEG data at a sampling frequency of 128 Hz. The EEG recordings have been divided into 25-second non-overlapping epochs. The goal has been to monitor short-term stress levels during activities such as relaxation, solving math problems, finding symmetrical mirror images, and the Stroop color-word test.

2) MAT Dataset [9]: : It comprises an EEG 23-channel system (Ukraine, XAI-MEDICA) recorded EEGs at a sample rate of 500 Hz per channel. Sixty-six volunteers, consisting of 47 women and 19 men, aged between 18 and 26 (average age = 18.6 years), have participated in the study. The researchers classified recordings into two conditions: resting state and mental counting. Every recording comprised artifact-free EEG segments, with 180 seconds for the resting state and 60 seconds for mental counting. The dataset was gathered to study the connection between EEG signals and cognitive activity during the difficult math exercise of serial subtraction.

B. Data Preprocessing

The raw EEG signals from the device contain unwanted noise and artifacts that require cleaning before analysis. The study utilizes preprocessed data provided in the dataset and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:50 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Related Work

Previous Work	Dataset	Subjects	Features Used	Model	Results	Limitations
Marthinsen et al. [10]	SAM40 [8]	40	Time and Frequency domain features	KNN, SVM MLP	KNN - 96.5% SVM - 94.50% MLP - 93.33%	Traditional machine learning relies on handcrafted feature engineering
Ahammed et al. [11]	MAT [9]	32	Entropy features	SVM	SVM - 90%	Accuracy can be improved by applying deep learning model and by applying more machine learning models.
Nirde et al. [12]	MAT [9]	32	Raw Data	ANN LSTM	ANN - 96.80% LSTM - 94%	The study only used one dataset, so the results may not apply to different situations or people.
Ajra et al. [13]	MAT [9]	12	Spectral-Temporal features	SNN with CNN LSTM, GoogleNet ResNet50	CNN - 90.68% GoogLeNet - 73.34% ResNet-50 - 72.77% LSTM - 67.76%	More feature domain can be explaored.

enhances the data by applying a Butterworth filter having a bandpass of 0–45 Hz.

The study normalizes the data using Min-Max scaling, which scales the values within the range of [0,1] to ensure consistent and comparable feature values [14]. Equation 1 defines normalization, where X' stands for normalized EEG data, where x denotes the original EEG data, $\min(x)$ denotes the EEG data's minimum value, and $\max(x)$ for the EEG data's maximum value. One benefit of the Min-Max normalization approach is that it maintains all the relationships within the data, ensuring smooth flow between the lines.

$$\text{X'} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

C. Devlopment of Models

1) VGG16-1D: The architecture of VGG16-inspired [15] model for processing of EEG signals uses 1D convolutional layers instead of the traditional 2D convolutional layers, which is shown in Fig. 1. This VGG16-inspired architecture, VGG16-1D, comprises five convolutional blocks. The initial two blocks each contain two 1D convolutional layers, while the subsequent blocks each consist of three 1D convolutional layers. Within these layers, the number of filters progressively increases, being 64, 128, 256, and 512. The network applies ReLU activation functions consistently and employs a kernel size of 3. The first three blocks use MaxPooling to reduce the dimensions, helping to capture and emphasize the most important features while reducing computational complexity [16]. The last fully connected layers consist of two hidden layers, each with 4096 units; following that, the network's final layer uses a sigmoid activation function that is appropriate for binary classification. The fully connected layers at the network's end incorporate a dropout of 0.5 to avoid overfitting.

2) ResNet18-1D: The original architecture of ResNet18 consists of 17 convolutional layers and one layer of the fully connected network [17]. Fig. 2 and Fig. 3 illustrates the modified ResNet18, referred to as ResNet18-1D with

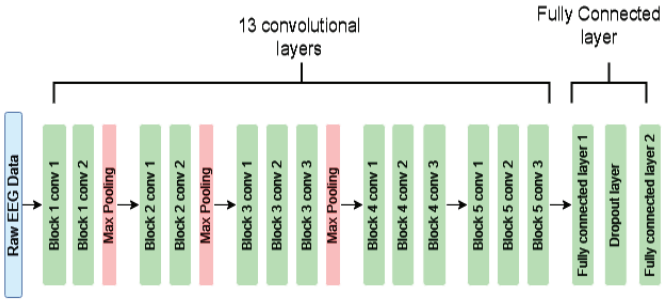


Fig. 1: VGG16-1D Architecture

modifications for a raw EEG classification task, incorporates several changes from the original design. The revised Residual block now incorporates two Conv1D layers, employing rectified linear unit (ReLU) activation, and applying batch normalization following every layer. A shortcut link preserves the input and adds it to the processed information to form the final output, facilitating the training of deep networks by mitigating the vanishing gradient problem. The initial Conv1D layer executes a convolutional operation utilizing 64 filters, with a kernel size of 7 and a stride of 2. It is a crucial layer for extracting basic features, which allows later layers to learn more advanced representations. The residual blocks, which comprise two Conv1D layers apiece, comprise the central component of the ResNet-18 design. The model then uses several residual blocks, each consisting of two convolutional layers designated as Conv Block 2, Conv Block 3, Conv Block 4, and Conv Block 5. The filters rise gradually (64, 128, 256, 512). Adding a dropout layer with a value of 0.5 improves generalization by mitigating overfitting. The model employs a GlobalAveragePooling1D layer to reduce dimensions [18]. The model concludes with a dense output layer comprising a single unit and a sigmoid activation function, making it suitable for binary classification tasks.

3) DenseNet-1D: DenseNet is known for its dense blocks, where every layer gets input from all the previous layers,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:50 UTC from IEEE Xplore. Restrictions apply.

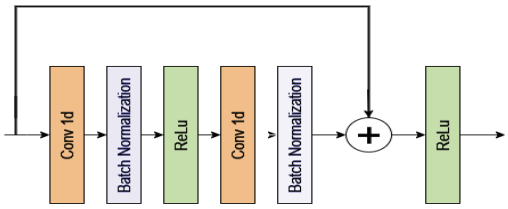


Fig. 2: Residual Block

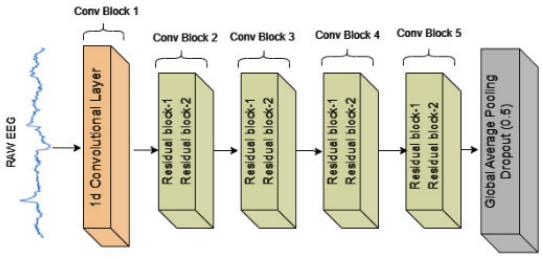


Fig. 3: ResNet18-1D Architecture

encouraging the reuse of features. The model comprises dense blocks scattered with transition blocks to manage the parameters and computational load [19]. The study improvises the DenseNet's performance by employing various modifications, known as DenseNet-1D, shown in Fig. 4. The model begins by extracting initial features using a Conv1D layer with 64 filters, a kernel size of 5, and a stride of 2. DenseNet-1D integrates four Dense Blocks into its architecture, forming the basis of the model. Every layer in a Dense Block gets input from every layer before it, encouraging feature reuse and helping the model pick up complex patterns quickly. The growth rate parameter, set to 32, determines the number of filters added to each layer within a Dense Block. This model captures features quickly with a higher growth rate. To control the model's computational burden, a Transition Block is used after every Dense Block. The 1x1 Conv1D layer, Batch Normalization, and ReLU activation with a compression factor of 0.5. The compression factor helps in reducing the number of filters after each Transition Block. It means that the number of filters halves with a compression value of 0.5. Finally, Global Average Pooling1D applies after the last dense block, resulting in a one-dimensional vector. For binary classification, the final layer is a dense layer with a sigmoid activation function.

D. Model Training

The study uses the SAM40 [8] dataset and further verifies the findings using the MAT [9] dataset. The dataset splits into training, test, and validation with an 80:10:10 ratio. The study selects the Adam optimizer with a learning rate of 0.0001 to optimize gradients [20] efficiently. The loss function is binary cross-entropy, quantifying the difference between real class labels and predicted probabilities. The training data shuffles before each epoch to ensure diverse input sequences and enhance the model's generalization ability. Implementation of early stopping prevents overfitting and enhances training

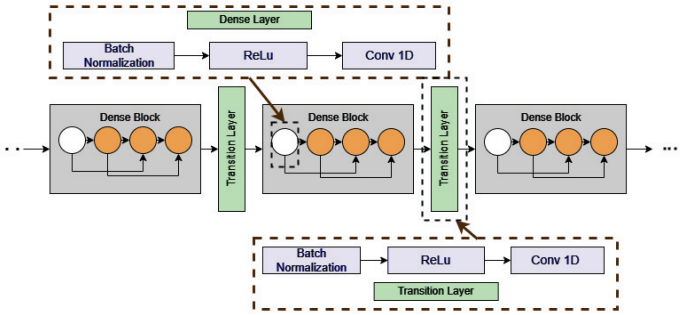


Fig. 4: DenseNet-1D Architecture

efficiency. This mechanism monitors the validation loss and halts training if it fails to decrease for a specific number of epochs. Validation loss is the metric that indicates how effectively the model extrapolates to new data. For the SAM40 dataset [8], the ResNet18-1D model trains for 70 epochs, the VGG16-1D model for 15 epochs, and the DenseNet-1D model for 12 epochs. For the MAT [9] dataset, the ResNet18-1D model trains for 100 epochs, the DenseNet-1D model for 20 epochs, and the VGG16-1D model for 28 epochs.

IV. Result and Discussion

This section discusses the comparative analysis of the classification accuracy of three models, VGG16-1D, ResNet18-1D, and DenseNet-1D, for stress classification. The results indicate that utilizing deep learning models on raw EEG signals provides a reliable approach, which solves the problem of past research requiring feature extraction and data preprocessing before applying the models.

This study utilizes various evaluation metrics for the deep learning-based stress classification models, such as Precision, Recall, Accuracy, and F1-score. Equations 2, 3, 4 and 5 calculate these metrics. The confusion matrix offers a comprehensive breakdown of the model's performance, categorizing predictions into four groups: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). True Positive represents correctly predicted stressed samples, False Positive indicates non-stressed samples wrongly classified as stressed, True Negative signifies correct identification of non-stressed samples, and False Negative denotes stressed samples misclassified as non-stressed. The metrics mentioned above thoroughly assess the model's ability to classify stress accurately. Precision assesses the precision of positive predictions, while Recall tells us how well the model finds all the relevant instances. Accuracy assesses overall correctness, and the F1 score balances Precision and Recall. Examining these metrics alongside the confusion matrix enables comprehension of the model's performance across different class predictions. It facilitates informed decisions regarding its ability to differentiate between stress and non-stress samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:50 UTC from IEEE Xplore. Restrictions apply.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

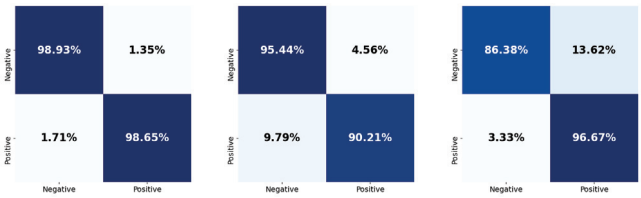
$$\text{F1-Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

Table II shows the accuracy metric on the SAM40 [8] and MAT dataset [9], revealing that the SAM40 dataset [8] ResNet18-1D achieves an impressive accuracy of 98.73%. Further, VGG16-1D obtains an accuracy of 98.51% while DenseNet-1D attains a slightly lower accuracy of 96.67%. Precision, representing the percentage of correctly identified positive cases out of all the cases that were predicted as positive, illustrates ResNet18-1D's dominance with a score of 98.64%, followed by VGG16-1D at 98.44%, and DenseNet-1D at 96.15%. Recall, indicating how many actual positive instances the model can capture or recall, also known as sensitivity, highlights ResNet18-1D's superior performance with a recall of 98.82%, indicating its ability to capture the most positive instances effectively. VGG16-1D achieves a recall of 98.60% and DenseNet-1D attains a recall of 97.27%. F1 scores of ResNet18-1D and VGG16-1D showcase comparable results of 98.73% and 98.52%, respectively, with DenseNet-1D achieving an F1 score of 96.70%. A high F1 score means the model makes few mistakes in identifying positives and negatives.

Similarly, ResNet18-1D performs better on the MAT [9] dataset as well, achieving 97.24% accuracy, 93.97% precision, 95.43% recall, and a 94.69% F1 score. DenseNet-1D follows it with 94.52% accuracy, 92.82% precision, 93.55% recall, and 94.39% F1 score. Subsequently, VGG16-1D follows with 91.54% accuracy, 87.72% precision, 96.67% recall, and 91.98% F1 score. The results show that ResNet18-1D is the most effective network across all metrics in both datasets.

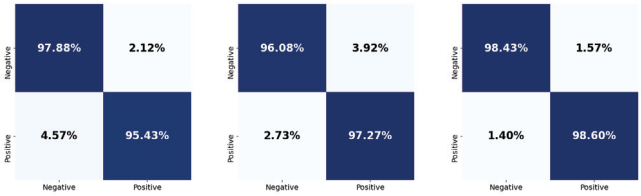
Table II also presents a comparative analysis of several models that utilize stress classification in mental arithmetic tasks. Nirde et al. [12] have employed Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) models, achieving an accuracy of 96.8% with ANN and 94% with LSTM. Ajra et al. [13] have utilized Convolutional Neural Network (CNN) and ResNet-50 models, obtaining an accuracy of 90.68% with CNN and 85.9% with ResNet-50. However, specific precision and recall values have not been provided for these models. In contrast, the ResNet18-1D model proposed in the study attained the highest accuracy of 98.73%, with precision and recall scores matching at 98.73%, resulting in an F1 score of 98.82%. The proposed ResNet18-1D model demonstrates superior performance compared to the other models concerning accuracy and overall performance metrics. It presents a promising solution for the given task.

Fig. 5 and 6 show the confusion matrix of all three models VGG16-1D, ResNet18-1D, DenseNet-1D from which it is clear that the Resnet-1D model, in both datasets, it exhibits a smaller percentage of false positives and false negatives and a higher number of true positives and true negatives, indicating a better degree of accuracy.



(a) ResNet18-1D (b) DenseNet-1D (c) VGG16-1D

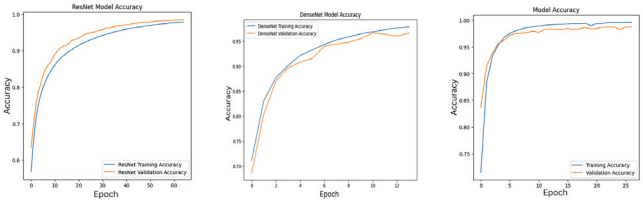
Fig. 5: Confusion Matrix for SAM40 Dataset [8]



(a) ResNet18-1D (b) DenseNet-1D (c) VGG16-1D

Fig. 6: Confusion Matrix for MAT Dataset [9]

Fig. 7 and Fig. 8 show the validation and accuracy curves of the three models, which are crucial for evaluating model performance. The training accuracy curve depicts the model's development overtime on the training data, demonstrating its improvement throughout the training process. In contrast, the validation accuracy curve helps assess its ability to generalize to unseen data and identify potential overfitting issues. It is clear from the graphs that the ResNet-1D model is the most effective, and both curves show a steady increase and remain close to each other, indicating effective learning without overfitting. The model typically exhibits high training and validation accuracy with a minimal gap between the curves.



(a) ResNet18-1D (b) DenseNet-1D (c) VGG16-1D

Fig. 7: Accuracy Graphs for SAM40 Dataset [8]

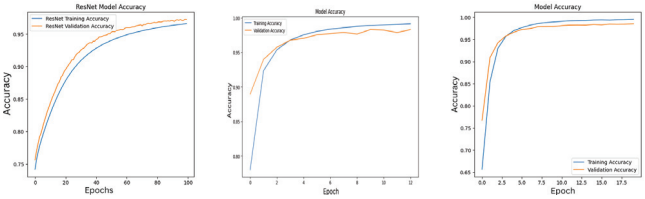
V. Conclusion

This study investigates deep learning models frequently used for image classification to analyze raw EEG data directly. Unlike previous studies that rely on pre-selected features from these signals, the examination focuses on how well the system could learn and classify stress-related patterns directly from the raw EEG data. This study suggests modifying established

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:50 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Results and comparison of MAT [9] and SAM40 [8] Dataset

	Feature	Model	Accuracy	Precision	Recall	F1score
MAT Dataset [9]						
Nirde et al. [12]	Raw Data	ANN	96.8	97	97	97
		LSTM	94	1.00	96	98
Ajra et al. [13]	Spectral temproral	CNN	90.68	-	-	90.21
		Resnet-50	85.9	-	-	84.86
Proposed work	**Raw Data**	**ResNet18-1D**	**97.24**	**93.97**	**95.43**	**94.69**
		DenseNet-1D	**94.52**	**92.82**	**93.55**	**94.39**
		VGG16-1D	**91.54**	**87.72**	**96.67**	**91.98**
SAM40 Dataset [8]						
Proposed work	**Raw Data**	**ResNet18-1D**	**98.73**	**98.64**	**98.82**	**98.73**
		DenseNet-1D	**96.67**	**96.15**	**97.27**	**96.7**
		VGG16-1D	**98.51**	**98.44**	**98.6**	**98.52**



(a) ResNet18-1D (b) DenseNet-1D (c) VGG16-1D

Fig. 8: Accuracy Graphs for MAT dataset [9]

architectures like ResNet18, VGG16, and DenseNet by replacing 2D kernels with 1D kernels. This adjustment allows the networks to effectively handle the one-dimensional nature of EEG signals.

This study tests the models on two publicly available raw EEG datasets, SAM40 [8] and MAT [9]. The models include ResNet-1D, DenseNet-1D, and VGG16-1D. Designing these models ensures they perform well with one-dimensional time series data like EEG signals. After evaluating the models on both datasets, ResNet18-1D consistently outperforms other models in all evaluation metrics. It achieved higher accuracy, with scores of 98.73% for SAM40 [8] and 97.24% for MAT [9] datasets. Its F1 score is also superior at 98.73% for SAM40 [8] and 94.69% for MAT [9]. Regarding precision, ResNet18-1D scores 98.64% for SAM40 [8] and 93.97% for MAT [9], while its recall rates are 98.82% for SAM40 [8] and 95.43% for MAT [9]datasets. ResNet18-1D better classifies stress-related EEG signals than the other proposed models. The success of models like ResNet18-1D, DenseNet-1D, and VGG16-1D in classifying stress-related EEG patterns without the need for extensive feature engineering highlights the potential of these models to simplify the EEG classification process.

REFERENCES

[1] J. M. Koolhaas, A. Bartolomucci, B. Buwalda, S. F. de Boer, G. Flügge, S. M. Korte, P. Meerlo, R. Murison, B. Olivier, P. Palanza *et al.*, "Stress revisited: a critical evaluation of the stress concept," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 5, pp. 1291–1301, 2011.

[2] M. E. Kemeny, "The psychobiology of stress," *Current directions in psychological science*, vol. 12, no. 4, pp. 124–129, 2003.

[3] G. A. Light, L. E. Williams, F. Minow, J. Sprock, A. Rissling, R. Sharp, N. R. Swerdlow, and D. L. Braff, "Electroencephalography (eeg) and event-related potentials (erps) with human participants," *Current protocols in neuroscience*, vol. 52, no. 1, pp. 6–25, 2010.

[4] A. Pedroni, A. Bahreini, and N. Langer, "Automagic: Standardized preprocessing of big eeg data," *NeuroImage*, vol. 200, pp. 460–473, 2019.

[5] A. Shoka, M. Dessouky, A. El-Sherbeny, and A. El-Sayed, "Literature review on eeg preprocessing, feature extraction, and classifications techniques," *Menoufia J. Electron. Eng. Res*, vol. 28, no. 1, pp. 292–299, 2019.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] K. H. Cheah, H. Nisar, V. V. Yap, C.-Y. Lee, G. Sinha *et al.*, "Optimizing residual networks and vgg for classification of eeg signals: Identifying ideal channels for emotion recognition," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[8] R. Ghosh, N. Deb, K. Sengupta, A. Phukan, N. Choudhury, S. Kashyap, S. Phadikar, R. Saha, P. Das, N. Sinha *et al.*, "Sam 40: Dataset of 40 subject eeg recordings to monitor the induced-stress while performing stroop color-word test, arithmetic task, and mirror image recognition task," *Data in Brief*, vol. 40, p. 107772, 2022.

[9] I. Zyma, S. Tukaev, I. Seleznov, K. Kiyono, A. Popov, M. Chernykh, and O. Shpenkov, "Electroencephalograms during mental arithmetic task performance," *Data*, vol. 4, no. 1, p. 14, 2019.

[10] A. J. Y. Marthinsen, "Detection of mental stress from eeg data using artificial intelligence," 2023.

[11] K. Ahammed and M. U. Ahmed, "Quantification of mental stress using complexity analysis of eeg signals," *Biomedical Engineering: Applications, Basis and Communications*, vol. 32, no. 02, p. 2050011, 2020.

[12] K. Nirde, M. Gunda, R. Manthalkar, and S. Gajre, "Eeg mental arithmetic task levels classification using machine learning and deep learning algorithms," in *2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP)*, 2023, pp. 1–8.

[13] Z. Ajra, B. Xu, G. Dray, J. Montmain, and S. Perrey, "Mental arithmetic task classification with convolutional neural network based on spectral-temporal features from eeg," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 52–55.

[14] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagation for classification," *International Journal of Computer Theory and Engineering*, vol. 3, no. 1, pp. 1793–8201, 2011.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] H. Gholamalinezhad and H. Khosravi, "Pooling methods in deep neural networks, a review," *arXiv preprint arXiv:2009.07485*, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:10:50 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384240837

The Evolution of Currency Regimes in Selected Countries from the Americas, Europe, Asia, and Africa: A Comparative Analysis from the Late 19th Century to 2023

Preprint · September 2024

DOI: 10.13140/RG.2.2.32959.93600

CITATIONS
0

READS
20

2 authors, including:



Sanjay Kumar
Dyal Singh College
6 PUBLICATIONS **2** CITATIONS

SEE PROFILE

All content following this page was uploaded by Sanjay Kumar on 22 September 2024.

The user has requested enhancement of the downloaded file.

The Evolution of Currency Regimes in Selected Countries from the Americas, Europe, Asia, and Africa: A Comparative Analysis from the Late 19th Century to 2023

Sanjay Kumar*[1] and Nand Kumar†[2]

[1,2]Department of Humanities, Delhi Technological University, New Delhi.India.

September 22, 2024

*Department of Humanities, Delhi Technological University, New Delhi.India. Dyal Singh College, Delhi University, New Delhi.India.sanjaykumar.eco@dsc.du.ac.in Orcid: http://orcid.org/0000-0003-4613-076X

†Department of Humanities, Delhi Technological University, New Delhi.India.nandkumar@dce.ac.in Orcid: http://orcid.org/0000-0002-5684-6078

Abstract

This paper presents a comparative analysis of the evolution of currency regimes in selected countries from the Americas, Europe, Asia, and Africa from the late 19th century to 2023. Using multiple classification methodologies, the study examines the key trends, patterns, and determinants of currency regime choices across different regions and stages of economic development. The findings reveal a diverse range of currency regime experiences, highlighting the complex interplay of economic, political, and institutional factors in shaping the evolution of exchange rate arrangements over time. The study suggests that the optimal currency regime for a country depends on its specific characteristics, vulnerability to external shocks, and level of economic development. It emphasises the importance of adopting a flexible and adaptive approach to currency regime management, as the appropriate exchange rate arrangement may change in response to evolving domestic and international conditions. The research findings have significant policy implications, providing insights into the factors policymakers should consider when choosing and managing exchange rate arrangements. The study underscores the need for careful assessment of a country's economic structure, political stability, institutional strength, and exposure to external shocks when determining the appropriate currency regime. This paper contributes to the literature by offering a comprehensive and comparative analysis of the historical evolution of currency regimes across a diverse set of countries, shedding light on the complex dynamics that shape exchange rate policy choices and their economic implications.

Keywords: Currency regimes, exchange rate arrangements, economic development, monetary policy, comparative analysis.

JEL Classification: E42, F33, F41, O57.

1 Introduction

Exchange rate regimes play a crucial role in shaping the economic performance and stability of countries around the world. The choice of a currency regime involves complex trade-offs between monetary policy autonomy, exchange rate stability, and the ability to absorb external shocks (Frankel, 1999; Obstfeld and Rogoff, 1995). Over the past century, countries have experimented with a wide range of exchange rate arrangements, from the classical gold standard to the Bretton Woods system of fixed exchange rates and the post-Bretton Woods era of floating rates and various intermediate regimes (Eichengreen, 2019). The evolution of currency regimes has been marked by significant shifts in the global monetary landscape, including the collapse of the gold standard in the interwar period, the rise and fall of the Bretton Woods system, and the emergence of the European Monetary Union (Bordo, 1993a). These changes have been driven by a combination of economic, political, and institutional factors, as countries have sought to adapt their exchange rate policies to changing domestic and international circumstances (Frieden, 2015). This paper aims to provide a comprehensive analysis of the evolution of currency regimes in selected countries from the Americas, Europe, Asia, and Africa over the period from the late 19th century to 2023. The countries included in the study are Brazil, Canada, and Venezuela from the Americas; France, Germany, Italy, Switzerland, and the United Kingdom from Eu-

rope; China, India, Japan, and South Korea from Asia; and South Africa from Africa. The selection of these countries is motivated by several factors. First, they represent a diverse set of economies with varying levels of development, economic structures, and historical backgrounds. This diversity allows for a rich comparative analysis of the factors shaping currency regime choices across different contexts. Second, the countries in the sample have played important roles in the global economy and have experienced significant shifts in their exchange rate policies over time. Examining their experiences can provide valuable insights into the drivers and consequences of currency regime changes. Finally, the inclusion of countries from four different continents ensures a broad geographical coverage and enables an analysis of regional patterns and differences in exchange rate arrangements. The paper contributes to the existing literature on currency regimes in several ways. First, it provides a detailed and up-to-date account of the historical evolution of exchange rate arrangements in a diverse set of countries, extending the analysis to 2023. Second, by using the IMF's Annual Report on Exchange Arrangements and Exchange Restrictions (AREAER) classification system (Fund, 2023) and other sources, the paper offers a consistent and comprehensive framework for comparing currency regimes across countries and over time. Third, the paper situates the analysis of currency regime choices within the broader context of economic, political, and institutional factors, shedding light on the complex interplay of forces shaping exchange rate policies. The rest of the paper is

structured as follows. Section 2 reviews the relevant literature on currency regimes, focusing on the theoretical and empirical work on the determinants and consequences of exchange rate arrangements. Section 3 gives the research objectives of the paper, section 4 describes the data and methodology used in the analysis, including the sources and classification systems employed. Section 5 presents the main findings of the paper, providing a detailed account of the evolution of currency regimes in the selected countries and discussing the key patterns and trends observed along with the triggers for change. Section 6 concludes the paper by summarising the main insights, section 7 gives policy prescriptions and section 8 gives limitations of this paper along with scope for future research.

2 Literature Review

The literature on currency regimes and their evolution is vast and diverse, spanning several decades and covering a wide range of countries and economic contexts. This section provides an overview of the key theoretical and empirical contributions to the study of currency regime choice and its implications for economic performance. One of the earliest and most influential contributions to the literature on currency regimes is the optimal currency area (OCA) theory, developed by Mundell (1961). The OCA theory provides a framework for understanding the conditions under which

countries should adopt fixed or flexible exchange rates, based on factors such as labor mobility, economic openness, and the similarity of economic shocks. McKinnon (1963) and Kenen (1969) further refined the OCA theory, emphasising the importance of economic size, product diversification, and fiscal integration in determining the optimal currency arrangement. The collapse of the Bretton Woods system in the early 1970s led to a proliferation of currency regime choices, ranging from hard pegs to free-floating arrangements. This period also saw the emergence of the "bipolar view" (also known as the "hollowing-out hypothesis"), which argues that countries should either adopt hard pegs or free-floating regimes, as intermediate arrangements are prone to speculative attacks and currency crises (Fischer, 2001; Eichengreen, 1994). Empirical evidence on the bipolar view is mixed, with some studies finding support for the hypothesis (Ghosh et al., 2002; Levy-Yeyati and Sturzenegger, 2005b) and others questioning its validity (Calvo and Reinhart, 2002; Reinhart and Rogoff, 2004). The "fear of floating" phenomenon, identified by Calvo and Reinhart (2002), suggests that many countries that claim to have floating exchange rates actually intervene heavily in foreign exchange markets to manage their currencies. This behavior is attributed to various factors, such as the desire to maintain trade competitiveness, the presence of large foreign currency-denominated liabilities, and the fear of inflationary pressures. Empirical studies have found evidence of fear of floating in both developing and developed countries (Hausmann et al., 2001; Levy-Yeyati and

Sturzenegger, 2005a). The role of institutional factors in shaping currency regime choices has also received attention in the literature. Frieden and Stein (2002) and Bernhard et al. (2002) argue that political economy considerations, such as the interests of key economic sectors and the balance of power between different societal groups, play a crucial role in determining exchange rate policies. Levy-Yeyati et al. (2016) find that countries with more democratic institutions are more likely to adopt flexible exchange rates, as they provide greater transparency and accountability in the conduct of monetary policy. Empirical studies have also examined the economic performance of countries under different currency regimes. Ghosh et al. (1997) find that countries with pegged exchange rates tend to have lower inflation rates but are more susceptible to currency crises. Levy-Yeyati and Sturzenegger (2005a) show that countries with floating exchange rates experience faster economic growth, particularly in the presence of well-developed financial markets. Husain et al. (2005) find that the benefits of different currency regimes vary depending on the level of economic development, with pegged regimes being more suitable for developing countries and floating regimes being more appropriate for advanced economies. Recent studies have focused on the evolution of currency regimes over time and across countries. Ilzetzki et al. (2019) provide a comprehensive database of exchange rate arrangements for 194 countries from 1946 to 2016, using a de facto classification methodology. They find that the global trend towards greater exchange rate flexibility has been uneven, with

many countries exhibiting "fear of floating" behavior. Chinn et al. (2014) analyze the determinants of currency regime choices in a large sample of countries, finding that trade openness, financial development, and institutional quality are key factors in the adoption of flexible exchange rates. The literature on the determinants of currency regime choice has also explored the role of economic and financial integration. Meissner (2003) argue that countries with higher levels of international trade and financial linkages are more likely to adopt fixed exchange rates, as they provide greater stability and predictability in international transactions. Klein (2002) find that countries with more open capital accounts are more likely to adopt floating exchange rates, as they provide greater flexibility in responding to external shocks. The relationship between currency regimes and economic performance has been a subject of debate in the literature. Ghosh et al. (2002) find that pegged exchange rates are associated with lower inflation rates but also with slower economic growth, particularly in developing countries. Levy-Yeyati and Sturzenegger (2005a) argue that floating exchange rates can act as shock absorbers, helping countries to adjust to external imbalances and maintain competitiveness. However, Aghion et al. (2009) find that the impact of exchange rate flexibility on economic growth depends on the level of financial development, with more flexible regimes being beneficial for countries with well-developed financial markets. The choice of currency regime has also been linked to the incidence of currency crises. Kaminsky and Reinhart (1999) find that fixed exchange rates are

more prone to speculative attacks and currency crises, particularly when they are not backed by sufficient foreign exchange reserves or when they are inconsistent with the underlying macroeconomic fundamentals. Obstfeld (1996) argue that the credibility of fixed exchange rates can be undermined by the potential for self-fulfilling currency crises, where investor expectations of a devaluation can lead to a speculative attack that forces the abandonment of the peg. The role of international financial institutions in shaping currency regime choices has also been examined in the literature. Dreher and Walter (2010) find that countries with higher levels of IMF lending are more likely to adopt fixed exchange rates, as they provide a signal of commitment to macroeconomic stability and reform. However, Edwards and Levy-Yeyati (2005) argue that the IMF's advocacy of flexible exchange rates in the aftermath of the Asian financial crisis may have contributed to the "fear of floating" phenomenon, as countries were reluctant to allow their currencies to fluctuate freely. Recent studies have also explored the implications of currency regime choices for the conduct of monetary policy. Obstfeld et al. (2005) argue that countries with fixed exchange rates face a "trilemma" between exchange rate stability, monetary policy autonomy, and financial openness, and that they must choose two out of these three objectives. Aizenman et al. (2016) find that countries with more flexible exchange rates have greater monetary policy autonomy and are better able to respond to external shocks, such as changes in global interest rates or commodity prices.

2.1 Currency Regime Taxonomy

The classification of currency regimes has been a central issue in the literature on international macroeconomics, as it has important implications for understanding the economic performance and policy choices of countries. Over time, various taxonomies have been developed to capture the diversity of exchange rate arrangements, reflecting the evolution of the global monetary system and the increasing complexity of currency regimes. One of the most widely used classification systems is the IMF's de jure taxonomy, which is based on countries' officially declared exchange rate policies. The IMF's system has undergone several revisions since its introduction in the 1970s, with the most recent version distinguishing between four main categories: hard pegs, soft pegs, floating regimes, and a residual category of other managed arrangements (Fund, 2019). Hard pegs include exchange rate arrangements with no separate legal tender, such as currency unions and dollarized economies, as well as currency board arrangements. Soft pegs encompass conventional pegged arrangements, pegged exchange rates within horizontal bands, crawling pegs, stabilized arrangements, and crawl-like arrangements. Floating regimes are divided into two subcategories: floating and free floating. The residual category of other managed arrangements includes all other managed arrangements that do not meet the criteria for the other categories. While the IMF's de jure classification provides a useful starting point, it has been criticized

for not accurately reflecting the actual behavior of exchange rates. In response to this concern, Reinhart and Rogoff (2004) developed an alternative "natural" classification system that aims to capture the true flexibility of exchange rate arrangements. Their taxonomy combines information on official exchange rate policies with data on parallel market exchange rates and the frequency of exchange rate changes. The Reinhart-Rogoff classification system includes fourteen categories, ranging from no separate legal tender to freely floating regimes. The categories are: no separate legal tender, pre-announced peg or currency board arrangement, pre-announced horizontal band that is narrower than or equal to ±2The Reinhart-Rogoff classification has been widely used in empirical research on currency regimes, as it provides a more nuanced picture of the actual behavior of exchange rates. For example, Ilzetzki et al. (2019) use the Reinhart-Rogoff classification to construct a new dataset of exchange rate arrangements covering 194 countries from 1946 to 2016. They find that the global trend towards greater exchange rate flexibility has been uneven, with many countries exhibiting "fear of floating" behavior, even when they are officially classified as having floating regimes. Other researchers have proposed alternative classification systems that seek to capture additional dimensions of currency regime choice. For example, Levy-Yeyati and Sturzenegger (2005b) develop a de facto classification based on the actual behavior of exchange rates and foreign exchange reserves, using cluster analysis to group countries into four categories: fixed, crawling peg, dirty float, and flexible.

Klein (2013) propose a classification system that takes into account both the exchange rate regime and the degree of capital account openness, arguing that the interaction between these two dimensions is crucial for understanding the economic implications of currency regime choices.

3 Research Objectives

- To investigate the evolution of currency regimes in selected countries from the Americas (Brazil, Canada, Venezuela), Europe (France, Germany, Italy, Switzerland, United Kingdom), Asia (China, India, Japan, South Korea), and Africa (South Africa) over a long time period, spanning from the late 19th century to 2023.

- To identify common trends, patterns, and determinants of currency regime choices across different regions and stages of economic development, by examining the experiences of a geographically diverse sample of countries.

- To employ multiple classification methodologies, such as the IMF's de jure and de facto classifications, Reinhart and Rogoff's "Natural Classification," and Levy-Yeyati and Sturzenegger's clustering approach, to provide a nuanced understanding of the factors that shape currency regime choices and their economic

implications.

- To analyze how regional and country-specific characteristics interact with global economic trends to influence currency regime decisions, by focusing on countries from four different continents.

- To provide policy recommendations based on the findings of the study, considering the specific characteristics and challenges faced by countries in different regions and at different stages of economic development.

4 Methodology and Data

This study employs a qualitative and descriptive approach to analyse the evolution of currency regimes in selected countries from the Americas, Europe, Asia, and Africa over the period from the 1930s to 2018. The research relies on a comprehensive dataset of exchange rate arrangements and historical information, compiled from various sources, including the International Monetary Fund (IMF), central banks, and academic publications.

4.1 Currency Regime Classification

To capture the diversity and complexity of exchange rate arrangements across countries and over time, we utilize multiple currency regime classification methodologies:

1. The IMF's de jure classification, which is based on countries' officially declared exchange rate policies (Fund, 2019). This classification categorizes regimes into hard pegs, soft pegs, floating arrangements, and a residual category of other managed arrangements.

2. The IMF's de facto classification, which focuses on the actual behavior of exchange rates and foreign exchange market interventions, rather than solely relying on official declarations (Fund, 2019). This approach accounts for the actual practice of exchange rate management, rather than just the officially stated policy.

3. Reinhart and Rogoff's "Natural Classification," which combines data on official exchange rate policies, parallel market exchange rates, and the frequency of exchange rate changes (Reinhart and Rogoff, 2004). This classification offers a more granular categorization, with fourteen distinct categories ranging from no separate legal tender to freely floating regimes.

4. Levy-Yeyati and Sturzenegger's clustering approach, which employs statistical

techniques to group countries into de facto currency regime categories based on the observed behavior of their exchange rates and foreign exchange reserves (Levy-Yeyati and Sturzenegger, 2005b).

By using these multiple classification methodologies, we aim to provide a comprehensive and nuanced understanding of the currency regime choices made by countries in our sample and to ensure the robustness of our findings to different classification schemes.

4.2 Data Sources and Variables

Our analysis draws on a wide range of data sources and variables to document the historical evolution of currency regimes and the factors influencing regime choices. The main data sources include:

- Official reports and publications from central banks and finance ministries of the selected countries, providing detailed information on the adoption and implementation of exchange rate policies and arrangements over time.
- Historical data on exchange rates, foreign exchange reserves, and other relevant economic indicators, sourced from databases such as the IMF's International Financial Statistics (IFS) and the World Bank's World Development Indicators

(WDI).

- Scholarly publications and historical accounts on the monetary and financial history of the selected countries, offering insights into the political, economic, and institutional context shaping currency regime decisions.

These data sources are used to construct a comprehensive dataset of currency regime classifications, historical exchange rate data, and other relevant variables for the selected countries over the period from the late 19th century to 2023.

4.3 Analytical Approach

To investigate the evolution of currency regimes and the factors influencing regime choices, we employ a descriptive and comparative approach, which involves:

- Tracing the historical changes in exchange rate arrangements adopted by each country over time, using the information gathered from the aforementioned data sources and currency regime classification methodologies.

- Identifying the key political, economic, and institutional factors that have shaped currency regime choices in each country, such as major political events, economic crises, and shifts in the international monetary system.

- Comparing and contrasting the currency regime experiences of countries across different regions and levels of economic development to identify common patterns and divergences in their exchange rate policy choices and outcomes.
- Assessing the implications of currency regime choices for economic performance and stability in each country, drawing on historical evidence and insights from the existing literature.

By employing this descriptive and comparative approach, we aim to provide a rich and contextualised analysis of the historical evolution of currency regimes, the factors shaping regime choices, and their economic consequences across a diverse sample of countries.

5 Research Findings

This section presents the historical evolution of currency regimes in selected countries from the Americas, Europe, Asia, and Africa, spanning from the late 19th century to 2023. The countries included in the analysis are Brazil, Canada, and Venezuela from the Americas; France, Germany, Italy, Switzerland, and the United Kingdom from Europe; China, India, Japan, and South Korea from Asia; and South Africa from Africa. The findings are based on data compiled from various sources and are

organized by region, with tables providing a comprehensive overview of the currency regime history for each country, along with the triggers for change.

Americas

Table 1: Evolution of Currency Regimes in Brazil, 1889-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1889-1930	Gold standard	Franco (2015)	Great Depression
1931-1964	Multiple exchange rates	Abreu (2005)	Economic instability
1964-1968	Crawling peg	Bacha (2003)	Macroeconomic reforms
1968-1986	Managed float	Bacha (2003)	Oil shocks, debt crisis
1986-1989	Crawling peg	Abreu (2005)	Cruzado Plan
1989-1993	Managed float	Abreu (2005)	Hyperinflation
1994-1998	Exchange rate anchor	Bacha (2003)	Real Plan
1999-2023	Dirty float	Ilzetzki et al. (2019); Fund (2023)	Currency crisis

Source: Authors' compilation based on the listed references.

Table 2: Evolution of Currency Regimes in Canada, 1870-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1870-1914	Gold standard	Bordo (1993b)	World War I
1914-1926	Floating exchange rate	Bordo (1993b)	Interwar instability
1926-1931	Gold standard	Bordo (1993b)	Great Depression
1931-1939	Floating exchange rate	Bordo (1993b)	World War II
1939-1950	Fixed exchange rate	Bordo (1993b)	Bretton Woods system
1950-1962	Fixed exchange rate	Bordo (1993b)	Bretton Woods system
1962-1970	Floating exchange rate	Bordo (1993b)	Bretton Woods tensions
1970-2023	Floating exchange rate	Ilzetzki et al. (2019); Fund (2023)	Post-Bretton Woods era

Source: Authors' compilation based on the listed references.

Table 3: Evolution of Currency Regimes in Venezuela, 1880-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1880-1914	Gold standard	Szlajfer (1990)	World War I
1914-1930	Floating exchange rate	Szlajfer (1990)	Great Depression
1930-1960	Fixed exchange rate	Edwards (1991)	Bretton Woods system
1960-1983	Fixed exchange rate with adjustments	Edwards (1991)	Economic challenges
1983-1989	Multiple exchange rates	Edwards (1991)	Economic crisis
1989-1993	Managed float	Edwards (1991)	Economic reforms
1993-2003	Crawling band	Ilzetzki et al. (2019)	Exchange rate policy
2003-2018	Fixed exchange rate	Ilzetzki et al. (2019)	Currency controls
2018-2023	Managed float	Fund (2023)	Economic crisis

Source: Authors' compilation based on the listed references.

Europe

Table 4: Evolution of Currency Regimes in France, 1880-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1880-1914	Gold standard	Bordo (1993c)	World War I
1914-1926	Floating exchange rate	Bordo (1993c)	Interwar instability
1926-1936	Gold standard	Bordo (1993c)	Great Depression
1936-1939	Managed float	Bordo (1993c)	Economic challenges
1939-1944	Exchange controls	Bordo (1993c)	World War II
1944-1948	Fixed exchange rate	Bordo (1993c)	Bretton Woods system
1948-1958	Multiple exchange rates	Bordo (1993c)	Economic recovery
1958-1973	Fixed exchange rate	Bordo (1993c)	Bretton Woods system
1973-1979	Managed float	Bordo (1993c)	Post-Bretton Woods era
1979-1998	European Monetary System	Ilzetzki et al. (2019)	European integration
1999-2023	Euro	Ilzetzki et al. (2019); Fund (2023)	European Monetary Union

Source: Authors' compilation based on the listed references.

Table 5: Evolution of Currency Regimes in Germany, 1880-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1880-1914	Gold standard	Wolf et al. (2008)	World War I
1914-1924	Floating exchange rate	Wolf et al. (2008)	Hyperinflation
1924-1931	Gold standard	Wolf et al. (2008)	Great Depression
1931-1945	Exchange controls	Wolf et al. (2008)	World War II
1945-1948	Fixed exchange rate	Wolf et al. (2008)	Post-war reconstruction
1948-1973	Fixed exchange rate	Wolf et al. (2008)	Bretton Woods system
1973-1979	Managed float	Wolf et al. (2008)	Post-Bretton Woods era
1979-1998	European Monetary System	Ilzetzki et al. (2019)	European integration
1999-2023	Euro	Ilzetzki et al. (2019); Fund (2023)	European Monetary Union

Source: Authors' compilation based on the listed references.

Table 6: Evolution of Currency Regimes in Italy, 1861-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1861-1914	Gold standard	Fratianni and Spinelli (2016)	World War I
1914-1927	Floating exchange rate	Fratianni and Spinelli (2016)	Interwar instability
1927-1936	Gold standard	Fratianni and Spinelli (2016)	Great Depression
1936-1943	Exchange controls	Fratianni and Spinelli (2016)	World War II
1943-1946	Floating exchange rate	Fratianni and Spinelli (2016)	Post-war transition
1946-1971	Fixed exchange rate	Fratianni and Spinelli (2016)	Bretton Woods system
1971-1979	Managed float	Fratianni and Spinelli (2016)	Post-Bretton Woods era
1979-1998	European Monetary System	Ilzetzki et al. (2019)	European integration
1999-2023	Euro	Ilzetzki et al. (2019); Fund (2023)	European Monetary Union

Source: Authors' compilation based on the listed references.

Table 7: Evolution of Currency Regimes in Switzerland, 1850-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1850-1914	Silver standard	Baltensperger (2007)	Adoption of gold standard
1914-1924	Floating exchange rate	Baltensperger (2007)	World War I
1924-1936	Gold standard	Baltensperger (2007)	Great Depression
1936-1973	Fixed exchange rate	Baltensperger (2007)	Bretton Woods system
1973-2000	Managed float	Baltensperger (2007)	Post-Bretton Woods era
2000-2011	Floating exchange rate	Baltensperger (2007); Ilzetzki et al. (2019)	Monetary policy shift
2011-2015	Crawling peg (Euro)	Ilzetzki et al. (2019)	Swiss franc appreciation
2015-2023	Floating exchange rate	Ilzetzki et al. (2019); Fund (2023)	Removal of Euro peg

Source: Authors' compilation based on the listed references.

Table 8: Evolution of Currency Regimes in the United Kingdom, 1816-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1816-1914	Gold standard	Dimsdale (1981)	World War I
1914-1925	Floating exchange rate	Dimsdale (1981)	Interwar instability
1925-1931	Gold standard	Dimsdale (1981)	Great Depression
1931-1939	Managed float	Dimsdale (1981)	Abandonment of gold standard
1939-1945	Exchange controls	Dimsdale (1981)	World War II
1945-1972	Fixed exchange rate	Dimsdale (1981)	Bretton Woods system
1972-1990	Managed float	Dimsdale (1981)	Post-Bretton Woods era
1990-1992	European Exchange Rate Mechanism	Ilzetzki et al. (2019)	European integration
1992-2023	Floating exchange rate	Ilzetzki et al. (2019); Fund (2023)	Exit from ERM

Source: Authors' compilation based on the listed references.

Asia

Table 9: Evolution of Currency Regimes in China, 1949-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1949-1952	Fixed exchange rate	Lu (2004)	Establishment of the People's Republic of China
1952-1973	Fixed exchange rate	Lu (2004)	Central planning
1973-1980	Fixed exchange rate	Lu (2004)	Economic reforms
1981-1984	Dual exchange rate system	Lu (2004)	Partial liberalization
1985-1993	Managed float	Lu (2004)	Further economic reforms
1994-2005	Crawling peg	Ilzetzki et al. (2019)	Unification of exchange rates 2005-2008
Crawling peg	Ilzetzki et al. (2019)	Gradual appreciation	
2008-2010	Stabilized arrangement	Ilzetzki et al. (2019)	Global financial crisis
2010-2015	Crawling peg	Ilzetzki et al. (2019)	Gradual appreciation
2016-2023	Managed float	Ilzetzki et al. (2019); Fund (2023)	Market-oriented reforms

Source: Authors' compilation based on the listed references.

Table 10: Evolution of Currency Regimes in India, 1947-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1947-1966	Fixed exchange rate	Joshi and Little (1998)	Independence, Bretton Woods system
1966-1975	Fixed exchange rate with devaluations	Joshi and Little (1998)	Balance of payments crisis, wars with Pakistan and China
1975-1979	Managed float	Joshi and Little (1998)	Partial liberalisation, oil price shocks
1979-1993	Crawling peg	Patnaik and Pundit (2015)	Managed depreciation, balance of payments crisis
1993-2008	Managed float	Patnaik and Pundit (2015)	Economic liberalisation, capital account convertibility
2008-2023	Floating exchange rate	Ilzetzki et al. (2019); Fund (2023)	Global financial crisis, inflation targeting framework

Source: Authors' compilation based on the listed references.

Table 11: Evolution of Currency Regimes in Japan, 1868-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1868-1897	Silver standard	Ito (2006)	Meiji Restoration
1897-1917	Gold standard	Ito (2006)	Adoption of gold standard
1917-1930	Floating exchange rate	Ito (2006)	World War I
1930-1931	Gold standard	Ito (2006)	Resumption of gold standard
1931-1941	Managed float	Ito (2006)	Abandonment of gold standard
1941-1949	Exchange controls	Ito (2006)	World War II
1949-1971	Fixed exchange rate	Ito (2006)	Bretton Woods system
1971-1973	Managed float	Ito (2006)	Collapse of Bretton Woods
1973-2023	Floating exchange rate	Ito (2006); Ilzetzki et al. (2019); Fund (2023)	Post-Bretton Woods era

Source: Authors' compilation based on the listed references.

Table 12: Evolution of Currency Regimes in South Korea, 1945-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1945-1961	Fixed exchange rate	Nam (1984)	Independence, Korean War
1961-1980	Fixed exchange rate with adjustments	Nam (1984)	Economic development
1980-1990	Multiple exchange rates	Nam (1984); Ilzetzki et al. (2019)	Economic liberalisation
1990-1997	Managed float	Ilzetzki et al. (2019)	Financial market development
1997-1999	Freely falling	Ilzetzki et al. (2019)	Asian financial crisis
1999-2023	Floating exchange rate	Ilzetzki et al. (2019); Fund (2023)	Post-crisis reforms

Source: Authors' compilation based on the listed references.

Africa

Table 13: Evolution of Currency Regimes in South Africa, 1910-2023

Period	Exchange Rate Regime	Source	Triggers for Change
1910-1914	Gold standard	Aron et al. (2000)	Establishment of the Union of South Africa
1914-1928	Floating exchange rate	Aron et al. (2000)	World War I
1928-1932	Gold standard	Aron et al. (2000)	Resumption of gold standard
1932-1939	Fixed exchange rate	Aron et al. (2000)	Abandonment of gold standard
1939-1979	Fixed exchange rate	Aron et al. (2000)	Bretton Woods system
1979-1996	Managed float	Aron et al. (2000)	Balance of payments pressures
1996-2023	Freely floating	Ilzetzki et al. (2019); Fund (2023)	Financial liberalization

Source: Authors' compilation based on the listed references.

5.1 Key Observations and Patterns in the Evolution of Currency Regimes

The data presented in the tables on the historical evolution of currency regimes in the selected countries reveals several notable observations and patterns:

Prevalence of the gold standard in the early periods: Many countries, including Brazil, Canada, France, Germany, Italy, Switzerland, and the United Kingdom, adhered to the gold standard in the late 19th and early 20th centuries. This reflects the widespread adoption of this exchange rate arrangement during that era. Shift towards floating exchange rates: Over time, particularly after the collapse of the Bretton Woods system in the early 1970s, many countries transitioned from fixed exchange rate arrangements to more flexible regimes, such as managed floating or freely floating exchange rates. This trend is evident in countries like Brazil, Canada, Japan, South Korea, and the United Kingdom. Adoption of intermediate regimes: Some countries, such as Brazil, China, and India, have experimented with intermediate exchange rate arrangements, such as crawling pegs or bands, at various points in their history. These regimes offer a balance between exchange rate stability and flexibility. Regional patterns and economic integration: The tables reveal regional patterns in the adoption of currency regimes. For example, many European countries, including France, Germany, and Italy, participated in the

European Monetary System and eventually adopted the euro, reflecting their economic integration and shared policy objectives. Impact of major economic events: The data suggests that major economic events, such as the Great Depression, the collapse of the Bretton Woods system, and the Asian financial crisis, have influenced currency regime choices. Countries have often adapted their exchange rate arrangements in response to these events and the changing global economic environment. Persistence of certain regimes: Some countries have maintained relatively stable currency regimes over long periods. For instance, Canada and Japan have primarily maintained floating exchange rates since the 1970s, while South Africa has had a freely floating regime since 1996.

5.2 Factors Shaping the Evolution of Currency Regimes

The analysis of currency regime choices across the selected countries reveals several key factors that have influenced the evolution of exchange rate arrangements over time. By examining the experiences of a geographically diverse sample of countries and employing multiple classification methodologies, this study provides a nuanced understanding of the determinants of currency regime choices and their economic implications. One of the primary factors shaping currency regime decisions is the stage of economic development. Countries at different levels of development often

face distinct challenges and priorities, which can influence their choice of exchange rate arrangement. For example, developing countries may prioritize exchange rate stability to promote trade and investment, while more advanced economies may place greater emphasis on monetary policy autonomy and the ability to respond to domestic economic conditions. Another important factor is the country's exposure to external shocks, such as fluctuations in global commodity prices, financial crises, or shifts in the monetary policies of major trading partners. Countries that are highly vulnerable to such shocks may opt for more stable exchange rate arrangements to mitigate the impact of external volatility on their domestic economies. Conversely, countries with more diversified economies and greater resilience to external shocks may be better positioned to adopt more flexible exchange rate regimes. The analysis also highlights the role of regional and country-specific characteristics in shaping currency regime choices. Countries within the same region may face similar economic challenges and opportunities, which can lead to the adoption of similar exchange rate arrangements. For instance, many European countries participated in the European Monetary System and eventually adopted the euro, reflecting their shared economic interests and the benefits of monetary integration. Similarly, countries in Asia and Latin America have often experimented with managed floating regimes or crawling pegs, in response to their specific economic circumstances and policy objectives. Global economic trends and

major events, such as the collapse of the Bretton Woods system or the Asian financial crisis, have also played a significant role in shaping currency regime choices. These events have often led to widespread changes in exchange rate arrangements, as countries have sought to adapt to new economic realities and challenges. The study's use of multiple classification methodologies, such as the IMF's de jure and de facto classifications, Reinhart and Rogoff's "Natural Classification," and Levy-Yeyati and Sturzenegger's clustering approach, helps to capture these shifts and provides a more comprehensive picture of the evolution of currency regimes over time. Furthermore, the analysis underscores the importance of institutional factors, such as central bank independence, the quality of governance, and the credibility of monetary and fiscal policies, in shaping the effectiveness and sustainability of different currency regimes. Countries with stronger institutions and more credible policy frameworks may be better able to maintain stable exchange rates or implement more flexible arrangements, depending on their specific economic circumstances and objectives. Overall, the research findings highlight the complex interplay of economic, political, and institutional factors that shape currency regime choices across different regions and stages of economic development. By providing a comprehensive analysis of these factors and their implications, this study contributes to a deeper understanding of the determinants and consequences of exchange rate arrangements in a diverse set of countries over a long historical period.

6 Summary and Conclusion

This study has investigated the evolution of currency regimes in selected countries from the Americas (Brazil, Canada, Venezuela), Europe (France, Germany, Italy, Switzerland, United Kingdom), Asia (China, India, Japan, South Korea), and Africa (South Africa) over a period spanning from the late 19th century to 2023. By examining the experiences of a geographically diverse sample of countries, the study has sought to identify common trends, patterns, and determinants of currency regime choices across different regions and stages of economic development. The analysis has employed multiple classification methodologies, including the IMF's de jure and de facto classifications, Reinhart and Rogoff's "Natural Classification," and Levy-Yeyati and Sturzenegger's clustering approach, to provide a nuanced understanding of the factors shaping currency regime choices and their economic implications. The findings have revealed that countries have adopted a wide range of currency regimes over time, from the classical gold standard to the Bretton Woods system of fixed exchange rates and the post-Bretton Woods era of floating rates and various intermediate regimes.

Ultimately, the choice of currency regime remains a critical decision for countries seeking to promote economic stability, growth, and resilience in an increasingly interconnected global economy. By shedding light on the historical evolution of

exchange rate arrangements and the factors shaping these choices, this study has provided a foundation for informed policymaking and further academic inquiry in this important area of international economics.

7 Policy Prescriptions

The findings of this study offer valuable insights for policymakers seeking to design and manage exchange rate arrangements effectively. While there is no one-size-fits-all approach to currency regime choice, the analysis highlights several key factors that should be considered when determining the most appropriate exchange rate policy for a given country. First, policymakers should carefully assess their country's economic characteristics, including its level of development, economic structure, and degree of integration with the global economy. Countries with more diversified economies, stronger institutions, and deeper financial markets may be better equipped to handle the challenges associated with floating exchange rates. In contrast, countries with less developed economies or significant external vulnerabilities may benefit from the stability provided by fixed or managed exchange rate arrangements, at least in the short to medium term. Second, the choice of currency regime should take into account the country's exposure to external shocks, such as fluctuations in global commodity prices, financial crises, or shifts in the

monetary policies of major trading partners. Countries that are highly vulnerable to such shocks may find that more stable exchange rate arrangements can help to mitigate the impact of external volatility on the domestic economy. However, policymakers should also be aware of the potential drawbacks of fixed exchange rates, such as the loss of monetary policy autonomy and the risk of unsustainable currency misalignments. Third, the study highlights the importance of institutional factors, such as central bank independence and the credibility of monetary and fiscal policies, in shaping the effectiveness of different currency regimes. Countries with stronger institutions and a track record of sound macroeconomic management may be better positioned to maintain flexible exchange rates and weather economic shocks. Policymakers should therefore prioritize reforms aimed at strengthening the institutional framework and enhancing the transparency and credibility of economic policies. Fourth, the analysis suggests that the optimal currency regime may change over time as a country's economic and institutional circumstances evolve. Policymakers should therefore adopt a flexible and adaptive approach to exchange rate management, being prepared to adjust their currency regime in response to changing domestic and international conditions. This may involve moving from fixed to more flexible exchange rate arrangements as a country's economy develops and becomes more resilient to external shocks, or temporarily adopting more stable arrangements during periods of heightened uncertainty or volatility. Finally, while

the choice of currency regime is an important policy decision, it is not a panacea for all economic challenges. Policymakers should recognize that exchange rate policy is just one element of a broader macroeconomic policy framework, and that its effectiveness depends on its coherence with other policies, such as monetary, fiscal, and structural reforms. A holistic approach to economic management, which takes into account the interlinkages between different policy areas and the potential trade-offs involved, is essential for promoting sustainable growth and stability.

8 Limitations and Future Research

While this study provides a comprehensive overview of the evolution of currency regimes in selected countries from the Americas, Europe, Asia, and Africa, it has some limitations that should be acknowledged and addressed in future research. First, the study relies on a relatively small sample of countries from each region, which may not fully capture the diversity of experiences within each continent. Future studies could expand the sample to include more countries and provide a more representative analysis of currency regime choices and their determinants across different regions and subregions. Second, the study primarily focuses on the historical evolution of currency regimes and the triggers for change, without delving deeper into the specific economic, political, and institutional factors that shape

these choices. Future research could conduct more in-depth case studies or employ quantitative methods to investigate the relationships between various economic, political, and institutional variables and currency regime decisions. Third, the study does not extensively examine the economic outcomes associated with different currency regimes, such as economic growth, inflation, trade, and financial stability. Future research could systematically analyze the impact of currency regime choices on these key macroeconomic variables and assess the trade-offs and challenges associated with different exchange rate arrangements. Fourth, the study relies on existing classification schemes, such as the IMF's de jure and de facto classifications, Reinhart and Rogoff's "Natural Classification," and Levy-Yeyati and Sturzenegger's clustering approach. While these methodologies provide valuable insights, they may not fully capture the nuances and complexities of currency regime choices in practice. Future research could refine these classification systems or develop new approaches that better reflect the actual behavior of exchange rates and monetary authorities. Fifth, the study focuses on the period up to 2023, which is the latest year for which comprehensive data on currency regimes is available. However, the global economic landscape is constantly evolving, and new challenges and opportunities may arise in the future that influence currency regime choices. Future research could extend the analysis to more recent periods and explore how countries adapt their exchange rate policies to emerging trends and shocks. Finally, while the

study identifies some common triggers for currency regime changes, such as global economic events and country-specific factors, it does not provide a comprehensive theory or framework for understanding the complex interplay of factors that shape these decisions. Future research could develop more integrated theoretical models that incorporate economic, political, and institutional factors and test their explanatory power using empirical data.

Acknowledgments

The authors would like to acknowledge that no funding was received for this research.

The authors declare that there are no competing interests involved in this study.

References

M. d. P. Abreu. The brazilian economy, 1930–1980. *The Cambridge Economic History of Latin America*, 2:283–394, 2005.

P. Aghion, P. Bacchetta, R. Ranciere, and K. Rogoff. Exchange rate volatility and productivity growth: The role of financial development. *Journal of monetary economics*, 56(4):494–513, 2009.

J. Aizenman, M. D. Chinn, and H. Ito. Managing exchange rate volatility and capital flows in the presence of financial market imperfections. *Handbook of Exchange Rates*, pages 481–506, 2016.

J. Aron, I. Elbadawi, and B. Kahn. The exchange rate in south africa. *South African Journal of Economics*, 68(4):282–311, 2000.

E. L. Bacha. Brazil's plano real: a view from the inside. In *Development Economics and Structuralist Macroeconomics*, pages 181–206. Edward Elgar Publishing, 2003.

E. Baltensperger. The swiss franc: a history of monetary policy. *Swiss National Bank*, pages 1–28, 2007.

W. Bernhard, J. L. Broz, and W. R. Clark. The political economy of monetary institutions. *International Organization*, pages 693–723, 2002.

M. D. Bordo. The bretton woods international monetary system: a historical overview. In *A retrospective on the Bretton Woods system: Lessons for international monetary reform*, pages 3–108. University of Chicago Press, 1993a.

M. D. Bordo. The gold standard, bretton woods and other monetary regimes: a historical appraisal. In *Dimensions of Monetary Policy: Essays in Honour of Anatol B. Balbach*, pages 123–191. Springer, 1993b.

M. D. Bordo. The bretton woods international monetary system: a historical overview. In *A retrospective on the Bretton Woods system: Lessons for international monetary reform*, pages 3–108. University of Chicago Press, 1993c.

G. A. Calvo and C. M. Reinhart. Fear of floating. *The Quarterly Journal of Economics*, 117(2):379–408, 2002.

M. D. Chinn, M. P. Dooley, and S. Shrestha. Once bitten, twice shy? financial crises and the choice of exchange rate regime. *Journal of International Money and Finance*, 52:102–126, 2014.

N. H. Dimsdale. British monetary policy and the exchange rate 1920–1938. *Oxford Economic Papers*, 33:306–349, 1981.

A. Dreher and S. Walter. On the choice of exchange rate regimes: Does politics matter? *Journal of International Economics*, 81(2):204–216, 2010.

S. Edwards. Exchange rate misalignment in developing countries. *The World Bank Research Observer*, 6(1):3–21, 1991.

S. Edwards and E. Levy-Yeyati. Flexible exchange rates as shock absorbers. *European Economic Review*, 49(8):2079–2105, 2005.

B. Eichengreen. *Globalizing capital: a history of the international monetary system*. Princeton University Press, 2019.

B. J. Eichengreen. International monetary arrangements for the 21st century. 1994.

S. Fischer. Exchange rate regimes: Is the bipolar view correct? *Journal of Economic Perspectives*, 15(2):3–24, 2001.

G. H. B. Franco. A framework for analyzing the process of financial integration. *OECD Journal: Financial Market Trends*, 2015(1):1–21, 2015.

J. A. Frankel. No single currency regime is right for all countries or at all times. *Essays in International Finance*, 215:1–49, 1999.

M. Fratianni and F. Spinelli. Italy in the gold standard period, 1861–1914. *A New Economic History of Argentina*, pages 69–98, 2016.

J. Frieden and E. Stein. The political economy of exchange rate policy in latin america: an analytical overview. *The currency game: Exchange rate politics in Latin America*, pages 21–50, 2002.

J. A. Frieden. *Currency politics: The political economy of exchange rate policy*. Princeton University Press, 2015.

I. M. Fund. *Annual Report on Exchange Arrangements and Exchange Restrictions 2019*. International Monetary Fund, 2019.

I. M. Fund. *Annual Report on Exchange Arrangements and Exchange Restrictions 2023*. International Monetary Fund, 2023.

A. R. Ghosh, A.-M. Gulde, J. D. Ostry, and H. C. Wolf. Does the nominal exchange rate regime matter? 1997.

A. R. Ghosh, A.-M. Gulde, and H. C. Wolf. Exchange rate regimes: Choices and consequences. Technical report, International Monetary Fund, 2002.

R. Hausmann, U. Panizza, and E. Stein. Exchange rates and financial fragility. *NBER Working Paper*, (7418), 2001.

A. M. Husain, A. Mody, and K. S. Rogoff. Exchange rate regime durability and performance in developing versus advanced economies. *Journal of Monetary Economics*, 52(1):35–64, 2005.

E. Ilzetzki, C. M. Reinhart, and K. S. Rogoff. Exchange arrangements entering the twenty-first century: Which anchor will hold? *The Quarterly Journal of Economics*, 134(2):599–646, 2019.

T. Ito. Japanese monetary policy: 1998–2005 and beyond. *Monetary Policy in Asia: Approaches and Implementation*, 31:105–132, 2006.

V. Joshi and I. M. D. Little. India's economic reforms: progress, problems, prospects. *Oxford Development Studies*, 26(3):333–351, 1998.

G. L. Kaminsky and C. M. Reinhart. The twin crises: the causes of banking and balance-of-payments problems. *American Economic Review*, 89(3):473–500, 1999.

P. B. Kenen. The theory of optimum currency areas: an eclectic view. *Monetary problems of the international economy*, pages 41–60, 1969.

M. W. Klein. Exchange rate regimes in the western hemisphere: a few comments. *The North American Journal of Economics and Finance*, 13(3):279–284, 2002.

M. W. Klein. Capital controls: Gates versus walls. *Brookings Papers on Economic Activity*, 2012(1):317–367, 2013.

E. Levy-Yeyati and F. Sturzenegger. Flexible exchange rates as shock absorbers. *European Economic Review*, 49(8):2079–2105, 2005a.

E. Levy-Yeyati and F. Sturzenegger. Classifying exchange rate regimes: Deeds vs. words. *European Economic Review*, 49(6):1603–1635, 2005b.

E. Levy-Yeyati, F. Sturzenegger, and P. A. Gluzmann. The political economy of exchange rate policy. *Handbook of exchange rates*, pages 547–577, 2016.

Q. Lu. Exchange rate regime choices of china: 1949–present. *Journal of World Trade*, 38(6):1057–1075, 2004.

R. I. McKinnon. Optimum currency areas. *The American economic review*, 53(4): 717–725, 1963.

C. M. Meissner. A new world order: explaining the international diffusion of the gold standard, 1870–1913. *Journal of international Economics*, 66(2):385–406, 2003.

R. A. Mundell. A theory of optimum currency areas. *The American economic review*, 51(4):657–665, 1961.

S.-W. Nam. Exchange rate policy in korea: a survey. *KDI Working Paper*, (8406), 1984.

M. Obstfeld. Models of currency crises with self-fulfilling features. *European Economic Review*, 40(3-5):1037–1047, 1996.

M. Obstfeld and K. Rogoff. The mirage of fixed exchange rates. *Journal of Economic Perspectives*, 9(4):73–96, 1995.

M. Obstfeld, J. C. Shambaugh, and A. M. Taylor. The trilemma in history: tradeoffs among exchange rates, monetary policies, and capital mobility. *Review of Economics and Statistics*, 87(3):423–438, 2005.

I. Patnaik and M. Pundit. Exchange rate regimes and macroeconomic performance in south asia. *International Journal of South Asian Studies*, 8(1):1–26, 2015.

C. M. Reinhart and K. S. Rogoff. The modern history of exchange rate arrangements: a reinterpretation. *the Quarterly Journal of economics*, 119(1): 1–48, 2004.

H. Szlajfer. Economic nationalism and the modernization of venezuela. *Latin American Research Review*, 25(3):63–98, 1990.

H. C. Wolf, A. R. Ghosh, H. Berger, and A.-M. Gulde. Currency boards in retrospect and prospect. 2008.

View publication stats

Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-024-34935-x

RESEARCH ARTICLE

The potential of magnesium oxide nanoparticles (MgONPs) in the transesterification of lipids produced by *Rhodotorula minuta*

Richa Srivastava[1] · Niti Srivastava[2] · SonamPreet Kaur[2] · Kamal Kant[2] · Kumar Gaurav[2]

Received: 29 January 2024 / Accepted: 4 September 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

The urgent need to address energy security risks and global warming has led to exploration of renewable energy sources. One such avenue is biodiesel specifically focusing on the potential of *Rhodotorula minuta*, a type of yeast known for producing lipids that can be used as a sustainable alternative for production of biodiesel. In the current study, this promising yeast was evaluated for its potential to produce lipids. The morphological characterization was carried out by scanning electron microscope (SEM), and intracellular detail was studied by transmission electron microscope (TEM). Changes in content and cellular biomass were monitored at time intervals with the highest biomass yield of 12.4 g/l and lipid content of 6.2 g/l achieved after 72 h. In the present work, magnesium oxide nanoparticles (MgO NPs) were synthesized and extensively characterized through Fourier transform infrared spectroscopy (FTIR), thermogravimetric analysis (TGA), Brunauer–Emmett–Teller (BET), SEM, TEM, and X-ray diffraction (XRD). By employing response surface methodology (RSM) with Box–Behnken design (BBD), optimal process conditions for transesterification could be determined. The best result achieved was a yield of 88.6% when the conditions were optimized, using methanol to oil ratio of 18:1 and 8% (w/w) amount of catalyst maintaining a reaction temperature of 55 °C and allowing the reaction to proceed for 120 min.

Keywords Oleaginous microorganism · Response surface methodology · Renewable energy · Heterogeneous catalysts · Electron microscope · Biodiesel

Introduction

In the wake of rising global energy needs, rapid urbanization, and persistent economic growth in today's world population, sustainable and renewable energy assumes greater relevance. Multiple use of fossil fuels results in critical challenges such as energy scarcity and global warming (Wang and Azam 2024; Gaurav et al. 2013). To confront these wide-reaching environmental problems, firm steps must be taken to encourage the use of renewable, biodegradable, and sustainable biofuels. Biodiesel has gained widespread popularity and attention from scientists around the globe as a viable alternative to petroleum diesel fuel (Ogunkunle and Ahmed 2019). Apart from being more productive than previous fossil fuel, it also has some extra positive features such as superior lubricity, non-toxicity, sulphur-free, environment-friendly, biodegradability, and acceptable fuel qualities (Vasić et al. 2020). Furthermore, the application of biodiesel reduces carbon emissions and greenhouse gas discharges, thus ensuring a cleaner and more sustainable atmosphere (Bhatia et al. 2021). It is the second largest produced liquid biofuel worldwide after bioethanol, reached a peak production of 48 billion litres by the end of 2021 (D'Amato et al. 2020). Initially, vegetable oil was the most frequently used feedstock for the manufacturing of biodiesel by transesterification process in the presence of a catalyst (Hoseini et al. 2019; Vakros 2018; Musa 2016 and Abdullah et al. 2017), which may be used directly in diesel engines with or without modification (Bhatia et al. 2017). However, as food prices rise and human needs increase, these primarily based edible oils are no longer used and non-edible oil such as Jatropha, Karanja, and Pongamia, and waste animal fats are used for biodiesel production, but the high content of

Responsible Editor: Ta Yeong Wu

✉ Kumar Gaurav
gauravbche@gmail.com

[1] Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

[2] Amity Institute of Biotechnology, Amity University Haryana, Gurugram 122413, India

Published online: 09 September 2024

fatty acid makes biodiesel more viscous and requires more hydrolysis and pretreatment process, which automatically raises the high cost (Tan et al.2015). Choice of feedstock also plays a significant role in total biodiesel production cost which is the bottle neck in its widespread use (Hoseini et al. 2019; Laurens et al. 2017). However, third-generation biofuel generated from oleaginous microorganisms has the potential to accumulate more lipids and serve as a renewable feedstock for biodiesel (Suparmaniam et al. 2019). These oleaginous microorganisms, yeast, bacteria, microalgae, and fungi have been found to produce microbial oil with a similar composition to vegetable oil. They can accumulate lipids that make up more than 20% (w/w) of their dry weight and occasionally even reach levels as high as 80% (w/w) (Ma et al. 2018; Abeln and Chuck 2021; Huang et al. 2013; Moyo et al. 2021). These little creatures are flexible in how they can use everything from pure glucose to waste but prefer waste because it can be cheaper and most abundantly available (Tomás-Pejó et al. 2021). These single cell oils (SCOs) exhibit similarities to vegetable oils but have unique advantages such as freedom from geographic restrictions, resilience to seasonal variations, scalable harvest schedules, flexibility, and scalability for industrial systems (IEA RIEA, Paris, 2022; Dong et al. 2016). These microbial oils are generated as lipid droplets during the production of fatty acids, TAGs (triacylglycerides). However, scaling-up biodiesel production from microbial oil faces challenges, including the required microbial culture conditions and the difficulties of both microbial harvesting and oil extraction (Abghari and Chen 2014). Some of the examples of typical oleaginous yeasts are *Yarrowia lipolytica*, *Cryptococcus albidus*, *Trichosporon pullulan*, and *Rhodotorula glutinis* (Busic et al. 2018). One such oleaginous yeast, *Rhodotorula minuta*, which may accumulate 40–70% of the total lipid content is better suited for biodiesel production than other oleaginous microorganisms. It may also be grown in low-cost carbon sources such as lignocellulosic biomass, sugarcane molasses, cheese whey, wheat, and rice straw, under nitrogen-limiting conditions, to produce more lipids for biodiesel generation (Spencer et al. 2002; Rebello et al. 2018). Heterogeneous catalysts like different metal oxides including CaO, MgO, ZnO, and TiO_2 have drawn people's interest recently due to their safe and stable nature in addition to improve biodiesel production and feasible recovery. MgO is particularly notable as a strong base hetero catalyst known for its chemical traits and versatility (Seffati et al. 2019; Chingakham et al. 2019; Julkapli and Bagheri 2016). It is chosen for biodiesel because it is more basic than other acid-based alternatives. Additionally, MgO is feasible to use, cheap, and does not have hazardous traits for the environment. Unlike similar metal oxides, one can get rid of it with an acid solution without damaging nature. The RSM is utilized to study the morphological aspects related to the Box–Behnken designed model. This research demonstrated the versatility of RSM in tuning biofuel production processes (Singh et al. 2020; Tamoradi et al. 2021; Gupta et al. 2022; Nazarpour et al. 2022; Manmai et al. 2021; Selemani and Kombe 2022). Further, it expanded the application of RSM by optimizing transesterification processes using Taguchi design. RSM allows for optimization of multiple responses. However, when dealing with responses, it becomes more complex to identify combinations of factors and levels. To address this challenge one practical alternative is using Box–Behnken design (BBD) which falls under response surface design category. BBD stands out for its prediction of first- and second-order coefficients while being less time-consuming and cost-effective compared with designs (Manohar et al. 2013). This optimization process aims to enhance the efficiency of the transesterification, thereby contributing to the potential applicability of *R. minuta* lipids in biodiesel production or other valuable processes. Moreover, the work explores the reusability of the catalyst, an essential consideration for sustainable and cost-effective processes. This aspect adds a practical dimension to the research, emphasizing its applicability beyond the laboratory setting.

In the present work, cell morphological analysis of the growth of *R. minuta* using glucose as a carbon source was done at different time intervals. The work deals with the comprehensive investigation of *R. minuta* cells, focusing on various aspects of its growth and lipid content. Further, the lipid content of *R. minuta* cells was investigated by employing techniques such as scanning electron microscope (SEM), transmission electron microscope (TEM), and fluorescence microscope (FM) to understand the details of lipid accumulation and distribution within the cells. Furthermore, the work focuses into the synthesis of nanoparticles (NP) utilizing green methods, showcasing an environmentally friendly approach. The characterization of these nanoparticles was carried out using Fourier transform infrared spectroscopy (FTIR), X-ray diffraction (XRD), SEM/EDAX, TEM, and thermogravimetric analysis (TGA). The synthesised nanoparticle is employed in transesterification of lipid from *R. minuta* for biodiesel production, and the study systematically optimizes various parameters using response surface methodology (RSM).

Materials and methods

Microorganism and culture conditions

The yeast, *R. minuta* MTCC 2518, was procured from Microbial Type Culture Collection, Institute of Microbial Technology, CSIR, Chandigarh, India. The YPD broth containing (20 g/L yeast extract, 20 g/L peptone, and 20 g/L dextrose) was adjusted at pH 5.6 and then autoclaved at 121 °C for

20 min. The freeze-dried strain was inoculated in YPD broth and incubated in rotary shaker at 28 °C for 3–5 days, and after, that strain was streaked on sterilized YPDA medium containing (10 g/L yeast extract, 20 g/L peptone, 20 g/L dextrose, 15 g/L agar) slant, was set at pH 5.6 and incubated at 28 °C for 3–5 days, and after the growth of yeast, stored at 4 °C for future use. A 100-ml seed culture was prepared using YM medium containing 3 g/L of yeast extract, 3 g/L of malt extract, 5 g/L of peptone, and 10 g/L of glucose. The pH was adjusted to 5.6 with 1 M NaOH/HCL before autoclave, and then 5% (v/v of inoculum) was transferred to 250-mL flask containing 100 mL of YM medium. Incubation was performed at 28 °C and under shaking (180 rpm) with a rotatory shaker for up to a period of 24 h. The considered production medium (100 ml) was formulated as follows: yeast extract 0.5 g/L, glucose 5.0 g/L, $(NH_4)_2SO_4$ (0.2 g/l), KH_2PO_4 (2.7 g/L), $Na_2HPO_4.12H_2O$ (2.4 g/L), $MgSO_4.7H_2O$ (0.1 g/L), and NaCl (0.005 g/L). The pH of medium was adjusted and then autoclaved. In the production, 5% (v/v) inoculum was added, and the flask was incubated at 28 °C and 180 rpm on rotary shaker for 3–5 days.

Measurement of biomass

After production, the culture broth was centrifuged at 5000 rpm for 15 min at 4 °C followed by washing twice before drying. The weight of the dried cells was calculated using gravimetric analysis.

Yield parameters

The biomass yield (Yx/s g/g) was calculated according to Eq. 1 where ΔX is the biomass concentration (g/L) and ΔS is the substrate consumed to generate the biomass at time t (hr) (Chiu et al. 2011; Park et al. 2012).

$$X_1 \quad (1)$$

The lipid yield concerning substrate utilization (Yp/s g/g) was calculated based on Eq. 2, where ΔP is the lipid concentration (g/L) and ΔS is the substrate consumed to generate the biomass at time t (hr):

$$X_2 \quad (2)$$

The lipid content concerning cell mass formation (Yp/x g/g) was calculated from Eq. 3:

$$X_3 \quad (3)$$

Lipid extraction and separation

Lipids were extracted from the dried yeast sample (0.5 g) using Bligh and Dyer technique (Bligh and Dyer 1959). The yeast sample was added in a mixture of methanol (10 mL, MeOH) and chloroform (5 mL, $CHCl_3$) in a ratio of 2:1 v/v. The mixture was stirred for 2 min, and further 5 mL of $CHCl_3$ was added and stirring continued for another 30 s. Afterward, it was transferred to a centrifuge tube and centrifuged at 2000 rpm for 10 min resulting in two phases. The upper phase contained MeOH, while the lower phase held $CHCl_3$ along with the extracted lipids. Solvent was evaporated to get the extracted lipids. Finally, lipid measurements were conducted through gravimetric methods, and the samples were preserved at − 20 °C in an inert atmosphere to safeguard against lipid oxidation. Solid phase extraction technique was employed for separation of the extracted lipids using a silica gel column that had been preconditioned with particle sizes ranging from 230 to 400 µm. The column was used to separate the neutral lipids (NLs) glycolipids (GLs), and phospholipids (PLs). NLs were eluted with $CHCl_3$, GLs by acetone and PLs by MeOH.

To analyse the fraction obtained from the solid phase extraction, thin layer chromatography was performed. The fraction which was dried and then dissolved in $CHCl_3$ was run alongside a triglyceride (TAG). The TLC plate was stained with 2,7 dichlorofluorescein and observed under UV light to visualize the TAG band. This band was identified, extracted using hexane, and subsequently dried.

Glucose estimation

Glucose content was estimated by the dinitrosalicylic acid (DNS) method described by Miller (Miller 1959).

Morphological analysis

Scanning *electron* microscopy (SEM)

To examine the morphology of *R. minuta*, cultures were grown in a production medium at a temperature of 28 °C. Two millilitres of cell solution was withdrawn daily and then subjected to centrifugation. The resulting cell pellets were treated with glutaraldehyde (2.5%) and paraformaldehyde (2%) in phosphate buffer in vials for 24 h. Afterwards, the samples were washed and air dried before being coated with a layer of gold (approximately 20 nm). Finally, the processed samples were visualized using SEM instrument (JEOL JSM 6100), at the Sophisticated Analytical Instruments Facility (SAIF), All India Institute of Medical Sciences (AIIMS), New Delhi.

Transmission *electron* microscopy (TEM)

The cells were cultivated for 1–3 days for TEM analysis. Samples were centrifuged at 5000 rpm for 10 min. The resulting cell pellets were fixed in a solution containing glutaraldehyde (2.5%) and paraformaldehyde

(2%) buffered at pH 7.4. The cell block was prepared after incubating the cells at a temperature of 4 °C for 24 h followed by rinsing with water. An ultramicrotome (Ultra cut E, Reichert Jung) was used to cut thin sections measuring between 60 and 80 nm, and these sections were then stained with alcoholic uranyl acetate and lead citrate for a duration of 10 min each. The stained grids were subsequently captured using a TEM (Philips, CM 100) operating at 60–80 kV. All TEM studies were carried out at the Sophisticated Analytical Instruments Facility (SAIF), All India Institute of Medical Sciences (AIIMS), New Delhi.

Analysis of microbial lipids using fluorescence microscopy

A fluorescence microscope model TS2FL, equipped with NIS Element software (version D.5.10), was used to study the triacylglycerols (TAG) accumulation in *R. minuta*. Nile red was used as a dye to specifically detect lipid droplets. The cells were dissolved in Nile red sodium chloride solution (0.01 M). The yeast cells were washed thrice with phosphate buffer saline and kept for 24 h. The stained cells along with the Nile solution were incubated in darkness for 30 min at room temperature to allow proper staining and interaction between the dye and intracellular lipids. After the incubation period, the cells were observed under a fluorescence microscope at 100 × magnification. These studies were conducted at Central Instrument Research Facility (CIRF), Amity University Haryana, Gurugram.

Green synthesis of MgO Nps

Magnesium oxide nanoparticles (MgO NPs) were synthesized using a green method with *Azadirachta indica* leaf extract. The leaves (5.0 g) were thoroughly washed with water and dried in an oven at 50 °C until they were completely dry. The dried leaves were mixed with water (200 mL) and boiled at 100 °C for 1 h. The resulting solution was then filtered using Whatman No.1 filter paper, and the filtrate was collected and stored in the refrigerator at 4 °C for further use. To synthesize MgO Nps, leaf extract (5 mL) was mixed with water (20 mL) and heated at temperature of 60 °C. Magnesium nitrate (5 g) was added to the solution while stirring it magnetically at a temperature of 80 °C, for 4 h. The resulting solid was separated using centrifugation at 10,000 rpm for 10 min. The material was then washed with distilled water and centrifuged at 5000 rpm for 10 min. The dried material was further treated in a muffle furnace at 600 °C to produce MgO nanoparticles.

Catalyst characterization

Fourier transform infrared spectroscopy (FTIR)

The functional groups of synthesized MgO Nps were determined by Fourier transform infrared spectroscopy (FTIR). The spectra were obtained via (Nicolet iS50) FTIR spectrometer in the wavelength range from 400 to 4000 cm^{-1}.

Thermogravimetric analysis (TGA)

Thermogravimetric analysis (TGA) is employed to investigate the stability and decomposition behaviour of MgO nanoparticles. The sample was heated with 10 °C/min under a nitrogen flow with temperature range of 29–876 °C using Perkin Elmer (4000 TGA analyzer).

BET analysis

The surface area, pore diameter, and pore volume were determined by BET method with aid of nitrogen gas adsorption–desorption using Quantachrome Autosorb Automated Gas Sorption System. Prior to BET analysis, the samples were degassed at 200 °C for overnight to remove any adsorbed molecules from pore and surface.

XRD analysis

XRD analysis of nanoparticle was carried out by X-ray diffraction (Rigaku Ultima IV, Ri). The average crystallite size (D) of MgO nanoparticle was calculated by Debye–Scherrer equation which is given as $D = K\lambda/\beta \cos\theta$, where K is the dimensional constant and has a value of about 0.9; λ is the wavelength of the X-ray radiation (1.5406 Å for Cu Kα); β is the full width at half-maximum (in radian); and θ is the position of the maximum of the diffraction peak. The measurement of 2θ angle was 10°–80°.

SEM/EDX analysis

The surface morphology of the nanoparticles was determined by scanning electron microscopy (SEM) using a Hitachi High Technologies instrument. Elemental identification and chemical composition of the MgO nanoparticles were determined by energy dispersive X-ray spectroscopy (EDAX). This analysis was performed in conjunction with the SEM to provide detailed information on the elemental identification and composition of the nanoparticles.

Transmission *electron* microscopy (TEM) analysis

The transmission electron microscopy (TEM) analysis was done on a FEI Philips Morgagni instrument. Synthesized

nanoparticles were finally prepared for sample preparation by setting apart the synthesized nanoparticle from its dispersion in ethanol, and then a small droplet of this suspension was pipetted onto carbon-coated copper grid and let solvent evaporate. High-resolution TEM imaging was performed on a FEI Tecnai G2 (Thermo Fisher Scientific) operating at 200 kV, with the bright-filed images collected to observe morphological information and nanoparticle size distribution.

Transesterification process

The transesterification process takes place in two steps: catalyst activation and subsequent transesterification reaction. The MgO NPs and methanol were taken in a round bottom flask and refluxed at 60 °C for 1–2 h. The preheated lipid was added to this mixture and stirred for 1–4 h. After completion of the reaction, clear biodiesel was obtained along with the glycerol as by product and catalyst at the bottom. The clear biodiesel was separated using separatory funnel and washed with distilled water.

Optimization of the reaction condition to produce biodiesel using MgO NPs by Box–Behnken design

Box–Behnken design (BBD) of RSM was used to optimize the process parameters for the production of biodiesel. RSM is one of the most significant statistical tools to study the relationship between the responses and the independent variables. The effect of process variables of methanol to oil ratio (X_1), catalyst amount (X_2), reaction temperature (X_3), and process time (X_4) on biodiesel yield (Y) was studied. The four variables of transesterification including X_1 (9:1–27:1), X_2 (1–15 wt%), X_3 (35–75 °C), and X_4 (60–180 min) were employed to design the experiments by RSM coupled with BBD. The percentage conversion of biodiesel was chosen as response. A total 27 experiments were designed using Minitab 13.2 software, and their coded levels are displayed in Table 1. The experimental data was analysed using regression and fitted to a second-order polynomial model to identify the possible interaction between the related factors with response function according Eq. 4.

$$X_4 \tag{4}$$

where Y was the yield of biodiesel; β_0 the intercept term; β_1, β_2, β_3, β_4 were linear coefficient; β_{11}, β_{22}, β_{33}, β_{44} were quadratic coefficient; β_{12}, β_{13}, β_{14}, β_{23}, β_{24}, β_{34} were interactive coefficient; and X_1, X_2, X_3, X_4 were independent variables.

The validity of design model was evaluated by the coefficient of determination (R^2) and analysis of variance (ANOVA). R^2 must be close to 1 indicating the better correlation between the experiment and the predicted values. ANOVA was used to check the efficiency of the regression model by calculating the F-value within a level of confidence of 95% and p-values of 0.05%. Contour plots were created to visualize the individual and the interaction between the effects of the independent variables towards the biodiesel yield.

Results and discussion

Cell mass, lipid content, and glucose consumption of *R. minuta* in production medium

The investigation into *R. minuta's* growth dynamics and lipid production in the production medium revealed intriguing patterns and substantial outcomes. Utilizing a 5% (v/v) inoculum yielded the highest biomass concentration of 12.4 g/L, accompanied by a biomass productivity of 0.155 g/L/d. Additionally, on the third day of growth, the specific growth rate and total lipid content reached 0.034 h^{-1} and 6.2 g/L, respectively. The growth profile illustrated distinct phases over the three days as shown in Fig. 1. During the initial growth phase, there was initially a rather swift surge in the cell mass concentration level, peaking at 0.11 g/L as observed in the rapid increase occurring at the outset. By the end of the second day, the concentration had amplified substantially to 8.76 g/L, achieving an ultimate maximum biomass concentration of 12.4 g/L which was attained during the prolonged growth across the third successive day. Analysing the biomass productivity revealed compelling insights. Between the initial and subsequent periods of development, the rate of productivity escalated quite remarkably from 0.136 g/L/d at the beginning to 0.216 g/L/d by the next interval of time. However, on the third day, it marginally declined to 0.155 g/L/d, likely due to decreased glucose availability in the production media. The lipid content demonstrated a significant increase throughout the fermentation period. Early-stage lipid production, linked to cell wall synthesis, is correlated with glucose consumption. Notably, during the initial growth phase, the lipid concentration reached 1.89 g/L, representing 36% of glucose consumption. During the second day of cultivation, the substantial surge in lipid content from 1.89 to 4.19 g/L demonstrated that significant escalation had occurred. Eventually, at the 72-h mark, the maximum lipid content of 6.2 g/L was attained, corresponding to 75% glucose consumption. In the present study, it is very evident that the characterized oleaginous yeast *Rhodotorula* is valuable source of single cell oil to produce biodiesel. In some of the studies, it is reported that the effect of pH also leads to substantial increase and decrease in lipid accumulation (Maza et al. 2020). The highest lipid content and total pigment content, measured after 3 days with a C/N ratio of 146, were 2.49 g/L and 0.31 g/L, respectively

Table 1 Experimental design and experimental value

Run order	Methanol-to-oil ratio	Catalyst amount	Temperature	Reaction time	Biodiesel yield
1	27	8	55	180	75.6
2	9	1	55	120	71.6
3	18	1	75	120	63.2
4	18	15	55	60	78.5
5	18	8	35	180	72.6
6	27	8	35	120	71.5
7	18	8	75	180	72.1
8	18	1	55	60	74.5
9	27	8	55	60	72.3
10	27	15	55	120	74.5
11	18	15	35	120	72.3
12	18	15	75	120	71.5
13	9	8	35	120	71.4
14	18	15	55	180	76.8
15	9	15	55	120	86.4
16	27	8	75	120	63.4
17	18	8	55	120	88.6
18	18	8	75	60	64.5
19	18	8	35	60	76.5
20	18	1	55	180	78.9
21	18	1	35	120	73.6
22	27	1	55	120	74.8
23	9	8	55	180	79.5
24	18	8	55	120	88.6
25	9	8	75	120	74.8
26	18	8	55	120	88.6
27	9	8	55	60	74.6

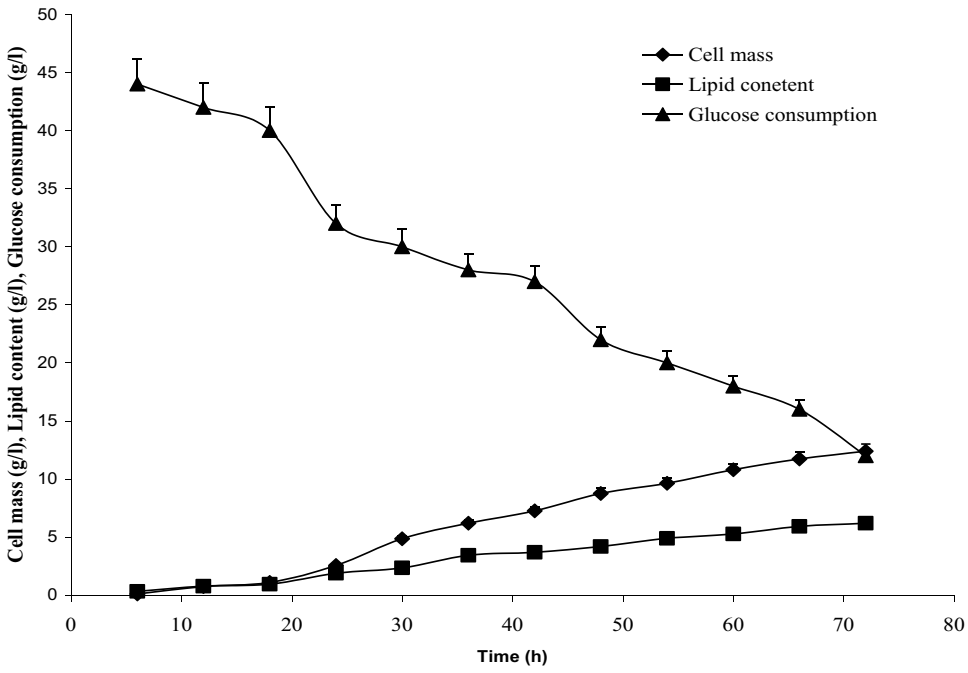


Fig. 1 Cell mass (black diamond) (g/L), lipid content (black square) (g/L) and glucose consumption (black triangle) (g/l) in different time interval (hours)

(Elfeky et al. 2019). The observed trend in the present work underscore *R. minuta* potential as an oleaginous yeast with substantial lipid production capacity. These findings suggest its viability as a candidate for biodiesel production. Further investigations into optimizing growth conditions could unlock its full potential in sustainable biofuel production.

Growth yield and product yield

The total value of biomass growth yield was ($Y_{x/s}$ 0.628 g/g), product (lipid) yield with cell mass formation ($Y_{p/\times}$ 29.02 g/g) and substrate consumption ($Y_{p/s}$ 14.06 g/g) are observed in 72 h of the fermentation. The values of growth yield ($Y_{x/s}$) and product yields ($Y_{p/x}$, $Y_{p/s}$) during different time intervals are shown in Fig. 2. During initial phase of growth (up to 6 h), the value of growth yield was 0.05 g/g. Whereas from 12 to 24 h of fermentation, the value of Yx/s of biomass increased to (0.078 to 0.081 g/g). The maximum biomass yield was achieved 0.115 g/g in 30 h. After 36 to 48 h, the Yx/s value decreased with the 0.06 to 0.053 g/L. From 54 to 72 h, the Yx/s value of biomass growth yield decreases from 0.02 to 0.017 g/L. The lipid yield ($Y_{p/s}$) (lipid produced/glucose consumed) increases in the beginning of the fermentation (6–24 h), i.e. 0.02 to 4.89 g/g, and then decreases until the end of the fermentation. The $Y_{p/s}$ value reached maximum value of 4.89 g/g within 24 h and then declined. Further analysis of lipid yield with respect to biomass ($Y_{p/x}$, g lipid produced/g of biomass) in fermentation broth at different time interval was studied. The lipid content with biomass formation increases from 6 to 18 h, and the value of $Y_{p/x}$ was 0.23 to 2.25 g/g. After 24 h, the $Y_{p/x}$ value was observed at 30 h, i.e. 4.89 g/g. In 36 h, decline in $Y_{p/x}$ value (1.22 g/g) indicates the lipid degradation, due to β-oxidation of the produced lipid by the yeast. After 66 to 72 h of fermentation, $Y_{p/x}$ value was 0.083 to 2.48 g/L.

SEM analysis

In *R. minuta*, cell wall is located on the outer surface of the cell and plays a crucial role in the transport of materials in and out of the cell. Therefore, analysis of the mechanisms forming the cell wall is significant in both fundamental and applied biology. The identification of cell walls and ultrastructure of *R. minuta* are revealed by the scanning electron microscope (SEM) after thin sectioning in identifying cell shape, cell size, and structure of *R. minuta* cell (Fig. 3). In SEM analysis, it revealed that *R. minuta* produce a noticeable change in shape and size of the cells in number of days. Notably, the selection of the production medium emerges as a determining factor for the *R. minuta* species. When it was grown in production medium using glucose as a carbon source at 28 °C for 1–3 days, cell become ovoid. In first day of the analysis, the cell size increases to 4.61 µm width 2.73 µm (Fig. 3a); on second day of the analysis, the cell size slightly decreases to 3.89 µm width 2.32 µm (Fig. 3b), while on third day, it decreases to 3.58 µm width 2.17 µm as shown in Fig. 3c. According to another report, when *R. minuta* species were grown in 5% malt extract agar at 28 °C for 6 days, cell become subglobose to ovoid, 1.5–3 × 3–8 µm, either occurs single or in pairs, and budding is predominantly polar. While in corn meal agar, the cell is poorly developed, pseudohyphae may be formed. Some cells are produced short with fine structure, i.e. (1 µm in diameter)

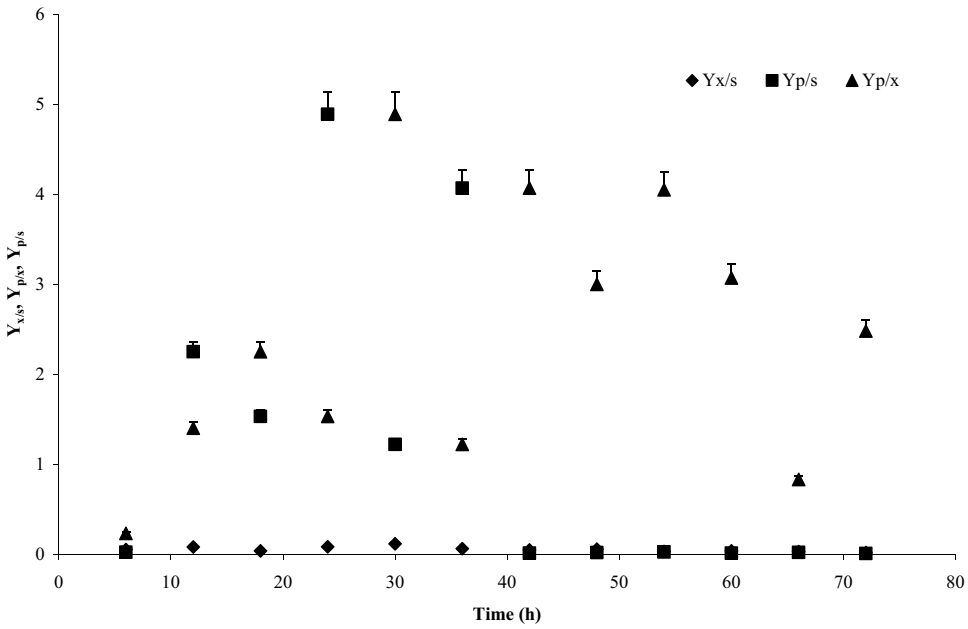


Fig. 2 Growth yield ($Y_{x/s}$) (black diamond), roduct yield with substrate utilization $Y_{p/s}$ (black square) and with cell mass formation $Y_{p/x}$ (black triangle) in different time intervals

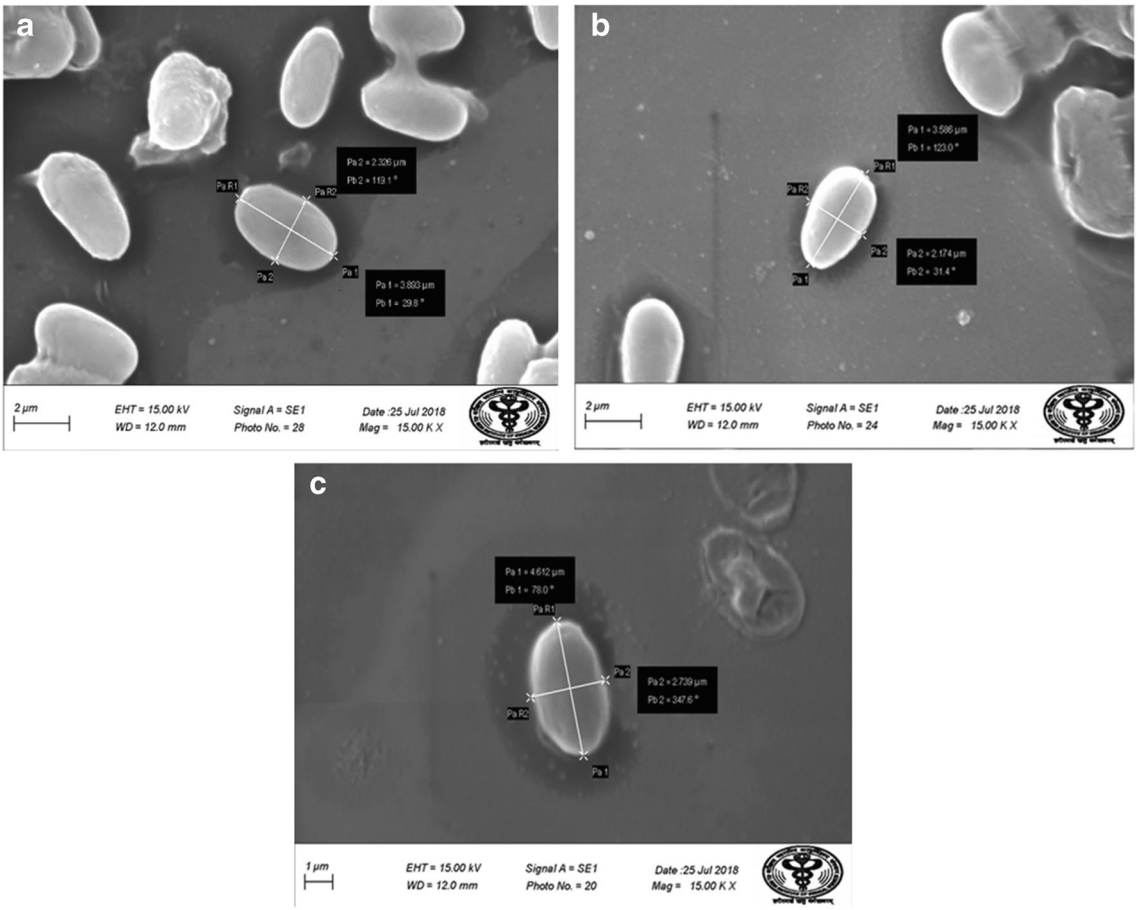


Fig. 3 SEM images of *R. minuta* showing different cell shape and size as the number of day increases **a** first day of growth on production media, glucose as a carbon source, **b** and **c** show second and third day of growth, respectively

with few individual cells and can be extended only up to 2 µm (Osumi et al. 1995). Figure 4 shows SEM images of *R. minuta* showing increase in population of cells per day of the growth at different magnification levels. The budding scar and birth scar were observed during the formation of new cell (Fig. 4a). On the first day of the growth, they are in an initial stage of multiplication. On second day of the growth, there is cell aggregation of yeast cell, shown in Fig. 4b and c. While on third day of the growth, there is enormous increase in cell number. It was observed that as the number of days increases, the population of the yeast cell increases. In yeast, the cells are colloids with each other forming hyphae (Fig. 5). First day of growth cells are binding with each other in less ratio (Fig. 5a). There are two type of hyphae, pseudohyphae and binding hyphae. In pseudohyphae, the cells are attached to each other, forming a single cells, while in binding hyphae, the number of cell is attached to each other (Fig. 5b), giving the appearance of every single cell in hyphae (Fig. 5c).

TEM analysis

In TEM, the internal structures of *R. minuta* cell are observed, i.e. cell wall, mitochondria, and lipid droplet. The numbers of lipid droplets were observed as the number of days increases. On first day of the growth, small size of lipid droplets, cell wall, and mitochondria were observed in Fig. 6a. Appearances of the surface layer are consistent with the assumption that lipid particles are surrounded by a monolayer membrane. On second day, the size (Fig. 6b) and number of lipid droplet increase (Fig. 6c), while on third day, lipid droplet size was increased (Fig. 6d).

Fluorescence analysis

Nile red stain, a fluorescent dye, was employed to detect neutral lipids stored intracellularly in the lipid droplets of *R. minuta.* After 1–3 days of growth, cells were analysed using fluorescence microscopy. The yeast cells exhibited fluorescence under different excitation conditions, such as FITC

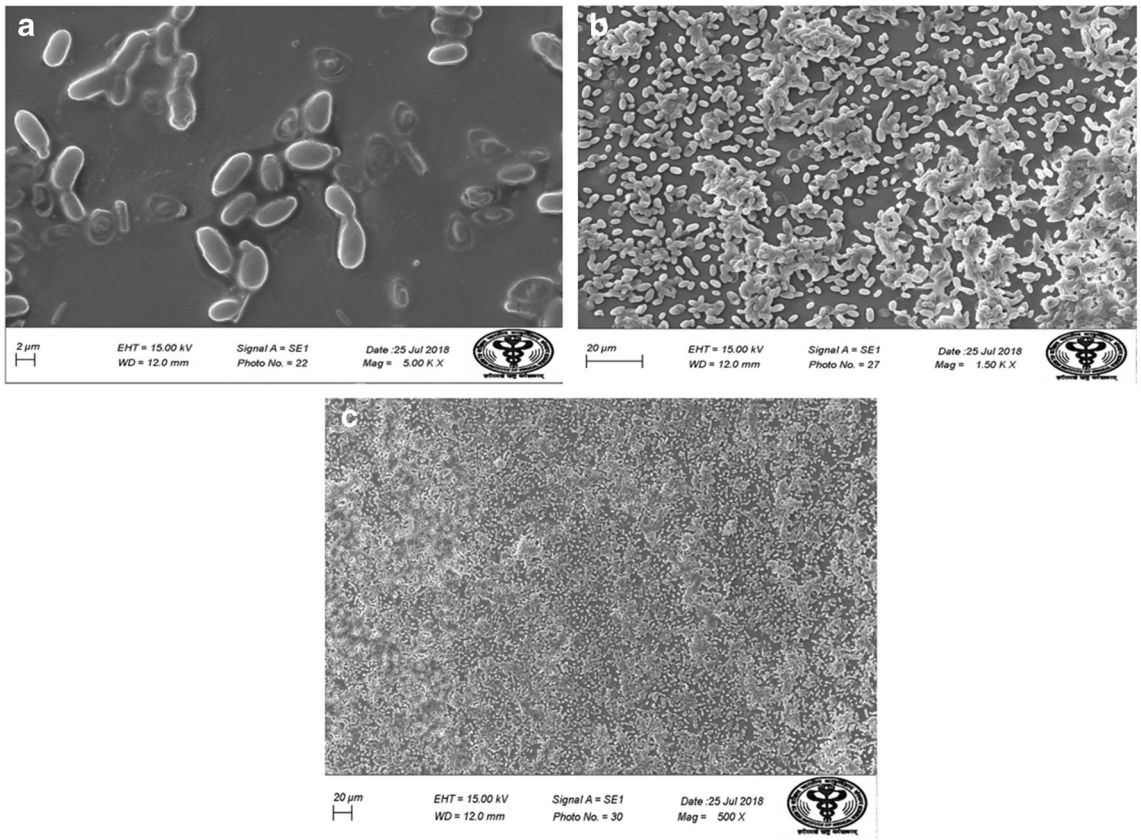


Fig. 4 SEM images of *R. minuta* showing increase in population of cells per day of the growth at different magnification levels: **a** budding of cells, **b** and **c** cell aggregation

excitation at 488 nm and emission at 520 nm, resulting in green fluorescence when observed under a microscope. While the strain displayed lipid bodies, there were variations in their shape, size, intensities, and the persistence of fluorescence emission. This indicated the accumulation of neutral intracellular lipids, as illustrated in Fig. 7. The number of liquid droplets increases from day 1 to day 3. In another study, strains of *Yarrowia lipolytica* were grown in glucose as a carbon source, displaying variable numbers of lipid bodies and cell sizes visualized by Nile red stain dye in a fluorescence microscope for all the strains (Katre et al. 2012). To observe neutral lipids and confirm the presence of lipid bodies, Maza et al. conducted studies on *Rhodotorula glutinis* R4 strains. The results indicated that all strains exhibited lipid bodies when glucose was used as a carbon source for cell growth, but differences were noted in the size, shape, and intensities of the strains (Maza et al. 2020).

Catalyst characterization

FTIR analysis

FTIR was used to find possible biomolecules which are responsible for capping and stabilization of the metal oxide (Dobrucka 2018; Suresh et al. 2018). Figure 8a presents the FTIR spectra of MgO NPs synthesized by using *A. indica* leaf extract. The broad absorption peak at 3480 cm^{-1} was due to the –OH stretching (Karthik et al. 2019). The peak at 1370 cm^{-1} corresponds to absorption peak of C = O. The peak at 550 cm^{-1} indicates the presence of Mg-O functional group (Ramanujam and Sundrarajan 2014).

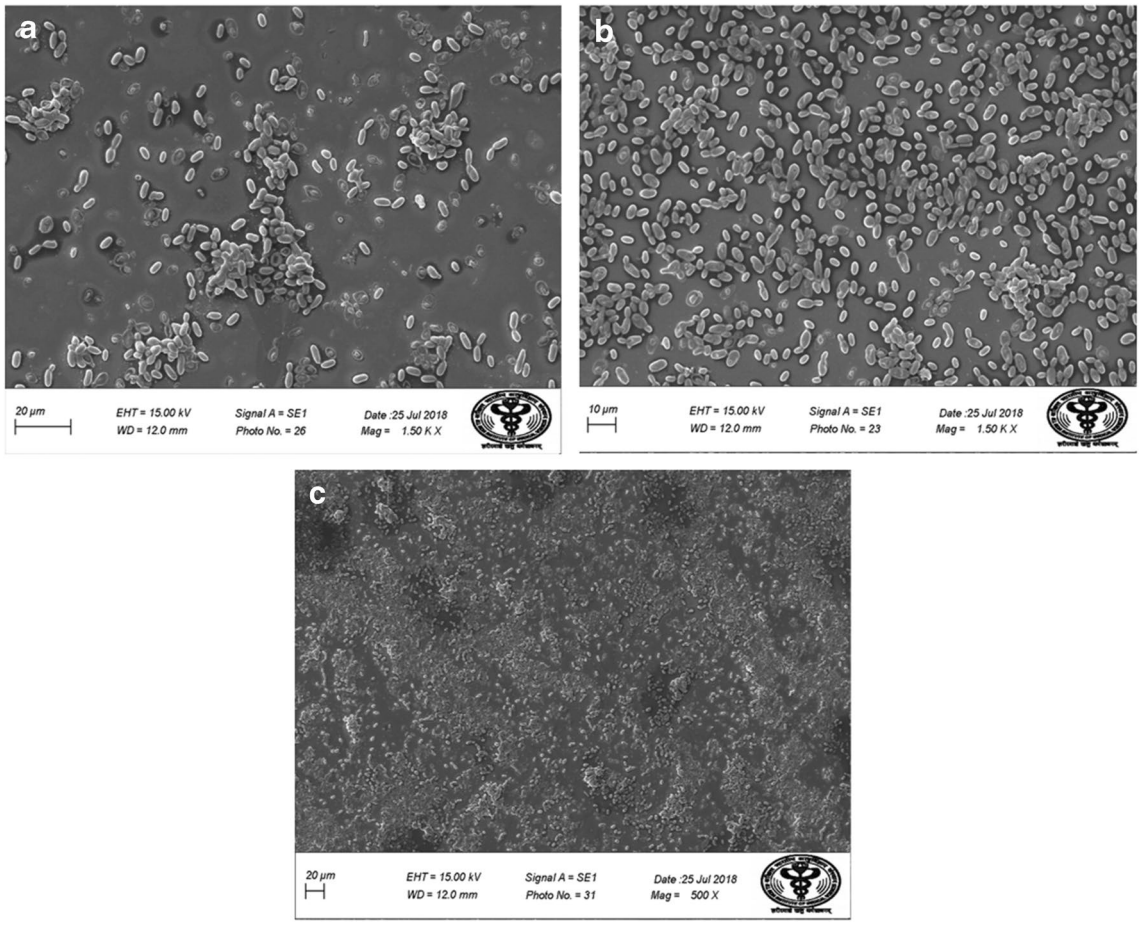


Fig. 5 SEM images of *R. minuta* showing the hyphae during 1–3 days of the growth

TGA analysis

Thermal transition of MgO NPs was investigated by TGA analysis. The resulting thermogram, Fig. 8b, provides insights into weight loss patterns and thermal transitions experienced by the nanoparticles during heating. It showed that the weight continuously decreases with increasing the temperature. The weight loss up to 100 °C was due to evaporation of water molecules at the surface of sample. The weight loss from temperature range 100 to 300 °C was due to oxidative decomposition of organic materials (Ouraipryvan et al. 2009). The significant weight loss observed from 300 to 600 °C was due to evaporation of organic and other residues of the sample. The total weight loss estimated by TGA curve was 23.31%.

BET analysis

The surface area of catalyst plays important role in catalytic activity. The results of BET test showed the surface area, average pore diameter, and pore volume of MgO nanoparticle which were 28.38 m^2/g, 2.6 nm, and 0.18 cm^3/g, respectively. Catalyst with the larger surface area facilitates better interaction between active site of catalyst and reactants.

XRD analysis

XRD analysis is widely used to determine the size and phase identification of nanoparticles. The XRD pattern of MgO NPs calcined at 600 °C was shown in Fig. 8c. The result shows that the structure of MgO NPs is face-cantered cubic phase. The XRD results obtained were matched with the JCPDF card number 89–7746, the peaks corresponding to the planes at 2θ = 36.86°, 42.82°, 62.17°, 74.52°, and 78.44° with corresponding (h k l) values (1 1 0), (2 0 0), (2 2 0), (3 1 1), and (2 2 2), respectively. The average crystallite size (D) of nanoparticle was 26.23 nm calculated by Debye–Scherrer equation, and these results are in accordance with those reported by researchers (Ramanujam and Sundrarajan 2014).

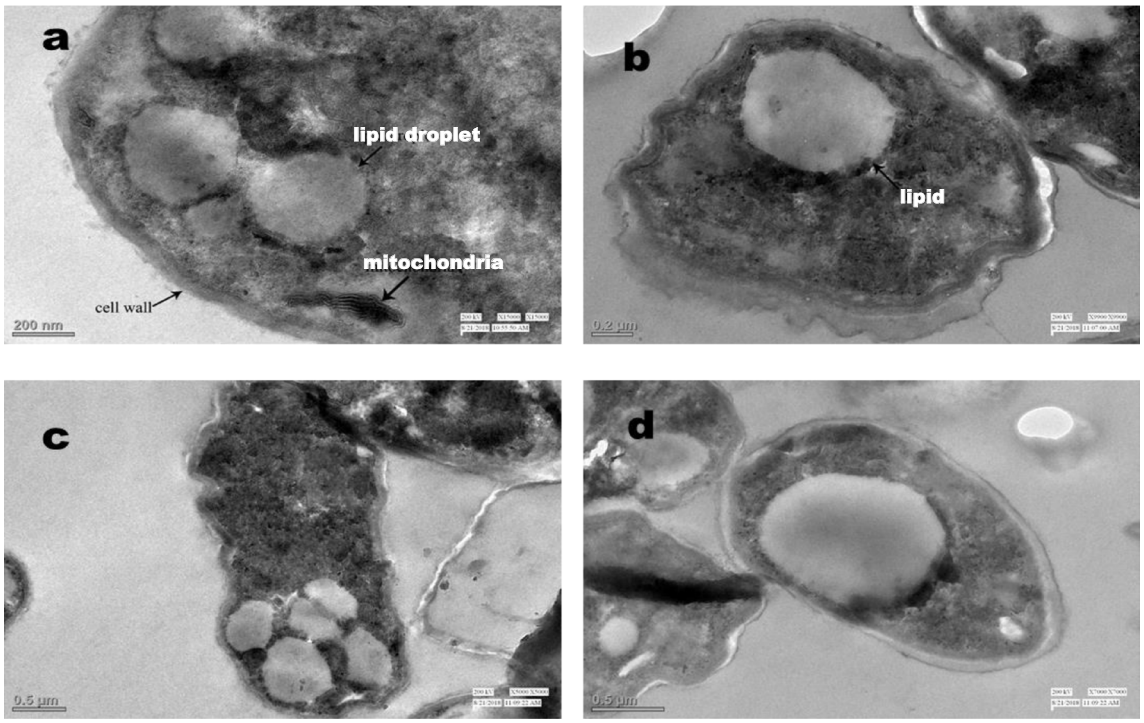


Fig. 6 Ultrathin structure of *R. minuta* observed by TEM

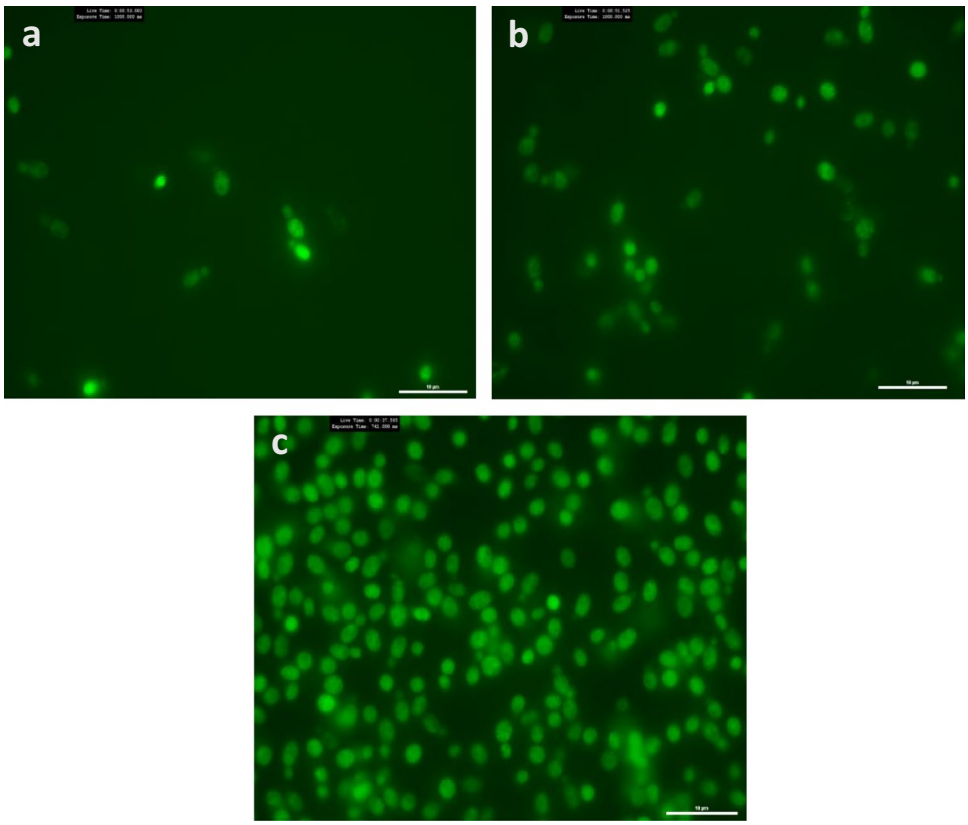


Fig. 7 Fluorescence microscopy images of *R. minuta*

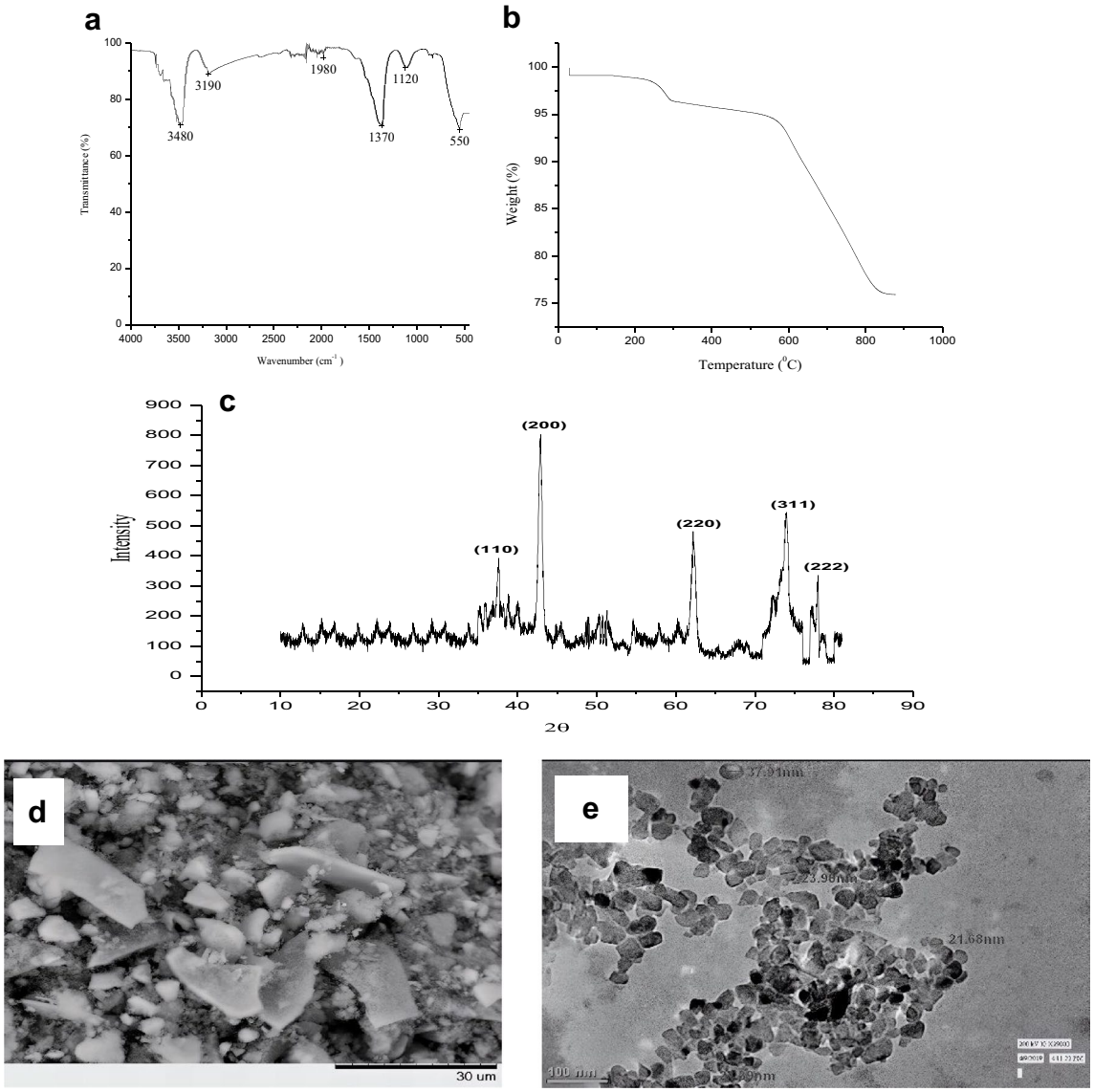


Fig. 8 Characterization details of MgO Np: **a** FTIR of MgO Nps, **b** TGA curve of MgO Nps, **c** XRD of MgO Nps, **d** SEM of MgO Nps, **e** TEM image of MgO Nps

SEM/EDAX analysis

Figure 8d depicts the SEM image of synthesized nanoparticle. The synthesized nanoparticle was agglomerated and form larger clusters. The agglomeration was due to electrostatic attraction and polarity of MgO Nps (Ramanujam and Sundrarajan 2014). It seems that particles shapes and surface were irregular. The additional evidence for the formation of MgO Nps was confirmed by EDAX profile. In the analysis of EDAX profile of MgO Nps, the presence of elemental metal signal (52.05% magnesium and 47.948%. oxygen content) was confirmed.

TEM analysis

TEM analysis is widely used to determine average particle size and shape of nanoparticle. The shape of nanoparticle was found to be cuboidal in shape as shown in Fig. 8e. It was found that size of spherical nanoparticle was 21–37 nm. The average size of nanoparticle was 26.22 nm. The TEM image was in good agreement with the size obtained from XRD data analysis.

Table 2 Estimated regression coefficients for optimization of biodiesel production using MgO NPs

Term	Coef	SE Coef	*T*	*P*
Constant	−58.62	13.3841	−4.380	0.001
X_1	4.12	0.5307	7.761	0.000
X_2	2.74	0.6050	4.526	0.001
X_3	3.03	0.2705	11.188	0.000
X_4	0.32	0.0796	3.996	0.002
X_1^2	−0.08	0.0102	−7.959	0.000
X_2^2	−0.12	0.0168	−7.019	0.000
X_3^2	−0.03	0.0021	−14.482	0.000
X_4^2	−0.00	0.0002	−7.049	0.000
X_1X_2	−0.06	0.0151	−3.968	0.002
X_1X_3	−0.02	0.0053	−3.022	0.011
X_1X_4	−0.00	0.0018	−0.420	0.682
X_2X_3	0.02	0.0068	2.522	0.027
X_2X_4	−0.00	0.0023	-1.603	0.135
X_3X_4	0.00	0.0008	3.022	0.011

Experimental design and regression model

In the present work, RSM model based on BBD with four variables has been formulated to access the performance of the catalyst for transesterification process. All the 27 experimental runs were conducted, and the results were analysed through multiple regression analysis. The coefficients were calculated using regression analysis and tested the significance based on *p*-value. The results of the experimental design and biodiesel yield in terms of actual values are presented in Table 1. The coefficient, *t*-value, and *p*-value at 95% significance level for linear, square, and interaction effects were given in Table 2. Based on the experimental results, quadratic regression equation was developed as given in Eq. 5:

$$X_1^2 \tag{5}$$

where Y was the biodiesel yield, X_1 methanol to oil ratio, X_2 catalyst amount, X_3 reaction temperature, and X_4 process time, respectively. The positive and negative signs of the regression coefficients indicate the synergistic and antagonistic effects of each variable towards the yield. The smaller *p*-value, higher significant is the corresponding coefficient. In this case, X_1, X_3, X_1^2, X_2^2, X_3^2, and X_4^2 were significant terms. From Table 2, it was observed that the coefficient of variables was highly significant ($p < 0.001$) on biodiesel yield.

The analysis of variance (ANOVA) was used to evaluate the statistical significance of model, and the result was summarized in Table 3. The model can be considered statistically significant according to *F*-test at 95% confidence level. Generally, the calculated *F*-value should be greater than the tabulated value if the model is a good prediction of the experimental results and estimated factor effects are real. Here, the computed *F*-value (22.84) was greater than the tabulated *F*-value, $F_{14,12} = 2.6464$ at α 0.05 level, so that the null hypothesis (H_0) is rejected. The values of *p*-value less than 0.05; thus, it can be inferred that treatment differences are statistically significant. To test the fit of the model, determination coefficient (R^2) were evaluated, and the value of R^2 was found to be 0.964 indicate that this model could explain 96.4% of the variability and 4.6% of variability in the response was not explained by this model.

Interaction effects of the process parameters on the biodiesel yield

Contour plots are the graphical representation of the regression equation and are plotted to visualize the interaction effects of the process parameters on biodiesel yield, as shown as in Fig. 9. The plots illustrate the relative effect of any two variables by keeping the third and fourth factor as constant. In general, it was expected that the biodiesel yield increases with increase in process parameters up to optimum value; however, the yield will decrease if the process parameters are increases beyond the optimum value. Figure 9a illustrates the impact of the methanol-to-oil ratio in relation to the amount of catalyst used. The highest biodiesel yield was achieved with an 8% (w/w) catalyst amount

Table 3 Analysis of variance for optimization biodiesel production using MgO NPs

Source	DF	Seq SS	Adj SS	Adj MS	*F*	*P*
Regression	14	1158.04	1158.043	82.717	22.84	0.000
Linear	4	187.81	516.189	129.047	35.64	0.000
Square	4	814.12	814.123	203.531	56.21	0.000
Interaction	6	156.11	156.110	26.018	7.19	0.002
Residual error	12	43.45	43.453	3.621		
Lack-of-fit	10	43.45	43.453	4.345		
Pure error	2	0.00	0.000	0.000		
Total	26	1201.50				

R^2 coefficient of determination = 96.4%, *SS* sum of square, *MS* mean of square, *DF* degree of freedom.

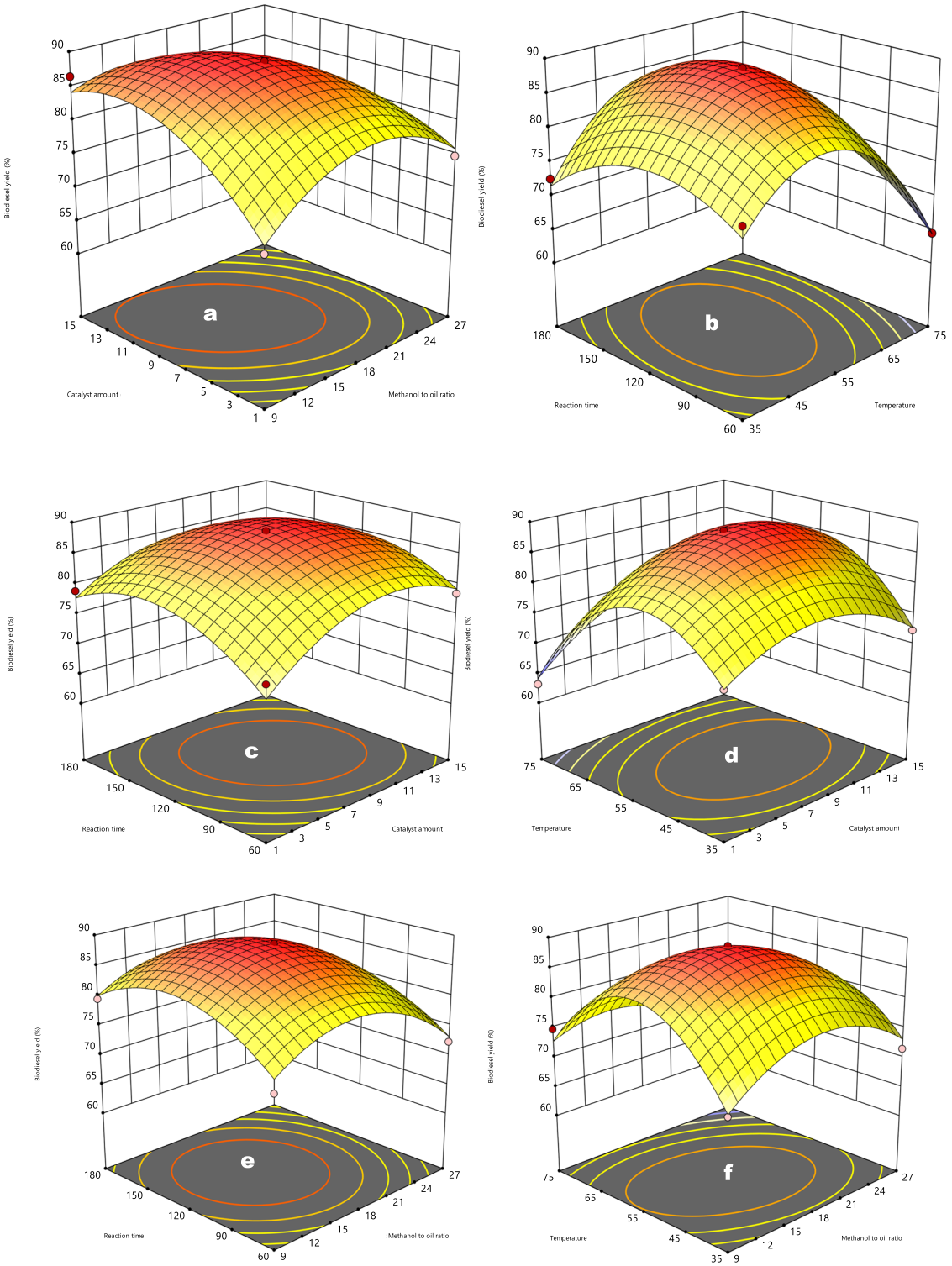


Fig. 9 Response contour plots showing the interaction effect of **a** methanol-to-oil ratio versus catalyst amount, **b** temperature versus reaction time, **c** catalyst amount versus reaction time, **d** catalyst amount versus temperature, **e** methanol-to-oil ratio versus reaction time, **f** methanol-to-oil ratio versus temperature on biodiesel yield

and an 18:1 molar ratio of methanol to oil. As the catalyst amount increases, the number of active sites in the catalytic particles also rises, facilitating surface-mediated heterogeneous catalytic reactions and resulting in a higher biodiesel yield. While beyond the limit, there may be saturation of catalyst particles inside the reaction mixture decreasing the contact area between methanol-to-oil molecules, which causes reduction in the production of biodiesel. Figure 9b depicts the effect of temperature versus reaction time on biodiesel yield at fixed catalyst amount (8%) and molar ratio of methanol to oil (18:1). Increasing the temperature while decreasing the reaction time positively influenced the biodiesel yield. The highest yield was achieved with an 8% (w/w) catalyst amount at 55 °C. Figure 9c presents the relationship between catalyst amount and reaction time on biodiesel yield. The optimal yield was obtained with an 8% (w/w) catalyst amount over 120 min, using a fixed 18:1 molar ratio of methanol to oil at 55 °C. This indicates that the biodiesel yield was positively affected by an increase in catalyst amount and a decrease in both reaction time and temperature. From Fig. 9d, the maximum biodiesel yield was obtained at catalyst amount 8% (w/w) at temperature 55 °C. Figure 9e depicts the effect of molar ratio of methanol to oil and reaction time. From Fig. 9e, the biodiesel yield sharply increases with the reaction time up to 120 min, and beyond that due to chemical changes occurred in transesterification process, the yield reduces significantly. Figure 9f shows that biodiesel production increased with molar ratio of methanol to oil from 9:1 to 18:1 and temperature 35 to 75 °C. The highest biodiesel yield was achieved with a methanol-to-oil molar ratio of 18:1 at 55 °C.

Reusability of catalyst

Reusability and recovery are important aspects from economic and environmental points of view for heterogeneous catalyst in biodiesel production (Conceicao et al. 2017). After the completion of reaction, catalyst was separated from the reaction mixture by filtration and washed thrice with *tert*-butyl alcohol and dried overnight at 100 °C. The catalyst was subjected to transesterification reaction under the optimum conditions (18:1 methanol-to-oil ratio, 8 wt% catalyst, 55 °C reaction temperature, and 120 min reaction time). Figure 10 elucidates a visual representation delineating the biodiesel yield. The graph shows an initial biodiesel yield of 88.6%, attesting to the catalyst's efficacy. By the sixth cycle, the biodiesel yield diminishes to 78.2%, marking an overall reduction of 10.04%. The first cycle, analogous to a scientific benchmark, remains impervious, showcasing the catalyst's sustained catalytic activity and stability. Further, from the second to the fourth, a gradual decrease in yield is observed from 85.4 to 81% and to 79.5% in fifth cycle. The decline in catalytic activity was attributed to the reduction in both the amount and strength of the catalyst's basic sites during regeneration.

Conclusion

The investigation into *R. minuta* growth dynamics and lipid production revealed promising outcomes, suggesting its potential as an oleaginous yeast for biodiesel production. Comparative analyses with other studies supported

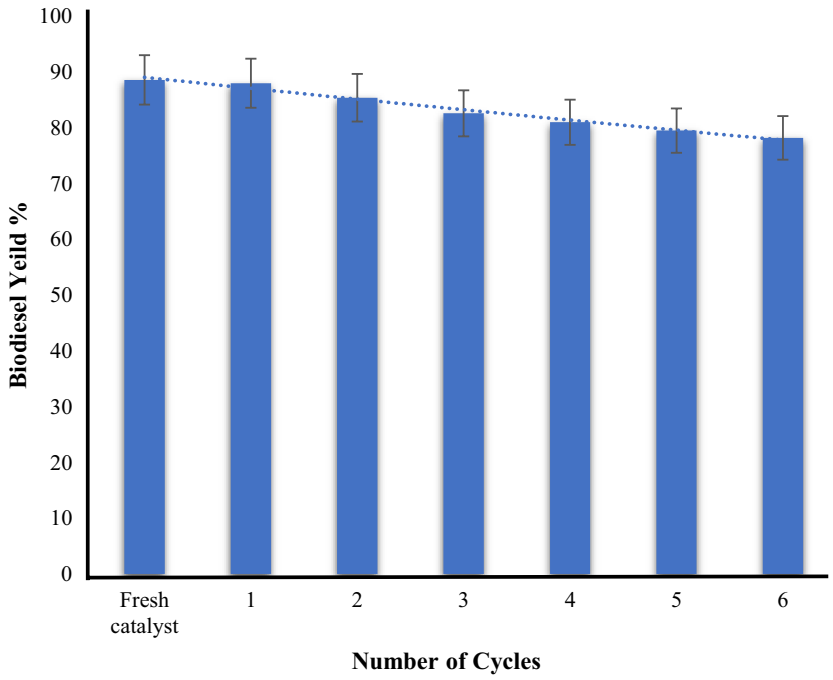


Fig. 10 Reusability of catalyst

R. minuta viability for biodiesel production. The study explored growth yield and product yield. Interaction effects of process parameters on biodiesel yield were illustrated through contour plots, emphasizing the roles of variables like methanol-to-oil ratio, catalyst amount, temperature, and process time. Cellular structure analyses through SEM and TEM provided insights into *R. minuta* morphology and lipid droplet evolution. Catalyst characterization using various analyses highlighted the significance of MgO nanoparticles' properties in catalytic activity. The reusability study showed the catalyst's effectiveness over multiple cycles, with a gradual decline in biodiesel yield, emphasizing its resilience and vulnerability in the biodiesel production process. In conclusion, *R. minuta* shows promise for biodiesel production, and further optimization could enhance its potential in sustainable biofuel production. The comprehensive analysis contributes to understanding the factors influencing biodiesel yield and the yeast's application in biofuel production. Thus, the overarching objective of the present work is to provide a holistic understanding of *R minuta* cells, encompassing their growth kinetics, lipid content, nanoparticle synthesis, and the utilization of extracted lipids in a transesterification process. This multidimensional approach aims to contribute valuable insights to both the fundamental understanding of *R. minuta* biology and its potential applications in environmentally friendly processes.

Acknowledgements The authors thankfully acknowledge the Amity Institute of Biotechnology, Amity University Haryana, and Delhi Technological University, Delhi, for facilitating the necessary infrastructure. The authors also acknowledge SAIF (sophisticated analytical instrumentation facility), AIIMS, New Delhi, for providing working in SEM (scanning electron microscopy) and TEM (transmission electron microscopy). The authors are also thankful to Indian Institute of Technology Kanpur, for BET analysis, and Indian Institute of Technology Delhi, for SEM/EDAX, XRD, and FTIR analysis.

Author contribution Kumar Gaurav and Richa Srivastava contributed to the study conception and design. The first draft of the manuscript was written by Richa Srivastava. Material preparation, data collection, and cell characterization were performed by Niti Srivastava, Sonampreet Kaur, and Kamal Kant. Analysis and discussion were performed by Richa Srivastava and Kumar Gaurav. All authors commented on previous versions of the manuscript and approved the final manuscript.

Data availability The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Abdullah SHYS, Hanapia NHMH, Azida A, Umara R, Juahira H, Khatoon H, Enduta A (2017) A review of biomass-derived heterogeneous catalyst for sustainable biodiesel production. Renew Sust Energ Rev 70:1040–1051. https://doi.org/10.1016/j.rser.2016.12.008

Abeln F, Chuck CJ (2021) The history, state of the art and future prospects for oleaginous yeast research. Microb Cell Factories 20:221. https://doi.org/10.1186/s12934-021-01712-1

Abghari A, Chen S (2014) Yarrowia lipolytica as an oleaginous cell factory platform for production of fatty acid-based biofuel and bioproducts. Front Energy Res 2. https://doi.org/10.3389/fenrg.2014.00021

Bhatia SK, Bhatia RK, Jeon JM, Pugazhendhi A, Awasthi MK, Kumar D, Kumar G, Yon JJ, Yang YH (2021) An overview on advancements in bio-based transesterification methods for biodiesel production: oil resources, extraction, biocatalysts, and process intensification technologies. Fuel 285:119117. https://doi.org/10.1016/j.fuel.2020.119117

Bhatia SK, Kim SH, Yoon JJ, Yang YH (2017) Current status and strategies for second generation biofuel production using microbial systems. Energy Convers Manag 148:1142–1156. https://doi.org/10.1016/j.enconman.2017.06.073

Bligh EG, Dyer WJ (1959) A rapid method of total lipid extraction and purification. Can J Biochem Physiol 37:911–917. https://doi.org/10.1139/o59-099

Busic A, Kundas S, Morzak G, Belskaya H, Mardetko N, Santek MI, Komes D, Novak S, Santek B (2018) Recent trends in biodiesel and biogas production. Food Technol Biotechnol 56:152–173

Chingakham C, Tiwary C, Sajith V (2019) Waste animal bone as a novel layered heterogeneous catalyst for the transesterification of biodiesel. Catal Lett 149:1100–1110. https://doi.org/10.1007/s10562-019-02696-9

Chiu SY, Kao CY, Huang TT, Lin CJ, Ong SC, Chen CD, Chang JS, Lin CS (2011) Microalgal biomass production and on-site bioremediation of carbon dioxide, nitrogen oxide and sulfur dioxide from flue gas using Chlorella sp. cultures. Biores Technol 102:9135–9142. https://doi.org/10.1016/j.biortech.2011.06.091

Conceicao LRVD, Giordani DS, Castro HFD (2017) Synthesis of biodiesel from macaw palm oil using mesoporous solid catalyst comprising 12-molybdophosphoric acid and niobia. Renew Ener 113:119–128. https://doi.org/10.1016/j.renene.2017.06.008

D'Amato D, Bartkowski B, Droste N (2020) Reviewing the interface of bioeconomy and ecosystem service research. Ambio 49:1878–1896. https://doi.org/10.1007/s13280-020-01374-0

Dobrucka R (2018) Synthesis of MgO nanoparticles using Artemisia abrotanum Herba extract and their antioxidant and photocatalytic properties. Iran J Sci Technol Trans Sci 42:547–555. https://doi.org/10.1007/s40995-017-0421-7

Dong T, Knoshaug EP, Pienkos PT, Laurens LML (2016) Lipid recovery from wet oleaginous microbial biomass for biofuel production: a critical review. Appl Energ 177:879–895. https://doi.org/10.1016/j.apenergy.2016.06.002

Elfeky N, Elmahmoudy M, Zhang Y, Guo J, Bao Y (2019) Lipid and carotenoid production by *Rhodotorula glutinis* with a combined cultivation mode of nitrogen, sulfur, and aluminum stress. Appl Sci 9(12):2444. https://doi.org/10.3390/app9122444

Gaurav K, Srivastava R, Singh R (2013) Exploring biodiesel: chemistry, biochemistry, and microalgal source. Int J Green Energy 10:775–796. https://doi.org/10.1080/15435075.2012.726673

Gupta S, Patel P, Mondal P (2022) Biofuels production from pine needles via pyrolysis: process parameters modeling and optimization through combined RSM and ANN based approach. Fuel 310:122230. https://doi.org/10.1016/j.fuel.2021.122230

Hoseini SS, Najafi G, Sadeghim A (2019) Chemical characterization of oil and biodiesel from common purslane (Portulaca) seed as novel weed plant feedstock. Ind Crops Prod 140:111582. https://doi.org/10.1016/j.indcrop.2019.111582

Huang C, Chen XF, Xiong L, Chen XD, Ma LL, Chen Y (2013) Single cell oil production from low-cost substrates: the possibility and potential of its Industrialization. Biotechnol Adv 31:129–139. https://doi.org/10.1016/j.biotechadv.2012.08.010

IEA RIEA, Paris (2022) https://iea.blob.core.windows.net/assets/ada7af90-e280-46c4-a577-df2e4fb44254/Renewables2022.pdf(accessed on 31 May 2023)

Julkapli NM, Bagheri S (2016) Magnesium oxide as a heterogeneous catalyst support. Rev Inorg Chem 36(1):1–41. https://doi.org/10.1515/revic-2015-0010

Karthik K, Dhanuskodi S, Gobinath C, Prabukumar S, Sivaramakrishnan S (2019) Fabrication of MgO nanostructures and its efficient photocatalytic, antibacterial and anticancer performance. J Photochem Photobiol B 190:8–20. https://doi.org/10.1016/j.jphotobiol.2018.11.006

Katre G, Joshi C, Khot M, Zinjarde S, Ravi Kumar A (2012) Evaluation of single cell oil (SCO) from a tropical marine yeast Yarrowia lipolytica NCIM 3589 as a potential feedstock for biodiesel. AMB Express 2(1):36. https://doi.org/10.1186/2191-0855-2-36

Laurens LML, Markham J, Templeton DW, Christensen ED, Wychen SV, Vadelius EW, Chen-Glasser M, Dong T, Davis R, Pienkos PT (2017) Development of algae biorefinery concepts for biofuels and bioproducts; a perspective on process-compatible products and their impact on cost-reduction. Energy Environ Sci 10:1716–1738. https://doi.org/10.1039/C7EE01306J

Ma Y, Gao Z, Wang Q, Liu Y (2018) Biodiesels from microbial oils: opportunity and challenges. Bioresour Technol 263:631–641. https://doi.org/10.1016/j.biortech.2018.05.028

Manmai N, Unpaprom Y, Ramaraj R (2021) Bioethanol production from sunflower stalk: application of chemical and biological pretreatments by response surface methodology (RSM). Biomass Convers Biorefin 11(5):1759–1773. https://doi.org/10.1007/s13399-020-01236-3

Manohar M, Joseph J, Selvaraj T, Sivakumar D (2013) Application of Box Behnken design to optimize the parameters for turning Inconel 718 using coated carbide tools. Int J Sci Eng Res 4(4):620–644

Maza DD, Vinarta SC, Su Y, Guillamon JM, Aybar MJ (2020) Growth and lipid production of Rhodotorula glutinis R4 in comparison to other oleaginous yeasts. J Biotechnol 310:21–31. https://doi.org/10.1016/j.jbiotec.2020.01.012

Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem 31: 426–428. https://pubs.acs.org/doi/10.1021/ac60147a030.

Moyo LB, Iyuke SE, Muvhiiwa RF, Simate GS, Hlabangana N (2021) Application of response surface methodology for optimization of biodiesel production parameters from waste cooking oil using a membrane reactor. S Afr J Chem Eng 35:1–7. https://doi.org/10.1016/j.sajce.2020.10.002

Musa IA (2016) The effects of alcohol to oil molar ratios and the type of alcohol on biodiesel production using transesterification process. Egypt J Pet 25(1):21–31. https://doi.org/10.1016/j.ejpe.2015.06.007

Nazarpour M, Taghizadeh-Alisaraei A, Asghari A, Abbaszadeh-Mayvan A, Tatari A (2022) Optimization of biohydrogen production from microalgae by response surface methodology (RSM). Energy 253:124059. https://doi.org/10.1016/j.energy.2022.124059

Ogunkunle O, Ahmed NA (2019) A review of global current scenario of biodiesel adoption and combustion in vehicular diesel engines. Energy Rep 5:1560–1579. https://doi.org/10.1016/j.egyr.2019.10.028

Osumi M, Yamada N, Yaguchi H, Kobori H, Nagatani T, Sato M (1995) Ultra-resolution low-voltage SEM reveals ultrastructure of the glucan network formation in fission yeast protoplast. J Electron Microsc 44:198–206. https://doi.org/10.1093/oxfordjournals.jmicro.a050957

Ouraipryvan P, Sreethawong T, Chavadej S (2009) Synthesis crystalline MgO nanoparticle with mesoporous-assembled structure via a surfactant-modified sol-gel process. Mater Lett 63:1862–1865. https://doi.org/10.1016/j.matlet.2009.05.037

Park KC, Whitney C, McNichol JC, Dickinson KE, MacQuarrie S, Skrupski BP, Zou J, Wilson KE, O'Leary SJB, McGinn PJ (2012) Mixotrophic and photoautotrophic cultivation of 14 microalgae isolates from Saskatchewan, Canada: potential applications for wastewater remediation for biofuel production. J Appl Phycol 24:339–348. https://doi.org/10.1007/s10811-011-9677-7

Ramanujam K, Sundrarajan M (2014) Antibacterial effects of biosynthesized MgO nanoparticles using ethanolic fruit extract of Emblica officinalis. J Photochem Photobiol B 141:296–300. https://doi.org/10.1016/j.jphotobiol.2014.10.003

Rebello S, Abraham A, Madhavan A, Sindhu R, Binod P, Babu AK, Aneesh EM, Pandey A (2018) Non-conventional yeast cell factories for sustainable bioprocesses. FEMS Microbiol Lett 365 https://doi.org/10.1093/femsle/fny222

Seffati K, Honarvar B, Esmaeili H, Esfandiari N (2019) Enhanced biodiesel production from chicken fat using CaO/$CuFe_2O_4$ nanocatalyst and its combination with diesel to improve fuel properties. Fuel 235:1238–1244. https://doi.org/10.1016/j.fuel.2018.08.118

Selemani A, Kombe GG (2022) Glycerolysis of high free fatty acid oil by heterogeneous catalyst for biodiesel production. Results Eng 16:100602. https://doi.org/10.1016/j.rineng.2022.100602

Singh RK, Sarkar A, Chakraborty JP (2020) Effect of torrefaction on the physicochemical properties of eucalyptus derived biofuels: estimation of kinetic parameters and optimizing torrefaction using response surface methodology (RSM). Energy 198:117369. https://doi.org/10.1016/j.energy.2020.117369

Spencer ARDSJ, Spencer J, Laluce C (2002) Non-conventional yeasts. Appl Microbiol Biotechnol 58:147–156. https://doi.org/10.1007/s00253-001-0834-2

Suparmaniam U, Lam MK, Uemura Y, Shuit SH, Lim JW, Show PL, Lee KT, Matsumura Y, Le PTK (2019) Flocculation of Chlorella vulgaris by shell waste-derived bioflocculants for biodiesel production: Process optimization, characterization and kinetic studies. Sci Total Environ 134995. https://doi.org/10.1016/j.scitotenv.2019.134995

Suresh J, Pradheesh G, Alexramani V, Sundrarajan M, Hong SI (2018) Green synthesis and characterization of hexagonal shaped MgO nanoparticles using insulin plant (Costus pictus D. Don) leave extract and its antimicrobial as well as anticancer activity. Adv Powder Technol 29:1685–1694. https://doi.org/10.1016/j.apt.2018.03.026

Tamoradi T, Kiasat AR, Veisi H, Nobakht V, Besharati Z, Karmakar B (2021) MgO doped magnetic graphene derivative as a competent heterogeneous catalyst producing biofuels via transesterification: process optimization through Response Surface Methodology (RSM). J Environ Chem Eng 9(5):106009. https://doi.org/10.1016/j.jece.2021.106009

Tan YH, Abdullah MO, Nolasco-Hipolito C, Taufiq-Yap YH (2015) Waste ostrich and chicken-eggshells as heterogeneous base catalyst for biodiesel production from used cooking oil: catalyst characterization and biodiesel yield performance. Appl Ener 160:58–70. https://doi.org/10.1016/j.apenergy.2015.09.023

Tomás-Pejó E, Morales-Palomo S, González-Fernández C (2021) Microbial lipids from organic wastes: Outlook and challenges. Bioresour Technol 323:124612. https://doi.org/10.1016/j.biortech.2020.124612

Vakros J (2018) Biochars and their use as transesterification catalysts for biodiesel production: a short review. Catalysts 8(11):562. https://doi.org/10.3390/catal8110562

Vasić K, Hojnik Podrepšek G, Knez Ž, Leitgeb M (2020) Biodiesel production using solid acid catalysts based on metal oxides. Catalysts 10:237. https://doi.org/10.3390/catal10020237

Wang J, Azam W (2024) Natural resource scarcity, fossil fuel energy consumption, and total greenhouse gas emissions in top emitting countries. Geosci Front 15:101757. https://doi.org/10.1016/j.gsf.2023.101757

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.

Transit Photometry for Estimating the Velocity of Exoplanets and specific Defence Applications

Harshita Bose
Department of Applied Physics
Delhi Technological University
Delhi, India
harshitabose_23ep041@dtu.ac.in

2024 IEEE Space, Aerospace and Defence Conference (SPACE) | 979-8-3503-6738-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/SPACE63117.2024.10667777

***Abstract*—The paper proposes a novel technique for the estimation of the velocities of exoplanets based on transit photometry using the slope of the transition light curve. An analytical expression is derived for the normalized velocity of the exoplanet as a function of the slope of the relative brightness. It is shown that the proposed technique is general, and works both for equatorial and non-equatorial crossings by the exoplanet. The theory is verified using transit data for the Trappist-1 system. The idea is extended to potential defence applications and velocity estimation of resident space objects.**

Keywords—Transit photometry, velocity estimation, exoplanets, resident space objects

I. Introduction

The study of exoplanets is an exciting area of active research involving a variety of techniques [1]. High contrast imaging (HCI) techniques have led to the discovery of new exoplanets and stimulated further studies on exoplanets [2-5]. HCI allows researchers to segregate light reflected from faint planets in the presence of their much-brighter host stars. Post-processing techniques are an integral part of HCI, and are used to improve the images obtained. Transit photometry is another well-known technique used for the discovery of new exoplanets. Transit photometry measures the variation in the amount of starlight received due to the presence of an exoplanet [6-9]. As the exoplanet passes in front of the star, it blocks a portion of the starlight that is observable, and this phenomenon is termed as transit. Alternatively, when the exoplanet passes behind the star, the light from the exoplanet is blocked, leading to a secondary eclipse. The analyses of the change in brightness, i.e., the transit light curve, have led to the discovery of new exoplanets. The most recent is the discovery of 55Cancri e by the James Webb Space Telescope (JWST), which orbits a sun-like star 12.6 parsecs from Earth [10]. In addition to the detection of the exoplanet, other parameters such as the transit duration, period, mass and radius of the exoplanet are also inferred from the data. The expression for the transit duration has been analytically derived in [11]. Researchers have also used machine learning techniques for the detection of exoplanets using HCI and transit photometry [12-13].

To estimate the velocity of exoplanets, the Doppler shift is typically used [14-15]. This paper proposes a novel technique to estimate the velocity of an exoplanet from the transition light curve obtained from transit photometry. In recent years, the number of space debris and Residential Space Objects (RSO) have dramatically increased. The results derived in this paper can be used to estimate the velocities of RSO when passing in front of a bright background.

The paper is organized as follows. Section II develops the theory for estimating the velocity of an exoplanet using transit photometry. Section III validates the theory developed through simulations using data for the Trappist-1 star system. Section IV extends the technique to certain defence and Space Situational Awareness (SSA) applications. Finally, the paper concludes in Section V.

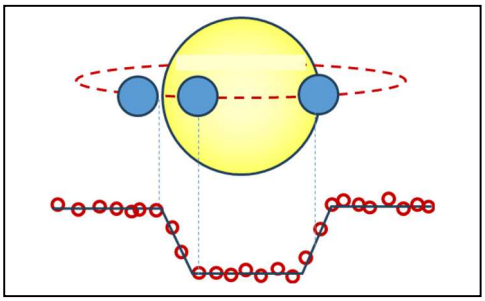


Fig. 1. A typical transition light curve corresponding to the blocking of the starlight due to the exoplanet.

II. Velocity Estimation

The transit photometry technique measures the amount of starlight that is blocked by the exoplanet and uses the transition light curve to estimate the radius of the exoplanet. In this Section an expression is derived linking the velocity of an exoplanet to the transition slope and its radius. A typical transition light curve is given in Fig.1. The radius of an exoplanet, R_P, can be determined using the radius of the star, R_S, and the transition depth, D obtained from the transition light curve, as follows [6]

$$R_P = R_S\sqrt{D} \tag{1}$$

In addition to the transition depth, the slope of the transition also contains information about the radius and the velocity of the exoplanet. Here, a method is presented to estimate the velocity of the exoplanet as a fraction of the velocity of the exoplanet closest to the star, which would have the highest velocity among the exoplanets.

Consider the period of transition pertaining to egress, i.e., when the exoplanet is in the process of blocking lesser amount of light as it is moving out, as depicted in Fig. 2. The

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:33 UTC from IEEE Xplore. Restrictions apply.

discussion in this section pertains only to this transition phase as the exoplanet is moving out. The relative brightness, B, is proportional to the area, A, that has moved out and is no longer blocking the starlight. Both the star and the exoplanet are modelled as circles. It is assumed that the exoplanet is opaque, and emits no light of its own. Since the radius of the star is much larger than that of the exoplanet ($R_S >> R_P$), the curvature of the star is approximated by a straight line. Therefore, the area, A, can be approximated by the segment, as depicted in Fig. 2. Let x represent the maximum width of the segment, $0 \le x \le 2R$. Here, the scenario $x = 0$ represents the case when the exoplanet is about to start the crossover along the equator, while the scenario $x = 2R$ represents the case when the exoplanet has finished the crossover and is no longer blocking the star. The relative brightness, B, is proportional to the segment area, A, i.e., $B = KA$, where K is a constant. The area of the segment, obtained by subtracting the area of the triangle from the sector, is given by

$$A = R_P^2 \cos^{-1}\left(1 - \frac{x}{R_P}\right) - (R_P - x)\sqrt{2R_P x - x^2} \quad (2)$$

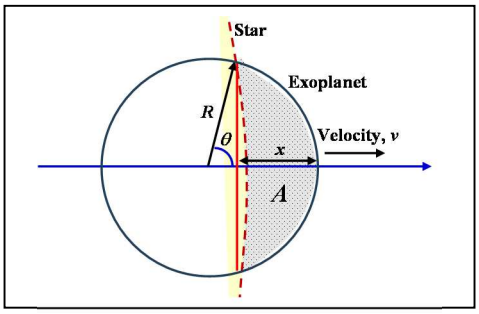


Fig. 2. Transition phase pertaining to egress as the exoplanet moves out. Here, A is the area that is no longer blocking the starlight.

Therefore, the relative brightness during the transition phase can be represented as

$$B = K\left(R_P^2 \cos^{-1}\left(1 - \frac{x}{R_P}\right) - (R_P - x)\sqrt{2R_P x - x^2}\right) \quad (3)$$

Based on (3), the plot of relative brightness during the transition period (egress) is given in Fig. 3(a), and can be compared with the shape of the transition period (egress) of Fig. 1. The slope of the curve for relative brightness during the transition is given by

$$\frac{dB}{dt} = \left(2K\sqrt{2R_P x - x^2}\right)\frac{dx}{dt} = \left(2K\sqrt{2R_P x - x^2}\right)v \quad (4)$$

where v is the velocity of the exoplanet along the x axis. The plot of the slope of relative brightness versus x/R is given in Fig. 3(b). From (4) it is evident that the slope of the relative brightness during the transition is dependent both on the radius and the velocity of the exoplanet. The slope is also a function of x because of the circular shape of the projection of the exoplanet. For constant velocity, the derivative of the slope should be zero, which is obtained for $x = R$, i.e., exactly at the midpoint during the crossover. At this point, the value of the slope reaches a maximum, S_m, and is given by

$$\left(\frac{dB}{dt}\right)_{max} = S_m = K_1 R_P v \quad (5)$$

where K_1 is another constant. Therefore, it is possible to estimate the velocity of the exoplanet, given the radius and the slope of the transition light curve. Consider the transition light curves for two exoplanets in the star system with radii R_{P1} and R_{P2} with (maximum) transition slopes S_1 and S_2. From (5),

$$\frac{v_1}{v_2} = \frac{R_{P2} S_1}{R_{P1} S_2} \quad (6)$$

Thus, the velocity of an exoplanet in the star system can be estimated in relation to the velocity of a reference exoplanet from the transition curve.

It is also possible to estimate the velocity directly from the transit duration. There are multiple definitions of transit duration, as explained in [11]. Let T_1 and T_2 be the transit durations for two exoplanets measured between the midpoints of ingress and egress. Then their velocities v_1 and v_2 can be simply related as

$$\frac{v_1}{v_2} = \frac{T_2}{T_1} \quad (7)$$

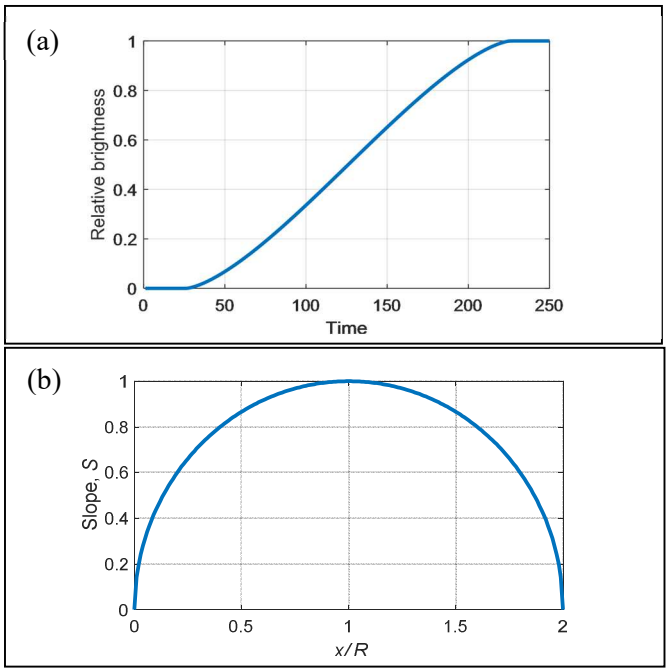


Fig. 3. (a) The plot of relative brightness versus x/R during the transition period (egress) as obtained from (3). (b) the slope of relative brightness versus x/R as obtained from (4).

There is an alternate method to determine the velocity of an exoplanet in a circular orbit. For an exoplanet revolving around the star of mass M_s and orbit radius R_O, its orbital velocity is given by [16]

$$v = \sqrt{\frac{GM_s}{R_O}} \quad (8)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:33 UTC from IEEE Xplore. Restrictions apply.

It should be noted that in (8) the radius refers to the orbital radius, while in (6) the radius refers to the exoplanet radius. Therefore, for two exoplanets in the star system with orbital radii R_{O1} and R_{O2}, and velocities v_1 and v_2, the ratio of the velocities can be expressed as

$$\frac{v_1}{v_2} = \sqrt{\frac{R_{O2}}{R_{O1}}} \tag{9}$$

Thus, the velocity of an exoplanet in the star system can also be estimated in relation to the velocity of a reference exoplanet using the ratio of their orbital radii.

So far it is assumed that the exoplanet crosses the star along the equator. However, that may not always be the case. Next, non-equatorial crossing of a star with radius R_S is considered, as shown in Fig. 4. Let the line of crossing be at a distance d, parallel to the equatorial line. From the figure,

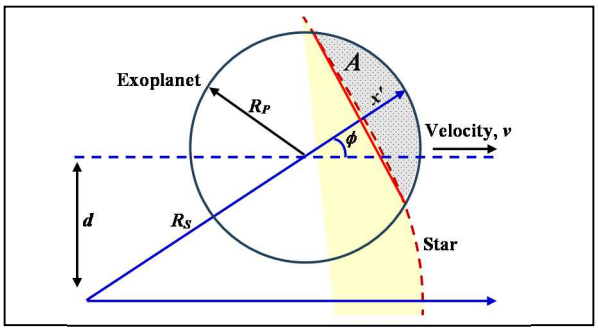


Fig. 4. Non-equatorial crossing of the exoplanet.

$x = x' \cos\phi$ and $v = v(x) = v(x')\cos\phi$, where $\cos\phi = \frac{\sqrt{R_S^2 - d^2}}{R_S} \approx \left(1 - \frac{d^2}{2R_S^2}\right)$ for $d \ll R_S$.

Observing the similarity between the shaded areas in Fig. 2 and Fig. 4, and making the necessary substitutions in (4) the following expression is obtained after some algebra

$$\frac{dB}{dt} \approx \left(2K\sqrt{2R_P x - x^2}\right)\left(1 + \frac{d^2}{R_S^2}\right)\left(1 + \frac{d^2}{2R_S^2}\left(\frac{R_P x - x^2}{2R_P x - x^2}\right)\right)v \tag{10}$$

For $x = R$, i.e., exactly at the midpoint during the crossover,

$$\left(\frac{dB}{dt}\right)_{max} \approx (2KR_P)\left(1 + \frac{d^2}{R_S^2}\right)v \tag{11}$$

Alternatively, it can be written as

$$\left(\frac{dB}{dt}\right)_{max} \approx K_2 R_P v \tag{12}$$

where K_2 is another constant. Thus, for the case of a non-equatorial crossing, the slope of the relative brightness during the transition phase (egress) increases by a factor of $\left(1 + \frac{d^2}{R_S^2}\right)$. Note that the velocity estimation using (6) remains valid even for the non-equatorial case because the ratios of the slopes are used which results in the cancelling out of the factor $\left(1 + \frac{d^2}{R_S^2}\right)$.

III. Validation using Trappist-1 System

In this section, the theory developed in Section II is verified using the data from the Trappist-1 system [6]. To validate (6), the ratios of radii of the exoplanets, as well, the slopes of the transition light curves for the exoplanets are required. The radius of an exoplanet is determined from the transition light curve using (1). The estimated values for the exoplanet radii are given in given in Tab. 1. The slope of the relative brightness during the transition is measured manually from the transition light curves and is also given in Tab. 1. Since the value of the slope varies in the transition phase, the maximum value of the transition slope is used, as suggested by (5). The transition duration required for (7) is measured manually from the transition light curves and is also given in Tab. 1. The measurement of the slope of the relative brightness during the transition and the transition duration can be automated for more accurate and faster estimation. The objective of this exercise is to validate the velocity ratios given by (6) and (7) with respect to (9). For determining the velocity ratio given by (9), the ratio of the orbital radii of the exoplanets is required. The semi-major axis has been used as a substitute for the orbital radii and is reproduced from [6] in Tab. 1. Since Trappist-1b is closest to the star, its velocity is the maximum from (8), and is therefore chosen as the reference velocity. The velocities of the other exoplanets in the star system are estimated as a fraction of the velocity of Trappist-1b.

Fig. 5 gives the plots for the normalized velocities of the exoplanets with respect to Trappist-1b as obtained from (6), (7) and (9). Parameters given in Tab. 1 are used to generate the plots. It can be observed from the figure that there is a fairly good match between the normalized exoplanet velocities obtained from (6), (7) and (9). This motivates us to look at potential applications of transit photometry for defence.

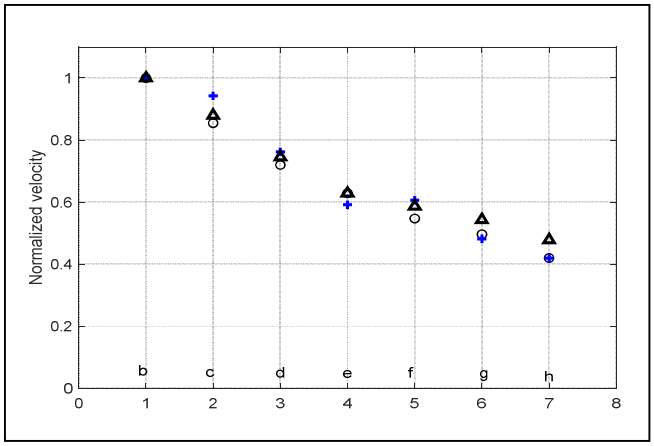


Fig. 5. Plots for the normalized velocities of the exoplanets (1b -1h) in the Trappist system with respect to Trappist-1b as obtained from (6) depicted by '+', (7) depicted by '**Δ**'and (9) depicted by '**o**'.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:33 UTC from IEEE Xplore. Restrictions apply.

TABLE I. EXOPLANET PARAMETERS DERIVED FROM [6]

	Trappist-1 System						
Exoplanets	***b***	***c***	***d***	***e***	***f***	***g***	***h***
Radius, R_P (R_{Earth})	1.09	1.06	0.77	0.92	1.05	1.13	0.76
Slope, S	2.4	2.2	1.3	1.2	1.4	1.2	0.7
Trans. Duration, T ($\times 10^{-3}$)	22	25	29	35	38	41	46

IV. DEFENCE APPLICATIONS

Space Situational Awareness (SSA) involves the detection and tracking of man-made objects in space. The detection of artificial satellites is critical for defence applications both from a strategic point of view and SSA [17-20]. There are multiple techniques for the detection of RSOs including LiDAR, RADAR and optical systems [21-22]. There are space-borne sensors too, but are constrained by their limited field of view.

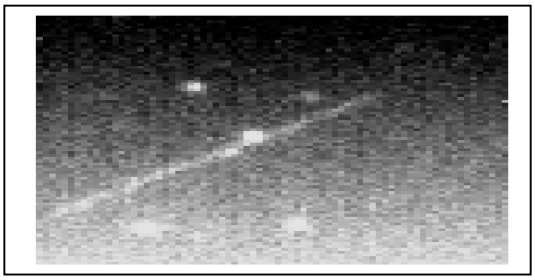


Fig. 6. A typical image showing a 'streak' caused by an artificial satellite.

The proposed technique in this paper can be used effectively for the detection of RSOs passing in front of a bright background, for example the sun, the moon or a distant star. Artificial satellites have a higher angular velocity than the background stars, and tend to appear as 'streaks'. A typical 'streak' is shown in Fig. 6. It is observed that, in this image, the streak passes in front of a bright star. Thus, the theory developed in this paper can be used to extract/ confirm the velocity and size related information about the satellite.

V. CONCLUSION

The paper presents a simple yet effective technique for the estimation of normalized velocities of exoplanets from their transition light curve. The technique focusses on the transition duration (ingress/ egress) and derives an analytical expression linking the velocity of the exoplanet with the slope of the transition light curve. Expressions are derived both for equatorial crossing and non-equatorial crossing, and the proposed velocity estimation technique works for both cases. The theory is validated using velocity estimations based on the orbital radii of the exoplanets in the Trappist-1 system. The theory developed in this paper can be used to estimate the size and velocity of RSOs and artificial satellites, and can be extended to other defence applications.

REFERENCES

[1] K. B. Follette, "An Introduction to High Contrast Differential Imaging of Exoplanets and Disks," *Publications of the Astronomical Society of the Pacific*, vol. 135, no. 1051, pp. 1, Sep. 18, 2023.

[2] M. Houllé, A. Vigan, A. Carlotti, É. Choquet, F. Cantalloube, et al., "Direct imaging and spectroscopy of exoplanets with the ELT/HARMONI high-contrast module," *Astronomy & Astrophysics*, vol. 652, no. A67, pp. 1-14, 2021.

[3] W. Thompson and C. Marois, "Improved Contrast in Images of Exoplanets Using Direct Signal-to-noise Ratio Optimization," *Astronomical Journal*, vol. 161, no. 5, pp. 236, April 26, 2021.

[4] D. Redding, E. Sidick, J. Z. Lou, B. Dube, and S. Basinger, "Integrated active control of coronagraph and telescope for high contrast imaging," in *Proceedings of Techniques and Instrumentation for Detection of Exoplanets XI*, vol. 12680, 5 October 2023.

[5] E. Sidick et al., "Design of Vector-Vortex Coronagraphs for Exoplanet Science using a Six Meter Space Telescope," in *Proceedings of IEEE Aerospace Conference*, Big Sky, MT, USA, 4-11 March 2023.

[6] M. Gillon, A. Triaud, B.-O. Demory, et al., "Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1," *Nature*, vol. 542, pp. 456-460, Feb. 2017.

[7] J. K. Barstow and P. G. J. Irwin, "Habitable worlds with JWST: transit spectroscopy of the TRAPPIST-1 system?" *Mon. Not. R. Astron. Soc.*, vol. 461, pp. L92-L96, Nov. 2016.

[8] A. E. Simon, G. M. Szabó, K. Szatmáry, and L. L. Kiss, "Methods for exomoon characterization: combining transit photometry and the Rossiter–McLaughlin effect," *Monthly Notices of the Royal Astronomical Society*, vol. 406, no. 3, pp. 2038-2046, Aug. 2010.

[9] D. A. Caldwell, "The Kepler Mission: Zeroing in on habitable Earths," *IVEC 2012*, Monterey, CA, USA, 2012, pp. 3-6.

[10] R. Hu, A. Bello-Arufe, M. Zhang, et al., "A secondary atmosphere on the rocky Exoplanet 55 Cancri e," *Nature*, May 8, 2024.

[11] D. M. Kipping, "Investigations of approximate expressions for the transit duration," *Mon. Not. R. Astron. Soc.*, vol. 407, pp. 301-313, 2010.

[12] C. A. Gomez Gonzalez, O. Absil, and M. Van Droogenbroeck, "Supervised detection of exoplanets in high-contrast imaging sequences," *Astron. & Astroph.*, vol. 613, no. A71, pp. 1-13, May 2018.

[13] A. Malik, B. P. Moster, and C. Obermeier, "Exoplanet detection using machine learning," *Monthly Notices of the Royal Astronomical Society*, vol. 513, no. 4, pp. 5505-5516, Dec. 2021.

[14] M. S. Khan, J. S. Jenkins, and N. B. Yoma, "Discovering new worlds: a review of signal processing methods for detecting exoplanets from astronomical radial velocity data," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. January, Jan. 2017.

[15] N. Jovanovic, N. Cvetojevic, O. Guyon, F. Martinache, and J. Lawrence, "Detecting exoplanets with extreme adaptive optics and a single-mode fibre fed spectrograph," in *2013 Conference on Lasers & Electro-Optics Europe & International Quantum Electronics Conference CLEO EUROPE/IQEC*, Munich, Germany, 2013, pp. 1-1.

[16] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 10th ed. New York: Wiley and Sons, 2014.

[17] B. Chen et al., "Spot the GEO Satellites: From Dataset to Kelvins SpotGEO Challenge," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, Nashville, TN, USA, 2021.

[18] John A Kennewell and Ba-Ngu Vo, "An overview of space situational awareness", *Proceedings of the 16th International Conference on Information Fusion*, 2013.

[19] J. Z. Gazak, M. Phelps, R. Swindle, C. Shaw, Z. Funke, and J. Fletcher, "The potential for satellite recognition from ground-based filter photometry," in *Proc. SPIE 12519, Algorithms, Technologies, and Applications for Multispectral and Hyperspectral Imaging XXIX*, vol. 125190A, 13 June 2023.

[20] R. Clark, Y. Fu, S. Dave, and R. S. K. Lee, "Resident space object (RSO) attitude and optical property estimation from space-based light curves," *Advances in Space Research*, vol. 70, no. 11, pp. 3271-3280, 2022, ISSN 0273-1177.

[21] V. Suthakar, A. A. Sanvido, R. Qashoa, and R. S. K. Lee, "Comparative Analysis of Resident Space Object (RSO) Detection Methods," *Sensors*, vol. 23, no. 24, pp. 9668, Dec. 7, 2023.

[22] R. Qashoa and R. Lee, "Classification of Low Earth Orbit (LEO) Resident Space Objects' (RSO) Light Curves Using a Support Vector Machine (SVM) and Long Short-Term Memory (LSTM)," *Sensors*, vol. 23, p. 6539, 2023.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:33 UTC from IEEE Xplore. Restrictions apply.

Water Air Soil Pollut (2024) 235:690
https://doi.org/10.1007/s11270-024-07486-7

Ultrasonic Assisted Synthesis of $CuFe_2O_4$-Ag infused Gum Hydrogels Nanocomposite for photocatalytic Degradation of Organic Dye from Wastewater

Meenu · Pratibha Sharma · Roli Purwar · Navita Sharma · Pooja Rawat · Seema R. Pathak · Atul Thakur · Hyunook Kim · Monu Verma · Sudip Majumder · Chandra Mohan Srivastava

Received: 1 May 2024 / Accepted: 1 September 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract The reusability of photocatalyst and bactericidal properties plays a crucial role in water pollutant management. Herein, copper ferrite-silver ($CuFe_2O_4$-Ag) nanomaterial incorporated Gum Arabic Acrylamide (GAA) hydrogel nanocomposites was fabricated via free radical polymerization method to enhance the photocatalytic efficiency and antimicrobial assay. The structural, morphological, compositional, thermal, optical, and physico-chemical properties of the synthesized hydrogel nanocomposite were analyzed using various techniques. A decrease in direct band gap value as 1.83 eV for $CuFe_2O_4$-Ag nanocomposite was observed which would further enhance the photocatalytic ability of nanocomposite. The results suggested that $CuFe_2O_4$-Ag nanocomposite was infused inside the matrix of GAA hydrogel. The swelling ability of fabricated hydrogel nanocomposite was found to be 1328%. The photocatalytic degradation followed a pseudo-first-order kinetic reaction model, achieving a 90% removal of methylene blue dye using the nanocomposite, with a rate constant value of 14.09×10^{-3} min^{-1} Moreover, photocatalyst maintain 71% degradation efficiency over three consecutive cycles. The antibacterial effects on *Escherichia coli* (*E. coli*) were also examined, and the $CuFe_2O_4$-Ag nanocomposite showed good efficacy, as indicated by the measured zones of inhibition. Hence, the nanocomposite-infused hydrogel demonstrates significant potential as an advanced material for the remediation of water pollution, effectively targeting both organic and biological contaminants.

Keywords $CuFe_2O_4$-Ag Hydrogel Nanocomposite · Dye Adsorption · Photocatalytic dye degradation · Reusability · Antimicrobial Activity

Meenu · P. Sharma · N. Sharma · P. Rawat · S. R. Pathak · C. M. Srivastava (✉)
Department of Chemistry, Biochemistry and Forensic Science, Amity School of Applied Sciences, Amity University Haryana, Gurugram 122413, India
e-mail: cms.srivastava@gmail.com

R. Purwar
Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

A. Thakur
Centre for Nanotechnology, Amity University Haryana, Gurugram 122413, India

H. Kim · M. Verma
Department of Environmental Engineering, Water-Energy Nexus Laboratory, University of Seoul, Seoul 02504, Republic of Korea

M. Verma
Department of Food Science and Technology, Graphic Era (Deemed to Be University), Dehradun, Uttarakhand 248002, India

S. Majumder
International Institute of Innovation and Technology, Kolkata 700156, India

Published online: 18 September 2024

1 Introduction

Today one of the major challenges for science is the treatment of polluted water containing industrial waste (Li et al., 2014). One of the main class of pollutants in industrial waste is colored effluents i.e., dyes. Industries like dyeing, textile, coloring, printing, paper mills, cosmetics, etc. are the main contributors of such colored effluents (Zhang et al., 2014). Dyes are organic compounds that are used to impart color (Benkhaya et al., 2020; Lanjwani et al., 2024). Due to resistance towards biodegradation and stability, effective elimination of organic dyes from industrial wastewater has been a major concern for researchers (Ji et al., 2015; Ma et al., 2014). Discharge of dye in water can affect the aquatic ecosystem due to its carcinogenic and mutagenic effects (Lellis et al., 2019). For wastewater management, several conventional methods have been used such as biological treatment, chemical treatment including ozone treatment, chemical oxidation, and physical treatment like coagulation-flocculation, ion exchange, photocatalysis, and adsorption (Bulut et al., 2015, Haleem et al., 2023). Out of these, the adsorption method is more efficient due to simplicity in design, cost effectiveness, and ease of operation (Bagheri et al., 2019; Kannaujia et al., 2019).

In the past few years, significant efforts have been made in the development of hydrogels as adsorbing material for dye (Yan et al., 2010). Adsorption of dye on the surface of hydrogel and their photocatalytic degradation is more efficient than other techniques (Dil et al., 2019a, 2019b; Kannaujia et al., 2019). Hydrogels are a three-dimensional polymeric material that shows the capability to swell and hold water (to a significant extent) within its structure without dissolving in water (Ahmed et al., 2015); Kasgoz et al., 2008). Due to electrostatic and hydrophilic interactions between cationic dye molecules and hydrogel, dye molecules are easily escaped into hydrogel matrices (Yan et al., 2010). In the recent past, 2-Acrylamide-2-methyl propane sulfonic acid sodium/clay Hydrogel, poly(acrylamide)/sepiolite-hydrogel composites, and lignin-based hydrogels have been used for the elimination of various dyes (Kasgoz et al., 2008; Ekici et al., 2006; Dominguez-Robles et al., 2018; Dil et al., 2016). Although hydrogels have better dye adsorption capability, but major drawbacks are their non-biodegradability as adsorbed dye poses toxicity in the environment when discarded because it only adsorbs the dye molecules and does not contribute to their degradation process. Hence, the degradation of adsorbed dye appears to be the optimal solution. In past few years, significant effort has been given to the development of nanomaterials like Fe_3O_4 NPs, $CuFe_2O_4$ NPs, TiO_2 NPs, Y_2O_3 NPs, Mn-doped CuO-NPs etc. for the degradation of dyes (Zhu et al., 2010; Mahmoodi, 2011; Abou-Gamra et al., 2015; Sharifpour et al., 2019). However, nanocomposites have been proven to be a better catalyst for dye removal with tunable band gap (Kanwal et al., 2023; Pethsangave et al., 2020). Ce-doped TiO_2 nanoparticles loaded on activated carbon were used for adsorption of Basic Red 46 dye (Dil et al., 2019a, 2019b), GO-Ag nanocomposite enhances catalytic activity towards MB dye (Kumari et al., 2020). Silver nanoparticles combined with yttrium oxide (Ag-Y_2O_3 nanocomposite) were used for the removal of heavy metal {Cu (II) and Cr (VI)} from aqueous solution (Pradhan et al., 2019). Hence, the hypothesis for this study was to fabricate a ternary nanocomposite where a nanocomposite with a tunable band gap will be ultrasonically distributed over the surface of hydrogel which will prove to exhibit excellent photocatalytic properties. Since there will be greater adsorption of dye on the surface of hydrogel and nanocomposite present will help in the degradation of adsorbed dye molecules into simpler molecules. Hence, photocatalytic degradation using such ternary nanocomposite will be more efficient than other techniques.

Here in this work, we have reported the adsorption and degradation properties of copper ferrite-silver ($CuFe_2O_4$-Ag) nanocomposite incorporated gum Arabic acrylamide (GAA) hydrogel for dye degradation using a selected dye model, methylene blue (MB) solution. Nano-crystalline copper ferrite ($CuFe_2O_4$), have been described to show significant dye degradation properties and can separate easily due to their magnetic nature (Priya et al., 2019). Similarly, Ag NPs due to their high surface area exhibit an enhanced reactivity towards the degradation of dye (Jemal et al., 2017). Further, to improve the efficacy of these nanomaterials, nanocomposites were prepared. In the present work, we are synthesizing a novel $CuFe_2O_4$-Ag nanocomposite doped hybrid gum Arabic acrylamide hydrogel for the degradation of MB dye from wastewater. Thus, it was thought worthwhile to explore the absorption

and degradation properties of copper ferrite-silver nanocomposite incorporated hybrid hydrogel for dye degradation using a selected dye model, methylene blue (MB).

2 Materials and Methods

$AgNO_3$, $CuCl_2.2H_2O$, $FeCl_3.6H_2O$, NaOH, urea, and glucose were purchased from Sigma Aldrich. Polyvinylpyrrolidone (PVP), MB dye, Acrylamide, Ammonium persulfate (APS) and N, N' methylene bisacrylamide (MBA) were purchased from Merck, India. All chemicals are used as such without any further purification.

2.1 Synthesis of Silver Nanoparticles (Ag NPs)

Silver nanoparticles were synthesized by reduction of $AgNO_3$ in polyvinylpyrrolidone solution where NaOH was used as a reaction accelerator (Wang et al., 2005). To the 50 ml of distilled water add an equal quantity (0.5 g) of glucose, PVP, and NaOH and heat over a magnetic stirrer at 60 ℃. To this solution, add already prepared silver nitrate solution (3.4 g of $AgNO_3$ in 10 mL of distilled water) dropwise. After the complete addition of this solution, the mixture was continued to stir for another 15 min. Finally, the mixture was centrifuged to separate the obtained precipitates which were further washed and dried in an oven (Kumari et al., 2020).

2.2 Synthesis of Copper ferrite Nanoparticles ($CuFe_2O_4$NPs)

A solution of $CuCl_2.2H_2O$ (2.796 g) and $FeCl_3.6H_2O$ (8.864 g) in 250 ml distilled water was prepared and stirred continuously. To this solution, added NaOH solution (11.5 g NaOH in 50 mL distilled water) dropwise over 20 min with constant stirring (Kannaujia et al., 2019). Brown precipitates appeared immediately showing the formation of $CuFe_2O_4$ NPs (Mahmoodi, 2011). The mixture was filtered using Whatman filter paper, washed with distilled water, and dried in an oven at 85 °C for 14 h. The prepared NPs were calcinated in a muffle furnace at 500 °C for 5 h.

2.3 Synthesis of Copper ferrite-silver Nanocomposite ($CuFe_2O_4$-Ag)

$CuFe_2O_4$-Ag nanocomposite was prepared by adding $AgNO_3$ (1.06 g), Urea (0.53 g) and prepared copper ferrite nanoparticles (1 g) in 40 ml distilled water (Solution 1) and stirred for 30 min on a magnetic stirrer at 70 ℃. Another solution of NaOH (0.5 g), PVP (0.5), and glucose (0.5 g) in 10 ml distilled water was prepared (solution 2). Solution 2 was added dropwise in solution 1 over 15 min with constant stirring. Black-colored precipitates appeared which were filtered using Whatman filter paper and washed several times with water. $CuFe_2O_4$-Ag nanocomposite was obtained after drying (Wang et al., 2019).

2.4 Synthesis of pure Acrylamide hydrogel

Pure Acrylamide hydrogel was prepared by taking 7.5 g of Acrylamide (monomer), 0.375 g of Ammonium persulphate (initiator) [5% of monomer], and 0.375 g of MBA (cross-linker) [5% of monomer] in 100 ml distilled water. The solution was heated at 60 ℃ in a water bath for 15 min with constant stirring. A gel-like product was formed which was cut into pieces on a polystyrene sheet and dried at 60 ℃ for 3 days.

2.5 Synthesis of Gum Arabic Acrylamide (GAA) Hydrogel

GAA hydrogel was prepared by taking 7.5 g of Acrylamide (monomer), 0.375 g of Ammonium persulfate (initiator) [5% of monomer], and 0.375 g of MBA (cross-linker) [5% of monomer] (Jafarigol et al., 2020) and incorporating different concentration of Acacia gum 10%, 20%, 30%, 40%(w/w), respectively in 100 ml distilled water. The solution was heated at 60℃ on a water bath for 1 h with constant stirring till all acacia gum was completely dissolved. Then the solution was allowed to stand for some time, a gel-like product was formed which was cut into pieces on a polystyrene sheet and dried at 60℃.

2.6 Synthesis of $CuFe_2O_4$-Ag Nanocomposite incorporated Gum Arabic Acrylamide Hydrogel

$CuFe_2O_4$-Ag nanocomposite (1% of monomer w/w) in 100 ml distilled water was sonicated for 30 min so that

nanoparticles were finely dispersed in water. Different concentrations of acacia gum (10%, 20%, 30%, 40% w/w) respectively were added to the nanocomposite-containing solution and heated in a water bath till all the acacia gum was completely dissolved. The solution containing 7.5 g of Acrylamide (monomer), 0.375 g of Ammonium persulfate (initiator) [5% of monomer], and 0.375 g of MBA (cross-linker) [5% of monomer] was prepared and poured into nanocomposite containing solution with constant stirring and ultra-sonication and allowed to stand for sometimes yielded a gel-like product which was cut into pieces on polystyrene sheet and dried at 60 ℃ (Purwar et al., 2014).

2.7 Swelling studies

The swelling study of hydrogel samples was evaluated as per the standard specified in ASTM D570 (Majumder et al., 2020; Zheng et al., 2021). The dried sample of pure Acrylamide hydrogel, GAA hydrogel [10%, 20%, 30%, 40% (w/w)] respectively and $CuFe_2O_4$-Ag incorporated GAA hydrogel [10%, 20%, 30%, 40%(w/w)] were firstly weighted *(WD)* and then immersed more than distilled water at room temperature for a definite period in different beakers and marked them. For measuring the absorption rate, after a fixed time interval (5, 10, 15, 30, and 60 min) samples were carefully removed from the water with the help of forceps, soaked on filter paper, and weighted (*WS*). The weight measurements were repeated until the weight remained constant (saturation point reached). The swelling index (Si) was calculated using the equation

$$\text{Si}\ (\%) = \frac{WS - WD}{WD} \times 100$$

2.8 Photocatalytic degradation of dye

Photocatalytic degradation is the method used for the elimination of environmentally hazardous pollutants from industrial water. Here, applications of copper ferrite nanoparticles ($CuFe_2O_4$ NPs), silver nanoparticles (Ag NPs), Copper ferrite-silver nanocomposite ($CuFe_2O_4$-Ag), and hydrogels as a photocatalyst under UV – Vis light for degradation of methylene blue (MB) dye solution have been reported. Batch experiments were performed for $CuFe_2O_4$ NPs, Ag NPs, $CuFe_2O_4$-Ag nanocomposite, GAA hydrogel, and $CuFe_2O_4$-Ag nanocomposite containing GAA hydrogel using 30 mL of MB dye solution (30 mg/L). Catalysts (nanoparticles and hydrogel) were added (1 mg/ml) to the dye solution and kept in a dark condition for half an hour to establish adsorption – desorption equilibrium. Afterward, solutions were exposed to UV – Visible radiation. At fixed time intervals aliquots (2 mL) were taken out, centrifuged, and recorded the absorbance using a UV–Vis spectrophotometer.

2.9 Reusability Experiment

The photocatalytic process becomes more economical when the photocatalyst can be reused without much loss of efficiency. To check the reusability of prepared catalysts, catalysts were recovered by centrifugation which separates the catalyst and adsorbed dye. The recovered sample was again utilized for the degradation process with afresh stock solution. The same experiments were repeated three times consecutively, every time with fresh stock solution and constant experimental conditions (Kanwal et al., 2023).

2.10 Antimicrobial assay

The disc diffusion assay was performed as per the protocol reported (AATCC 30) earlier against *E. coli* (Naseer et al., 2020). The agar gel plate was formed by pouring a hot and sterile solution of agar and Luria broth into 90 mm diameter Petri dishes and left untouched till cooling. A sterile paper disc was impregnated with $CuFe_2O_4$-Ag and $CuFe_2O_4$-Ag doped hydrogel nanocomposites were placed over the agar gel plate which had been cultured with *E. coli* and pressed gently to ensure close contact with the agar gel plate and sample impregnated paper disc. These gel plates were incubated at 37 °C in an oven for 12 h.

3 Characterization

Structural properties of synthesized samples were analyzed by PXRD using Bruker Discover-D8 high-resolution diffractometer with 10 mm Cu slit and

using a Cu Kα radiation of 1.54 Å wavelengths. Samples were scanned from 2θ value 5 to 80 degrees. The average particle size of the prepared samples was calculated using Debye-Scherer's equation, $D = K\lambda/\beta\cos\theta$, where D is the average crystallite size, λ is the X-ray wavelength, and β is line broadening in radians (full width at half maximum of the peak), and θ is Bragg's angle. K is constant (Geometric factor = 0.94). For FTIR analysis, the prepared particles were dried, crushed with KBr pellets, and examined using the Perkin-Elmer FTIR spectrometer model 2000. HR-TEM analysis of synthesized samples was carried out using an FEI Technai G2-20 electron microscope operating at 200 kV. SEM–EDS analysis was carried out by Zeiss EVO-50 and FE-SEM Quanta-200 FEG microscope. The direct band gap of samples was evaluated by UV–Vis lambda 365 Perkin Elmer. through dynamic light scattering technique through Malvern zeta sizer ver.7.11. Thermal analysis of samples was evaluated using Simultaneous Thermal Analyzer (Perkin Elmer, STA-8000) as per the standard specified in ASTM E1877, 793 (Singh et al., 2012). The values of Zeta potential and surface charge of synthesized particles were assessed through Zetasizer Nano ZS (Malvern-RDET 48125). A suspension of 1 mg/mL in milli-Q water was prepared and sonicated on probe a sonicator to determine the zeta potential value. Photocatalytic degradation of dye was carried out on an Agilent double beam UV − Visible spectrophotometer run at 1 nm resolution from 200 to 800 nm.

3.1 Statistical analysis

Data are reported as means and standard deviations (SDs).

4 Results and Discussion

4.1 Powder X-ray Diffraction (PXRD) Analysis

PXRD pattern of synthesized $CuFe_2O_4$NPs, Ag NPs, and $CuFe_2O_4$-Ag nanocomposite are shown in Fig. 1. The PXRD pattern $CuFe_2O_4$ nanoparticles shows a strong peak at 2θ = 35.16° value that corresponds to the plane (311). Additionally, many other peaks of $CuFe_2O_4$ at 2θ = 30.1°, 38.0°, 45.4°, 53.6°, 56.6° and 62.38° were observed which can be attributed to the (220), (222), (400), (422), (511) and (440) plane, respectively. PXRD pattern shows that prepared $CuFe_2O_4$ NPs were well matched with cubic spinel $CuFe_2O_4$ (JCPDS No. 25–0283) (Priya et al., 2019). The PXRD pattern of the synthesized Ag NPs exhibited four distinct diffraction peaks at 2θ values of 38.37°, 44.4°, 64.83° and 77.70°, respectively, corresponding to the (111), (200), (220) and (311) crystalline planes of cubic Ag (JCPDS cards 04–0783) (Jemal et al., 2017). This data confirmed that the synthesized Ag NPs consisted of pure silver with high crystallinity and face-centered cubic structure. The PXRD spectrum of $CuFe_2O_4$-Ag nanocomposite as presented in Fig. 1 revealed peaks of both the copper ferrite and Ag NPs. This spectrum confirms the successful synthesis of nanocomposite. Interestingly, the peaks of $CuFe_2O_4$ in the Cu ferrite-silver nanocomposite have a much lower intensity than the bare $CuFe_2O_4$ nanoparticles due to the Ag NPs coating on the surface of copper ferrite. This spectrum confirms the support of Ag NPs on the surface of copper ferrite nanoparticles which was further confirmed using SEM–EDS analysis in the subsequent sections (Kooti et al., 2013).

The crystallite size of all the synthesized nanoparticles was calculated using the Debye–Scherrer equation from the most intense peak in the respective sample as reported in Table 1. Moreover, the strain of synthesized samples was calculated using the Stokes-Wilson equation $\varepsilon_{str} = \beta/4\tan\theta$ and reported in Table 1. It was observed that strain within the lattice decreased with an increase in the size of the crystal lattice which indicates lattice expansion. It is well fact that lattice deformation due to increased strain resulted in peak broadening and hence decreased crystallite size (Sharma et al., 2021). Additionally, d-spacing was calculated by using Braggs equation, $n\lambda = 2d\sin\theta$ which also varies according to the crystallite size.

4.2 Fourier Transform Infrared Spectroscopy (FTIR) Analysis

Figure 2 displayed the FT-IR spectrum for all the fabricated samples. The broad absorption band of all spinel ferrite was observed in the range of 600–500 cm^{-1}. Hence, bands in this region correspond to Cu–O and Ag–O vibrations for all the samples (Kooti et al., 2013). The FT-IR spectrum of copper ferrite nanoparticles displayed peak positions

between 570 cm^{-1} and 848 cm^{-1} corresponding to the vibration of the metal–oxygen bond in the tetrahedral region (Jemal et al., 2017). The peak position around 1631 cm^{-1} corresponds to the stretching vibration of the C=O group and H–O–H bond. The peak at 2370 cm^{-1} was due to the stretching mode of C-O arising from the adsorption of atmospheric CO_2 on the surface of the nanoparticles (Jemal et al., 2017). The spectrum for the nanocomposite exhibits all the peaks from both Ag and $CuFe_2O_4$ nanoparticles.

4.3 High-resolution Transmittance Electron Microscopy (HRTEM) Analysis

Morphological analysis of all the synthesized nanoparticles was done using HRTEM as reported in Fig. 3. From Fig. 3a, it was observed that $CuFe_2O_4$ NPs are tetragonal in shape (Jemal et al., 2017), and due to their magnetic nature, particles are agglomerated as represented by darker parts in the image. A histogram of $CuFe_2O_4$ NPs shows the average particle size of nanoparticles is 62 nm (as calculated using ImageJ software). Figure 3b indicates that Ag NPs were agglomerated and spherical with an average particle size of 77 nm. Figure 3c displayed the micrograph of $CuFe_2O_4$-Ag nanocomposite that has confirmed the successful synthesis of nanocomposite where spherical Ag NPs were attached to the surface of tetragonal $CuFe_2O_4$ NPs which are in good agreement with PXRD data. SAED pattern as reported in Fig. 3d-f also confirmed the crystalline nature of all the synthesized samples with d-spacing with d-spacing as reported in Table 1 corresponds to (111), (311), and (200) plane for Ag NPs, $CuFe_2O_4$ NPs, and $CuFe_2O_4$-Ag nanocomposite which was calculated using ImageJ software matched well with the d-spacing as calculated using PXRD results.

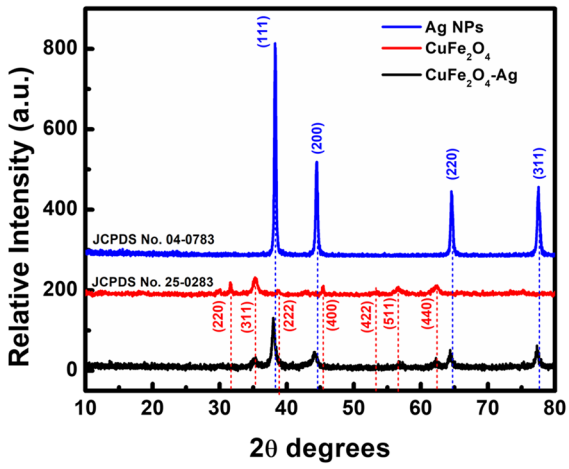


Fig. 1 Comparative PXRD patterns for $CuFe_2O_4$NPs, Ag NPs, and $CuFe_2O_4$-Ag nanocomposite

4.4 SEM–EDS Analysis

The topographical analysis of $CuFe_2O_4$-Ag nanocomposite and $CuFe_2O_4$-Ag nanocomposite ultrasonically distributed over GAA hydrogel were analyzed using SEM–EDS color mapping as reported in Fig. 4. It was observed from the figure that synthesized nanocomposite was spherical with partial agglomeration which decreased when nanocomposite was ultrasonically distributed in GAA hydrogel which was observed from Fig. 4e where smaller particles of nanocomposite were uniformly distributed over the larger particles of the polymer. This observation was further confirmed by EDS color mapping images as reported in Fig. 4b and 4f where different colors corresponding to each element were uniformly distributed over the entire surface. Hence, the ultrasonic method was efficient in uniformly distributing nanocomposite over the hydrogel surface for its better functionalization for practical applicability as discussed in a later section of this paper. Moreover, the composition of both samples was confirmed using EDS where peaks for each element present in the samples were obtained that confirmed the successful synthesis of nanoparticles (Fig. 4c and 4g). The quantitative elemental analysis is also reported in Fig. 4d and 4h which showed the maximum weight % of Ag in nanocomposite which was also related to the PXRD data of nanocomposite where intensity of the Ag peaks is more than $CuFe_2O_4$ peaks. On the other hand, the weight % of oxygen was found to be maximum in Fig. 4h due to the presence of GAA hydrogel.

Table 1 Structural parameters for all synthesized nanoparticles

Structural Parameters	Ag NPs	$CuFe_2O_4$	$CuFe_2O_4$-Ag nanocomposite
Crystalline size (nm)	35.35	11.96	15.89
Strain	0.0033	0.0095	0.0067
d-spacing (Å)	2.35	2.54	2.36
Particle size (nm)	77	62	57
d-spacing (Å) using TEM	2.34	2.66	2.13

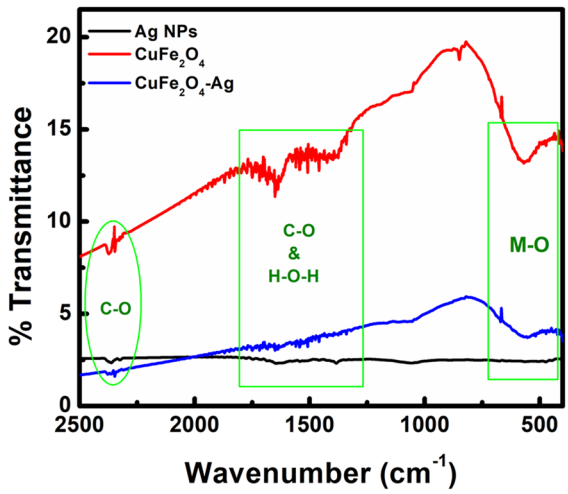


Fig. 2 FTIR Spectra of $CuFe_2O_4$ NPs, Ag NPs, and $CuFe_2O_4$-Ag nanocomposites

4.5 Zeta Potential

Zeta potential is a measure of charge on the surface of nanoparticles. Zeta potential values for $CuFe_2O_4$NPs, Ag NPs, and $CuFe_2O_4$-Ag nanocomposite were -11.0 mV, -8.1 mV, and -14.0 mV, respectively. The higher negative zeta potential value of nanocomposite suggested that $CuFe_2O_4$-Ag nanocomposite exhibits better colloidal stability than $CuFe_2O_4$ NPs and Ag NPs. This could be explained on basis of decrease in the agglomeration of $CuFe_2O_4$-Ag nanocomposite providing colloidal stability (Serrano-Lotina et al., 2023). Moreover, a higher negative zeta potential value supports lower particle size which helps to discrete the nanoparticles to avoid agglomeration. This was also confirmed from PXRD and SEM–EDS which clearly indicates lower crystalline size and reduced agglomeration, respectively.

4.6 Swelling Studies

Figure 5a illustrates the % Swelling of pure Acrylamide hydrogel (AH) and GAA hydrogel [10%, 20%, 30%, 40% (w/w)], respectively with time. As the concentration of acacia gum in hydrogel increases, the water-absorbing capability of hydrogel also enhanced which was due to presence of hydrophilic groups of acacia gum biopolymer (Singh et al., 2017). Due to the presence of hydrophilic groups, cross-linked density increases, and the porosity of material also increases. Hydrogel which contains 40% of acacia gum shows maximum swelling percentage as compared to other concentrations of acacia gum. Figure 5b depicted the % swelling of $CuFe_2O_4$-Ag nanocomposite incorporated gum Arabic acrylamide hydrogel [10%, 20%, 30%, 40% (w/w)] respectively with time. This graph also follows the same trend i.e., with an increase in concentration swelling increases, that was due to presence of nanocomposite which makes the material more porous and hence the water holding capability of material become enhanced. The maximum swelling for 40% GAA hydrogel and 40% GAA hydrogel containing $CuFe_2O_4$-Ag nanocomposite was found to be 1208% and 1328%, respectively.

4.7 Thermo gravimetric Analysis (TGA) and Differential Scanning Calorimetry (DSC)

TGA and DSC thermograms of GAA hydrogel and $CuFe_2O_4$-Ag nanocomposite incorporated GAA hydrogel were shown in Fig. 6a and b, respectively. TGA thermogram shows that both the hydrogels exhibit weight loss in two different ranges of temperature. The first weight loss was in the vicinity of 110 ℃ which corresponds to the evaporation of water. Due to hydrophilicity and supramolecular structure water molecules were encapsulated into natural polymer-based hydrogels, even after drying (Paulino et al., 2012). The second weight loss was very prominent and occured in the temperature range of 200–310 ℃ due to the breakage of glycosidic bonds and decomposition (Rusu et al., 2015). After the incorporation of $CuFe_2O_4$-Ag nanocomposite into GAA hydrogel, the nature of the thermogram was slightly changed. The first weight loss was at 115℃ due to the evaporation of water. The second weight loss was in the temperature range of 200–325 ℃ indicating that the thermal stability of hydrogel was slightly improved by the incorporation of $CuFe_2O_4$-Ag nanocomposite. Similarly, the DSC curve of Gum Arabic acrylamide (GAA) hydrogel and $CuFe_2O_4$-Ag nanocomposite incorporated GAA hydrogel was shown in Fig. 6a and b, respectively. DSC curve of GAA hydrogel & $CuFe_2O_4$-Ag nanocomposite incorporated GAA hydrogel display a prominent endothermic transition peak at 330 ℃ and 332 ℃ respectively, which were due to thermal decomposition of polymeric material. After analyzing the DSC curve, it was shown that the thermal stability of

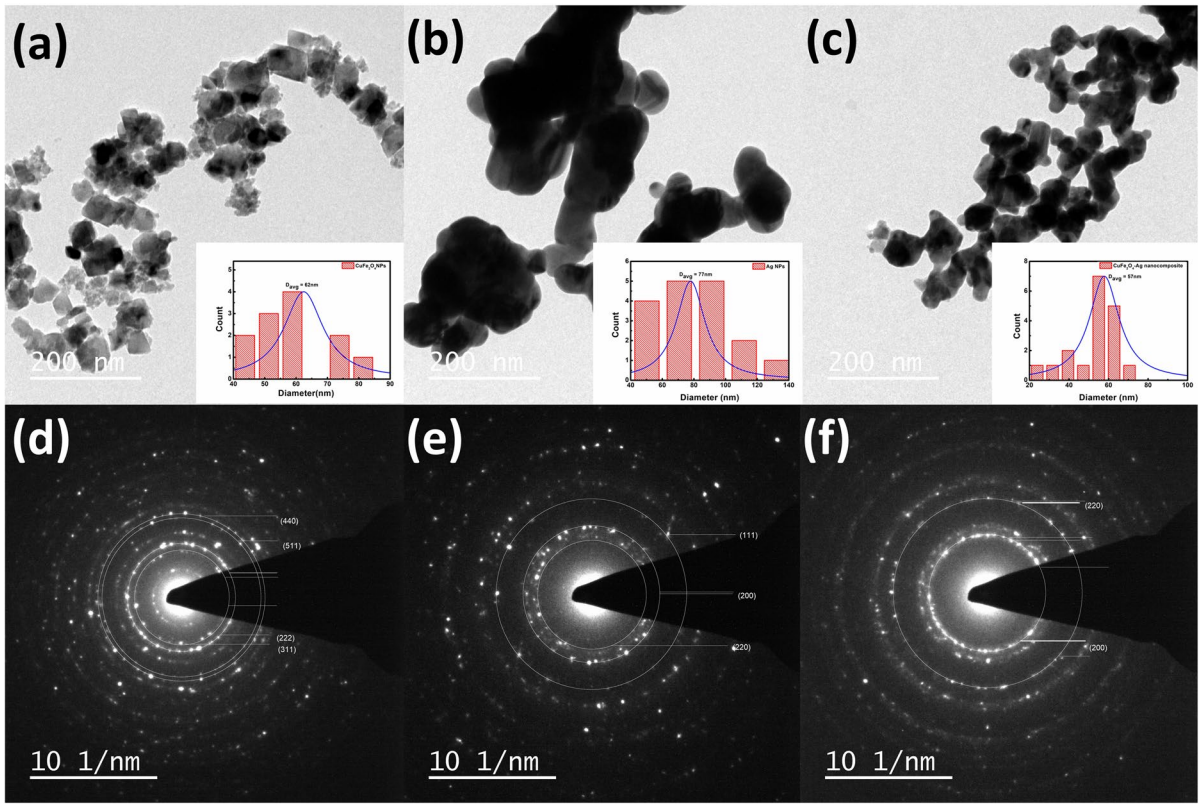


Fig. 3 HR-TEM micrographs and SAED pattern of (**a**, **d**) $CuFe_2O_4$ NPs, (**b**, **e**) Ag NPs, and (**c**, **f**) $CuFe_2O_4$-Ag nanocomposite

hydrogel was slightly increased after the incorporation of the nanocomposite.

4.8 Diffuse Reflectance Spectroscopy (DRS)

The optical properties of the synthesized Ag NPs, $CuFe_2O_4$ NPs, and $CuFe_2O_4$-Ag nanocomposite were studied using diffuse reflectance spectroscopy where the direct band gap was determined using Tauc plot as reported in Fig. 7 which was plotted by converting the reflectance spectra into absorption spectra using Kubelka–Munk Function

$$F(R\infty) = \frac{K}{S} = \frac{\left(1 - R_\infty\right)^2}{2R_\infty}$$

where, $R_\infty = \frac{R_{sample}}{R_{standard}}$ is the reflectance, and K and S are the absorption and scattering coefficients, respectively. Hence, the Tauc plot was plotted using the relation $(F(R\infty).h\upsilon)^{1/\gamma} = A(h\upsilon - E_g)$, where, h is Planck's constant, υ is photon frequency, A is constant, and E_g is band gap with γ factor is ½ for the direct transitions, respectively. On extrapolation of this curve to the x-axis provide the direct band gap values as 2.19 eV, 2.07 eV, and 1.83 eV for Ag NPs, $CuFe_2O_4$ NPs, and $CuFe_2O_4$-Ag nanocomposite, respectively (Fig. 7). A decrease in the band gap values was observed for the nanocomposite which would further enhance the ability of nanocomposite to be a better photocatalyst as proved under Section 4.9.

4.9 Photocatalytic Degradation of Methylene Blue Dye

To demonstrate the practical applicability of all the fabricated samples, photocatalytic degradation of methylene blue dyes under UV–visible radiation was carried out. Methylene Blue, a heterocyclic aromatic compound is widely used in textile industries (Vidhu et al., 2014). In the UV–visible spectrum absorption band of MB dye normally appears at 665 nm corresponding to the n-π* transition (Wang et al., 2019).

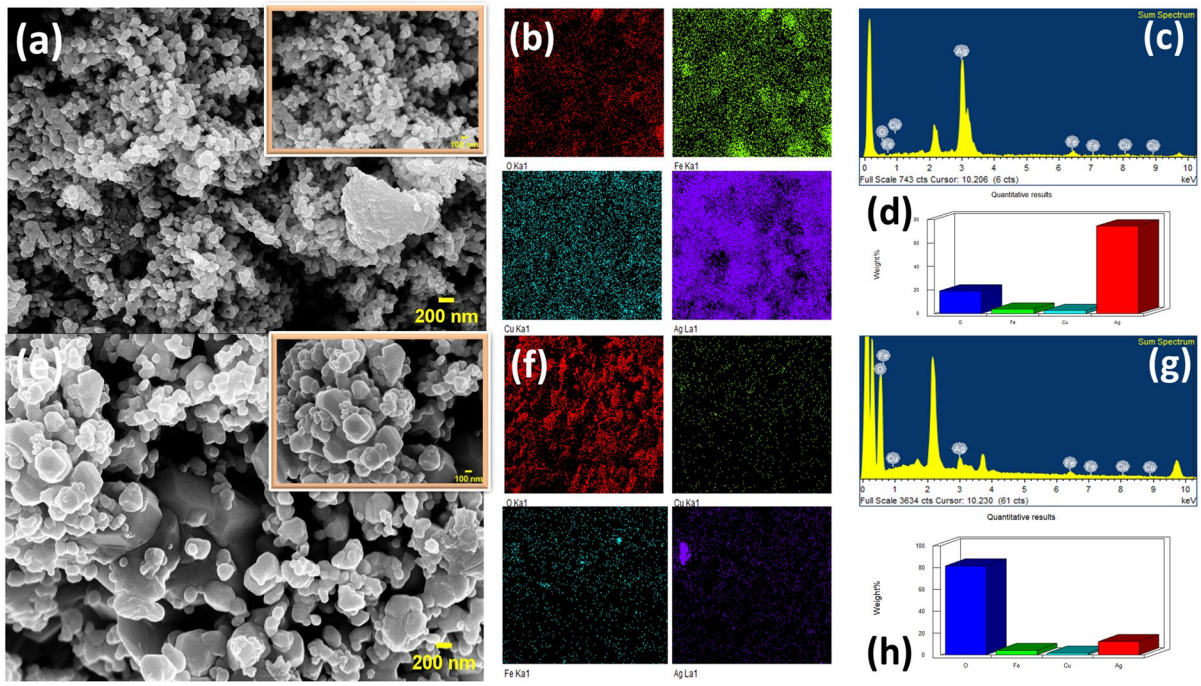


Fig. 4 Morphological analysis using SEM–EDS for $CuFe_2O_4$-Ag nanocomposite (**a**, **b**, **c** and **d**) and $CuFe_2O_4$-Ag ultrasonically distributed into GAA hydrogel (**e**, **f**, **g**, and **h**). High resolution images are shown within the inset of image (**a**) and (**e**)

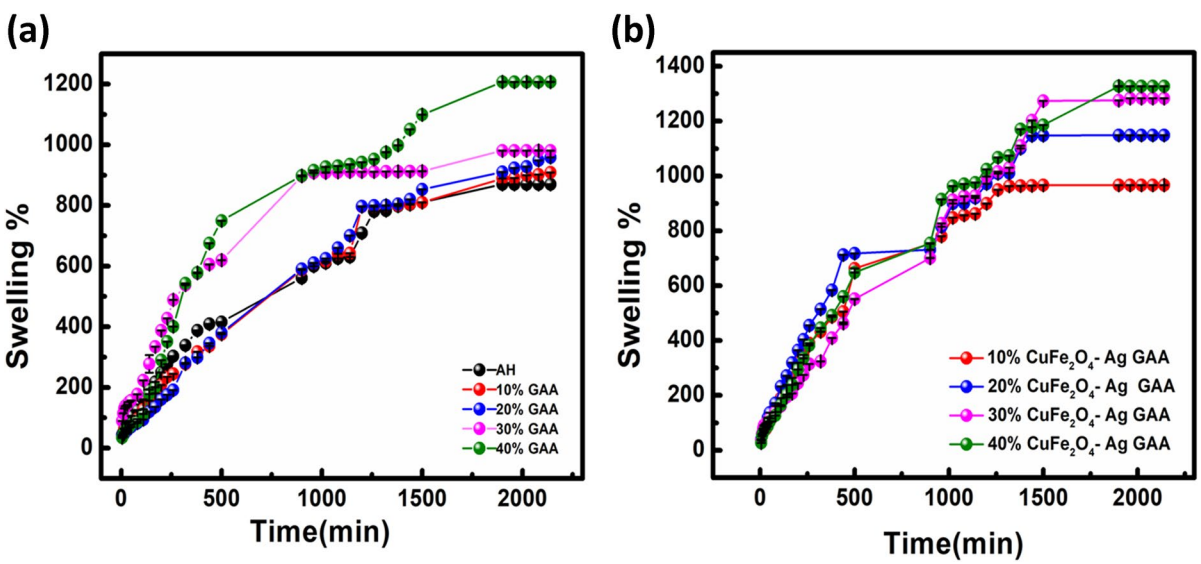


Fig. 5 Swelling (%) graph for (**a**) Pure Acrylamide Hydrogel (AH) & Gum Arabic Acrylamide (GAA) hydrogel (10%, 20%, 30%, 40%(w/w)) (**b**) $CuFe_2O_4$-Ag containing GAA hydrogel (10%, 20%, 30%, and 40%(w/w))

The photodegradation ability of synthesized samples was evaluated by examining successive lowering of the intensity of absorption over a while. Photolysis of pure dye solution does not display any significant change in their absorbance spectra. The percentage degradation of MB dye concerning time using

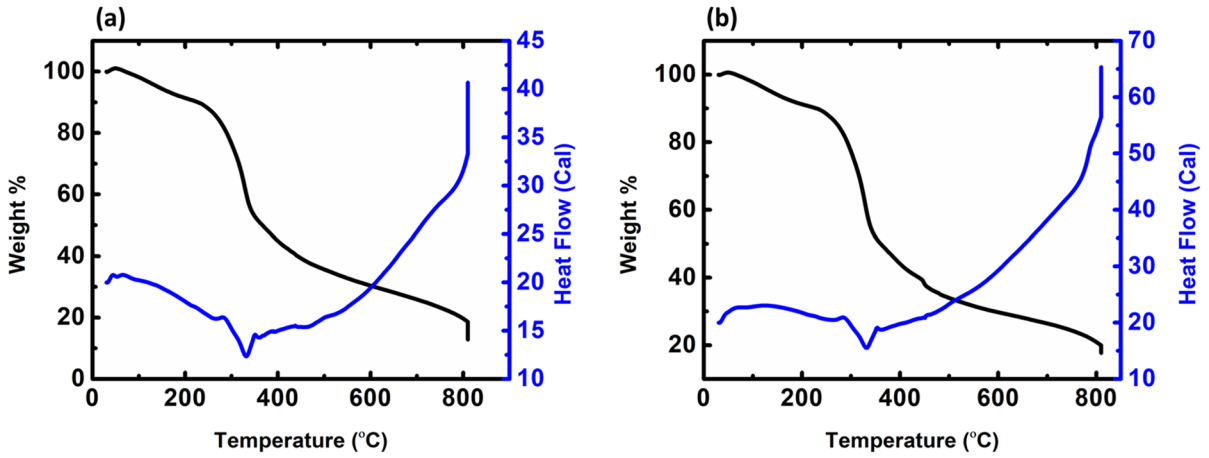


Fig. 6 TGA and DSC analysis of (**a**) GAA hydrogel (**b**) $CuFe_2O_4$-Ag containing GAA hydrogel

synthesized samples as photocatalyst was evaluated as shown in Fig. 8a and reported in Table 2. After the addition of the catalyst, the graph does not show any appreciable change in absorbance under dark for almost 30 min, which proves that they were more reactive in light. After the UV–visible light exposure, a decrease in the absorption value of MB dye was observed. Figure 8b shows the C_t/C_o plot concerning time, for $CuFe_2O_4$-Ag nanocomposite containing GAA hydrogel degrade the MB dye up to 90% while individual degradation of GAA hydrogel, $CuFe_2O_4$-Ag nanocomposite, $CuFe_2O_4$NPs, and Ag NPs was up to 83%, 75%, 63%, and 58%, respectively.

Kinetic studies were also performed to measure the photocatalytic ability of the synthesized samples. The kinetic plot of ln (C_o/C_t) versus time as shown in Fig. 9 displayed the linear behavior, which confirms pseudo-first-order kinetics with rate constants as reported in Table 2 for all the samples. Hence, it was observed from rate constant values that $CuFe_2O_4$-Ag nanocomposite containing GAA hydrogel displays maximum catalytic activity among others at a faster pace. Our results concord with the finding reported in literature (Özçelik et al., 2023; Miri et al., 2018). Copper ferrites degrade the 55.4% dye, and pure Ag degrade up to 70% MB dye, but when their nanocomposite have been incorporated into hydrogel photocatalytic efficiency enhanced up to 90%. Table 3 summarizes dye degradation outcomes reported in literature, offering a reference point for evaluating the performance of our prepared samples. This comparative analysis of different photocatalyst underscore the effectiveness of different approaches and contextualizes our results.

4.10 Reusability Experiment

The photocatalytic process becomes more economical when the photocatalyst can be reused without much loss of efficiency. The results of reusability experiments are presented in Fig. 10a which showed that the efficiency of $CuFe_2O_4$-Ag nanocomposite for dye removal was greater than $CuFe_2O_4$ and Ag nanoparticles after three successive cycles. The result indicated

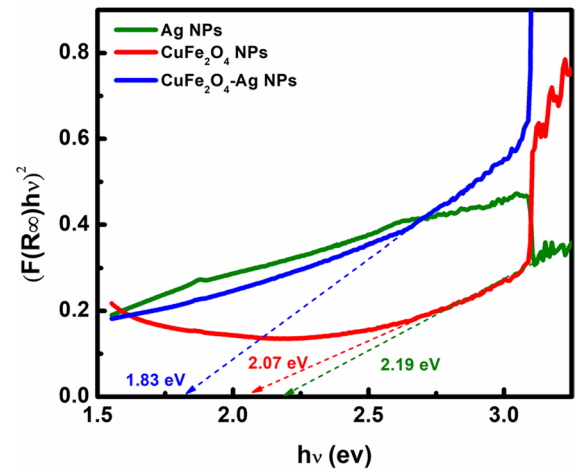


Fig. 7 Tauc plot showing the direct band gap of the synthesized samples using Kubelka–Munk function

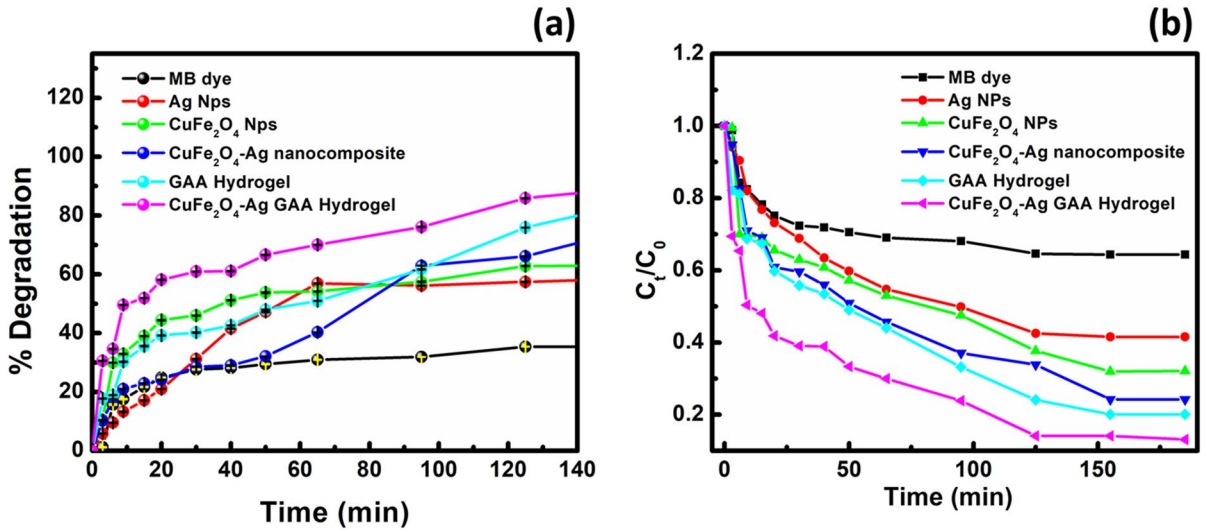


Fig. 8 (**a**) (%) Dye degradation versus time (min), and (**b**) C_t/C_0 versus time (min) for all the synthesized nanoparticles

that synthesized adsorbents, as a reusable material, can be continually applied for the removal of dye. Photocatalytic degradation using nanocomposite was found to be 71% after three successful cycles. Figure 10b depicts a similar PXRD pattern for recycled $CuFe_2O_4$ NPs, Ag NPs, and $CuFe_2O_4$-Ag nanocomposite showing that there was no structural change observed in recovered samples.

4.11 Mass Spectrometry (MS) Analysis

Figure 11a, b, and c depict the Total Ion Chromatogram (TIC) of $CuFe_2O_4$ NPs, Ag NPs, and $CuFe_2O_4$-Ag nanocomposite. The main aim of mass spectroscopy was to find out the reaction intermediates of dye using different literature that explains fragments of MB dye under photodegradation (Gnaser et al., 2005). There were no peaks observed at m/z 314 amu as the parent component which indicates that there was no compound that remained intact. Figure 11a showed the peaks at m/z 284 amu (MB cation), 256 amu (Azure A), 226 amu, 214 amu, 158 amu, 125 amu, 90 amu, and 84 amu that correspond to the photodegraded fragments as shown in Scheme 1. In addition to the main peaks, there were several low-mass peaks which were either due to organic contamination or fragmentation. Figure 11b and c showed the peaks at m/z 284 amu, 214 amu, 158 amu, 141 amu, 125 amu, 110 amu, 84 amu, 284 amu, 214 amu, 125 amu, 110 amu, and 84 amu, respectively that corresponds the photodegraded fragments as shown in Scheme 2 and 3. The change in residual MB concentration on reaction with photocatalyst

Table 2 Some important parameters for practical applicability of all the fabricated nanoparticles for photocatalytic dye degradation and antimicrobial assay using disc diffusion method

Sample	% degradation	Rate constant (min^{-1})	R^2	Zone of inhibition (mm) against *E. Coli*
Methylene Blue dye	$35 \pm 1.6 \times 10^{-2}$	$2.63 \times 10^{-3} \pm 1.94 \times 10^{-4}$	0.93	-
Ag NPs	$58 \pm 2.2 \times 10^{-1}$	$5.42 \times 10^{-3} \pm 4.30 \times 10^{-4}$	0.92	11
$CuFe_2O_4$ NPs	$63 \pm 5.7 \times 10^{-2}$	$5.64 \times 10^{-3} \pm 6.18 \times 10^{-4}$	0.88	6
$CuFe_2O_4$-Ag Nanocomposite	$75 \pm 8.8 \times 10^{-2}$	$8.63 \times 10^{-3} \pm 3.44 \times 10^{-4}$	0.98	12.2
GAA hydrogel	$83 \pm 5.5 \times 10^{-2}$	$10.58 \times 10^{-3} \pm 3.75 \times 10^{-4}$	0.98	0
$CuFe_2O_4$-Ag-GAA hydrogel	$90 \pm 1.07 \times 10^{-1}$	$14.09 \times 10^{-3} \pm 4.83 \times 10^{-4}$	0.99	10

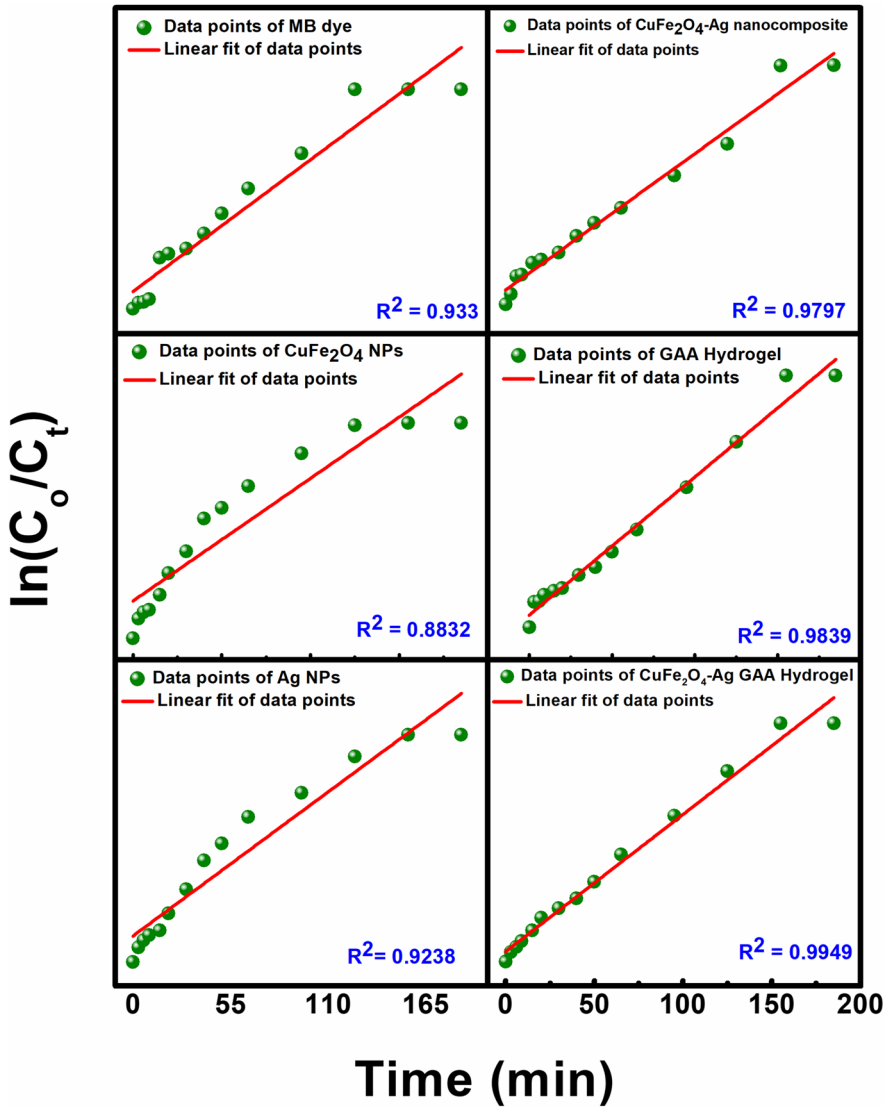


Fig. 9 Pseudo first order plots for MB dye using $CuFe_2O_4$ NPs, Ag NPs, $CuFe_2O_4$-Ag nanocomposite, GAA hydrogel and $CuFe_2O_4$-Ag nanocomposite containing GAA hydrogel

Table 3 A comparative study table of catalytic dye degradation through nanoparticles

Catalyst	UV/non-UV	Dye	% Degradation	References
Copper ferrite NPs	UV	Methylene Blue (MB)	55.4%	Özçelik et al., 2023
AgNPs	UV	Methylene Blue (MB)	70%	Miri et al., 2018
Rosemary and Cherry Laurel modified PVA nanocomposite film	Non—UV	Methylene Blue (MB)	61.3%	Yontar et al., 2022
polyacrylamide/TiO_2 (PAM/TiO_2) nanocomposites	UV	MG dye	76%	Ali et al., 2020
zero valent iron @guar gum *crosslinked* soya lecithin nanocomposite hydrogel	Non-UV	Methyl Violet	81%	Sharma et al., 2018
TiO_2-containing Hydrogel	UV	Acid Orange 7	71%	Bahram et al., 2017

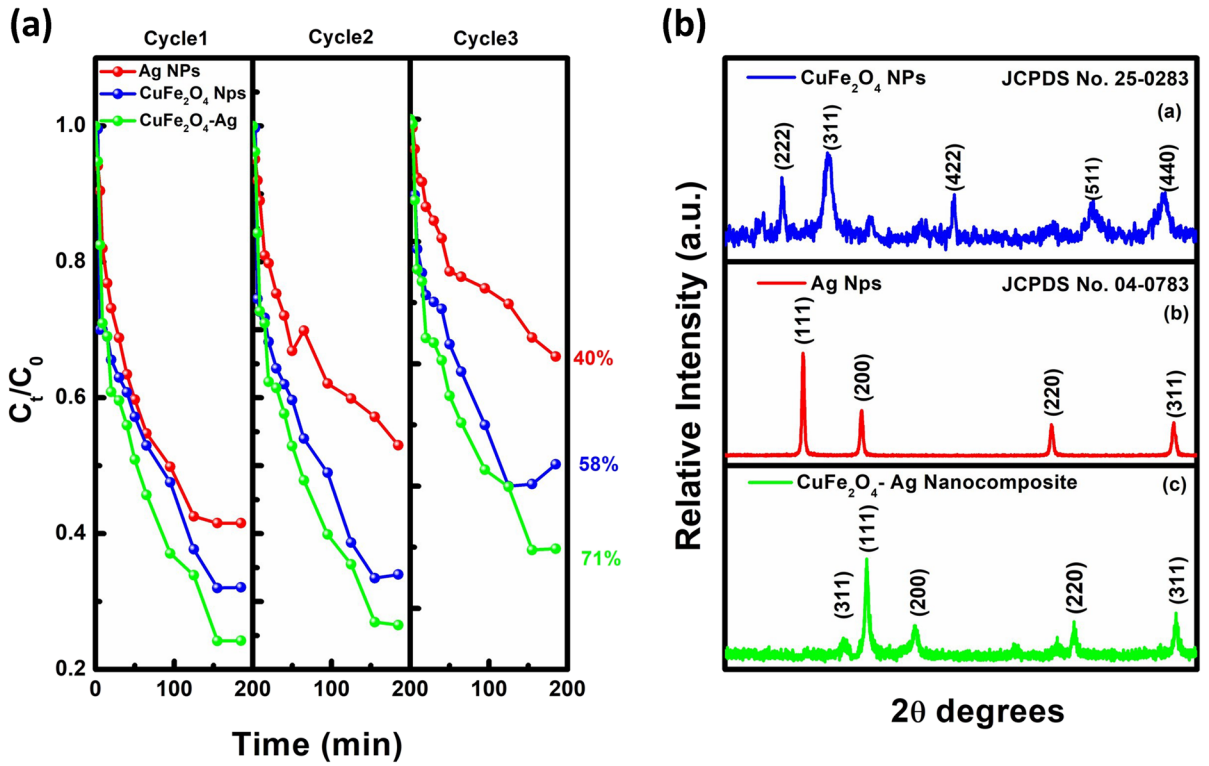


Fig. 10 Reusability graph (**a**) and comparative PXRD (**b**) of Ag NPs, $CuFe_2O_4$ NPs and $CuFe_2O_4$-Ag nanocomposite for three successive cycles for degradation of methylene blue dye

was measured by UV–Vis spectroscopy. Based on the degradation pathway, we can conclude that UV–Vis irradiation caused a significant amount of methylene blue to degrade at the surface of photocatalysts.

4.12 Antimicrobial Analysis

The antimicrobial activity of $CuFe_2O_4$-Ag and $CuFe_2O_4$-Ag doped hydrogel nanocomposite was further evaluated by disc diffusion assay and results

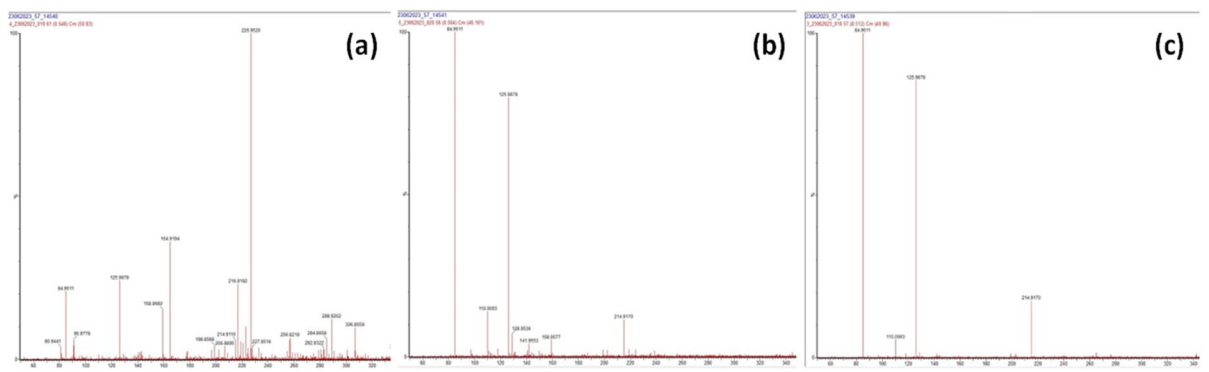
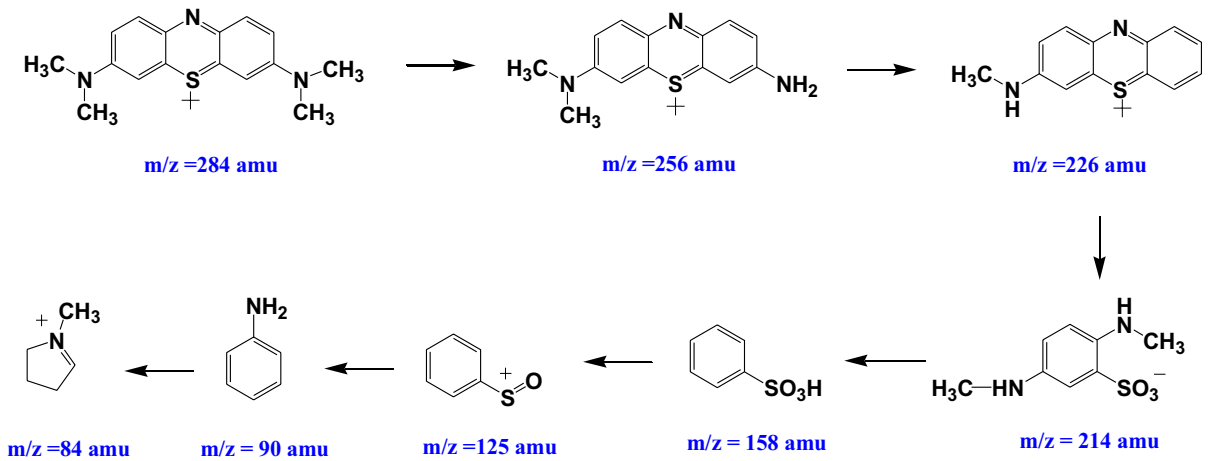
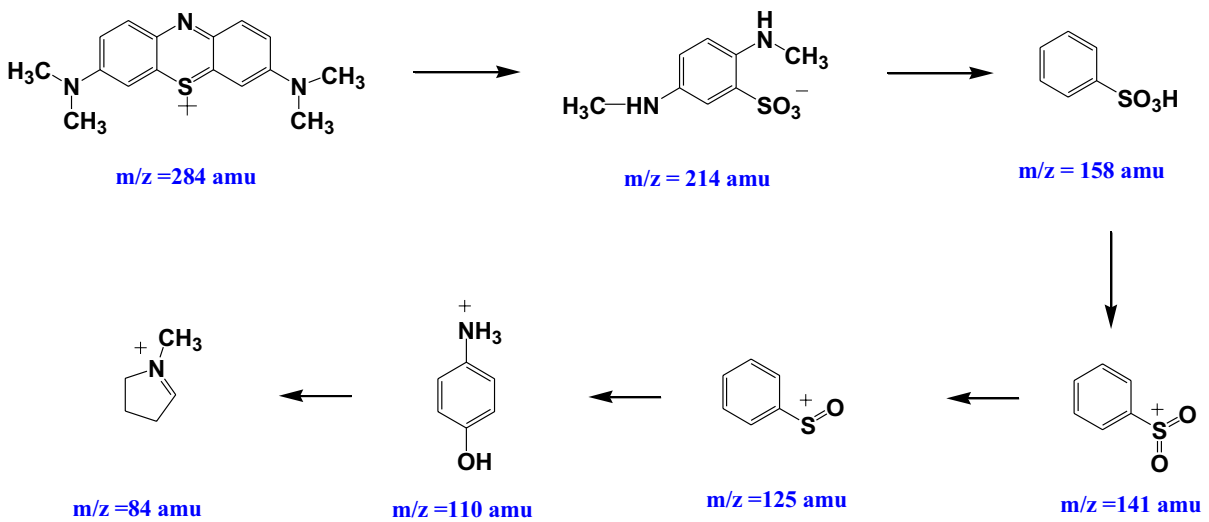


Fig. 11 Mass Spectra of methylene blue dye after the degradation by (**a**) $CuFe_2O_4$ NPs (**b**) Ag NPs (**c**) $CuFe_2O_4$-Ag Nanocomposite

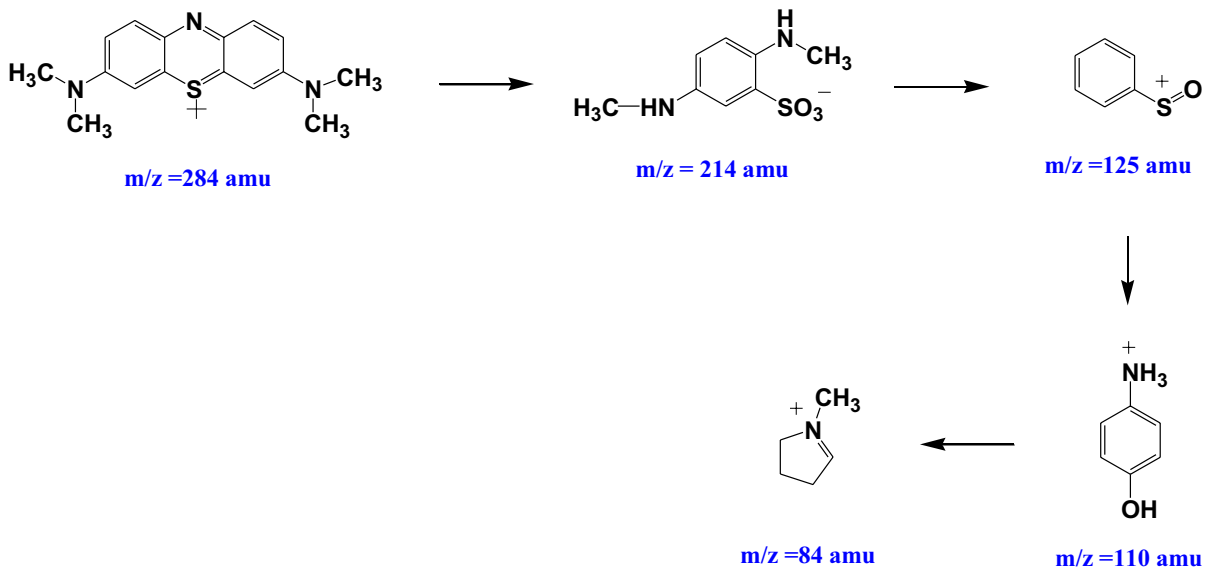


Scheme 1 Possible reaction pathway of photodegradation of Methylene Blue with $CuFe_2O_4$ NPs

are depicted in Fig. 12. From Fig. 12, $CuFe_2O_4$-Ag nanocomposite shows good antimicrobial activity with a prominent (12 mm) zone of inhibition as compared to $CuFe_2O_4$-Ag doped hydrogel nanocomposite with a 10 mm zone of inhibition (Table 2). There was not any zone of inhibition formed for the pure hydrogel sample. The higher antimicrobial activity of $CuFe_2O_4$-Ag nanocomposite over the doped hydrogel nanocomposite was because nanomaterial came in direct contact with the bacterial cell wall and disintegrated it either forming pit or by altering the permeability of the cell wall membrane but hydrogel nanocomposite the nanomaterial presents only on the surface of hydrogel interact with bacterial membrane either by direct contact or by leaching. The bactericidal activity of $CuFe_2O_4$-Ag doped hydrogel nanocomposite materials would be beneficial in treating the contaminated industrial effluent having organic dye and pathogenic microorganism.



Scheme 2 Possible reaction pathway of photodegradation of Methylene Blue with Ag NPs



Scheme 3 Possible reaction pathway of photodegradation of Methylene Blue with $CuFe_2O_4$-Ag nanocomposite

5 Conclusions

In this work, $CuFe_2O_4$ NPs, Ag NPs, and Copper ferrite-silver ($CuFe_2O_4$-Ag) nanocomposite have been synthesized. Additionally, acacia gum containing hydrogel (GAA) and Copper ferrite-silver ($CuFe_2O_4$-Ag) nanocomposite incorporated GAA hydrogel have been prepared by conventional methods to improve the adsorption and photocatalytic performance of the nanocomposite. Results suggested that all the synthesized nanoparticles are crystalline in nature and nanocomposite was successfully synthesized where Ag NPs are embedded over the $CuFe_2O_4$ NPs as confirmed from PXRD, FTIR, SEM,

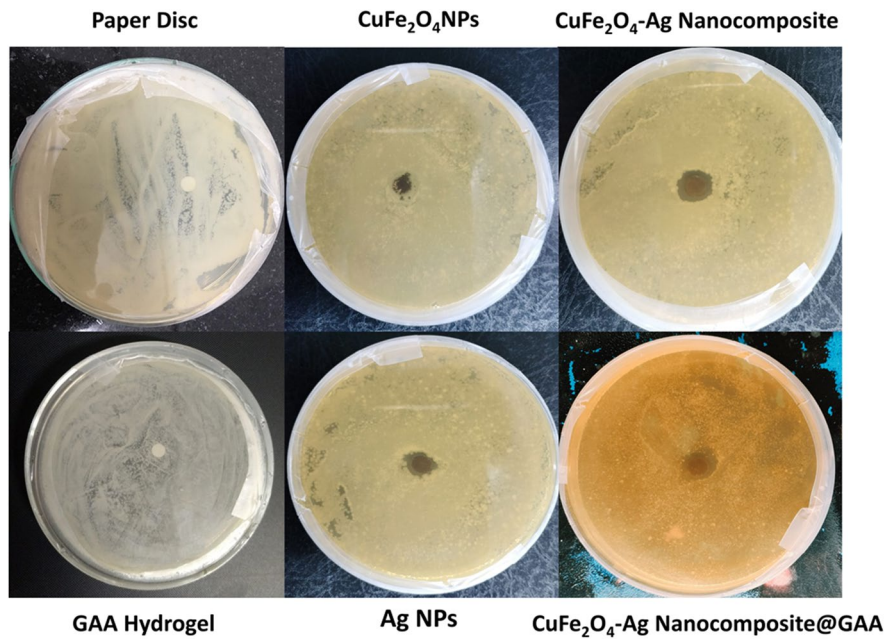


Fig. 12 Evaluation of antimicrobial activity of using disc diffusion method against *E.Coli* bacteria

and TEM. Moreover, the $CuFe_2O_4$-Ag nanocomposite is uniformly dispersed over the surface of gum-Arabic-acrylamide hydrogel which was confirmed using EDS color mapping. Zeta potential values confirmed the good colloidal stability of nanocomposite as compared to individual NPs. Thermal stability and mechanical resistance of GAA hydrogel increased with the incorporation of nanocomposite. With the increase in the concentration of biopolymer (gum Arabic) swelling ability of hydrogel was increased which was further enhanced with the incorporation of the nanocomposite. $CuFe_2O_4$-Ag nanocomposite incorporated GAA hydrogel was found to be an efficient photocatalyst among other samples which not only adsorbs the dye but also degrades it upto 90%. Moreover, $CuFe_2O_4$-Ag nanocomposite exhibits the highest efficiency of dye degradation even after three cycles of reusability. These nanocomposite-incorporated hydrogels could be a promising material for the remediation of the toxic dyes as well as the pathogenic bacteria present in industrial effluents. Polymer nanocomposites may associate with few limitations towards water treatment on large scale like challenges in achieving high photocatalyst loading, aggregation of nanomaterials in matrix of hydrogel and complexity in recycling of crosslinked hydrogel nanocomposites. Future research should focus on improving performance of material, evaluating environmental and economic impacts, expanding the capabilities to degrade multiple pollutants, and scaling production for industrial use.

Authors Contributions **Meenu:** Synthesis, Analysis, Data curation, Investigation, Methodology, original manuscript draft writing; **Pratibha Sharma:** Analysis, Validation, Review and Editing; **Roli Purwar:** Visualization and validation; **Navita Sharma:** Investigation, Data curation; **Pooja Rawat:** Validation and Analysis; **Seema R Pathak:** Final Drafting and formatting; **Atul Thakur:** Instrumentation; **Hyunook Kim:** Instrumentation; **Monu Verma:** Instrumentation; **Sudip Majumder:** Final review and editing; **Chandra Mohan Srivastava:** Design of the work, Conceptualization, Project administration, Final review and editing.

Funding The authors declare that no fund, grants, or other support were received during the preparation of this manuscript.

Data Availability The authors confirm that all the data which support the finding of this study are available in the manuscript.

Declarations

Ethics Approval This work is original and not been published elsewhere in any form or language. This manuscript has been submitted to this journal solely for consideration.

Consent to Participate The authors give their consent to participate.

Consent to Publish The authors give their consent to publish.

Conflict of Interest The authors declare that there was no conflict of interest in this work.

References

Abou-Gamra, Z. M., & Ahmed, M. A. (2015). TiO_2 Nanoparticles for Removal of Malachite Green Dye from Waste Water. *Advances in Chemical Engineering and Science, 5*, 373–388. https://doi.org/10.4236/aces.2015.53039

Ahmed, E. M. (2015). Hydrogel: Preparation, Characterization and Applications a review. *Journal of Advanced Research, 6*(2), 105–121. https://doi.org/10.1016/j.jare.2013.07.006

Ali, M. A. M., Alsabagh, A. M., & Sabaa, M. W. (2020). Polyacrylamide hybrid nanocomposites hydrogels for efficient water treatment. *Iranian Polymer Journal, 29*, 455–466. https://doi.org/10.1007/s13726-020-00810-y

Bagheri, R., Ghaedi, M., Asfaram, A., Alipanahpour Dil, E., & Javadian, H. (2019). RSM-CCD design of malachite green adsorption onto activated carbon with multimodal pore size distribution prepared from Amygdalus scoparia: Kinetic and isotherm studies. *Polyhedron, 171*, 464–472. https://doi.org/10.1016/j.poly.2019.07.037

Bahram, M., Salami, S., Moghtader, M., Moghadam, P. N., Fareghi, A. R., Rasoli, M., & Salimpoor, S. (2017). Photocatalytic Degradation of Anionic Azo Dyes Acid Orange 7 and Acid Red 88 in Aqueous Solutions Using TiO_2-containing Hydrogel. *Analytical and Bioanalytical Chemical Research, 4*, 53–63. https://doi.org/10.22036/abcr.2017.41098

Benkhaya, S., Harfi, A. E., & M'rabet, S. (2020). A review on classifications, recent synthesis and applications of textile dyes. *Inorganic Chemistry Communications, 115*, 107891. https://doi.org/10.1016/j.inoche.2020.107891

Bulut, Y., & Karaer, H. (2015). Adsorption of methylene blue from aqueous solution by cross linked chitosan/ Bentonite composite. *Journal of Dispersion Science and Technology, 36*, 61–67. https://doi.org/10.1080/01932691.2014.888004

Dil, E. A., Ghaedi, M., Asfaram, A., Mehrabi, F., Bazrafshan, A. A., & Tayeb, L. (2019a). Synthesis and application of Ce-doped TiO2 nanoparticles loaded on activated carbon for ultrasound-assisted adsorption of Basic Red 46 dye. *Ultrasonics Sonochemistry, 58*, 104702. https://doi.org/10.1016/j.ultsonch.2019.104702

Dil, E. A., Ghaedi, M., Asfaram, A., Mehrabi, F., & Sadeghfar, F. (2019b). Efficient adsorption of Azure B onto CNTs/Zn: ZnO@Ni2P-NCs from aqueous solution in the presence of ultrasound wave based on multivariate optimization. *Journal of Industrial and Engineering Chemistry, 74*, 55–62. https://doi.org/10.1016/j.jiec.2018.12.050

Dil, E. A., Ghaedi, M., Ghezelbash, G. R., Asfaram, A., Ghaedi, A. M., & Mehrabi, F. (2016). Modeling and optimization of Hg2+ ion biosorption by live yeast Yarrowia lipolytica 70562 from aqueous solutions under artificial neural network-genetic algorithm and response surface methodology: Kinetic and equilibrium study. *RSC Advances, 6*(59), 54149–54161. https://doi.org/10.1039/c6ra11292g

Dominguez-Robles, J., Peresin, M. S., Tamminen, T., Rodriguez, A., & Larraneta, E. (2018). Lignin based hydrogels with super swelling capacities for dye removal. *International Journal of Biological Macromolecules, 115*, 1249–1259. https://doi.org/10.1016/j.ijbiomac.2018.04.044

Ekici, S., Isikver, Y., & Saraydin, D. (2006). Poly(acrylamide-sepiolite) composite hydrogel; Preparation, swelling and dye adsorption properties. *Polymer Bulletin, 57*(2), 231–241. https://doi.org/10.1007/s00289-006-0552-0

Gnaser, H., Savina, M. R., Calaway, W. F., Tripa, C. E., Veryovkin, I. V., & Pellin, M. J. (2005). Photocatalytic degradation of methylene blue on nanocrystalline TiO_2: Surface mass spectrometry of reaction intermediates. *International Journal of Mass Spectrometry, 245*, 61–67. https://doi.org/10.1016/j.ijms.2005.07.003

Haleem, A., Shafiq, A., Chen, S., & Nazar, M. (2023). A Comprehensive Review on Adsorption, Photocatalytic and Chemical Degradation of Dyes and Nitro-Compounds over Different Kinds of Porous and Composite Materials. *Molecules, 28*(03), 1081. https://doi.org/10.3390/molecules28031081

Jafarigol, E., Salehi, M. B., & Mortaheb, H. R. (2020). Synergetic effects of additives on structural properties of acrylamide-based hydrogel. *Journal of Dispersion Science and Technology, 42*(6), 1–10. https://doi.org/10.1080/01932691.2020.1721012

Jemal, K., Sandeep, B. V., & Pola, S. (2017). Synthesis, characterization, and evaluation of the antibacterial activity of Allophylus serratus leaf and derived callus extracts mediated silver nanoparticles. *Journal of Nanomaterials, 2017*, 4213275. https://doi.org/10.1155/2017/4213275

Ji, K. M., Deng, J. G., Zang, H. J., Han, J. H., Arandiyan, H., & Dai, H. X. (2015). Fabrication and high photocatalytic performance of noble metal nanoparticles supported on 3DOM $InVO_4$-$BiVO_4$ for the visible-light-driven degradation of rhodamine B and methylene blue. *Applied Catalyst B-Environment, 165*, 285–295. https://doi.org/10.1016/j.apcatb.2014.10.005

Kannaujia, R., Prasad, V., Sapna, Rawat, P., Rawat, V., Thakur, A., Majumdar, S., Verma, M., Rao, G. K., Gupta, A. P., Kumar, H., & Srivastava, C. M. (2019). Facile synthesis of $CuFe_2O_4$ doped polyacrylic acid hydrogel nanocomposite and its application in dye degradation. *Material letters, 252*,198-201. https://doi.org/10.1016/j.matlet.2019.05.094

Kanwal, A., Shahzadi, T., Riaz, T., Zaib, M., Khan, S., Habila, M. A., & Sillanpaa, M. (2023). Photocatalytic Degradation Studies of Organic Dyes over Novel Cu/Ni Loaded Reduced Graphene Oxide Hybrid Nanocomposite: Adsorption. *Kinetics and Thermodynamic Studies. Molecules, 28*(18), 6474. https://doi.org/10.3390/molecules28186474

Kasgoz, H., & Durmus, A. (2008). Dye removal by a novel hydrogel- clay nanocomposite with enhanced swelling properties. *Polymers for Advanced Technologies, 19*, 838–845. https://doi.org/10.1002/pat.1045

Kooti, M., Saiahi, S., & Motamedi, H. (2013). Fabrication of silver- coated cobalt ferrite nanocomposite and study of its antibacterial activity. *Journal of Magnetism and Magnetic Materials, 333*, 138–143. https://doi.org/10.1016/j.jmmm.2012.12.038

Kumari, S., Sharma, P., Yadav, S., Kumar, J., Vij, A., Rawat, P., Kumar, S., Sinha, C., Srivastava, C. M., & Majumdar, S. (2020). A novel synthesis of the graphene oxide-silver (GO-Ag) nanocomposite for unique physiochemical applications. *ACS Omega, 5*(10), 5014–5047. https://doi.org/10.1021/acsomega.9b03976

Lanjwani, M. F., Tuzen, M., Khuhawar, M. Y., & Saleh, T. A. (2024). Trends in photocatalytic degradation of organic dye pollutants using nanoparticles: A review. *Inorganic Chemistry Communications, 159*, 111613. https://doi.org/10.1016/j.inoche.2023.111613

Lellis, B., & Polonio, J. C. (2019). Effect of textile dyes on health and the environment and bioremediation potential of living organisms. *Biotechnology Research and Innovation, 3*(2), 275–290. https://doi.org/10.1016/j.biori.2019.09.001

Li, D., Wang, F. X., Zhang, X. Y., Liang, L., & Sun, J. M. (2014). The efficient removal of organic pollutant over magnetic mesoporous. *Polymer Journal of Porous Materials, 21*, 811–817. https://doi.org/10.1007/s10934-014-9829-0

Ma, G., Zhu, Y. M., Zhang, Z. Q., & Li, L. C. (2014). Preparation and characterization of multi-walled carbon nanotube/TiO_2 composites: Decontamination organic pollutant in water. *Applications of Surface Science, 313*, 817–822. https://doi.org/10.1016/j.apsusc.2014.06.079

Mahmoodi, N. M. (2011). Photocatalytic ozonation of dyes using copper ferrite nanoparticle prepared by co-precipitation method. *Desalination, 279*, 332–337. https://doi.org/10.1016/j.desal.2011.06.027

Majumder, S., Dahiya, U. R., Yadav, S., Sharma, P., Ghosh, D., Rao, G. K., Rawat, R., Kumar, G., Kumar, A., & Srivastava, C. M. (2020). Zinc Oxide Nanoparticles Functionalized on Hydrogel Grafted Silk Fibroin Fabrics as Efficient Composite Dressing. *Biomolecules, 10*(5), 710. https://doi.org/10.3390/biom10050710

Miri, A., Shahraki Vahed, H. O., & Sarani, M. (2018). Biosynthesis of silver nanoparticles and their role in photocatalytic degradation of methylene blue dye. *Research on Chemical Intermediates, 44*, 6907–6915. https://doi.org/10.1007/s11164-018-3529-3

Naseer, M., Aslam, U., Khalid, B., & Chen, B. (2020). Green route to synthesize Zinc Oxide Nanoparticles using leaf extracts of Cassia fistula and Melia azadrach and their antibacterial potential. *Scientific Reports, 10*, 9055. https://doi.org/10.1038/s41598-020-65949-3

Özçelik, S. (2023). Copper Ferrite Nanoparticles: Structural, Magnetic, Optical, Photocatalytic Activity and Blood

Compatibility Properties. *Bio Nano Science, 13*, 958–972. https://doi.org/10.1007/s12668-023-01130-0

Paulino, A. T., Pereira, A. G. B., Fajardo, A. R., Erickson, K., Kipper, M. T., Muniz, E. C., Belfiore, L. A., & Tambourgi, E. B. (2012). Natural polymer-based magnetic hydrogels: Potential vectors for remote-controlled drug release. *Carbohydrate Polymers, 90*, 1216–1225. https://doi.org/10.1016/j.carbpol.2012.06.051

Pethsangave, D. A., Khose, R. V., Wadekar, P. H., & Kulal, D. K. (2020). One-pot synthetic approach for magnetically seperable graphene nanocomposite for dye degradation. *Chemistry Select, 5*(4), 1516–1525. https://doi.org/10.1002/slct.201903966

Pradhan, S. K., Pareek, V., Panwar, J., & Gupta, S. (2019). Synthesis and characterization of ecofriendly silver nanoparticles combined with yttrium oxide (Ag-Y_2O_3) nanocomposite with assorted adsorption capacity for Cu (II) and Cr (VI) removal: A mechanism perspective. *Journal of Water Process Engineering, 32*, 100917. https://doi.org/10.1016/j.jwpe.2019.100917

Priya, R., Stanly, S., Anuradha, R., & Sagadevan, S. (2019). Evaluation of photocatalytic activity of copper ferrite nanoparticles. *Material Research Express, 6*, 095014. https://doi.org/10.1088/2053-1591/ab2d15

Purwar, R., Rajput, P., & Srivastava, C. M. (2014). Composite wound dressing for drug release. *Fibers and PolymErs, 15*, 1422–1428. https://doi.org/10.1007/s12221-014-1422-2

Rusu, A. G., Popa, M. L., Lisa, G., & Verestiuc, L. (2015). Thermal behavior of hydrophobically modified hydrogels using TGA/FTIR/MS analysis technique. *Thermochimica Acta, 613*, 28–40. https://doi.org/10.1016/j.tca.2015.05.018

Serrano-Lotina, A., Portela, R., Baeza, P., Alcolea-Rodriguez, V., Villarroel, M., & Avila, P. (2023). Zeta potential as a tool for functional materials development. *Catalysis Today, 423*, 113862. https://doi.org/10.1016/j.cattod.2022.08.004

Sharifpour, E., Alipanahpour Dil, E., Asfaram, A., Ghaedi, M., & Goudarzi, A. (2019). Optimizing adsorptive removal of malachite green and methyl orange dyes from simulated wastewater by Mn-doped CuO-Nanoparticles loaded on activated carbon using CCD-RSM: Mechanism, regeneration, isotherm, kinetic, and thermodynamic studies. *Applied Organometallic Chemistry*, e4768. https://doi.org/10.1002/aoc.4768

Sharma, G., Kumar, Sharma, A., Shweta, Al-Muhtaseb, A., Naushad, M., Ghfar, A. A., Ahamad, T., & Stadler, F. J. (2018). Fabrication and characterization of novel FeO@ Guar gum-crosslinked-soya lecithin nanocomposite hydrogel for photocatalytic degradation of methyl violet dye. *Separation and Purification Technology, 211,* 895–908. https://doi.org/10.1016/j.seppur.2018.10.028

Sharma, P., Kumari, S., Ghosh, D., Yadav, V., Vij, A., Rawat, P., Kumar, S., Sinha, C., Saini, S., Sharma, V., & Hassan, M. I. (2021). Capping agent-induced variation of physicochemical and biological properties of α-Fe_2O_3 nanoparticles. *Materials Chemistry and Physics, 258*, 123899. https://doi.org/10.1016/j.matchemphys.2020.123899

Singh, B., Sharma, S., & Dhiman, A. (2017). Acacia gum polysaccharide-based hydrogel wound dressings: Synthesis, characterization, drug delivery and biomedical properties. *Carbohydrate Polymers, 165*, 294–303. https://doi.org/10.1016/j.carbpol.2017.02.039

Singh, G., Bhunia, H., Bajpai, P. K., & Choudhary, V. (2012). Thermal degradation and physical aging of linear low-density polyethylene and poly (l-lactic acid) blends. *Journal of Polymer Engineering, 32*(1), 59–66. https://doi.org/10.1515/polyeng-2011-0106

Vidhu, V. K., & Philip, D. (2014). Catalytic degradation of organic dye using biosynthesized silver nanoparticles. *Micron, 56*, 54–62. https://doi.org/10.1016/j.micron.2013.10.006

Wang, H., Qiao, X., Chen, J., & Ding, S. (2005). Preparation of silver nanoparticles by chemical reduction method. *Colloids and Surfaces A, 256*, 111–115. https://doi.org/10.1016/j.colsurfa.2004.12.058

Wang, Y., Tian, D., Chu, W., Li, M., & Lu, X. (2019). Nanoscaled magnetic $CuFe_2O_4$ as an activator of peroxymonosulfate for degradation of antibiotics norfloxacin. *Separation and purification technology, 212*, 536–544.

Yan, C. Q., & Pochan, D. J. (2010). Rheological properties of peptide-based Hydrogels for biomedical and other applications. *Chemical Society Reviews, 39*, 3528–3540. https://doi.org/10.1039/B919449P

Yontar, A. K., Avcioğlu, S., & Çevik, S. (2022). Nature-based nanocomposites for adsorption and visible light photocatalytic degradation of methylene blue dye. *Journal of Cleaner Production, 380*(2), 135070. https://doi.org/10.1016/j.jclepro.2022.135070

Zhang, Y. R., Shen, S. L., Wang, S. Q., Huang, J., Su, P., Wang, Q. R., & Zhao, B. X. (2014). A dual function magnetic nanomaterial modified with lysine for removal of organic dyes from water solution. *Chemical Engineering Journal, 239*, 250–256. https://doi.org/10.1016/j.cej.2013.11.022

Zheng, M., Guo, J., Li, Q., Yang, J., Han, Y., Yang, H., Yu, M., Zhong, L., Lu, D., Li, L., & Sun, L. (2021). Syntheses and characterization of anti-thrombotic and anti-oxidative Gastrodin-modified polyurethane for vascular tissue engineering. *Bioactive Materials, 6*(2), 404–419. https://doi.org/10.1016/j.bioactmat.2020.08.008

Zhu, H. Y., Jiang, R., Xiao, L., & Li, W. (2010). A novel magnetically separable _-Fe_2O_3/crosslinked chitosan adsorbent: Preparation, characterization and adsorption application for removal of hazardous azo dye. *Journal of Hazardous Materials, 179*, 251–257. https://doi.org/10.1016/j.jhazmat.2010.02.087

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 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) | 979-8-3503-9156-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSTSN61422.2024.10670939

Unveiling Anomalies: A Review of Anomaly Detection Through Lens of Explainable AI

Shikhar Asthana
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
shikhar28.shikhar28@gmail.com

Anurag Goel
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
anurag@dtu.ac.in

***Abstract*—In today's transitional world, anomaly detection (AD) has made itself a major foothold application that spans across multiple domains, specifically in domains where detecting abnormal behavior or events is of paramount importance. The onset of deep learning techniques brings along with it a surge in the advancement of anomaly detection methods with the capability to handle complex and high-dimensional data. However, with these advancements, a major concern is in the explainability of these approaches – to overcome the 'black box' nature of these approaches. This review paper provides a detailed exploratory overview of recent advancements in AD techniques, concentrating on how and where these fall in the domain of explainable AI. Key insights from the surveyed literature include the importance of interpretability in AD systems, the versatility of deep learning architectures for AD tasks, and the emerging trends of graph-based AD using deep learning techniques. The review also discusses evaluation metrics, challenges, and future directions in the field, providing researchers and practitioners with valuable insights for advancing AD research and applications.**

***Index Terms*—Anomaly detection, Explainable AI, Deep learning, Interpretability, Graph-based anomaly detection, Model-specific interpretation.**

I. Introduction

The ever-changing and dynamic world is full of patterns and sequences waiting to be detected. With these comes the opportunity to observe and detect anomalies in these patterns which play a very vital role in the world. The field of Artificial Intelligence (AI) which is concerned with identification and observation of detecting anomalies is called anomaly detection (AD). AD often is also coupled with a great opportunity for business or industrial impacts like – pre-emptive detection of potential threats or irregularities, better planning for inventory-based operations, and many others, which are key in domains such as cybersecurity, finance, healthcare, and industrial systems [1] [2] [3] [4]. Modern datasets have increasing levels of complexity and dimensionality which are problem areas for the traditional AD approaches, which leads to a growing interest in finding ways to leverage deep learning techniques for more effective AD. However, the widespread adoption of deep learning approaches in AD raises concerns about the interpretability and transparency of these models, often tokenized as "black box" problem. As a result, there is a pressing need for research development, innovations and development efforts centered around improving the interpretability of deep learning-based AD systems. This review paper endeavors to offer an exhaustive examination of recent progressions in AD methodologies, with a specific emphasis on the convergence of deep learning and explainable AI. By synthesizing insights from existing literature, this review seeks to elucidate the importance of interpretability in AD systems and explore emerging trends and challenges in the field.

As part of the introduction, first developing an understanding of a brief overview of traditional methods is crucial. Traditional AD methods, usually, relied upon statistical-based techniques including, but not limited to, clustering, density estimation, distance-based methods, and many more, to search out specific occurrences that diverge significantly from the majority data occurrences. [5] While these approaches tend to have a very high explainability in the sense that the approaches are easy to interpret and understand, but these only work well in certain restricted scenarios. These approaches, often struggle to cope with the increasing complexity and high dimensionality of the modern datasets which are currently being used in the industries, where anomalies may manifest themselves in subtle or more complex ways. Moreover, traditional based AD involves a lot of manual interventions in the machine learning pipeline. These interventions can range far and wide from manual data pre-processing, and annotation, to manual algorithm selection, hyperparameter tuning, and other manual interventions which can also result in human errors creeping into the AD performance.

In recent times, with the trifactor rise in data, computational power, and better deep learning algorithms, industry has seen a huge boost in the adoption of deep learning techniques and how these techniques have revolutionized the field of AD [6]. This revolution was brought by offering more sophisticated and data-driven approaches to scrutinize and search out anomalies in complex datasets [7]. Deep learning models, over traditional AD methods, have demonstrated, in multiple scenarios, how remarkable their capabilities are in automatically learning from unprocessed data, the hierarchical based representations, allowing them to seize baroque patterns and relationships that may not be apparent if viewed from the lenses of the traditional methods [8]. These techniques are also more versatile in terms of data modalities on which AD can be performed – including images, time series, text, and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:04 UTC from IEEE Xplore. Restrictions apply.

TABLE I
RESEARCH QUESTIONS & MOTIVATIONS

RQ#	Research Question	Motivation
RQ1	What are the current state-of-the-art techniques for anomaly detection?	To identify the latest advancements in anomaly detection methods across diverse domains.
RQ2	How can explainable AI techniques improve the interpretability of anomaly detection models?	To explore methods for enhancing the transparency and trustworthiness of anomaly detection systems.
RQ3	What are the key challenges and limitations of deep learning-based anomaly detection?	To uncover the obstacles hindering the widespread adoption of deep learning in anomaly detection.
RQ4	What are the emerging trends and future directions in anomaly detection research, including the role of graph-based neural networks?	To anticipate upcoming challenges and opportunities in anomaly detection, with a focus on graph-based neural networks.

graph-structured data.

Real-world applications of social networks, biological networks, cybersecurity networks, and other such emerging fields have garnered considerable attention recently, which brings with it a mounting heap of graph-structured data and a crucial need for specialized AD methodology for these datasets – Graph Anomaly Detection (GAD), a specialized area within AD is the answer [9]. The reason for a specialized approach is because traditional AD techniques fail to consider inherent structural information and relationships among entities which causes them to ineffectively handle graph data. Graph Neural Networks (GNN), a deep learning based approach, have opened up new avenues for AD in graph-based data [10]. GNNs can recognize complex structural dependencies and node interactions in graphs, making them well-suited for AD tasks in graph-structured data. This intersection of deep learning and graph anomaly detection holds great promise for addressing challenges in anomaly detection in complex networked systems.

A. Need for explainable AI based review

As briefed in the above paragraphs, deep learning based AD has risen to become a widely applied application in the industry – like for diagnostic applications in the healthcare domain, fraud and other abnormal behavior detections in the financial domain, intrusion and other attack detections in the cybersecurity domain, and many others [1] [2] [11] [12] [3] [13] [4] [14]. Thus, deep learning based AD systems are not only impacting aspects of people's daily lives but also playing a vital role in industry with huge turnovers, extreme competition, and high impact. In such a situation, the 'black box' nature of many deep learning based AD systems can be an obvious red flag for any industrial application. Basic questions regarding trust on the system, validation of the rationale of the system, identification of causal relationships derived by the system, and other similar aspects become extremely important in the decision to adopt the AD system or not. Thus, there is a dire need for not only understanding deep learning based AD systems that are accurate but which also provide a high level of interpretability and explainability. This paper will dwell into which explainable metrics have been used, in detail, in the later sections.

B. Research Questions

Research questions (RQ) being explored have been derived from the objective of providing a comprehensive analysis of recent advancements in anomaly detection techniques, with a priority focus on the role of explainable AI. By synthesizing insights from current literature, this review, and its corresponding research questions, aim to elucidate the vital nature of interpretability in AD, while at the same time, exploring the emerging challenges and trends in the field with the hopes of identifying opportunities for future research and application. It is the endeavor of this paper to offer valuable insights to researchers, practitioners, and stakeholders involved in anomaly detection across various domains. Table I highlights the RQs and the motivation behind selecting and focusing on them.

II. SOTA TECHNIQUES FOR ANOMALY DETECTION

Through State-of-the-Art (SOTA) Techniques for AD, the paper's objective is to procure an exhaustive overview for 18 prominent industrial AD methods. These techniques include Z-Score (ZS) [15], Grubbs Test (GT) [16] [17], K Nearest Neighbors (KNN) [18], Local Outlier Factor (LOF) [19], K Means Clustering (KMC) [20] [21], DBSCAN [22], Variational Autoencoder (VAE) [23] [24], Sparse Autoencoder (SAE) [24] [25], DeepConvLSTM (DCLSTM) [26], Capsule Network (CapsNet) [27], LSTM [28], GRU [29], GANomaly (GANM) [30], Skip-GANomaly (SGANM) [31], NetWALK [32], ResGCN [33], DeepSphere (DS) [34], and DeepFD [35]. Table II and Table III contain detailed information about each technique, including objective function formulas, measures, and outputs, offering valuable insights into their mechanisms and performance characteristics.

III. EXPLAINABLE AI FOR ANOMALY DETECTION

Anomalies have a critical role in many domains like finance, cybersecurity, healthcare, and manufacturing, emphasizing the need for interpretability and transparency in AD models. Explainable AI (XAI) parameters offer a solution by providing insights into model decisions, enhancing trust, and acceptance among stakeholders. XAI also aids in result validation, bias identification, regulatory compliance, and ethical considerations. By examining XAI parameters in various AD techniques, this paper aims to provide valuable insights into mechanisms for enhancing understanding and trust in AD systems.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:04 UTC from IEEE Xplore. Restrictions apply.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:04 UTC from IEEE Xplore. Restrictions apply.

TABLE II
KNN, KMC, DBSCAN DISTANCE PARAMETERS AND OBJECTIVE FUNCTIONS CAN CHANGE BASED ON IMPLEMENTATION. SSA - STATISTICALLY SIGNIFICANT ANOMALIES. LES - LOCATION IN EMBEDDING SPACE. MED - MEDIUM
AD APPROACHES THROUGH EXPLAINABLE AI

Method	Technique	Approach	Objective Function	Measures	Outputs	INT	FI	MC	VIZ	HRO	RP
Traditional AD Methods	Statistical	ZS [15]	$z_i = \frac{x_i - \mu}{\sigma}$	Z-Score List	SSA	High	High	Low	No	Yes	Med
		GT [16] [17]	$G = \frac{\lvert x_i - \bar{x}\rvert}{s}$	G List	SSA	Med	Med	Low	No	Yes	Med
	Distance	KNN [18]	$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$	Distance to KNN	Flags Anomalies	High	Med	Low	Yes	Yes	Low
		LOF [19]	$\frac{\sum_{o \in N(p)} \frac{LRD(o)}{LRD(p)}}{\lvert N(p)\rvert}$	LOF Score	SSA	Med	High	Med	No	Yes	Med
	Clustering	KMC [20] [21]	$\sum_{i=1}^{k} \sum_{x \in C_i} \lvert\lvert x - \mu_i\rvert\rvert^2$	Cluster Members	Flags Anomalies	Med	Med	Low	Yes	Yes	Low
		DBSCAN [22]	$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$	Density Cluster Measure	Cluster Assignments	Med	High	Med	Yes	Yes	High
Graph AD Methods	Anoumalous Node	NetWALK [32]	$\gamma L_{AE} + L_{Clique} + \lambda \lvert\lvert W\rvert\rvert_F^2 + \beta KL$	Anomaly Score	Nearest Distance to Cluster Center	Med	High	High	No	Yes	High
		ResGCN [33]	$(1 - \alpha)E_S + \alpha E_A$	LES	Anomalous Label	Low	High	High	Yes	Yes	High
	Anomalous Edge	DS [34]	$L_h + \lambda L_{res}$	LES	Anomalous Label	Med	High	High	Yes	Yes	High
		DeepFD [35]	$L_{recon} + \alpha L_{sim} + \gamma L_{reg}$	LES	Anomalous Label	Low	High	High	Yes	Yes	High

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:04 UTC from IEEE Xplore. Restrictions apply.

TABLE III
DCLSTM, LSTM, GRU CAN HAVE VARYING OBJECTIVE FUNCTIONS BASED ON THE APPLICATION.
AD APPROACHES THROUGH EXPLAINABLE AI CONTD.

Method	Technique	Approach	Objective Function	Measures	Outputs	INT	FI	MC	VIZ	HRO	RP
Deep Learning AD Methods	Autoencoder	VAE [23] [24]	$-\mathbb{E}_{z\sim q(z\|x)}[\log p(x\|z)]$ $+\text{KL}(q(z\|x)\|\|p(z))$	Anomaly Score	Flags Anomaly	Med	Med	Med	Yes	Yes	Med
		SAE [24] [25]	$L_{recon} + L_{sparsityReg}$	Anomaly Score	Flags Anomaly	Med	Med	Med	Yes	Yes	Med
	CNN	DCLSTM [26]	$\frac{1}{N}\sum_i y_i \log(\hat{y}_i)$	Prediction Loss	Predictions	Low	High	High	Yes	Yes	Med
		CapsNet [27]	$T_c \max(0, m^+ - \|\|v_c\|\|)^2 +$ $\lambda(1-T_c)\max(0, \|\|v_c\|\| - m^-)^2$	Margin Loss	Re - constructions	Low	High	High	Yes	Yes	Med
	RNN	LSTM [28]	$-\frac{1}{N}\sum_i y_i \log(\hat{y}_i)$	Prediction Loss	Predictions	Med	High	High	Yes	Yes	Low
		GRU [29]	$-\frac{1}{N}\sum_i y_i \log(\hat{y}_i)$	Prediction Loss	Predictions	Med	High	High	Yes	Yes	Low
	GAN	GANM [30]	$W_1 * L_{adv} + W_2 * L_{con} + W_3 * L_{enc}$	Anomaly Score	Flags Anomaly	Med	Med	High	Yes	Yes	High
		SGANM [31]	$W_1 * L_{adv} + W_2 * L_{con} + W_3 * L_{lat}$	Anomaly Score	Flags Anomaly	Med	Med	High	Yes	Yes	High

Interpretability (INT): in anomaly detection models is vital for fostering trust and collaboration between stakeholders and the AI system. Techniques with high INT offer transparent explanations, while medium INT strikes a balance between transparency and complexity. Low INT poses challenges, leading to skepticism. In critical domains like healthcare and finance, clear explanations are essential. Enhancing interpretability in AD models is crucial for ensuring trust and acceptance.

Feature Importance (FI): elucidates the contribution of features to AD outcomes. High FI techniques provide clear measures of feature importance, aiding stakeholders in understanding influential features and their impact on AD. Medium FI offers insights, while low FI hinders transparency. FI is critical in cybersecurity and fraud detection for effective risk mitigation.

Model Complexity (MC): impacts stakeholders' understanding and trust in model decisions. High MC methods, like deep learning, can be intricate and challenging to comprehend. Low MC techniques, such as traditional statistics, offer simplicity and transparency. Medium MC strikes a balance, capturing complex patterns while remaining accessible. Balancing sophistication and transparency is crucial for effective anomaly detection.

Visualization (VIZ): is pivotal for explainable AI in AD, providing stakeholders with intuitive representations of complex data patterns. Techniques with VIZ capabilities enable effective exploration and understanding of model outputs through interactive dashboards, heatmaps, or scatter plots. Lack of VIZ capabilities hinders explainability, limiting stakeholders' ability to interpret and validate model outputs. VIZ plays a vital role in enhancing trust and collaboration in anomaly detection systems.

Human-Readable Outputs (HRO): are essential for explainable AI in AD, providing stakeholders with comprehensible insights about detected anomalies. Techniques with HRO outputs, such as textual descriptions or visualizations, enhance interpretability and facilitate collaboration. Conversely, methods lacking HRO hinder understanding and may lead to ambiguity.

Robustness to Perturbations (RP): is critical to ensure the dependability of AD models in actual world applications. High RP techniques demonstrate resilience against variations in input data, maintaining consistent performance across diverse datasets. Conversely, low RP methods exhibit sensitivity to variations, resulting in unreliable performance. Medium-level techniques offer moderate stability, reliably detecting anomalies under typical conditions but may struggle with severe perturbations.

IV. Challenges & Future Directions

While AD is vital across domains, deep learning based methods still face persistent challenges. Through this section, an exploration of these challenges and future directions for advancing anomaly detection research and applications will be undertaken.

Lack of Explainable Approaches: Deep learning models have a reputation of working as black boxes, which in turn make it challenging to understand their decision-making process in AD. Moreover, as evident from Tables II and III, none of the prevalent approaches are a one-stop solution which is positive across all parameters. Addressing this challenge is crucial for enhancing anomaly detection models' transparency and explainability, aligning with the research question's objectives on the importance of explainable AI techniques.

Data Imbalance and Labeling Issues: AD datasets often exhibit class imbalance, with normal instances outnumbering anomalies, as observed across the reviewed papers and models. This imbalance poses challenges for deep learning models, hindering their power of learning through limited anomaly examples. While models like GANomaly and Skip-GANomaly partly address this issue, effectively managing data imbalance remains a critical need in the AD domain.

Anomaly Generalization: Deep learning models, particularly supervised ones, often struggle to generalize to all possible anomalies, as these models tend to overfit to the anomalies present in labeled datasets. Although semi-supervised models like GANomaly and Skip-GANomaly may offer partial solutions, these models may sacrifice feature importance and struggle with complex anomalies. Developing techniques to mitigate overfitting and enhance generalization is essential for improving the robustness of deep learning-based AD systems, and addressing research questions on limitations of these approaches.

Addressing these challenges opens avenues for future research directions. This includes enhancing explainability in AD methods through improving existing approaches or developing novel, more transparent methods. Additionally, efforts can focus on mitigating data imbalances through modifications to existing approaches, proposing new methods, or introducing improved metrics. Moreover, research may explore new approaches to enhance anomaly generalization capabilities.

V. Conclusion

In conclusion, the paper provides a comprehensive overview of industry prevalent AD approaches, their challenges, and future directions in deep learning-based AD, with a focus on the role of explainable AI techniques. Through an analysis of SOTA methods and emerging trends, this paper has highlighted the importance of interpretability, model complexity, model robustness, and the integration of graph structures for enhancing anomaly detection performance. Continued progress and innovation in the field are anticipated by exploring future directions, including the adoption of explainable AI techniques and the evolution of graph-based approaches. Ultimately, this paper aims to contribute to advancing AD research and facilitate the improvement of more robust and interpretable AD systems qualified for addressing complex challenges across diverse domains.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:04 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.

[2] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, vol. 55, pp. 278–288, 2016.

[3] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[4] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*, pp. 146–157, Springer, 2017.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[6] C. C. Aggarwal and S. Sathe, "Theoretical foundations and algorithms for outlier ensembles," *Acm sigkdd explorations newsletter*, vol. 17, no. 1, pp. 24–47, 2015.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[9] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, pp. 626–688, 2015.

[10] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[11] D. Bogdoll, M. Nitsche, and J. M. Zöllner, "Anomaly detection in autonomous driving: A survey," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4488–4499, 2022.

[12] S. Fuertes, G. Picart, J.-Y. Tourneret, L. Chaari, A. Ferrari, and C. Richard, "Improving spacecraft health monitoring with automatic anomaly detection techniques," in *14th international conference on space operations*, p. 2430, 2016.

[13] E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, and S. Silvia, "A survey on outlier explanations," *The VLDB Journal*, vol. 31, no. 5, pp. 977–1008, 2022.

[14] A. Ukil, S. Bandyoapdhyay, C. Puri, and A. Pal, "Iot healthcare analytics: The importance of anomaly detection," in *2016 IEEE 30th international conference on advanced information networking and applications (AINA)*, pp. 994–997, IEEE, 2016.

[15] P. J. Rousseeuw and M. Hubert, "Anomaly detection by robust statistics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 2, p. e1236, 2018.

[16] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[17] F. E. Grubbs, *Sample criteria for testing outlying observations*. University of Michigan, 1949.

[18] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[19] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

[20] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[21] C. C. Aggarwal and K. Subbian, "Anomaly detection in large datasets: A survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 15, 2016.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, pp. 226–231, 1996.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[24] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[26] R. Murugesan, E. Mishra, and A. H. Krishnan, "Deep learning based models: Basic lstm, bi lstm, stacked lstm, cnn lstm and conv lstm to forecast agricultural commodities prices," 2021.

[27] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Advances in neural information processing systems*, vol. 30, 2017.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] K. Cho and e. a. Van Merriënboer, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[30] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 622–637, Springer, 2019.

[31] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[32] W. Yu, W. Cheng, C. C. Aggarwal, K. Zhang, H. Chen, and W. Wang, "Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2672–2681, 2018.

[33] Y. Pei, T. Huang, W. van Ipenburg, and M. Pechenizkiy, "Resgcn: Attention-based deep residual modeling for anomaly detection on attributed networks," *Machine Learning*, vol. 111, no. 2, pp. 519–541, 2022.

[34] X. Teng, M. Yan, A. M. Ertugrul, and Y.-R. Lin, "Deep into hypersphere: Robust and unsupervised anomaly discovery in dynamic networks," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[35] H. Wang and e. a. Zhou, "Deep structure learning for fraud detection," in *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 567–576, IEEE, 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 23,2024 at 11:16:04 UTC from IEEE Xplore. Restrictions apply.

2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) | 979-8-3503-8592-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONECCT62155.2024.10677188

Unveiling Toxic Tendencies of Small Language Models in Unconstrained Generation Tasks

Lakshay Chandra, Seba Susan*, Dhruv Kumar, and Krishan Kant
Department of Information Technology, Delhi Technological University, Delhi 110042, India
* seba_406@yahoo.in

***Abstract*—The prevalence of toxicity online presents a significant challenge for platforms and publishers alike. Recent studies conducted on Small Language Models (SLMs) have identified the inherent toxicity that dwell in these models. In this work, we study and benchmark the extent to which SLMs can be prompted to generate toxic language. The following SLMs are evaluated for their toxicity levels: GPT-2 Large, Gemma-2B, Mistral-7B, Falcon-7B, and Llama 2-13B. We go a step closer to understanding the correlation between toxicity and the intrinsic parameters of the state-of-the-art SLMs. Next, we study the efficacy of a basic word-filtering approach to controlled text generation. Following this, we proceed to establish a mathematical ground for computing the weighted toxicity of continuations with respect to the toxicity of prompts by treating toxicity as a fuzzy metric. Finally, we extend our analysis to examine the unexpected toxicity levels of generated continuations when prompted with non-toxic inputs.**

Keywords—toxicity analysis, small language models, language generation, deep learning

I. INTRODUCTION

Large language models (LLMs) have made substantial progress in natural language generation through rapid breakthroughs in pre-training and few-shot learning [1]. LLMs are rapidly replacing the statistical n-gram language models [2], that predict the likelihood of the next word based on the preceding words, for language generation tasks. One reason is that LLMs learn long-term dependencies across the entire text, unlike the statistical language models that capture context from a fixed-size window. There is increasing awareness of the potential risks of the misuse of LLMs, and there is a need to devise a system to evaluate LLMs from the perspective of their impact on society [3].

Small Language Models (SLMs) are a significant development in the current generation of language models, providing a small footprint, but potent solutions, to various linguistic problems [4]. To put it simply, SLMs are LLMs on a smaller scale. Compared to LLMs with hundreds of billions of parameters, SLMs have a great deal fewer; usually, a few million to a few billion parameters. In general opinion, SLMs have less than 10 billion parameters, as opposed to LLMs that have now got "bigger" and tend to have more than 100 billion parameters. GPT-2, Gemma, Mistral, Falcon and Llama 2 are a few examples of state-of-the-art SLMs [5]. SLMs are becoming more appealing for organizations to design and deploy because they offer greater control, such as fine-tuning for specific domains and data security. They are also cost-effective to operate.

Although SLMs have immense potential, their openness has undermined AI safety and created new risks of toxic behaviour [6]. Several researchers have highlighted the problem in recent times. Gehman et al. [7] introduced RealToxicityPrompts, a testbed of 100K prompts. They assess the toxicity of various pre-trained language models and the effectiveness of methods for detoxifying generations. Their study also involved a toxicity analysis of the GPT-2 pretraining corpus, to better understand the underlying reasons for toxic generations. Welbl et al. [8] explored the difficulty of evaluating the toxicity of language models and the effects of automatic toxicity mitigation, utilizing both simple and traditional methodologies. They emphasized the disparity between traditional toxicology measures and human perceptions. He et al. [9] conducted the first evaluation of prompt learning with customized prompts (prompt tuning) to counter harmful content. The study compared prompt tuning's performance in three tasks (toxicity categorization, toxic span detection, and detoxification) to state-of-the-art baselines. Among other findings, they discovered that prompt adjustment can result in equivalent or even superior performance to baselines. A recent work [10] highlighted the vital role of prompts in the detoxification procedure. Specifically, the authors propounded instance-level positive prefixes for prompts to counteract the negative prefixes, as an alternative to fine-tuning LLMs for detoxification.

Language models are trained using large amounts of text from the Internet, which contains a lot of harmful information [11]. Naturally, language models pick up on potentially toxic patterns during training. When presented with an already toxic prompt, the most intuitive behaviour of SLM is to perpetuate the toxic behaviour as the prompt continues, rather than trying to mitigate it, which poses a huge blot on AI safety [12]. In this paper, we present a detailed toxicity analysis report for five state-of-the-art SLMs, namely, GPT-2 Large, Gemma-2B, Mistral-7B, Falcon-7B, and Llama 2-13B.

The organization of the paper is as follows. Section II presents the methodology followed in the evaluation of the toxicity levels of SLMs. Section III describes the experimentation and the results obtained in detail. Finally, section IV concludes the paper.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:07:50 UTC from IEEE Xplore. Restrictions apply.

II. Methodology

In our research, we study and compare the effect of the intrinsic parameters of SLMs on the toxicity level of the text generated. In this section, we describe the background and experimental settings needed for evaluating the toxicity levels of the SLMs under study. We prompt the model with an incomplete sentence **p** which has a measured toxicity of t_i to generate a continuation whose measured toxicity is t_o. We use the RealToxicityPrompts dataset for our study. Each prompt of this dataset is associated with probability scores for different toxicity levels. A limited sample set of 781 prompts is carefully extracted for the study. Our sampled dataset **D** has a fair proportion of the following prompts.

- *toxic* $t_i \geq 0.5$
- *slightly toxic* $0.1 \leq t_i < 0.5$
- *nontoxic* $t_i < 0.1$

Perspective API: Perspective API is a machine learning model that can identify toxic or abusive comments by returning a probability score. We have used the Perspective toxicity score in our analysis.

Models: We have used the following SLMs to analyse and benchmark toxicity output for the prompts in **D**. The models are (in increasing order of parameter size): GPT-2 Large, Gemma-2B, Mistral-7B, Falcon-7B, and Llama 2-13B. We have opted for 4-bit quantization during inference from Falcon, Mistral, and Llama 2. Fig. 1 elucidates the methodology for evaluating the toxicity of the prompt (t_i) and the continuation (t_o).

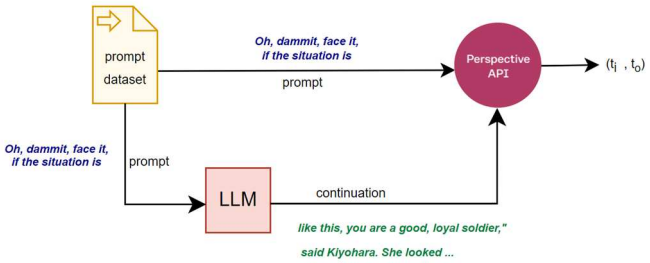


Fig. 1. Methodology for evaluating t_i and t_o

Toxicity ratio: We define toxicity ratio t(z), at a threshold or cutoff z, as the fraction of samples with toxicity score $\geq z$. The toxicity ratio curve is the plot of z vs t(z), $\forall\, z \in [0,1]$, step-size = 0.01. It is useful in visualizing the variation in the probability of perceiving a text as toxic, as the definition of toxicity becomes less stringent.

Consider two hypothetical models $\mathbf{H_1}$ and $\mathbf{H_2}$ for which the toxicity ratio curves are plotted in Fig. 2. As observed from Fig. 2, the toxicity ratio curve for $\mathbf{H_2}$ lies above that of $\mathbf{H_1}$. This implies that model $\mathbf{H_2}$ is more toxic than $\mathbf{H_1}$, since for any threshold z in the given range, there are a greater number of continuations having $t_o \geq z$ in $\mathbf{H_2}$ than in $\mathbf{H_1}$. However, in real-world models, such behaviour is rarely observed; the curves overlap at various values of z, and variation in toxicity is erratic. For some z: $t(z)_{H1} > t(z)_{H2}$, and vice versa.

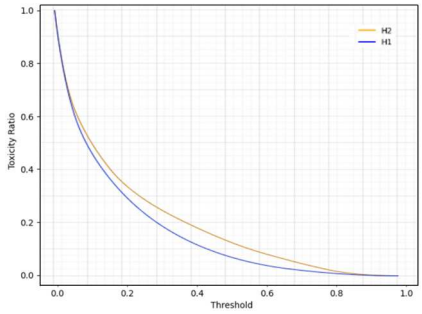


Fig. 2. Toxicity Ratio curves for two hypothetical SLMs $\mathbf{H_1}$ and $\mathbf{H_2}$.

Hence, to quantify the toxicity, we obtain the toxicity ratios $\forall\, z \in [0.1, 0.9]$, step-size = 0.01, and take their weighted mean score τ, where the weight is defined by the logistic function $\sigma(x)$ for a given value of steepness j, and $x = t(z)$.

$$\sigma(x) = \frac{1}{1+e^{-j(x-0.5)}} \tag{1}$$

We discard the ratios in the extreme intervals [0, 0.1) and (0.9, 1.0]. Using a threshold > 0.9 reduces τ, as all the toxicity ratios tend towards zero. Moreover, 0.9 is an adequate threshold to identify the content as toxic. Cutoffs < 0.1 are too low to contribute to τ. The weight function in (1) is shown plotted in Fig. 3.

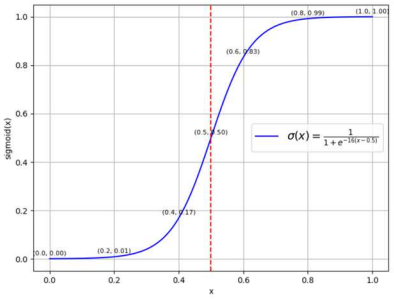


Fig. 3. Weight function y = $\sigma(x)$, where $x = t(z)$

We note that a higher weight should be given to $x > 0.5$ as the perceived toxicity is closer to the actual. Choosing the steepness parameter j as 16 gives a fair estimation of the model toxicity, as per our observation after many empirical trials conducted to determine the optimal value of j. In this paper, we use the terms, toxicity and τ interchangeably. We divide the proposed methodology into a series of tasks described below.

(i) TASK 1: Analysing the effect of internal parameters- Transformers have a Boolean parameter called *do_sample*. When set to true, the model can sample from the distribution of the next likely tokens. The sampling population is controlled by a parameter *top_k*. The randomness is controlled by another parameter called *temperature*. When *do_sample* is set to false, the model returns the most probable token every time (greedy decoding). We analyse the toxicity of the model in both cases i.e. *do_ sample* set to true and false. In the former case, we study the effect of *temperature*. In the latter case, we study the effect of *repetition_penalty*.

(ii) TASK 2: Analysing the efficacy of Word Filtering Detoxification- To mitigate toxicity, we use a simple word filtering technique that prevents the model from choosing a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:07:50 UTC from IEEE Xplore. Restrictions apply.

forbidden list of words (bad_words) as the next token (Fig. 4). We examine the efficacy of this method in detoxification for two SLMs: GPT-2 Large and Mistral-7B.

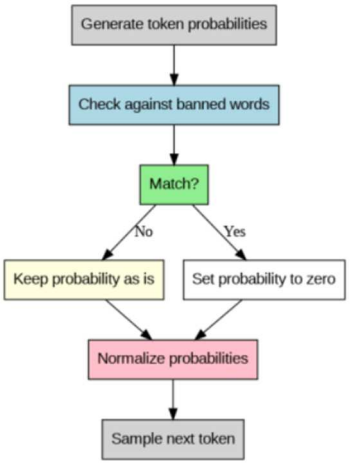


Fig. 4. Token Filtering at decoding time

(iii) TASK 3: Studying the correlation between the toxicity of generations subject to the toxicity of the prompts- We use the principles of fuzzy logic to find the correlation between the toxicity of generations subject to the toxicity of the prompts. We define a fuzzy universal set **T** that comprises the universe of toxic sentences. Next, we compose perspective toxicity and logistic function to obtain a discrete fuzzy membership function. This is followed by computing the memberships of responses conditioned to the membership of prompts. We study these scores across all SLMs and analyse the effect of word filtering on the scores for GPT-2 Large and Mistral-7B.

(iv) TASK 4: Studying the unexpected toxicity of models relative to non-toxic inputs- We analyse the toxicity of the generated continuation t_o for prompts that are almost free from toxicity. We create a sample set **S** of ~500 prompts each having $t_i < 0.2$, we make sure that these prompts are not toxic, and devoid of any hateful vocabulary. Next, we analyse the continued generations obtained by the SLMs for toxicity and benchmark the toxicity t_o relative to **S**.

III. Results

In this section, we describe in detail the results obtained and analyse the toxicity levels of a variety of SLMs, namely, GPT-2 Large, Gemma-2B, Mistral-7B, Falcon-7B, and Llama 2-13B, from different perspectives. PyTorch was used for all the implementations. We describe the results of all four tasks listed in the methodology section of the paper, in a sequential manner.

A. Analysing the effect of internal parameters

a) Sampling: Consider the GPT-2 model. The toxicity ratio curves are shown plotted in Fig. 5 for *do_sample = False* (greedy-decoding) and *do_sample = True* (top_k = 200). We observe from Fig. 5 that $t(z)_{sampling} < t(z)_{no\text{-}sampling}, \forall\, z > 0.2$. The curves give a clear visual description of varying toxicity ratios as a function of z. $\tau_{no\text{-}sampling} = 0.10$ (10%) and $\tau_{sampling} = 0.05$ (5%). This means that continuations generated without sampling are 5% more toxic than those generated with sampling, thus proving that sampling is beneficial for SLMs.

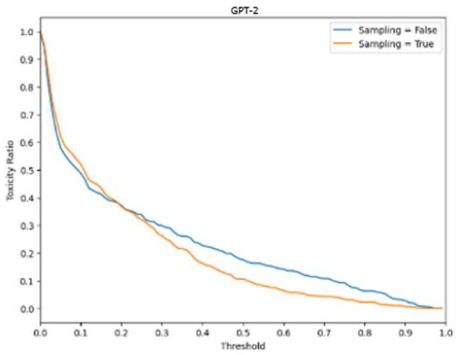
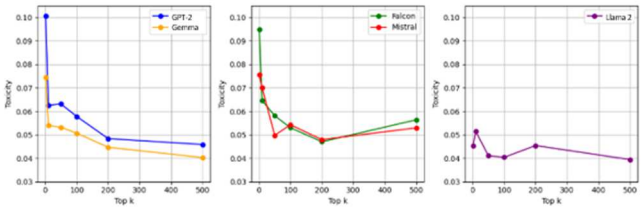


Fig. 5. Toxicity ratio curves for *do_sample=True* and *do_sample=False*.

b) Toxicity vs top_k: We plot the toxicity τ for each model as a function of top_k (at default temperature = 1.0) in Fig. 6. For a fair study, we plot τ vs top_k for similar-sized models together: (GPT-2, Gemma-2B) and (Falcon-7B, Mistral-7B). We study the plot of Llama 2-13B separately. For each plot, we have obtained the τ scores for top_k in the set {1, 10, 50, 100, 200, 500}. It is noted that top_k = 1 is equivalent to greedy decoding.

Consider the first plot for GPT-2 (blue) and Gemma (orange) in Fig. 6 (a). We observe that a growing top_k leads to a nearly continuous decrease in the toxicity τ. There is a considerable initial drop in τ between greedy decoding (top_k = 1) and sampling at top_k = 10. The decrease becomes negligible beyond top_k = 200. Hence for smaller language models, choosing top_k as 200 gives satisfactory results.



(a) (b) (c)

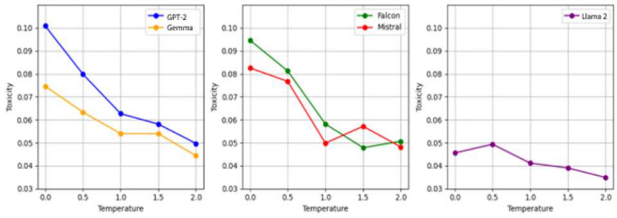
Fig. 6. Plot of τ vs top_k for different SLMs

The following analysis can be drawn from the second plot for Falcon (green) and Mistral (red) in Fig. 6 (b). In the case of Falcon, we find that τ reduces up to top_k = 200, after which irregularity sets in. Regarding Mistral, the erratic behaviour is more noticeable and τ hits its optimum at top_k = 200. Additionally, Falcon has a wider range of toxicity ($\tau_{max} - \tau_{min}$) than Mistral (4.5% vs 2.5%). It should be noted that there is still a considerable drop in τ between greedy decoding and sampling at top_k = 10. Therefore, sampling is still recommended, and selecting top_k = 200 yields the best results. In the case of the plot for Llama 2 in Fig. 6 (c), we do not observe a monotonic pattern. The plot is unstable; τ may occasionally rise with an increase in top_k and vice versa. But we may conclude that τ is nearly constant ($\tau_{max} - \tau_{min} = 1\%$). Thus, it can be said that for larger models such as Llama 2-13b, the top_k effect is not very profound, and for most tasks, we may resort to the default value (top_k = 50). An interesting observation to note is that for most top_k values, Gemma not only outperforms GPT-2 and Falcon but even Mistral. Notably, GPT-2 and Falcon exhibit higher toxicity levels as compared to the other models. Llama 2-13B

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:07:50 UTC from IEEE Xplore. Restrictions apply.

consistently maintains its position as the least toxic model across all examined top_k values. Interestingly, there is a convergence among all models, with τ ~ 5% at top_k = 200.

c) Toxicity vs temperature: We plot the toxicity τ for each model as a function of temperature (at default top_k = 50) in Fig. 7. We plot τ vs temperature for similar-sized models: (GPT-2, Gemma-2B) and (Falcon-7B, Mistral-7B). We study the plot of Llama 2-13B separately.

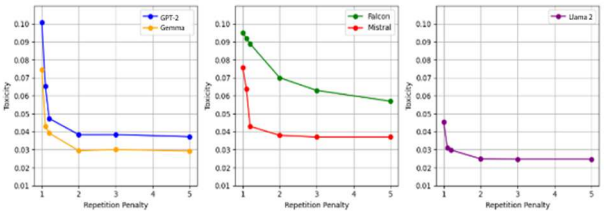


(a) (b) (c)

Fig. 7. Plot of τ vs temperature for different SLMs

From the plot for GPT-2 (blue) and Gemma (orange) in Fig. 7(a) we infer that the toxicity τ decreases continuously with an increase in temperature. It is relatively high for temperature < 1.0 and beyond this, the reduction in toxicity slows down. There is a negligible reduction for temperatures between 1.0 and 1.5, and τ reaches its optimum at the highest temperature of 2.0. Moreover, the toxicity of GPT-2 is much higher than Gemma at t = 0 (~3%). The plot for Falcon (green) and Mistral (red) in Fig. 7(b) suggests that for larger models, the toxicity significantly reduces for temperatures between 0 to 1.0 but beyond that, the behaviour is erratic and toxicity remains almost constant (±1%). τ reaches the near optimum at temperature = 2.0 for both models. From the plot for Llama in Fig. 7(c), we can infer that toxicity nearly decreases with an increase in temperature (except between [0, 0.5]), but the change is negligible. At all values of temperature, Llama 2-13B performs reasonably well with toxicity of ~ 4%. Again, the highest temperature accounts for the least toxicity (3.5%). Hence, we may conclude that for all the SLMs under study, the toxicity reduces with an increase in temperature between 0 to 1.0, beyond which the reduction is negligible and the minimum toxicity prevails at the highest temperature.

d) Toxicity vs repetition penalty: We plot the toxicity τ for each model as a function of repetition penalty for the default setting of no-sampling scenario, in Fig. 8. We plot τ vs temperature for similar-sized models: (GPT-2, Gemma-2B) and (Falcon-7B, Mistral-7B). We study the plot of Llama 2-13B separately. For each plot, we have obtained the τ scores for repetition penalty in the set {1, 1.1, 1.2, 2, 3, 5}. A common observation across all plots is that the toxicity τ continually decreases with an increase in repetition_penalty. There is a considerable initial drop in τ as we go from repetition penalty 1.0 to 1.1 in all models except the Falcon model. In the case of GPT-2 and Mistral, τ drops considerably again as we go from 1.1 to 1.2. For other models, this drastic drop is observed at 2.0. We make the fair assumption that toxicity decreases drastically as the repetition penalty rises above 1.0; this decrease is more pronounced in the beginning, and τ becomes more or less constant beyond 2.0.



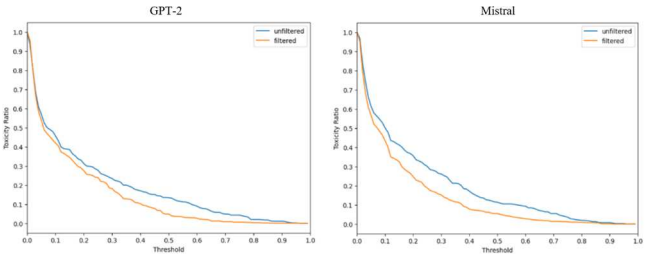
(a) (b) (c)

Fig. 8. Plot of τ vs repetition penalty for different SLMs

Hence for all models, choosing a repetition penalty greater than 1 is always preferred to avoid high toxicity. A close observation of Fig. 8 suggests that Falcon despite its size is considerably more toxic than the rest of the models, even GPT-2. Its plot (green) lies above all other plots, and even at high repetition penalties, τ is considerably high. On the opposite end, we have Llama 2-13B with its curve lying at the very bottom and it exhibits very low toxicity beyond repetition-penalty 1.5 (~ 2%). Next is Gemma-2B which outperforms the remainder models for all values of repetition penalty. We may conclude that Falcon must always be sampled to avoid toxic completions as it is more probable to generate toxic continuations. For other models, we may either choose sampling with temperature ≥ 1 or greedy decoding with a repetition penalty of 1.2 or above.

B. *Analysing the efficacy of Word Filtering Detoxification*

Word filter detoxification is the method of applying a filter that removes a set of "bad words" from the vocabulary of the model. It ensures that no words (tokens) in the bad words list are generated during inference. By passing an appropriate blocklist we can reduce the toxicity at the decoding time. To experiment with word filtering, we create a blocklist of common "bad words" along with their synonyms and inflexions. We experiment with GPT-2 and Mistral-7B, to capture the efficacy of this detoxification in small and large models. Fig. 9 shows the toxicity ratio curves for a) GPT-2 and b) Mistral-7B, with and without word filtering (*do_sample* = True, temperature = 1.5).



(a) (b)

Fig. 9. Toxicity ratio curves for (a) GPT-2 and GPT-2 filtered (b) Mistral and Mistral filtered

In each plot, we observe that $t(z)_{SLM\ filtered} \leq t(z)_{SLM}$, validating the usefulness of applying word filtering detoxification on the responses generated by SLMs. In Fig. 10 we show the plot for toxicity (τ) of the filtered and unfiltered models. Interestingly, the detoxification effect is similar in both models. There is a considerable reduction in toxicity from 5% to 2% in both GPT-2 and Mistral. We will explore more about the filtered continuations in the subsequent sections.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:07:50 UTC from IEEE Xplore. Restrictions apply.

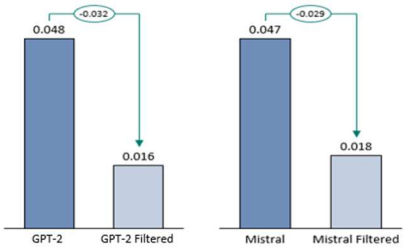


Fig. 10. (a) Bar plot for $\tau_{GPT\text{-}2}$ and $\tau_{GPT\text{-}2\ filtered}$
(b) Bar plot for $\tau_{Mistral}$ and $\tau_{Mistral\ filtered}$

C. *Studying the correlation between the toxicity of generations subject to the toxicity of the prompts*

In this section, we apply the principles of fuzzy logic to find the correlation between the toxicity of generations subject to the toxicity of the prompts. We define a universal fuzzy set **T** that represents toxic language. We can map any toxic sentence *s* to this set by using the membership function μ as shown in (2).

$$\mu(s) = w(\kappa(s)) \tag{2}$$

where κ is the function that maps *s* to its perspective toxicity score. w is the logistic function centred at 0.5 with steepness = 8 as shown in (3).

$$w(\kappa(s)) = \frac{1}{1+e^{-8(\kappa(s)-0.5)}} \tag{3}$$

Now, let **P** and **R** denote the crisp sets of perspective toxicity scores of prompts and responses, respectively.

$$P[i] = \kappa(p[i]) \text{ for i in range(length(}\mathbf{P}\text{))} \tag{4}$$

$$R[i] = \kappa(r[i]) \text{ for i in range(length(}\mathbf{R}\text{))} \tag{5}$$

We define the following fuzzy memberships.

$$\mu(P[i]) = \text{membership of P[i] IN } \mathbf{T} \tag{6}$$

$$\underline{\mu}(P[i]) = \text{membership of P[i] NOT IN } \mathbf{T} \tag{7}$$

$$\mu(R[i]) = \text{membership of R[i] IN } \mathbf{T} \tag{8}$$

$$\underline{\mu}(R[i]) = \text{membership of R[i] NOT IN } \mathbf{T} \tag{9}$$

Now we compute the following weighted memberships of the responses **R** with respect to the prompts **P**.

$$\alpha_{T,T} = \frac{\sum(\mu(P[i]) \times \mu(R[i]))}{\sum \mu(P[i])} \tag{10}$$

$$\alpha_{NT,T} = \frac{\sum(\underline{\mu}(P[i]) \times \mu(R[i]))}{\sum \underline{\mu}(P[i])} \tag{11}$$

$\alpha_{T,T}$ provides a measure of how much the membership of **P** IN **T**, influences the membership of **R** IN **T.** Table I shows $\alpha_{T,T}$ and $\alpha_{NT,T}$ scores for all unfiltered models (at temperature = 1.5). Table I suggests that the membership scores denoting the level of toxicity are the lowest for the Llama 2 and Gemma models.

TABLE I. FUZZY MEMBERSHIP SCORES FOR SMALL LANGUAGE MODELS DENOTING THE TOXICITY LEVEL IN THE RESPONSE GIVEN THE TOXICITY (T) OR NON-TOXICITY (NT) OF THE PROMPT

Model	$\alpha_{T,T}$	$\alpha_{NT,T}$
GPT-2 Large	0.25	0.08
Gemma-2B	0.24	**0.07**
Falcon-7B	0.23	0.10
Mistral-7B	0.24	0.09
Llama 2-13B	**0.22**	0.08

We next plot $\alpha_{T,T}$ and $\alpha_{NT,T}$ for GPT-2 (left bar) and GPT-2 Filtered (right dotted bar) in Fig. 11. We analyse the toxicity scores obtained for GPT-2 and GPT-2 Filtered.

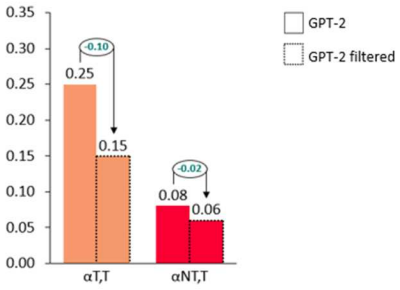


Fig. 11. Comparing the bar plots of GPT-2 and GPT-2 filtered across $\alpha_{T,T}$, and $\alpha_{NT,T}$

(i) $\alpha_{T,T}$ = 0.25. This means that the membership of continuations in **T** given the membership of prompts in **T** is 0.25. This can be reiterated as: *for the set of prompts that are identified as toxic, 25% of them have their completions identified as toxic.*

(ii) The lower the value of $\alpha_{T,T}$, the better the model is at shielding itself from toxic prompts. Hence a drop of 10% for GPT-2 Filtered suggests that the Word filtering strategy is effective in shielding the model against toxic prompts.

(iii) $\alpha_{NT,T}$ = 0.08. This means that the membership of continuations in **T** given the membership of prompts NOT in **T** is 0.08. This can be reiterated as: *for the set of prompts that are identified as non-toxic, 8% of them have their completions identified as toxic.*

(iv) Undoubtedly, $\alpha_{NT,T}$ must be as low as possible for any model as it should not output toxic generations for non-toxic inputs. We observe $\alpha_{NT,T}$ is already low for GPT-2 (8%) and gets only slightly lowered (6%) for GPT-2 Filtered. Hence both models effectively steer clear of generating toxic completions for non-toxic inputs, but word filtering is only negligibly effective in improving this unexpected toxicity score.

(v) A similar analysis can be drawn by observing the scores for Mistral. We conclude that word filtering has a uniform effect across models and helps shield them against toxic inputs. However, it is unable to improve the unexpected toxicity when the inputs are non-toxic.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:07:50 UTC from IEEE Xplore. Restrictions apply.

D. *Analysing the unexpected toxicity of SLMs relative to non-toxic inputs*

Let t_o be the perspective score of prompt continuation and t_i be the perspective score of the prompt itself. We define the unexpected toxicity ratio $t_{THR}(z)$ as the fraction of continuations having $t_o \geq z$ ($z \in [a, 1]$ step-size = 0.01), $\forall$ prompts in sample **S** having $t_i \leq$ THR, where THR is a pre-determined threshold for the toxicity score. This metric is generally useful when:

i) **S** is large, and

ii) a is sufficiently larger than THR

We may add stricter requirements to **S** to customize $t_{THR}(z)$, such as, no words in "bad words list" should be contained in prompts belonging to **S**. For our analysis, we choose THR = 0.2 and sample ~500 prompts each having $t_i < 0.2$. Additionally, we make sure to choose those prompts that are devoid of any hateful vocabulary. We generate the continuations for all models at temperature = 1.5 and obtain the unexpected toxicity ratio curve in Fig. 12.

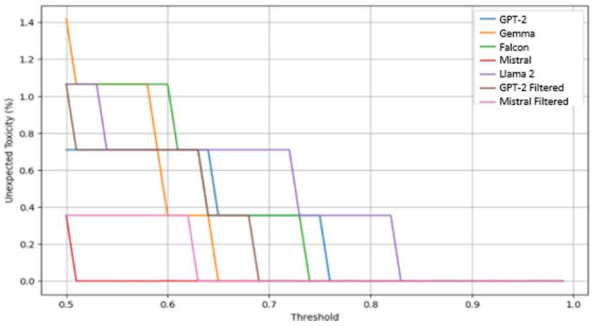


Fig. 12. Benchmarking the unexpected toxicity percentage for all models

Fig. 12 shows the thresholds [0.5, 1] on the x-axis and the unexpected toxicity (percentage) on the y-axis. We can observe that all models exhibit very low unexpected toxicity < 1.5%. This is a good indication that the models are not naturally inclined to produce toxic continuations to non-toxic inputs. A close look at the curve shows the effect of filtering has a negligible effect on the unexpected toxicity for both Mistral and GPT-2 and this is not surprising as word filtering has limited effectiveness in avoiding toxicity that does not arise from the explicit use of toxic language.

IV. Conclusion

We introduce the toxicity ratio curve as a graphical visualization of toxicity in Small Language Models (SLMs) and propose the weighted average toxicity score as a realistic measure for gauging the toxic content of text generated by SLMs. We study the effect of internal parameters: *do_sample*, *top_k*, *temperature*, and *repetition_penalty* on the toxicity of the text generated, and propose optimal parameters for the SLMs under study. Our results on word filter detoxification suggest that it is indeed a crude but effective strategy in mitigating toxicities by shielding the models from toxic inputs. However, it is limited in its effect as the seemingly clean continuations may still turn out to be toxic without the use of toxic vocabulary. Our definition of unexpected toxicity ratio captures the instances when the model produces toxic completions to seemingly nontoxic inputs. We conclude that all SLMs have low unexpected toxicity which does not improve with the use of word filter detoxification. We also introduce the concept of fuzzy toxicity following the principles of fuzzy logic and examined the correlation between the toxicity of generations subject to the toxicity of the prompts. The analysis of weighted fuzzy toxicities establishes a similar profile across all models and an improvement in the filtered models.

References

[1] Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

[2] Choudhry, Arjun, Seba Susan, Anmol Bansal, and Anubhav Sharma. "Tlmote: A topic-based language modelling approach for text oversampling." In *The International FLAIRS Conference Proceedings*, vol. 35. 2022.

[3] Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. "A survey on evaluation of large language models." *ACM Transactions on Intelligent Systems and Technology* (2023).

[4] Schick, Timo, and Hinrich Schütze. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2339-2352. 2021.

[5] Li, Shiyao, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. "Evaluating Quantized Large Language Models." *arXiv preprint arXiv:2402.18158* (2024).

[6] Ousidhoum, Nedjma, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. "Probing toxic content in large pre-trained language models." 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. 4262-4274. 2021.

[7] Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356-3369. 2020.

[8] Welbl, Johannes, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. "Challenges in Detoxifying Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2447-2469. 2021.

[9] He, Xinlei, Savvas Zannettou, Yun Shen, and Yang Zhang. "You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content." In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 61-61. IEEE Computer Society, 2023.

[10] Yi, Xin, Linlin Wang, Xiaoling Wang, and Liang He. "Fine-Grained Detoxification via Instance-Level Prefixes for Large Language Models." *arXiv preprint arXiv:2402.15202* (2024).

[11] Wen, Jiaxin, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. "Unveiling the Implicit Toxicity in Large Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1322-1338. 2023.

[12] Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng et al. "Ethical and social risks of harm from language models." *arXiv preprint arXiv:2112.04359* (2021).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:07:50 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/384240872

Volatility Clustering and Long-Memory -a FIGARCH Analysis of Selected Currency Pairs

Preprint · September 2024

DOI: 10.13140/RG.2.2.24466.47049

CITATIONS
0

READS
26

2 authors, including:



Sanjay Kumar
Dyal Singh College
6 PUBLICATIONS **2** CITATIONS

SEE PROFILE

All content following this page was uploaded by Sanjay Kumar on 22 September 2024.

The user has requested enhancement of the downloaded file.

Volatility Clustering and Long- Memory -a FIGARCH Analysis of Selected Currency Pairs

Sanjay Kumar*[1] and Dr. Nand Kumar†[2]

[1,2]Department of Humanities, Delhi Technological University, New Delhi.

September 22, 2024

*M. Phil in Economics. Research Scholar, Department of Humanities, Delhi Technological University, New Delhi. Dyal Singh College, Delhi University, New Delhi.sanjaykumar.eco@dsc.du.ac.in Orcid: http://orcid.org/0000-0003-4613-076X

†Ph.D. in Economics.Professor, Department of Humanities, Delhi Technological University, New Delhi.nandkumar@dce.ac.in Orcid: http://orcid.org/0000-0002-5684-6078

Abstract

This paper employs the Fractionally Integrated GARCH (FIGARCH) model to analyze the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD, using data from January 4, 1999, to August 6, 2021. Before model fitting, the squared-innovation series is examined for the presence of long memory through informal and formal methods. While the Hurst Exponent (H) does not indicate long memory, the Local Whittle Estimator and GPH estimator suggest evidence of long memory for all pairs except JPY-USD.

The findings of the FIGARCH study reveal significant short-term volatility clustering and persistence in volatility, with negligible average returns for the currency pairs. These results contribute to the existing literature on financial time series analysis and have practical implications for market participants. Understanding the presence of long memory in volatility can help traders, risk managers, and policymakers anticipate and manage risks associated with extreme market events and volatility clustering, leading to more informed decision-making in currency markets. The study highlights the importance of considering long memory in volatility when analysing and modelling exchange rate dynamics.

Keywords: FIGARCH, long memory, volatility clustering, exchange rates, risk management.

JEL Classification: C22, C58, F31, G15

1 Introduction

The complex dynamics of financial time series, particularly in foreign exchange markets, have long fascinated researchers and practitioners alike. Three interrelated and stylized facts about financial time series resulting from autocorrelation-induced dependency are long memory, volatility clusters, and fat tails (Cont, 2001). Accurately modeling and forecasting the volatility of exchange rates is crucial for effective risk management, asset allocation, and derivative pricing (Andersen et al., 2000).

The seminal works of Engle (1982) and Bollerslev (1986) introduced the ARCH and GARCH models, respectively, which have become the standard tools for capturing volatility clustering in financial time series. However, these models fail to account for the long memory property often observed in the volatility of asset returns (Baillie et al., 1996). Long memory, also known as long-range dependence or persistence, refers to the slow decay of autocorrelations at long lags, implying that past volatility has a long-lasting impact on future volatility (McElroy and Politis, 2020).

To address this shortcoming, Baillie et al. (1996) introduced the Fractionally Integrated GARCH (FIGARCH) model, which extends the GARCH framework by allowing for fractional integration in the volatility process. The FIGARCH model

has gained popularity due to its ability to capture both short-term volatility clustering and long memory in financial time series (Davidson, 2004; Belkhouja and Boutahary, 2011).

Despite the growing body of literature on FIGARCH models, the understanding of long memory in the volatility of foreign exchange rates remains limited, particularly for emerging market currencies. Moreover, the practical implications of long memory for market participants, such as traders, risk managers, and policymakers, have not been fully explored.

This paper aims to fill these gaps by employing the FIGARCH model to analyse the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD. By examining the presence of long memory in the volatility of these currency pairs, we contribute to the existing literature on financial time series analysis and provide valuable insights for market participants.

Our study is motivated by the need for a better understanding of the long-term dynamics of exchange rate volatility and its potential implications for risk management and decision-making in currency markets. The inclusion of an emerging market currency (INR) alongside major developed market currencies (EUR, GBP, JPY) allows for a comparative analysis and sheds light on the differences in volatility characteristics across markets.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature on FIGARCH models and their applications in finance. Section 3 describes the data and methodology employed in our analysis. Section 4 presents the empirical results and discusses their implications. Finally, Section 5 concludes the paper and offers suggestions for future research.

2 Literature Review

The application of fractionally integrated models to capture long memory in financial time series has gained significant attention in the literature. Granger (1980) introduced the concept of long memory and its implications for economic and financial time series. Hosking (1981) and Geweke and Porter-Hudak (1983) further developed the theory and estimation methods for fractionally integrated processes.

In the context of volatility modeling, Ding et al. (1993) found evidence of long memory in the absolute and squared returns of the SP 500 index. Andersen and Bollerslev (1997) and Baillie et al. (1996) confirmed the presence of long memory in the volatility of foreign exchange rates using high-frequency data and the FIGARCH model, respectively.

Since the introduction of the FIGARCH model by Baillie et al. (1996), numerous

studies have applied the model to various financial assets. Bollerslev and Mikkelsen (1996) used the FIGARCH model to analyze the volatility of exchange rates and found evidence of long memory in the Deutsche Mark-US Dollar and Japanese Yen-US Dollar rates. Tse (1998) applied the FIGARCH model to stock market indices and demonstrated its superiority over the GARCH model in capturing long memory.

Beine et al. (2002) investigated the impact of central bank interventions on the volatility of exchange rates using the FIGARCH model. They found that accounting for long memory and structural breaks improves the fit of the model. Conrad and Haag (2006) examined the persistence of volatility in the FIGARCH model and derived the impulse response function for the conditional variance.

More recently, Kiliç (2011) proposed a smooth transition FIGARCH (ST-FIGARCH) model to capture both long memory and nonlinearity in the volatility of exchange rates and stock market indices. Belkhouja and Boutahary (2011) employed a time-varying FIGARCH model to study the volatility of crude oil prices and found evidence of long memory and structural changes.

In the context of emerging markets, Goudarzi and Ramanarayanan (2010) applied the FIGARCH model to the Indian stock market and found strong evidence of long memory in volatility. Oberholzer and Venter (2015) used the FIGARCH model to

analyze the volatility of the South African Rand against major currencies and found that long memory is more pronounced in the Rand-US Dollar exchange rate compared to other currency pairs.

Several studies have compared the performance of FIGARCH with other long memory volatility models. Davidson (2004) introduced the hyperbolic GARCH (HYGARCH) model and showed that it outperforms the FIGARCH model in terms of goodness-of-fit and forecasting ability. Yalama and Sevil (2008) compared the forecasting performance of FIGARCH, HYGARCH, and fractionally integrated APARCH (FIAPARCH) models for the Turkish stock market and found that the FIAPARCH model provides the best out-of-sample forecasts.

3 Research Objectives

The primary objective of this study is to analyze the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD, using the Fractionally Integrated GARCH (FIGARCH) model. Specifically, we aim to:

1. Examine the presence of long memory in the volatility of the selected currency pairs using informal methods, such as the autocorrelation function and

periodogram, and formal tests, including the Hurst Exponent, Local Whittle Estimator, and GPH estimator.

2. Estimate the FIGARCH model for each currency pair to capture short-term volatility clustering and long-term persistence in volatility.

3. Assess the significance of the FIGARCH model parameters, particularly the fractional differencing parameter (d), which indicates the degree of long memory in volatility.

4. Investigate the practical implications of the findings for market participants, such as traders, risk managers, and policymakers, in terms of understanding and managing risks associated with extreme market events and volatility clustering.

The relevance of this study lies in its focus on four currency pairs that play a significant role in global financial markets. By examining the presence of long memory in the volatility of these pairs, we contribute to a deeper understanding of how volatility dynamics manifest in these specific markets. The inclusion of an emerging market currency (INR) alongside major developed market currencies (EUR, GBP, JPY) allows for a comparative analysis and provides insights into potential differences in volatility characteristics across markets.

Furthermore, this research aims to bridge the gap between theoretical models and practical applications by discussing the implications of long memory in volatility for market participants. Understanding the persistence of volatility can help investors, risk managers, and policymakers make more informed decisions when dealing with currency risk, portfolio optimisation, and market stability.

Overall, this study contributes to the existing literature on financial time series analysis by providing a comprehensive examination of long memory in the volatility of key currency pairs using the FIGARCH model, while also highlighting the practical significance of the findings for stakeholders in the foreign exchange market.

4 Methodology

4.1 The FIGARCH Process

Let t be an index of time and $P_t, t \in T$ be the observed price of an asset. Let $r_t = \ln P_t - \ln P_{t-1}$ be its return r_t such that its parameters θ are the maximum likelihood ($L(\theta)$) estimates of an appropriate mean model with pre-chosen density function (f_θ). Then we have:

$$L(\theta; r_1, r_2, \ldots, r_T) = \prod_{t=1}^{T} \log(f_\theta(r_t)) \tag{4.1}$$

The log-likelihood will be given by:

$$l(\theta; r_1, r_2, \ldots, r_T) = \sum_{t=1}^{T} \log(f_\theta(r_t)) \tag{4.2}$$

To get the estimator of θ, we maximize $l(\theta)$. Then, the innovations ϵ_t are:

$$\epsilon_t = r_t - \theta(B) r_t \tag{4.3}$$

where B is a backshift operator, $\theta(B) = \sum_{i=1}^{n} \theta_i B^i$, and ϵ_t is white noise with

$\mathrm{E}(\epsilon_t) = 0$ and $\mathrm{E}(\epsilon_t^2) = \sigma_t^2$.

The GARCH(p, q) model (Bollerslev, 1986) can be given as:

$$\sigma_t^2 = \omega + \alpha(B)\epsilon_t^2 + \beta(B)\sigma_t^2 \tag{4.4}$$

where $\omega, \alpha, \beta > 0$, and conditional variance is stationary for $\alpha + \beta < 1$.

$\alpha(B) = \sum_{i=1}^{q} \alpha_i B^i$ and $\beta(B) = \sum_{j=1}^{p} \beta_j B^j$. For $q = 0$, it is an ARCH(p) process,

and for $p = q = 0$, it becomes a white noise process. After rearranging, we get:

$$(1 - \alpha(B) - \beta(B))\epsilon_t^2 = \omega + (1 - \beta(B))(\sigma_t^2 - \epsilon_t^2) \tag{4.5}$$

Hence,

$$(1 - B)\phi(B)\epsilon_t^2 = \omega + (1 - \beta(B))(\sigma_t^2 - \epsilon_t^2) \tag{4.6}$$

Here, $\phi(B) = \sum_{i=1}^{m-1} \phi_i B^i$. In the fractionally integrated model, $(1 - B)$ is replaced by $(1 - B)^d$ for $0 < d < 1$. This is done to capture the slow hyperbolic decay of memory. Thus, we can write the FIGARCH(p, d, q) model equation as:

$$(1 - B)^d\phi(B)\epsilon_t^2 = \omega + (1 - \beta(B))(\sigma_t^2 - \epsilon_t^2) \tag{4.7}$$

We have:

$$(1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \tag{4.8}$$

Solving Equation (7) and truncating the lag polynomial at 1000 for operational

purposes (Baillie et al., 1996), we have the FIGARCH(p, d, q) model with $0 < d < 1$ (allowing for slow decline of volatility after volatility shocks, a phenomenon called long memory) as:

$$\sigma_t^2 = (\omega - \bar{\epsilon}^2) + \sum_{j=1}^{q} \alpha_j(\epsilon_{t-j}^2 - \bar{\epsilon}t - j^2) + \sum j = 1^p \beta_j(\sigma_{t-j}^2 - \epsilon_{t-j}^2) \tag{4.9}$$

Thus, GARCH(1, 1) and GARCH(1, 1, 1) can be taken as special cases of the FIGARCH(1, d, 1) model with $d = 0$ and $d = 1$, respectively.

4.2 Data and Model Estimation

The data set consists of daily exchange rates for four currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD. The sample period spans from January 4, 1999, to August 6, 2021, resulting in 5,669 observations for each currency pair.

We first calculate the logarithmic returns for each currency pair as $r_t = \ln P_t - \ln P_{t-1}$, where P_t is the exchange rate at time t. Following Baillie et al. (1996), we fit the mean model $r_t = \mu + \epsilon_t$ to all four return series.

Before estimating the FIGARCH model, we examine the squared returns for the presence of long memory using informal methods, such as the autocorrelation

function (ACF) and periodogram, and formal tests, including the Hurst Exponent (H), Local Whittle Estimator, and GPH estimator.

Finally, we estimate the FIGARCH(1, d, 1) model for each currency pair using the maximum likelihood method, assuming normal and Student's t distributions for the standardized innovations. The model parameters, including the fractional differencing parameter (d), are assessed for statistical significance, and the implications of the findings are discussed.

5 Research Findings

In this section, we present the results of the FIGARCH analysis of the squared returns of four selected currency pairs: Euro-USD, GBP-USD, JPY-USD, and INR-USD. The FIGARCH model, an extension of the GARCH model, is used to capture volatility and long memory, specifically to model volatility clustering. The model is specified in terms of a univariate conditional mean model, a variance model, a distribution model, and a fractional integration parameter. Before fitting the model to the data, we examine the squared-innovation data for the presence of long memory using the autocorrelation function and periodogram. Additionally, we analyze the data using the Hurst Exponent (H) (Hurst, 1956), Local Whittle

Estimator (Robinson, 1995), and GPH estimator (Geweke and Porter-Hudak, 1983).

The ADF test, tau statistics value (with p-values indicated in parentheses) for error terms of the four mean models with constant is -15.4397 (1.858e-036), -17.5335 (1.858e-036), -27.3391 (1.539e-051), and -76.5754 (0.0001), respectively. Very low p-values indicate that we can reject the null that the four series have unit root. Table 1 shows the summary statistics of the error terms of the mean models of all the four daily return series. All the four series are fat-tailed as indicated by excess kurtosis values. All the popular tests of normality point to the return series being non-normally distributed.

Table 1: Summary Statistics, Returns, 1999-21

Statistics/Currency Pairs	**EUROUSD**	**GBPUSD**	**INRUSD**	**JPYUSD**
Mean	-1.1643e-019	-1.1781e-020	-9.8762e-020	5.1043e-018
Median	-7.6303e-007	8.5727e-005	-9.7861e-005	-0.0012322
Maximum	0.044380	0.044380	0.039285	6.9988
Minimum	-0.081662	-0.081662	- - -0.037658	-0.053389
Std. Dev.	0.0058991	0.0058991	0.0042961	0.093181
Skewness	-0.099739	-0.65764	0.20986	74.763
Ex-Kurtosis	2.4824	10.688	10.810	5612.9
Jarque-Bera Test	1465	27390.6	27643.1	7.44703e+009
P Value	0	0	0	0
Doornik-Hansen Test	808.679	5119.98	6182.23	3.58164e+007
P Value	2.49787e-176	0	0	0
Shapiro-Wilk Test	0.978187	0.946218	0.879669	0.0138291
P Value	1.24335e-028	22.44678e-041	9.7656e-055	6.37173e-099
Lilliefors Test	0.048886	0.0446517	0.109872	0.4295
P Value	0	0	0	0

Source: Author's calculation.

Bai and Ng (2005), with the help of their test specially designed for time series,

tested the normality of 21 time series out of these six-time series are found to be non-symmetric and all of them are financial time series. In Table 2, we have presented the result of the Bai & Ng Normality Test for error terms of our mean models for all the four series as most popular tests for normality may not be applicable to financial time series. All return series seem to be non-normally distributed.

Table 2: Bai & Ng Long Run Normality Test, Returns 1999-21

Currency Pairs	π_3	$\pi_3/\sqrt{5}$	π_4	π_{34}
EuroUSD	1.187239	2.213913	3.498400	19.98223
	(0.117567)	(0.013417)	(0.000234)	(4,58E-05)
GBPUSD	0.878083	0.952704	1.343738	3.781008
	(0.189949)	(0.170370)	(0.089517)	(0.150996)
INRUSD	0.306107	0.458222	3.031943	10.32863
	(0.379762)	(0.323397)	(0.001215)	(0.005717)
JPYUSD	1.440571	2.335931	2.854371	17.10611
	(0.074853)	(0.009747)	(0.002156)	(0.000193)

Source: Author's calculation.

Note: p-values are in parentheses. As per Bai and Ng (2005), $\pi_3, \pi_4 \rightarrow N$, $\pi_{34} \rightarrow \chi^2$, 5% critical value is 1.96 for π_3, π_4 and 5.99 for π_{34} at 2df.

In Figure 2, we have plotted error square (ϵ^2) of the mean model of the FIGARCH process of Baillie et al. (1996).

An examination of Figure 2 reveals that all (ϵ^2) series, except JPY-USD series, have GARCH effect as well as long memory. In the case of Euro-USD, GBP-USD, and

INR-USD currency pairs, we have volatility and clustering around the 2008 global financial crisis. For GBP-USD, we have high volatility and clustering around the Brexit period, and in the case of Euro-USD, we also have high volatility clustering during the Brexit period. For autocorrelated series with k lag, we have $\epsilon^2 \propto \sum_k^T \rho_k$, i.e., clusters in the return series will be caused if lagged observations are significantly correlated, individually or jointly. Individual dependencies can be gauged by autocorrelation function at different lags and joint dependencies can be gauged by portmanteau tests. Long memory is indicated by a slow or hyperbolic decline of autocorrelation as opposed to fast or geometric decline in case of short-term memory. Figure 3 presents ACF of (ϵ^2) of all the four currency pairs. The autocorrelation function (ACF) at lag k is given by the power rule $\rho_k = c_\rho k^{2d-1}$ with $k \to \infty$, where c_ρ is a constant and d is a memory parameter. We have a sample size of 5669 for our series, hence 95% confidence interval for autocorrelations, as shown by the horizontal bars in Figure 3, using Bartlett standard error will be given by $\pm\frac{1.96}{\sqrt{5669}} = \pm 0.026$. The volatility clustering phenomenon seen in Figure 1 is confirmed by ACFs given in Figure 2 . The (ϵ^2) autocorrelations for very long lags frequently exceed the two 95% Bartlett (1946) confidence bands except for JPY-USD series. Also, the Ljung and Box (1978) portmanteau test for the joint significance of autocorrelations is quite significant for all lags with p-value equal to zero and this is for all the (ϵ^2) series except for

JPY-USD series for which we do not have any significant Q statistics]. This apparent persistence in the autocorrelation function is substantially reduced if we take the first difference of the (ϵ^2) series. This is indicated by reduced values for ACF though these ACF values are still significant at quite distant lags and the overall significance shown by Q statistics is still very high. The frequency domain definition of long memory states that the spectral density function is unbounded at some frequency in the interval [0,). Most of the empirical literature has concentrated on the case where the singularity, or pole, in the spectrum takes place at the zero frequency (Gil-Alana and Hualde, 2009). We present periodogram, with Bartlett window length of 150, of (ϵ^2) for all the four series in Figure 4. The figure reveals long memory in the case of all the four return series.

Following Engle and Bollerslev (1986), we will use fractionally differenced return series to capture the long memory in the series. Taylor (2008) studied power transformation of the return from a number of time series and found that squared returns are more likely to be autocorrelated than returns. For a formal examination of long memory, if any, in ϵ^2, we use Hurst Exponent (H) (Hurst et al., 1965), Local Whittle Estimator (Robinson, 1995), and GPH estimator (Geweke and Porter-Hudak, 1983) of memory parameter d. The results are reported in Table 3 and Table4, respectively. We do not find evidence of long memory in any of the series on basis of Hurst Exponent (H), but on the basis of Local Whittle Estimator

and GPH estimator, we find evidence of long memory. So we don't have a conclusive outcome here](Conclusion). Following Weron and Weron (2002), we estimate Hurst Exponent (H) = 0.5 + Empirical Hurst Exponent – Anis-Lloyd corrected Hurst Exponent = 0.5 + slope of (R/S)n – E(R/S)n. Here (R/S)n is empirical Hurst Exponent estimated by (R/S)n = CHn and E(R/S)n is the expected or theoretical Hurst Exponent at size n (Anis and Lloyd, 1976). There is a lot of debate in literature on the asymptomatic property of Hurst Exponent, original R/S analysis gave only the true value of the parameter H but it did not provide its probability distribution. Some researchers have shown that the original Hurst Exponent is normally distributed. Couillard and Davison (2005) studied 10,000 Hurst exponents from independent time series of 10,000 observations each and found them to be normally distributed. In this paper, we use 95% confidence interval suggested by Weron and Weron (2002) for Anis-Lloyd corrected H (R/S-AL). The estimated value of Empirical Hurst Exponent, Theoretical Hurst Exponent, and Anis-Lloyd corrected H (R/S-AL) Exponent, along with 95% confidence interval for Anis-Lloyd corrected H (R/S-AL), is reported in Table 3 .

GPH test, $H_0 : d \geq 0.5$, no presence of long memory; $H_1 : d < 0.5$ presence of long memory. LW test, $H_0 : d \geq 0.5$, no presence of long memory; $H_1 : d < 0.5$ presence of long memory, Level of Significance 5%.

Table 3: Hurst Exponent Test for Long Memory in Square Returns

	1.INR USD, Sample Size (T) = 5669	**2.EUROUSD, Sample Size (T) = 5669**	**3.GBPUSD, Sample Size (T) = 5669**	**4.JPYUSD, Sample Size (T) = 5669**
Empirical Hurst Exponent H(R/S)n	0.7771236	0.8956424	0.8956424	0.8229069
Theoretical Hurst Exponent H(E(R/S)n)	0.5287461	0.5287461	0.5287461	0.5287461
Anis-Lloyd Corrected Hurst Exponent H(R/S-AL) (= 0.5 + H(R/S)n - H(E(R/S)n))	0.7483775	0.8668963	0.8604084	0.7908433
95% Confidence Interval for Anis-Lloyd Hurst Exponent	-0.49931 <R/S-AL< 1.408454	-0.49931 <R/S-AL< 1.408454	-0.49931 <R/S-AL< 1.408454	-0.49931 <R/S-AL< 1.408454
H_0	Not rejected	Not rejected	Not rejected	Not rejected
Conclusion	No Long memory	No Long memory	No Long memory	No Long memory

Source: Author's calculation.

Table 4: Frequency Domain Long Memory Estimators

	1.INR USD, Sample Size (T) = 5669	**2.EUROUSD, Sample Size (T) = 5669**	**3.GBPUSD, Sample Size (T) = 5669**	**4.JPYUSD, Sample Size (T) = 5669**
Local Whittle Estimator (Robinson, 1995)				
m	177	177	177	177
d	0.291183	0.452571	0.298797	0.344853
Standard Error	0.0375823	0.0375823	0.0375823	0.0375823
Test statistic: z =	7.74786	12.0421	7.95048	9.17595
p-value	0.0000	0.0000	0.0000	0.0000
Conclusion	Reject the null. Long memory	Reject null. Long memory	Reject null. Long memory	Reject null. Long memory
GPH Estimator				
m	177	177	177	177
d	0.293947	0.518973	0.297206	0.40157
Standard Error	0.0465883	0.0502571	0.0393567	0.0489204
Test statistic: t(175) =	6.30947	0.0502571	7.55159	8.20865
p-value	0.0000	0.0000	0.0000	0.0000
Conclusion	Reject the null. Long memory	Reject the null. Long memory	Reject the null. Long memory	Reject the null. Long memory

Source: Author's calculation.

The estimated values of the FIGARCH(1, d, 1) model with normal and Student's t distributions for the four currency pairs are presented in Table 5 and Table 6, respectively. The mean return (μ) is estimated to be very close to zero for all currency pairs, suggesting negligible average returns. The constant term (ω) in the volatility equation is not statistically significant for any of the pairs, indicating that the models do not require a constant mean or variance component.

However, both past squared shocks (α_1) and long memory (δ) have significant impacts on current volatility for all currency pairs, suggesting the presence of short-term volatility clustering and strong persistence in volatility. The Student's t distribution parameter (ν) is also significant for all currency pairs, indicating the presence of heavy tails in the standardized residuals.

Table 5: FIGARCH(1, d, 1) Mean Model: ARFIMA(0, 0, 0) Distribution Model: Normal. Optimal Parameters

Currency Pairs	μ	ω	α_1	β_1	δ
EuroUSD	-0.000036	0.000000	0.051781	0.937948	0.887184
	(0.57868)	(0.37205)	(0.00000)	(0.00000)	(0.00000)
GBPUSD	0.000031	0.000001	0.242715	0.654901	0.446753
	(0.641479)	(0.007183)	(0.000000)	(0.000000)	(0.000000)
INRUSD	0.000016	0.000000	0.151041	0.914590	0.956082
	(0.626094)	(0.012456)	(0.000000)	(0.000000)	(0.000000)
JPYUSD	0.000016	0.000000	0.151041	0.914590	0.956082
	(0.626094)	(0.012456)	(0.000000)	(0.000000)	(0.000000)

Source: Author's calculation.

These findings have several important implications for market participants. The presence of long memory in volatility suggests that extreme market events and

Table 6: FIGARCH(1, d, 1) Mean Model: ARFIMA(0, 0, 0) Distribution Model: Student's t distribution. Optimal Parameters

Currency Pairs	μ	ω	α_1	β_1	δ	ν
EuroUSD	-0.000030	0.000000	0.052874	0.937948	0.883961	8.385051
	(0.634173)	(0.144420)	(0.000005)	(0.00000)	(0.00000)	(0.000000)
GBPUSD	0.000055	0.000000	0.004340	0.951721	0.984915	8.313940
	(0.39247)	(0.19537)	(0.82997)	(0.000000)	(0.000000)	(0.00000-)
INRUSD	-0.000037	0.000000	0.110330	0.888670	0.956082	4.250676
	(0.080559)	(0.934861)	(0.000000)	(0.000000)	(0.000000)	(0.00000)
JPYUSD	0.000016	0.000000	0.151041	0.914590	0.956082	8.310940
	(0.626094)	(0.012456)	(0.000000)	(0.000000)	(0.000000)	(0.00000)

Source: Author's calculation.

volatility clustering are more likely to occur than in a standard GARCH setting, and market participants should be prepared for prolonged periods of high or low volatility. The heavy tails of the standardized residuals also highlight the importance of accounting for extreme events in risk models and decision-making processes.

6 Summary and Conclusion

In this study, we employed the Fractionally Integrated GARCH (FIGARCH) model to analyze the volatility of daily USD returns in four major currency pairs: Euro-USD, GBP-USD, INR-USD, and JPY-USD, using data from January 4, 1999, to August 6, 2021. Our primary objectives were to examine the presence of long memory in volatility, estimate the FIGARCH model to capture short-term volatility clustering and long-term persistence, and investigate the practical implications of

our findings for market participants.

In summary, our FIGARCH analysis of the daily USD returns for four major currency pairs (Euro-USD, GBP-USD, INR-USD, and JPY-USD) reveals several important findings. The preliminary analysis shows that the return series are stationary but exhibit fat tails and non-normality. The squared returns display volatility clustering and long memory, particularly for the Euro-USD, GBP-USD, and INR-USD pairs, as evidenced by the autocorrelation function and periodogram.

The formal tests for long memory provide mixed results, with the Hurst Exponent not indicating long memory for any of the series, while the Local Whittle Estimator and GPH estimator suggest evidence of long memory for all pairs except JPY-USD. The FIGARCH(1, d, 1) model estimates confirm the presence of short-term volatility clustering and strong persistence in volatility for all currency pairs, as indicated by the significant past squared shocks (α_1) and long memory (δ) parameters. The heavy tails of the standardized residuals are also evident from the significant Student's t distribution parameter (ν).

Our study contributes to the existing literature on financial time series analysis by providing a comprehensive examination of long memory in the volatility of key currency pairs using the FIGARCH model. The inclusion of an emerging market currency (INR) alongside major developed market currencies (EUR, GBP, JPY)

allows for a comparative analysis and provides insights into potential differences in volatility characteristics across markets.

7 Policy Prescription

The findings of our study have several important implications for policymakers, particularly central banks and financial regulators. Understanding the presence of long memory in the volatility of major currency pairs can help inform policy decisions related to exchange rate stability, financial market regulation, and risk management.

First, central banks should take into account the persistence of volatility when designing and implementing exchange rate policies. The presence of long memory in volatility suggests that extreme market events and prolonged periods of high or low volatility are more likely to occur than in a standard GARCH setting. This implies that central banks may need to be prepared for more frequent and longer-lasting interventions in foreign exchange markets to maintain stability and prevent excessive fluctuations in exchange rates.

Second, financial regulators should consider the implications of long memory in volatility when setting capital requirements and other risk management standards

for financial institutions. The heavy tails of the standardized residuals in our FIGARCH model estimates highlight the importance of accounting for extreme events in risk models. Regulators may need to require financial institutions to hold more capital or maintain higher liquidity buffers to cushion against potential losses arising from prolonged periods of high volatility.

Third, policymakers should promote the development and use of more sophisticated risk management tools that take into account the presence of long memory in volatility. This could include encouraging the adoption of long memory volatility models, such as FIGARCH, HYGARCH, or FIAPARCH, by financial institutions and market participants. Policymakers could also support research and education initiatives aimed at improving the understanding and application of these models in practice.

Fourth, given the evidence of long memory in volatility, policymakers should be cautious about relying too heavily on short-term measures of volatility, such as implied volatility indices or short-term historical volatility, when making policy decisions. These measures may not fully capture the long-term persistence of volatility and could lead to suboptimal policy choices. Instead, policymakers should consider using longer-term measures of volatility and incorporating the insights from long memory models when formulating policies.

Finally, our findings highlight the importance of international cooperation and information sharing among policymakers and regulators. The presence of long memory in volatility in major currency pairs suggests that volatility shocks in one market can have long-lasting effects on other markets. This underscores the need for policymakers to work together to monitor and manage risks arising from cross-border financial market linkages.

8 Limitations and Future Research

While our study provides valuable insights into the presence of long memory in the volatility of major currency pairs and its implications for market participants and policymakers, it is important to acknowledge the limitations of our analysis and identify areas for future research.

One limitation of our study is the mixed results from the formal tests for long memory. While the Local Whittle Estimator and GPH estimator suggested evidence of long memory for all currency pairs except JPY-USD, the Hurst Exponent did not indicate long memory for any of the series. These mixed results suggest that further research may be needed to fully understand the nature of long memory in these currency pairs and to develop more robust tests for detecting long memory in

financial time series.

Another limitation is that our analysis focuses on a specific time period (1999-2021) and may not capture potential changes in volatility dynamics over time. Future research could extend our analysis by considering longer time periods or by using rolling window estimation techniques to examine how the degree of long memory in volatility evolves over time. This could provide insights into how changes in economic conditions, financial market structure, or policy regimes affect the persistence of volatility in foreign exchange markets.

Our study also focuses on a limited set of currency pairs (Euro-USD, GBP-USD, INR-USD, and JPY-USD). While these pairs represent a mix of major developed and emerging market currencies, future research could extend the analysis to a broader set of currency pairs to provide a more comprehensive understanding of long memory in foreign exchange market volatility. This could include examining the presence of long memory in the volatility of other major currency pairs, such as USD-CHF or USD-CAD, as well as in the volatility of currency pairs involving other emerging market currencies, such as USD-BRL or USD-ZAR.

Another area for future research is to explore the use of alternative long memory volatility models, such as the Hyperbolic GARCH (HYGARCH) or the Fractionally Integrated Asymmetric Power ARCH (FIAPARCH) models, to assess the robustness

of our findings. These models may provide additional insights into the nature of long memory in volatility and its implications for risk management and policy design.

Future research could also investigate the economic factors driving the differences in long memory across currency pairs. This could involve examining how differences in macroeconomic fundamentals, financial market development, or institutional factors affect the persistence of volatility in different currency markets. Understanding the underlying drivers of long memory in volatility could help market participants and policymakers better anticipate and manage risks arising from volatility persistence.

Finally, future research could explore the implications of long memory in volatility for other financial markets, such as commodities or derivatives. Long memory in volatility may have important implications for the pricing and risk management of financial derivatives, such as options or futures contracts, as well as for the hedging strategies used by market participants. Examining the presence and implications of long memory in volatility across different asset classes could provide a more comprehensive understanding of the role of long memory in financial market dynamics.

References

T. G. Andersen and T. Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2-3):115–158, 1997.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Exchange rate returns standardized by realized volatility are (nearly) gaussian. *Multinational Finance Journal*, 4(3/4):159–179, 2000.

A. A. Anis and E. H. Lloyd. Expected value of the adjusted rescaled hurst range of independent normal summands. *Biometrika*, 63(1):111–116, 1976.

J. Bai and S. Ng. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics*, 23(1):49–60, 2005.

R. T. Baillie, T. Bollerslev, and H. O. Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1):3–30, 1996.

M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41, 1946.

M. Beine, A. B'enassy-Qu'er'e, and C. Lecourt. Central bank intervention and

foreign exchange rates: new evidence from figarch estimations. *Journal of International Money and Finance*, 21(1):115–144, 2002.

M. Belkhouja and M. Boutahary. Modeling volatility with time-varying figarch models. *Economic Modelling*, 28(3):1106–1116, 2011.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

T. Bollerslev and H. O. Mikkelsen. Modeling and pricing long memory in stock market volatility. *Journal of Econometrics*, 73(1):151–184, 1996.

C. Conrad and B. R. Haag. Inequality constraints in the fractionally integrated garch model. *Journal of Financial Econometrics*, 4(3):413–449, 2006.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

M. Couillard and M. Davison. Hurst exponents and the prediction of stock market crashes. *Studies in Economics and Finance*, 2005.

J. Davidson. Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *Journal of Business Economic Statistics*, 22(1):16–29, 2004.

Z. Ding, C. W. Granger, and R. F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106, 1993.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, pages 987–1007, 1982.

R. F. Engle and T. Bollerslev. Modelling the persistence of conditional variances. *Econometric reviews*, 5(1):1–50, 1986.

J. Geweke and S. Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4):221–238, 1983.

L. A. Gil-Alana and J. Hualde. Fractional integration and cointegration: An overview and an empirical application. *Palgrave Handbook of Econometrics*, pages 434–469, 2009.

H. Goudarzi and C. Ramanarayanan. Modeling and estimation of volatility in the indian stock market. *International Journal of Business and Management*, 5(2): 85–98, 2010.

C. W. Granger. Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14(2):227–238, 1980.

J. R. Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 1981.

H. E. Hurst. Methods of using long-term storage in reservoirs. *Proceedings of the Institution of Civil Engineers*, 5(5):519–543, 1956.

H. E. Hurst, R. Black, and Y. Simaika. Long-term storage: an experimental study. *London*, 1965.

R. Kiliç. Long memory and nonlinearity in conditional variances: A smooth transition figarch model. *Journal of Empirical Finance*, 18(2):368–378, 2011.

G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

T. McElroy and D. N. Politis. Properties of the figarch and related processes. *Econometric Theory*, 36(6):1266–1289, 2020.

N. Oberholzer and P. Venter. Univariate garch models applied to the jse/ftse stock indices. *Procedia Economics and Finance*, 24:491–500, 2015.

P. M. Robinson. Log-periodogram regression of time series with long range dependence. *The annals of Statistics*, 23(3):1048–1072, 1995.

S. J. Taylor. *Modelling financial time series*. world scientific, 2008.

Y. K. Tse. The conditional heteroscedasticity of the yen–dollar exchange rate. *Journal of Applied Econometrics*, 13(1):49–55, 1998.

R. Weron and A. Weron. Hurst analysis in matlab. *HSC Research Report, Wrocław University of Technology*, 2002.

A. Yalama and G. Sevil. Forecasting world stock markets volatility. *International Research Journal of Finance and Economics*, 15:159–174, 2008.

Appendix: Figures

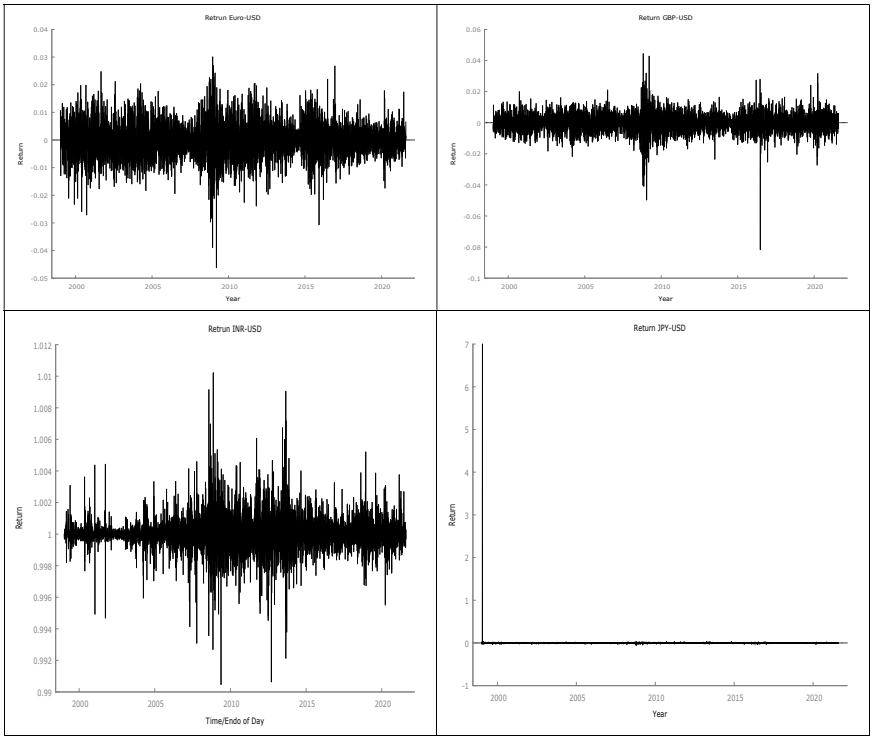


Figure 1: Time Series Plot, Daily Returns Selected Currency Pairs

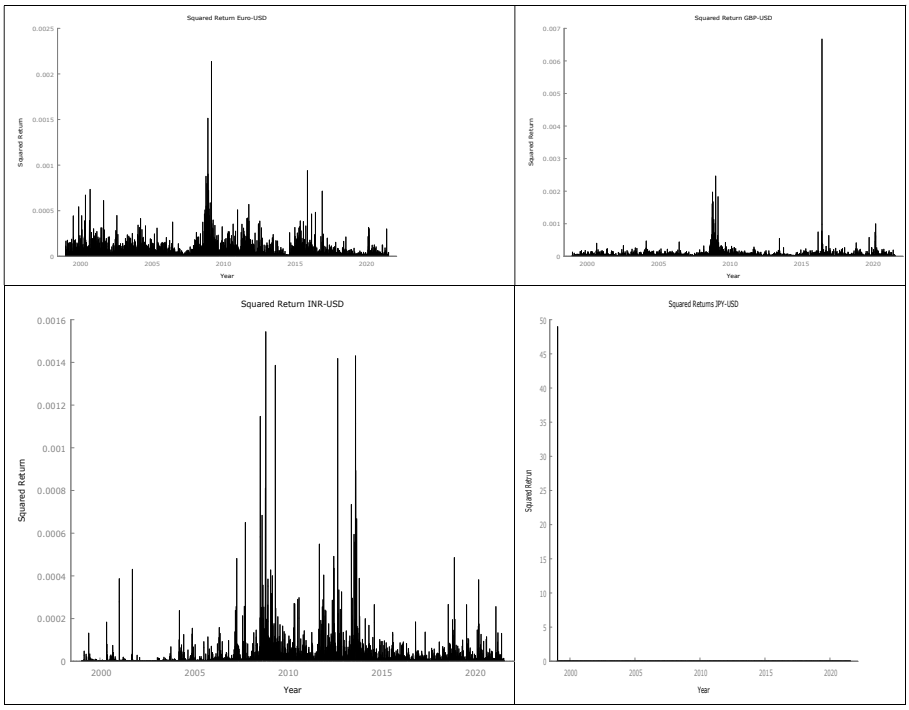


Figure 2: Square of Daily-Rate-of-Return Selected Currency Pairs

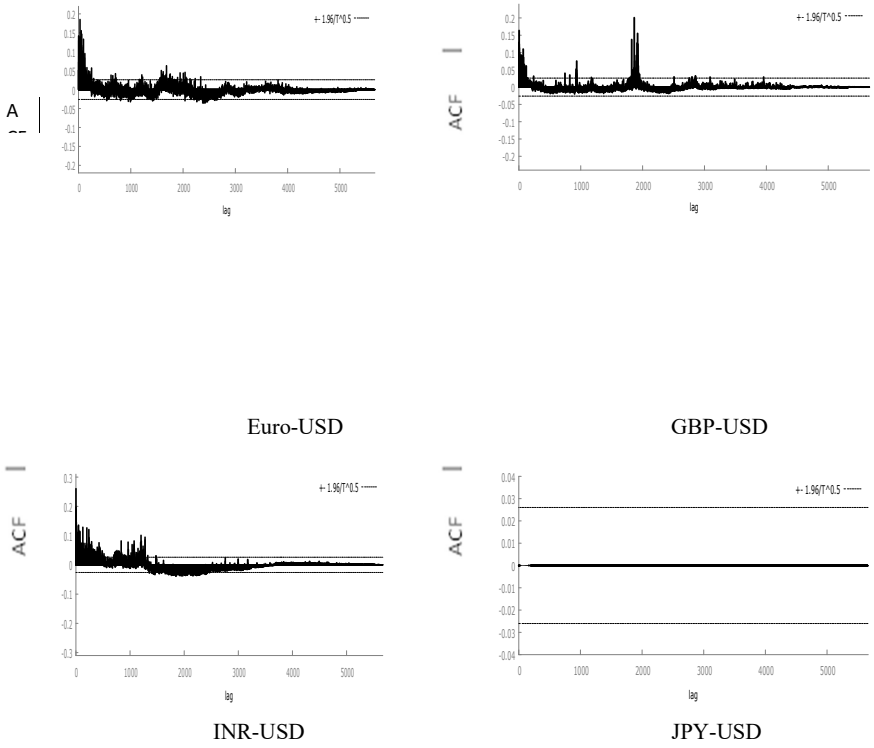


Figure 3: ACF, Squared Returns

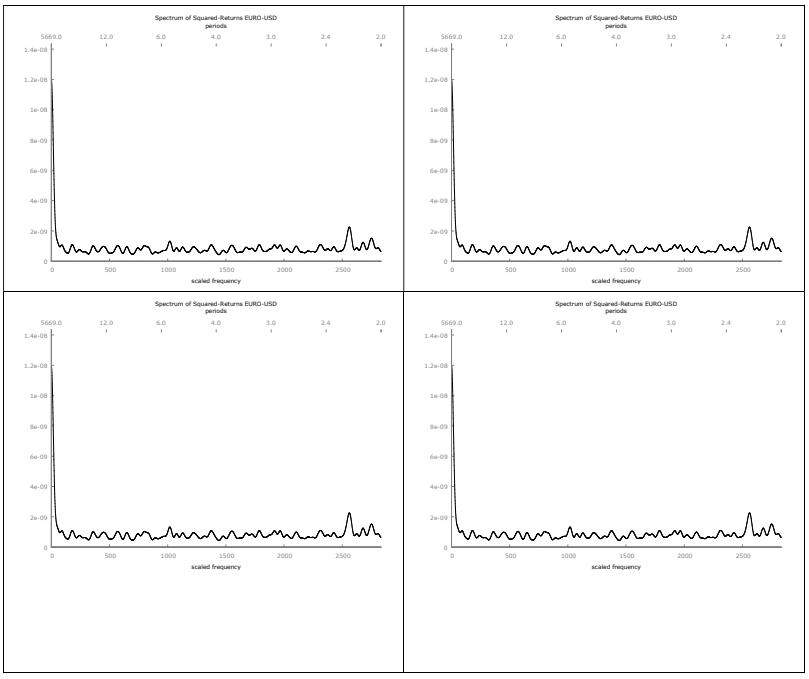


Figure 4: Periodogram, Squared Returns

View publication stats

AUTISM IN ADULTHOOD
Volume 00, Number 00, 2024
© Mary Ann Liebert, Inc.
DOI: 10.1089/aut.2024.0109

Open camera or QR reader and scan code to access this article and other resources online.

"You're Just So Weird and Crazy": Recounting My Lived Experiences as a PhD Scholar in India

Elixir Sagar

Downloaded by Harvard University Countway Library of Medicine from www.liebertpub.com at 09/26/24. For personal use only.

Loud, angry chirps and the sound of an empty bowl falling to the bottom of the cage wake me up as I fumble to reach for my phone. I force my bleary eyes to focus on the phone screen. Half past seven it reads. I sit up bolt upright and rush to get ready and prepare breakfast for my bird as I contemplate the possibility of reaching my university on time dwindling with each passing moment.

At exactly half past eight, I grab my handbag and a water bottle from the dining table and rush out of my house. I can feel my heart pounding as I narrowly escape being run over by oncoming traffic on my way to the metro station. Once inside, I am overwhelmed by the cacophony of morning commuters. I frantically search for and fish out my noise cancellation headphones as I settle into the reserved ladies' seat in Delhi Metro.

After a 2-hour commute that feels like an eternity to me, I finally reach the metro station nearest to my university. I exit the metro station to be greeted by a deluge of sights and sounds from everyday commuters, street food vendors, and the loud honking of impatient car drivers stuck in traffic. This cacophony of diverse sensory inputs threatens to overwhelm me as I focus my attention on crossing the busy road outside my university. I barely manage to escape being run over by a bus as I manage to cross the road to enter the university premises.

Once inside, I quickly make my way to the Department of Management Studies and heave a sigh of relief as I put my handbag on my desk. My attention goes to the Christmas card stuck to the pinboard at my workstation. It was given to me by H, my co-scholar, last Christmas. H made hand-painted cards describing each of his co-scholar's peculiar traits. While he described our co-scholar A as childlike, R as a foodie, and S as a sass queen, he described me as weird.

I mentally grimace as I reach my hand out to touch the card and feel the texture of the five embossed letters "WEIRD" under my fingers. A deluge of emotions, predominantly anger and sadness, threatens to overwhelm me as I recall the many instances from childhood to adulthood when people close to me, including friends, family, and teachers, used this word to describe me.

My reverie is broken as harsh footsteps in the alley outside my room bring me back to the present. H enters the scholar room and hollers at me. I put on my best normal face and exchange pleasantries with him, taking care to keep my tone polite and friendly. I get back to my workstation and open my laptop to resume work on my PhD research topic of mindfulness at the workplace.

I immerse myself in reviewing the relevant literature for the next hour. I become so engrossed in my research that I don't seem to notice when H asks me to stop shaking my legs as it's distracting him. I mumble an apology to him and get up to have a sip of water. As I settle back into my workstation, my mind starts to wander, and I find it difficult to refocus on my research work.

I decide to break the monotony of work. I open my social media account in a new tab on my laptop and begin scrolling through it. As I distractedly scroll through my Facebook account, I ruminate over my close call with the bus while en route to university in the morning and how disastrous the outcome could have been for me in a parallel universe. I open another tab and Google "Why do I find it difficult to cross the road?" In less than a second, Google comes up with a billion answers to my question. I scroll through the long list of search results, but nothing really resonates with me until I stumble upon a blog post on autistic women and their difficulties navigating busy roads. My curiosity piques as I read through the blog post. I click on one of the links to a book titled *Aspergirls*. I spend the next few hours reading the book and gasping with disbelief at the many similarities between my personality traits and Asperger's women.

Like most fellow Indians, I have a very rudimentary understanding of autism spectrum disorders. Up until that moment, I never entertained the possibility that I too could be autistic. I recall an instance from many years ago when I was referred to a government-run physiotherapy and occupational therapy clinic for treatment of a slipped disc injury. While I was getting treatment for my physical ailment there, I came across another patient, a child of about 10 years old, who had an unusual gait and would cry out loud, bang his head on the clinic walls, or shout at the nurses for even minor inconveniences to him during his physical therapy. His mother revealed to me that he was autistic and expressed worry over his bleak future and educational prospects as very few Indian schools

Delhi School of Management, Delhi Technological University, Delhi, India.

Downloaded by Harvard University Countway Library of Medicine from www.liebertpub.com at 09/26/24. For personal use only.

are equipped or willing to admit a special needs child. I can't help but compare the stark contrast of my life as a seemingly "normal", academically "gifted" Doctoral Researcher with that of the autistic child I met in the occupational therapy clinic.

Yet again, my reverie is broken by another co-scholar of mine, R, who unbeknownst to me has quietly walked up behind me and started reading through the blog on autistic women. I can barely hide my embarrassment at the open search tab on my laptop as I fumble to shut it. R is from the Punjabi/Sikh community with a boisterous and exuberant personality. He often uses humor to deflect from emotionally charged situations. Sensing my discomfort as I look at him with pleading eyes, he guffaws loudly and offers me a slice of pizza that he'd just ordered. I feel a deep sense of gratitude toward him as he seems to understand my predicament silently. Instead of probing me further regarding the article on Asperger's women, he spends the next few minutes engaging in lighthearted banter with me.

After wrapping up my research and teaching responsibilities for the day, I brace myself to once again endure the 2-hour-long commute on Delhi Metro back to my home. However, taking the metro in the evening time affords me the small mercy of not having to deal with a rush of commuters. The space and relative anonymity of commuting with strangers allow me the opportunity to once again pull out my laptop and browse through the plethora of information on autism spectrum disorders. The next 2 hours of my commute back home seem to go by in a flash as I read through articles on stimming behaviors and draw parallels between autistic stimming and my "oddly" soothing habit of shaking my legs and chewing on my lower lip whenever I am faced with a stressful situation. I also chance upon blog posts from autistic authors who have suffered from autistic burnout. As I read through the symptoms of autistic burnout, the mysterious physical ailments that made me unable to function and forced me to drop one semester during my MBA start making sense to me.

Once back home, I unwind by getting into my soft satin pj's and having a quiet meal in my room. As I lie down in my bed to sleep, I ruminate over the events of the day. I spend my time tossing and turning as I recall the many small and big moments in my life where I found myself overwhelmed and unable to fit into the stereotypical role of a "normal" Indian woman. I learned to playact my way through all these situations. However, pretending to be someone I was not to fit into the societal mold of "normal" behavior left me feeling empty and drained over the years. It also made me feel like I was somehow defective or less than the other women who seemed to effortlessly conform to their traditional gender roles as dictated by Indian sociocultural norms.

The confusion and bewilderment of nighttime ruminations are replaced by the stark clarity of my thoughts the next morning. The realization that I am autistic (and had most likely always been autistic) dawns upon me with the crystal clarity of the clear blue skies visible from my apartment window.

Over the next few days, I find myself grappling with a deep sense of loss for the life that I could have lived if I were "normal" coupled with intense fear and shame as I attempted to navigate a toxic work environment and an emotionally abusive supervisor while keeping my autistic identity hidden from my supervisors and colleagues. However, my strong sense of justice and deep sense of responsibility toward myself and other autistic women who might find themselves in a similar predicament force me to change the entire trajectory of my PhD research. I abandon my research on mindfulness to fully immerse myself into researching the lived experiences of late-diagnosed autistic Indian adults who have been at the receiving end of bullying, harassment, and unwarranted hatred because of their variant neurology.

Through the course of my PhD research, I get in touch with exceptional autistic Indians, each with their own unique tale of the myriad challenges they face in their personal and professional lives and their journey toward self-diagnosis through a broken system. One such story belongs to my childhood friend who, despite her double master's degree and exceptional knowledge of geoinformatics, still struggles to retain her job because of sensory issues. Yet another autistic acquaintance recalls having to hide his autistic identity at the workplace for fear of losing his job. Listening to these experiences prompts me to focus my PhD research on examining the lived experiences of late-diagnosed autistic women in India, the efficacy of workplace accommodation strategies that allows autistic individuals to work remotely, stimming behaviors of autistic employees, and Indian managers' perspectives on employing neurodiverse talent in Indian business organizations.

Although my research ideas are praised by many acclaimed academicians at national and international conferences, I struggle to gain the same acceptance and support from my non-autistic PhD supervisors. My research ideas are labeled "too novel" and "not practically feasible" by them as they threaten to stop my PhD funding. Additionally, my fear of being labeled crazy by my ignorant supervisors further prevents me from revealing to them that my insights into the autistic experience come from my own lived experience as an autistic adult.

I find the next few months to be extremely emotionally taxing as the hostile attitude of my PhD supervisors, coupled with the countless rejections of my research manuscripts from acclaimed international journals in the field of autism research, plunges me into a year of self-doubt and depression. Until one cold, foggy November morning, when I refresh my email to see an acceptance letter from a globally acclaimed journal for my research paper "'I wish they'd just let us be' experiences of Indian autistic individuals around stimming behaviors at the workplace."[1] I feel like life has come full circle for me at that moment, and my silent crusade to fight for the rights of autistic individuals has not been for nothing.

This small yet significant win proved to be a silver lining through the dense cloud of uncertainty and self-doubt that had been plaguing me the past year. I realize that my foray into autism research gives me an incredible opportunity to not only make sense of my own life experiences, which were equal parts isolating and painful, but also use my experiences to advocate for the rights of other autistic Indians. In doing so I envision a place for autistic academicians in Indian universities wherein they no longer fear ostracization or job loss because of their divergent neurology.

Reference

1. Sagar E, Khera SN, Garg N. "I Wish They'd Just Let Us Be." Experiences of Indian autistic individuals around stimming behaviors at the workplace. Autism in Adulthood. 2023 Dec. 19.

Address correspondence to:
Elixir Sagar
Delhi School of Management
Delhi Technological University
P4X9+Q8X
Bawana Rd
Delhi Technological University
Shahbad Daulatpur Village
Rohini
New Delhi
Delhi 110042
India

E-mail: crastoelixir@gmail.com

Downloaded by Harvard University Countway Library of Medicine from www.liebertpub.com at 09/26/24. For personal use only.

2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) | 979-8-3503-8592-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONECCT62155.2024.10677026

ZCS Switched-Capacitor Cell Balancing Circuit with Bidirectional Buck-Boost Charging

Archit Chauhan
Department of Electrical Engineering
Delhi Technological University
Delhi, India
chauhanarchit208@gmail.com

Sudarshan K. Valluru
Department of Electrical Engineering
Delhi Technological University
Delhi, India
sudarshan_valluru@dce.ac.in

Vanjari Venkata Ramana
Department of Electrical Engineering
Delhi Technological University
Delhi, India
venkat.vr90@gmail.com

***Abstract*— Recent battery technology advancements have made it essential to boost overall battery charging efficiency as well as shortens the charging time. This paper aims to investigate and evaluate the current switched-capacitor (SC) cell balancing and system's performance in conjunction with a bi-directional battery charger circuit. The switches used in the balancing circuit desires zero-current switching (ZCS) during transition state. SC-ZCS does not require large magnetic components or a complex monitoring and control mechanism. This method has the advantage that its balancing time is independent of both the initial imbalance in cell voltage distribution and the quantity of battery cells. The proposed bidirectional converter is controlled by the closed-loop PI controller; it functions as a boost converter during draining and as a buck converter during CC-CV mode of charging. MATLAB-Simulink is used for validating the system outcomes.**

Keywords*— Resonant tank, Switched-Capacitor (SC), Zero Current Switching (ZCS), Constant Current (CC)- Constant Voltage (CV), Closed loop PI control

I. INTRODUCTION

Numerous real-world applications, such as electric cars and bikes, bicycles and wheelchairs, household appliances, healthcare instruments, defense equipment and telecom energy systems require battery cells to be connected in series. For these demanding applications, the best choice is to use lithium-ion batteries as they excel with their long life, slow self-discharge, superior energy density, light in weight and relatively less maintenance [1].

Lithium-ion battery charging and discharging are restricted to lessen the risk of mishaps, such as slight overcharging may overheat and explode the battery and over discharging can increase impedance and decrease its lifetime. Thus, charging and discharging must be carefully regulated for optimal battery performance. Furthermore, when series cells work together, their non-uniform individual cell characteristics cause a minor imbalance that emerges as inconsistent voltages across the cells throughout the charging and draining out phases. The significance of cell balancing lies in the fact that individual cell voltages will diverge significantly over a period of time. Moreover, when the battery pack is in use, its total capacity will run out faster, leading to the failure of the battery system [2]-[3]. As a matter of fact, the best way to solve this issue employs appropriate power electronics topologies—also referred to as battery equalization—to stop cell imbalance. There have been numerous algorithms and circuits developed for cell balancing. These balancing techniques are classified into two types: passive and active equalization. Switched-capacitor (SC) circuits have emerged as a popular choice for cell balancing because of their compact size, affordability and ease of control. In particular, this method is relied upon switching of a suitable capacitor in parallel with each battery cell, so that charge could be transferred from cells with high voltage to those with low voltage [4]. Two complementary signals are required to operate an entire system at a specific frequency and duty ratio in SC-based cell balancing circuits, thus eliminating the requirement for large magnetic components [5]-[6]. Conversely, the designs and control schemes of active cell balancing systems, which are based on switched-mode power converters (SMPC), are more intricate. Furthermore, to charge the battery pack in CC-CV mode, majority battery chargers rely on the basic idea of closed loop control. Because of the battery's nonlinear characteristics, efficient charging algorithms, high dependability, efficiency, compact size, and low cost are necessary for fully charging a battery pack [7]-[8].

This research paper aims to analyze a SC cell balancing method based on resonance that causes all switches to operate under the condition of ZCS. The second objective is to apply the core concept of closed loop control to construct a bidirectional charger which maintains the charging current and voltage. The remaining portions of the paper are arranged as follows: Section II delineates the cell balancing circuit's operational modes and system topology, Section III offers a thorough explanation of the battery charger operation, Section IV covers the design and control of the cell balancing and charging stage, Section V evaluates the system's results and performance, and Section VI contains the paper's conclusion.

II. OPERATING MODES AND ANALYSIS OF CELL BALANCING CIRCUIT

A. Circuit Description

Slow balancing time and large switching losses occurring in the SC structure are its primary disadvantage. Addressing this limitation, Fig.2 showcases a SC- ZCS cell balancing circuit which works on the principle of resonance.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:08:55 UTC from IEEE Xplore. Restrictions apply.

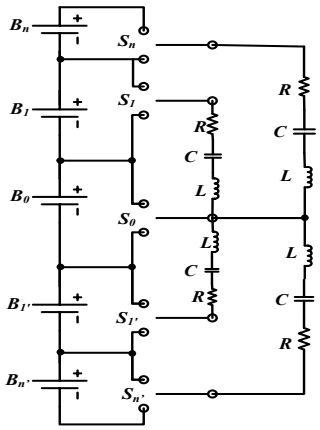


Fig.1. ZCS cell balancing circuit

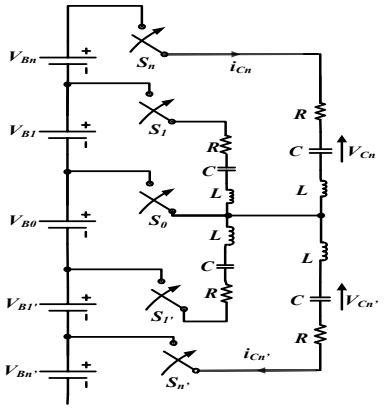


Fig.2. Cell balancing circuit in State I

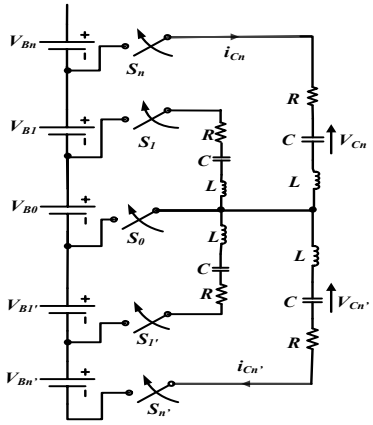


Fig.3. Cell balancing circuit in State II

A resonant inductor-capacitor (LC) tank is created by adding an inductor *L*. By connecting (n–1) LC tank circuit in parallel with each pair of nearby battery cells, the architecture enables energy sharing across all of the series connected n battery cells. This LC tank makes a resonant loop with one capacitor, one inductor, and two switches, the total parasitic resistance is represented by *R*. The values *C* and *L* of each capacitor and inductor utilized in the circuit are the same. A single capacitor is used to change the charges of the two adjacent cells. By continuously flipping the capacitor back and forth, the unequal charge is dispersed until the two cell voltages are precisely equal. The loop's capacitance is substantially less than the battery's capacity and the cells' internal resistances are negligible, so it is disregarded. The two complementing square wave switching pulses governs all switches to work under ZCS. The higher voltage cells transfer their extra energy to the lower voltage cells. By doing this, the issue of cell voltage imbalance brought on by battery chemistry is solved and in addition the switching losses are also minimized.

B. Analysis of ZCS operation

In **State I** as illustrated in the Fig.2 when the switches $\mathbf{S_n}$ & $\mathbf{S_0}$ connects the RLC tank from the top half of the circuit to cells $\mathbf{B_n}$ & $\mathbf{B_1}$, the capacitor current $\boldsymbol{i_{Cn}}$ gradually increase in a sinusoidal manner. Similarly, when the RLC tank from the bottom half connects to cells $\mathbf{B_{(n-1)'}}$ & $\mathbf{B_0}$ through corresponding switches $\mathbf{S_{n'}}$ & $\mathbf{S_0}$, capacitor current $\boldsymbol{i_{Cn'}}$ rise in a similar manner.

Taking one resonant loop in Laplace domain,

$$I(s) = \frac{V_B/_s}{\left(sL+\frac{1}{sC}+R\right)} \tag{1}$$

$$I(s) = \frac{V_B}{L\left[\left(s+\frac{R}{2L}\right)^2+\left\{\sqrt{\frac{1}{LC}-\left(\frac{R}{2L}\right)^2}\right\}^2\right]} \tag{2}$$

Let, $\omega_r = \sqrt{\frac{1}{LC}-\left(\frac{R}{2L}\right)^2}$ and $\rho = \frac{R}{2L}$ (3)

$$I(s) = \frac{V_B * \omega_r}{\omega_r L\left[\left(s+\frac{R}{2L}\right)^2+{\omega_r}^2\right]} \tag{4}$$

Taking inverse Laplace transform we get,

$$i_c(t) = \frac{V_B}{\omega_r L} e^{-\rho t} \sin \omega_r t \tag{5}$$

If minimum voltage across capacitor was assumed to be V_{c_min} and not zero then,

$$i_c(t) = \frac{(V_B - V_{c_min})}{\omega_r L} e^{-\rho t} \sin \omega_r t \tag{6}$$

Generalizing it for $\boldsymbol{n}$ cells we get,

$$i_{cn}(t) = \sum_{i=1}^{n} \frac{(V_{Bi} - V_{cn_min})}{\omega_r L} e^{-\rho t} \sin \omega_r t \tag{7}$$

Similarly for bottom loops,

$$i_{cn'}(t) = \sum_{i=0}^{n-1} \frac{(V_{Bi'} - V_{cn'_min})}{\omega_r L} e^{-\rho t} \sin \omega_r t \tag{8}$$

Now we know, $V_L = L\frac{di}{dt}$

$$V_L(t) = \frac{(V_B - V_{c_min})}{\omega_r} [\{e^{-\rho t} \cos \omega_r t * \omega_r\} + \{e^{-\rho t} \sin \omega_r t * (-\rho)\}] \tag{9}$$

and $$V_R(t) = R * i(t) = \frac{(V_B - V_{c_min})}{\omega_r L} * R * e^{-\rho t} \sin \omega_r t \tag{10}$$

Multiplying and dividing by 2 and forming the expression of $\rho = \frac{R}{2L}$ we will get,

$$V_R(t) = 2\frac{(V_B - V_{c_min})}{\omega_r} * \beta * e^{-\rho t} \sin \omega_r t \tag{11}$$

also, $V_C = V_B - (V_L + V_R)$

therefore,

$$V_C(t) = V_B - \frac{(V_B - V_{c_min})}{\omega_r} e^{-\rho t} * [\rho \sin \omega_r t + \omega_r \cos \omega_r t] \tag{12}$$

Generalizing it for $\boldsymbol{n}$ cells we get,

$$V_{Cn}(t) = \sum_{i=1}^{n} V_{Bi} - \sum_{i=1}^{n} \frac{(V_{Bi} - V_{cn_min})}{\omega_r} e^{-\rho t} [\rho \sin \omega_r t + \omega_r \cos \omega_r t] \tag{13}$$

Similarly for bottom loops,

$$V_{Cn'}(t) = \sum_{i=0}^{n-1} V_{Bi'} - \sum_{i=0}^{n-1} \frac{(V_{Bi'} - V_{cn'_min})}{\omega_r} e^{-\rho t} [\rho \sin \omega_r t + \omega_r \cos \omega_r t] \tag{14}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:08:55 UTC from IEEE Xplore. Restrictions apply.

The capacitor currents $i_{cn}(t)$ and $i_{cn'}(t)$ return to zero when half of the resonant period is over at $t = {}^{\pi}/_{\omega_r}$, and the switches turn OFF in accordance with ZCS condition.

At this time period all the capacitor charges up to their maximum voltage value given by V_{Cn_max} and $V_{Cn'_max}$, *i.e.,*

$$V_{Cn_max} = \left(1 + e^{-\rho\pi/\omega_r}\right) \sum_{i=1}^{n} V_{Bi} - V_{Cn_min} e^{-\rho\pi/\omega_r} \quad (15)$$

$$V_{Cn'_max} = \left(1 + e^{-\rho\pi/\omega_r}\right) \sum_{i=0}^{n-1} V_{Bi'} - V_{Cn'_min} e^{-\rho\pi/\omega_r} \quad (16)$$

In **State II** as shown in the Fig.3 the capacitor currents i_{Cn} and $i_{Cn'}$ rise again in sinusoidal manner from zero, but in opposite direction when the RLC tank in top half of the circuit completes its connection in parallel to the cell $\mathbf{B_0}$ & $\mathbf{B_{n-1}}$ through corresponding switches $\mathbf{S_0}$ & $\mathbf{S_n}$. Similarly, the switches $\mathbf{S_0}$ & $\mathbf{S_n}$' in bottom half connects RLC tank to cells $\mathbf{B_1}$' & $\mathbf{B_n}$' respectively.

Equations of capacitor current and voltage remains same except the limits of summation.

$$i_{cn}(t) = \sum_{i=0}^{n-1} \frac{(V_{Bi} - V_{cn_max})}{\omega_r L} e^{-\rho t} \sin \omega_r t \quad (17)$$

$$i_{cn'}(t) = \sum_{i=1}^{n} \frac{(V_{Bi'} - V_{cn'_max})}{\omega_r L} e^{-\rho t} \sin \omega_r t \quad (18)$$

$$V_{Cn}(t) = \sum_{i=0}^{n-1} V_{Bi} - \sum_{i=0}^{n-1} \frac{(V_{Bi} - V_{cn_max})}{\omega_r} e^{-\rho t} * [\rho \sin \omega_r t + \omega_r \cos \omega_r t] \quad (19)$$

$$V_{Cn'}(t) = \sum_{i=1}^{n} V_{Bi'} - \sum_{i=1}^{n} \frac{(V_{Bi'} - V_{cn'_max})}{\omega_r} e^{-\rho t} * [\rho \sin \omega_r t + \omega_r \cos \omega_r t] \quad (20)$$

The capacitor currents $i_{cn}(t)$ and $i_{cn'}(t)$ return to zero when the resonant period gets over, at $t = {}^{2\pi}/_{\omega_r}$

The capacitor voltages reach to their minimum values V_{Cn_min} and $V_{Cn'_min}$, i.e.,

$$V_{Cn_min} = \left(1 + e^{-\rho\pi/\omega_r}\right) \sum_{i=0}^{n-1} V_{Bi} - V_{Cn_max} e^{-\rho\pi/\omega_r} \quad (21)$$

$$V_{Cn'_min} = \left(1 + e^{-\rho\pi/\omega_r}\right) \sum_{i=1}^{n} V_{Bi'} - V_{Cn'_max} e^{-\rho\pi/\omega_r} \quad (22)$$

Voltage ripple ΔV_{Cn} is given by,

$$\Delta V_{Cn} = V_{Cn_max} - V_{Cn_min} \quad (23)$$

$$\Delta V_{Cn} = \left(1 + e^{-\rho\pi/\omega_r}\right)\left(\sum_{i=1}^{n} V_{Bi} - \sum_{i=0}^{n-1} V_{Bi}\right) + e^{-\rho\pi/\omega_r}\left(V_{Cn_max} - V_{Cn_min}\right) \quad (24)$$

$$\Delta V_{Cn} - \Delta V_{Cn} e^{-\rho\pi/\omega_r} = \left(1 + e^{-\rho\pi/\omega_r}\right)(V_{Bn} - V_{B0}) \quad (25)$$

Finally, the expression becomes

$$\Delta V_{Cn} = \frac{\left(1 + e^{-\rho\pi/\omega_r}\right)}{\left(1 - e^{-\rho\pi/\omega_r}\right)} (V_{Bn} - V_{B0}) \quad (26)$$

Similarly,

$$\Delta V_{Cn'} = \frac{\left(1 + e^{-\rho\pi/\omega_r}\right)}{\left(1 - e^{-\rho\pi/\omega_r}\right)} (V_{B0} - V_{Bn'}) \quad (27)$$

Assuming an RLC tank from the top half, the quantity of charge that travels from battery cell B_0 via B_n during one switching cycle, will be

$$Q = C * \Delta V_{Cn} \quad (28)$$

Therefore, the average current during one switching cycle flowing from B_n to B_0 is given by,

$$I_{Bn} = f_s * C * \Delta V_{Cn} \quad (29)$$

Similarly, $$I_{Bn'} = f_s * C * \Delta V_{Cn'} \quad (30)$$

Substituting the values of ΔV_{Cn} & $\Delta V_{Cn'}$ we get,

$$I_{Bn} = f_s * C * \frac{\left(1 + e^{-\rho\pi/\omega_r}\right)}{\left(1 - e^{-\rho\pi/\omega_r}\right)} (V_{Bn} - V_{B0}) \quad (31)$$

$$I_{Bn'} = f_s * C * \frac{\left(1 + e^{-\rho\pi/\omega_r}\right)}{\left(1 - e^{-\rho\pi/\omega_r}\right)} (V_{B0} - V_{Bn'}) \quad (32)$$

The expression further reduces to,

$$I_{Bn} = \frac{(V_{Bn} - V_{B0})}{R_{eq}} \;\&\; I_{Bn'} = \frac{(V_{B0} - V_{Bn'})}{R_{eq}} \quad (33)$$

Where, $$R_{eq} = \frac{\left(1 - e^{-\rho\pi/\omega_r}\right)}{C\, f_s \left(1 + e^{-\rho\pi/\omega_r}\right)}$$

It is evident from Eq. (33) that energy can be transferred between any two cells having a charge imbalance in just two steps. The current will flow from the cells with higher voltage to central cell B_0 where they will subsequently enter the cells with lower voltage.

Rewriting the expression of R_{eq} in terms of quality factor Q, total parasitic resistance R, switching frequency f_s and resonant frequency f_r we get,

$$R_{eq} = \frac{2\,\rho\, Q^2 R}{f_s} * \tanh\left(\frac{\pi}{2\sqrt{4Q^2 - 1}}\right) \quad (34)$$

Taking R in denominator the final expression becomes,

$$\frac{R_{eq}}{R} = 4\pi * \frac{f_r}{f_s} * \frac{Q^2}{\sqrt{4Q^2 - 1}} * \tanh\left(\frac{\pi}{2\sqrt{4Q^2 - 1}}\right) \quad (35)$$

This represents the corresponding resistance ratio in terms of switching frequency f_s, resonant frequency f_r, quality factor Q.

The ratio of $\frac{R_{eq}}{R}$ is plotted against quality factor Q shown in Fig.4

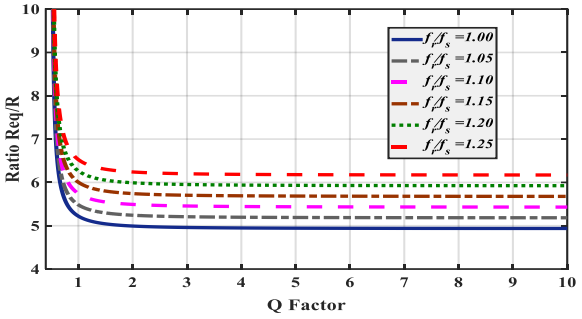


Fig.4. Graph of the ratio $\frac{R_{eq}}{R}$ versus the Q factor.

These plots depict a major drawback at $Q < 1$, then the ratio $\frac{R_{eq}}{R}$ shoots up sharply which is not desirable. However, this negative influence disappears when the quality factor is chosen to be more than 1.5

To reduce the equivalent resistance and subsequently losses, all component parasitic resistances should be as little as feasible. After that, the frequency of switching ought to be nearer to the resonant frequency because a wider gap between

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:08:55 UTC from IEEE Xplore. Restrictions apply.

the two frequencies will eventually lead to a higher $\frac{R_{eq}}{R}$ ratio which leads to an additional loss in the system.

III. Bi-Directional Battery Charger Circuit

A. Circuit Description

Fig. 5(a) depicts the bidirectional DC-DC converter. This topology comprises of a DC voltage source V_S, the bus capacitor C_{Bus} with parasitic resistance R_{Bus}. Two complementary MOSFET switches S_1 & S_2, provide the switching action, whose outputs are routed to an LC filter with values L_f and C_f and parasitic resistances R_L and R_C respectively. The battery pack that needs to be charged is connected to the filter in parallel.

B. Operational Modes

The converter has two modes of operation. In mode I it acts like a buck converter. An input voltage V_S is applied across inductor L, causing the current to flow through it to increase linearly while maintaining a constant output current i_{Batt}. At this point, the battery pack and the load R_{Load} begin to charge while switch S_1 is turned ON and switch S_2 is switched OFF. In mode II, the battery pack begins to discharge through the inductor, providing current to the load- an action akin to that of a boost converter. The switch S_1 is turned OFF while the switch S_2 is turned ON. Figures 5(b) and 5(c) depict the mode I and mode II, respectively.

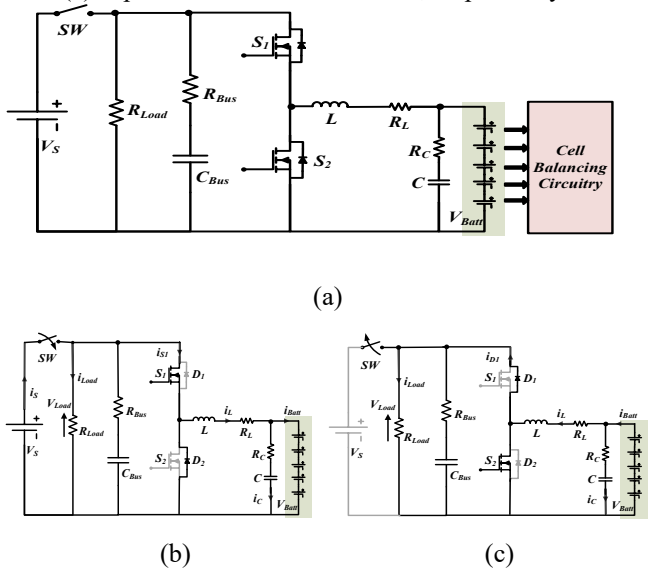


Fig.5. (a) Bi-directional battery charger circuit. (b) Mode I. (c) Mode II.

IV. Design And Control Method

A. Cell balancing circuit

To reduce the overall parasitic resistance R, and thus losses, MOSFET on-state resistance (R_{ON}) should be as little as practical. Conversely, the MOSFET's gate-to-source is inversely correlated with the on-state resistance. This implies that a smaller R_{ON} leads to a greater C_{GS}, which raises the gate driving losses to some extent. Therefore, the R_{ON} cannot be very low. There is a trade-off between the switching frequency and the on-state resistance of MOSFET since, as per Eq. (35), a higher switching frequency is required to reduce the ratio $\frac{R_{eq}}{R}$, but doing so would increase the gate driving losses. Keeping in mind the curves from Fig.4, the equivalent resistance R_{eq} value should be chosen to be approximately 5–6 times the total parasitic resistance, as this is the lowest ratio that will minimize the losses.

In the cell balancing circuit, the number capacitors to be used is always one less than the total number of battery cells utilized. The maximum voltage stress that the capacitors can withstand is represented as $n \times V_{B_max}$, where n represents the number of battery cells. The value of each capacitor in the top and bottom RLC tank, capacitance C must satisfy the Eq (36).

$$\frac{\Delta V_{C_max}}{2} = \frac{\Delta V_{B_max}}{2CR_{eq}f_s} < V_{B_min} \tag{36}$$

Where, V_{B_min} is the lowest cell voltage among all series connected cells, ΔV_{B_max} is the largest voltage differential between all the cells in the battery pack and ΔV_{C_max} is the highest voltage ripple across the capacitor. Inductors used in all RLC tanks should have the same value, just like capacitors, to guarantee that the resonant frequency stays constant throughout each loop. As shown in Fig. 4, when Q factor is more than 1.5, the ratio of $\frac{R_{eq}}{R}$ stays constant across a broad range. Which suggests that the circuit's inductors simply have to be big enough to make certain that the Q factor is more than 1.5. The design specifications of ZCS circuit is given in Table I.

TABLE I. Design Specifications of ZCS Circuit

Parameter	Value
Switching frequency (f_s)	50KHz
Resonant frequency (f_r)	50KHz
MOSFET on state resistance (R_{ON})	0.96mΩ
Tank capacitance (C)	10μF
Tank inductance (L)	1μH
Tank resistance (R)	0.18Ω
Quality factor (Q)	1.756

B. Battery charger circuit

The bi-directional battery charger circuit is modeled similarly to a buck converter because, when charging, it lowers down the input voltage and operates like a buck converter.

In order to retain practicality, parasitic resistances are also taken into consideration, and the inductor is operated in continuous conduction mode (CCM) to simplify the analysis. Hence, inductor value is calculated as follows.

$$L_{min} = \frac{(1-D)(R)}{2f} = 4.5 \text{ mH} \tag{37}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:08:55 UTC from IEEE Xplore. Restrictions apply.

With the aim that the converter operates in continuous conduction mode (CCM) mode, the inductor value is set to be 28% greater than the minimum.

$$L = 1.28\, L_{min} = (1.28)(4.5\text{ mH}) = 5.76\text{ mH} \quad (38)$$

Output capacitor value is found as

$$C = \frac{(1-D)}{8L\left(\Delta V_0/V_0\right)f^2} = 10.85\ \mu\text{F} \quad (39)$$

The design parameters of bi-directional converter used in proposed charging system are listed in TABLE II.

TABLE II. DESIGN SPECIFICATIONS OF BI-DIRECTIONAL CONVERTER

Parameter	Value
Input DC voltage (V_s)	48V
Output DC voltage (V_0)	24V
Switching frequency (f)	10KHz
Bus capacitor (C_{Bus})	1000 µF
Parasitic bus resistance (R_{Bus})	0.1 m Ω
Filter inductance (L)	5.76 mH
ESR of inductor (R_L)	0.05 Ω
Filter capacitance (C)	10.85 µF
ESR of capacitor (R_C)	0.005 Ω
Load resistance (R_{Load})	180 Ω
Output voltage ripple (ΔV_0)	1%
Battery nominal voltage	24V
Battery rated capacity	50Ah
Initial SOC of battery	54%

To achieve the desired charging current and voltage as well as smooth charging, closed-loop control must be used. This paper uses a proportional-integral (PI) controller to regulate a bi-directional converter in both voltage and current modes shown in Fig.6. Using this control algorithm, the battery voltage V_{Batt} and current I_{Batt} is compared with a reference voltage V_{Batt}* and current I_{B_ref} respectively and the error that results is fed to a PI controller that produces PWM pulses for two switches S_1 and S_2. The discharging situation, when the reference voltage is V_{Load}*, employs the same technique. The closed-loop voltage and current mode control algorithm performance is assessed by Simulink environment.

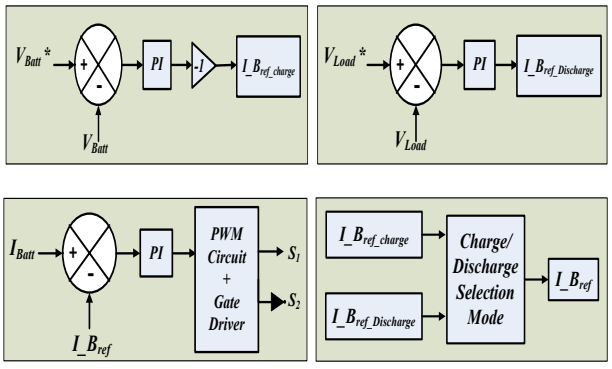


Fig.6. Control algorithm of battery charging circuit

V. SIMULATION RESULTS AND ANALYSIS

The simulation result with various cell voltages [2.684V, 2.764V, 2.646V, 2.795V, 2.736V] is displayed in Fig. 7. All of the cell voltages were balanced at 2.725V after the balancing procedure took roughly 1.6 seconds, which is the average of the initial cell voltages. The capacitor current in the resonant tank and the current flowing through both switches are shown in Fig. 8. Following a sinusoidal pattern, the capacitor current waveform rises from zero and falls back to zero at the end of each half-switch cycle. This is extremely advantageous for minimizing switching losses. Over time, the voltage differential between the cells begins to reduce as the cells get more balanced. Gradually, this leads to a drop in the voltage and current of the capacitor.

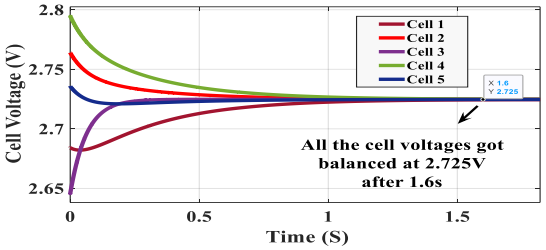


Fig.7. Cell balancing results

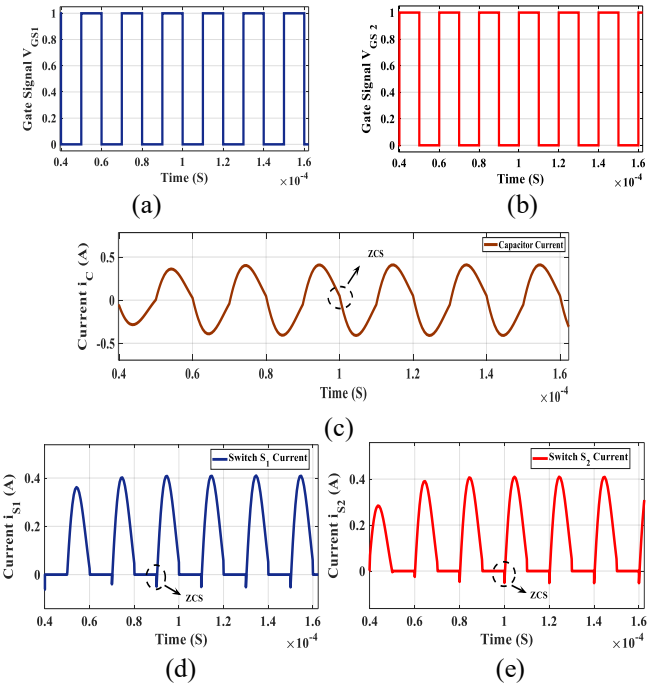


Fig.8.Waveforms of gate signal to (a) Switch S_1. (b) Switch S_2. (c) Capacitor current. (d) Current through switch S_1 (e) Current through switch S_2

Fig.9 shows the waveforms of battery SOC, charging current I_{Batt}, battery voltage V_{Batt} and the load voltage V_{Load} in context with a battery charging circuit. It is evident that the proposed control technique is effective and that the battery current waveform tracks the reference current. The maximum battery current in this instance is 22A.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:08:55 UTC from IEEE Xplore. Restrictions apply.

The converter functions like a buck converter during the charging process. When the battery begins to charge, the charging current is quite high initially but gradually starts to decrease. The battery voltage steadily rises while battery current is maintained constant, up to 80% of SOC battery will be charged in CC mode. The battery enters CV mode, requiring the voltage to remain constant while the current starts to decrease, when the battery voltage reaches 25.98V, or 80% SOC. Throughout the charging process, load voltage is constant up to 48V as it is continuously supplied by the source voltage V_s.

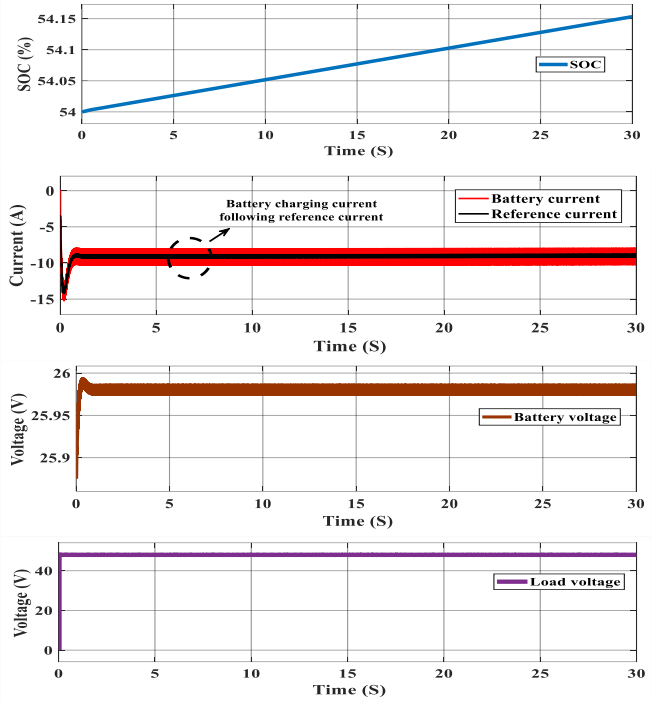
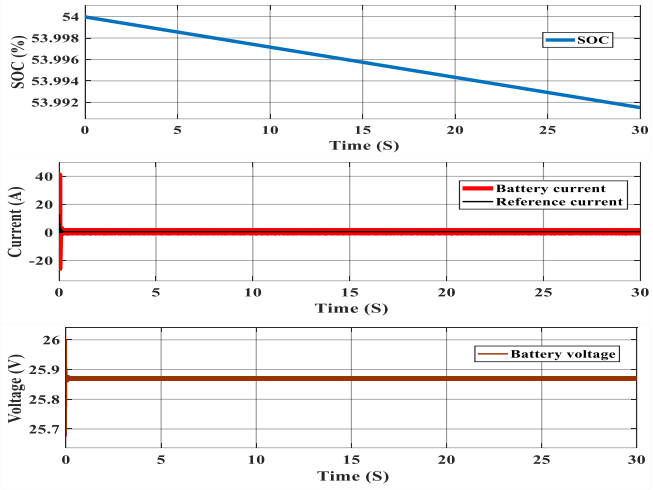


Fig.9. Battery charging waveforms

The battery pack can be charged or discharged in either direction with this DC-DC converter. In the discharging process as shown in Fig.10, the converter has a boost action to maintain constant voltage across the load.



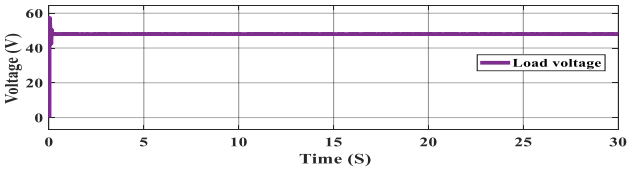


Fig.10. Battery discharging waveforms

VI. Conclusion

The ZCS switched capacitor cell balancing topology with a bidirectional battery charging circuit is investigated and this topology tends to be smaller and lighter and do not require magnetic components. The research gaps and limitations of SC-based cell balancing systems is addressed by this topology, which exhibits if the number of battery cells increases, hence the balancing time decreases. The balancing time can be shortened by as much as 1.6 seconds when compared to how long it takes in a normal SC topology to balance the same number of cells with the same voltage imbalance. Furthermore, the validated results confirms that this topology works better with batteries which have a greater number of series-connected cells because of its less balancing time and helps in reducing the losses occurring in the system.

References

[1] Zhang, Yongzhi, Rui Xiong, Hongwen He, and Weixiang Shen. "Lithium-ion battery pack state of charge and state of energy estimation algorithms using a hardware-in-the-loop validation." IEEE Transactions on Power Electronics 32, no. 6 (2016): 4421-4431.

[2] Cao, Jian, Nigel Schofield, and Ali Emadi. "Battery balancing methods: A comprehensive review." In 2008 IEEE Vehicle Power and Propulsion Conference, pp. 1-6. IEEE, 2008.

[3] Moo, Chin S., Yao Ching Hsieh, and I. S. Tsai. "Charge equalization for series-connected batteries." IEEE Transactions on Aerospace and Electronic Systems 39, no. 2 (2003): 704-710.

[4] Pascual, Cesar, and Philip T. Krein. "Switched capacitor system for automatic series battery equalization." In Proceedings of APEC 97-Applied Power Electronics Conference, vol. 2, pp. 848-854. IEEE, 1997.

[5] Yuanmao, Ye, Ka Wai Eric Cheng, and Y. P. B. Yeung. "Zero-current switching switched-capacitor zero-voltage-gap automatic equalization system for series battery string." IEEE transactions on power electronics 27, no. 7 (2011): 3234-3242.

[6] Shang, Yunlong, Qi Zhang, Naxin Cui, and Chenghui Zhang. "A cell-to-cell equalizer based on three-resonant-state switched-capacitor converters for series-connected battery strings." Energies 10, no. 2 (2017): 206.

[7] Ye, Yuanmao, and Ka Wai Eric Cheng. "Analysis and design of zero-current switching switched-capacitor cell balancing circuit for series-connected battery/supercapacitor." IEEE Transactions on Vehicular Technology 67, no. 2 (2017): 948-955.

[8] Singh, Kushank, and Vanjari Venkata Ramana. "Design of 50 kW Two Stage off-Board EV Charger using CC-CV Algorithm." In 2023 IEEE 3rd International Conference on Sustainable Energy and Future Electric Transportation (SEFET), pp. 1-7. IEEE, 2023.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2024 at 04:08:55 UTC from IEEE Xplore. Restrictions apply.