SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(100th Edition)

APRIL 2021

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 **Hundred** 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 **April, 2021**. 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 Comprehensive Analysis of Deep Learning Techniques for Documentation Classification, ***3.Vinod Kumar***, ***8.Ayushdeep Dabas***, ***8.Ishika Jain*** and ***8.Kartik Pawar***, CSE, DTU

2. A Hybrid Machine Learning and Data Mining Based Approach to Network Intrusion Detection, ***8.Abhinav Singhal***, ***8.Akash Maan***, ***8.Daksh Chaudhary*** and ***3.Dinesh Vishwakarma***, IT, DTU

3. A Novel Approach to Scene Graph Vectorization, ***3.Vinod Kumar***, ***8.Deepanshu Aggarwal***, ***8.Vinamra Bathwal*** and ***8.Saurabh Singh***, CSE, DTU

4. A NOVEL CLASS OF STARLIKE FUNCTIONS, ***3.S. SIVAPRASAD KUMAR*** AND ***6.SHAGUN BANGA***, Mathematics, DTU

5. A novel method of water remediation of organic pollutants and industrial wastes by solution- route processed CZTS nanocrystals, Pooja Semalti, ***6.1.Vikash Sharma*** and Shailesh Narain Sharma, Physics, DTU

6. A review on deep learning based pneumonia detection systems, ***3.Dr Shailender Kumar***, ***8.Pankaj Singh*** and ***8.Mohit Ranjan***, CSE, DTU

7. A review on machine learning based student's academic performance prediction systems, ***3.Rahul Katarya***, ***8.Jalaj Gaba***, ***8.Aryan Garg*** & ***8.Varsha Verma***, CSE, DTU

8. A Stochastic Model for Characterizing Fluctuations in Chemical Sensing, ***8.1.Abhishek Grover*** and ***7.1.Brejesh Lall***, Electronics, DTU

9. A Survey on Analysis of Fake News Detection Techniques, ***3.Shailender Kumar***, ***8.Subodh Kumar***, ***8.Pooja Yadav*** and ***8.Meghna Bagri***, CSE, DTU

10. A Unified Audio Analysis Framework For Movie Genre Classification Using Movie Trailers, ***8.Aditya Sharma***, ***8.Mayank Jindal***, ***8.Ayush Mittal*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

11. Advances in applications of Non-Destructive Testing (NDT): A review, ***8.Mridul Gupta***, ***6.Muhsin Ahmad Khan***, ***3.Ravi Butola*** & ***3.Ranganath M. Singari***, Mechanical, DTU

12. An Empirical Investigation of the Disaggregated Import Demand Function: Non-linear ARDL Framework, ***6.Khyati Kathuria*** and ***3.Nand Kumar***, Humanities, DTU

13. An intelligent approach for predicting the strength of geosynthetic-reinforced subgrade soil, Muhammad Nouman Amjad Raja, ***3.Sanjay Kumar Shukla*** and Muhammad Umer Arif Khan, Civil, DTU

14. An Intelligent Approach to Active and Reactive Power Control in a Grid-Connected Solar Photovoltaic System, Ibrahim Alsaidan, ***6.Priyanka Chaudhary***, Muhannad Alaraj and ***3.Mohammad Rizwan***, Electrical, DTU

15. Analysing the impact of various incentives on solar tariff in India, ***6.Abhis***hek Sharma, ***3.Priya Mahajan*** and ***3.Rachana Garg***, Electrical, DTU

16. Analysis and Design of Series LC Partial-Resonance-Pulse Based ZCS Current-Fed Push-Pull Converter, ***7.1.Swati Tandon*** and Akshay Kumar Rathore, Electrical, DTU

17. Analysis of a Novel Nanoscale Vacuum Channel TF-FinFET, ***8.Mridul Prakash Kashyap***, ***8.Sanmveg Saini*** and ***3.Rishu Chaujar***, Applied Physics, DTU

18. Analysis of Carbon Fiber, ***8.Siddharth Garg***, ***8.Tanishq Gupta*** and ***3.Mukesh Dagdeb***, Mechanical, DTU

19. Brayton-Moser Passivity Based Controller for Electric Vehicle Battery Charger, ***7.1.Kumari SHIPRA***, Rakesh MAURYA, & Shambhu N. SHARMA, Electrical, DTU

20. Catabolic enzyme activities during biodegradation of three-ring PAHs by novel DTU-1Y and DTU-7P strains isolated from petroleum-contaminated soil, ***6.Sakshi***, ***3.S. K. Singh*** and ***3.A. K. Haritash***, Environmental, DTU

21. Clickbait in Social Media : Detection and Analysis of the Bait, ***8.Mini Jain***, ***8.Peya Mowar***, ***8.Ruchika Goel*** and ***3.Dinesh K. Vishwakarma***, IT, DTU

22. Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of COVID-19 fake tweets, ***8.Rachit Bansal***, William Scott Paka, Nidhi, Shubhashis Sengupta and Tanmoy Chakraborty, DTU

23. Consequences of COVID-19 pandemic on global maritime trade industry, ***3.Shailender Kumar*** and ***6.Aman Jolly***, CSE, DTU

24. Control of Grid Integrated Photovoltaic system using new Variable Step Size Least Mean Square adaptive filter, ***6.Avdhesh kumar***, ***3.Rachana Garg*** and ***3.Priya Mahajan***, Electrical, DTU

25. Covid Detection from X-RAY and CT Scans using Transfer Learning – A Study, ***3.Shailender Kumar***, ***8.Naman***, ***8.Mohd Talib*** and ***8.Praduman Verma***, CSE, DTU

26. COVID-19 vaccine on board ships: current and future implications of seafarers, Gopi Battineni, ***3.Shailender Kumar***, Mamta Mittal and Francesco Amenta, CSE, DTU

27. Crowd anomaly detection with LSTMs using optical features and domain knowledge for improved inferring, ***6.Mohammad Sabih*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

28. Design and Analysis of LFSR based Content Addressable Memory, ***8.Vyom Garg*** and ***8.Parangat Mittal***, Electronics, DTU

29. Determination of bioactive phytochemicals, antioxidant and anti-inflammatory activity of Colchicum autumnale L. (Suranjanshireen), ***6.Temesgen Hailu***, Ratnika Sharma, Sonia Mann, Promila Gupta, ***3.Rajinder K. Gupta*** and ***3.Archna Rani***, Chemistry, DTU

30. Developments in Machine Tool Spindle Technology, ***8.Himanshu Bhatt*** and ***8.Himanshu Bist***, Mechanical, DTU

31. Drinking water quality evaluation for groundwater of Delhi, India using GIS techniques, ***7.1.Ayushi Raghav*** and ***3.S.K. Singh***, Environmental, DTU

32. ECG Denoising Using Artificial Neural Networks and Complete Ensemble Empirical Mode Decomposition, ***3.Rajesh Birok***, ***3.Rajiv Kapoor*** and ***3.Mahipal Singh Choudhry***, Electronics, DTU

33. Effect of Fiber Length and Content on the Mechanical Properties of Pineapple Leaf Fiber Reinforced-Epoxy Composites, ***6.Mohit Mittal*** and ***3.Rajiv Chaudhary***, Mechanical, Production, and Industrial Engineering, DTU

34. Energy efficient removal of COD from landfill leachate wastewater using electrocoagulation: parametric optimization using RSM, K. Faheem1 · S. U. Khan, M. Washeem and ***6.S. U. Khan***, Civil, DTU

35. Enhanced Third-Order Generalized Integrator-Based Grid Synchronization Technique for DC-Offset Rejection and Precise Frequency Estimation, ***6.Hemant Saxena***, ***3.Alka Singh*** and ***3.Jitendra Nath Rai***, Electrical, DTU

36. EnPSO: An AutoML Technique for Generating Ensemble Recommender System, ***6.Garima Gupta*** and ***3.Rahul Katarya***, CSE, DTU

37. Evaluation of urban transport-environment sustainable indicators during Odd–Even scheme in India, ***6.Ravindra Kumar***, ***3.Rajeev Kumar Mishra***, ***6.Satish Chandra*** and ***6.Asif Hussain***, Environmental, DTU

38. Evolving Intelligent System for Trajectory Tracking of Unmanned Aerial Vehicles, ***8.1.Rupam Singh*** and ***3.Bharat Bhushan***, Electrical, DTU

39. Experimental investigations of performance and smoke emission of Karanja biodiesel blends & diesel in a diesel engine, Sanjay Mohite, Dr. Sudhir Kumar, ***3.Dr. Sagar Maji*** and ***3.Dr. Amit Pal***, Mechanical, DTU

40. Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Muhammad Nouman Amjad Raja and ***3.Sanjay Kumar Shukla***, Civil, DTU

41. Fractional Order SinusoidalOscillator Using only a Single Operational Trans-resistance Amplifier, ***3.Manoj Kumar***, ***3.D. R. Bhaskar*** and ***3.Pragati Kumar***, Electrical, DTU

42. Gold nanomaterials for optical biosensing and bioimaging, Peng Si, Nasrin Razmi, Omer Nur, Shipra Solanki, ***3.Chandra Mouli Pandey***, Rajinder K. Gupta, ***3.Bansi D. Malhotra***, Magnus Willander and Adam de la Zerda, Chemistry, DTU

43. Harmonics Mitigation Techniques in Grid Integrated PV based Microgrid: A Comparative Analysis, ***6.Avdhesh Kumar***, ***3.Rachana Garg*** and ***3.Priya Mahajan***, Electrical, DTU

44. High-quality reversible data hiding scheme using sorting and enhanced pairwise PEE, Gurjinder Kaur, Samayveer Singh, Rajneesh Rani, ***3.Rajeev Kumar*** and Aruna Malik, CSE, DTU

45. Hybrid Dynamic CML with Modified Current Source (H-MDyCML): A Low-Power Dynamic MCML Style, ***8.Riya JAIN***, Kirti GUPTA and ***3.Neeta PANDEY***, Electronics, DTU

46. Improving Traditional Stock Market Prediction Algorithms using Covid-19 Analysis, ***3.Rajni Jindal***, ***Nikhil Bansal***, ***Nitin Chawla*** and ***8.Sanskriti Singhal***, CSE, DTU

47. Influence of Walnut (Juglans L.) Shell Particles Addition on the Mechanical Properties of Epoxy Composites, ***6.Mohit Mittal***, Kanchan Phutela, and ***3.Rajiv Chaudhary***, Mechanical, Production, and Industrial Engineering, DTU

48. Interaction between Picric Acid and 5-aminoquinoline: A Theoretical Study, Nupur Pandey, ***3.Mohan Singh Mehata*** and Sanjay Pant, Applied Physics, DTU

49. IoT based prediction of water quality index for farm irrigation, ***3.Rajesh Kumar Yadav***, ***8.Adarsh Jha*** and ***8.Aditya Choudhary***, CSE, DTU

50. Kinetic theory of effect of dust charge fluctuations on the parametric decay of lower hybrid wave instability by relativistic runaway electrons in tokamak, Jyotsna Sharma, ***3.Suresh C. Sharma*** and Ajay Gahlot, Applied Physics, DTU

51. Learning of classification tasks with an array of uniform-mode spin Hall nano-oscillators, ***8.1.Utkarsh Singh***, Neha Garg, Saurabh Kumar, Pranaba Kishor Muduli and Debanjan Bhowmik, Electronics, DTU

52. Low bandwidth data hiding for multimedia systems based on bit redundancy, Neeraj Kumar, ***3.Rajeev Kumar***, Aruna Malik and Samayveer Singh, CSE, DTU

53. Machine Learned Classifiers for Trustworthiness Assessment of Web Information Contents, ***3.Priyanka Meel*** and ***3.Dinesh kumar Vishwakarma***, IT, DTU

54. Modeling and Finite Element Analysis of Universal Joint, ***8.Kanishk Yadav*** and ***8.Harshit Jain***, Mechanical, DTU

55. Modeling gully erosion susceptibility in Phuentsholing, Bhutan using deep learning and basic machine learning algorithms, Sunil Saha, ***3.Raju Sarkar***, Gautam Thapa and Jagabandhu Roy Civil, DTU

56. Object tracking and counting in a zone using YOLOv4, DeepSORT and TensorFlow, ***3.Dr. Shailender Kumar***, ***8.Vishal***, ***8.Pranav Sharma*** and ***8.Nitin Pal***, CSE, DTU

57. On Certain Generalizations of S∗ (Ã), ***6.Kamaljeet Gangania*** and ***3.S. Sivaprasad Kumar***, Applied Mathematics, DTU

58. ON OBLIQUE DOMAINS OF JANOWSKI FUNCTION, ***3.S. SIVAPRASAD KUMAR*** AND ***6.POOJA YADAV***, Applied Mathematics,, DTU

59. Optimal DG Allocation Using Particle Swarm Optimization, ***3.Prem Prakash***, Electrical, DTU

60. Partial Series Resonance Pulse Commutated Current-Fed Three-phase Current Sharing DC/DC Converter: ZCS Analysis, Design and Experimental Results, ***7.1.Swati Tandon***, Akshay Kumar Rathore and Jose Rodriguez, Electrical, DTU

61. PERFORMANCE EVALUATION OF VERMIFILTRATION TECHNOLOGY FOR THE TREATMENT OF WASTEWATER USING DETOXIFYING EARTHWORMS: A CASE STUDY, ***3.Prof. S. K Singh*** and ***7.Twinkle Bhatia***, Environmental, DTU

62. Prediction of an optimum engine response based on various input parameters on common rail direct injection diesel engine- A Response Surface Methodology Approach, ***8.1.Manish Kumar***, ***3.Naushad Ahmad Ansari***, Abhishek Sharma, ***6.Varun Kumar Singh***, ***3.Raghvendra Gautam*** and Yashvir Singh, Mechanical, DTU

63. Quaternions-based detection of power quality events, ***3.Rajiv Kapoor*** and ***8.Swapnil Verma*** and Aarchishya Kapoor, DTU

64. Resilient Protection of Medium Voltage DC Microgrids against Cyber Intrusion, ***8.1.Vaibhav Nougain***, Sukumar Mishra and Sushree Subhasmita Jena, Electrical, DTU

65. Review of biomarker systems as an alternative for early diagnosis of ovarian carcinoma, ***3.A. Srivastava***, ***6.A. Gupta*** and ***3.S. Patidar***, Biotechnology and CSE, DTU

66. Scene Text recognition in the wild with motion deblurring using deep networks, ***8.Sukhad Anand***, ***3.Seba Susan***, ***8.Shreshtha Aggarwal***, ***8.Shubham Aggarwal*** and ***8.Rajat Singla***, DTU

67. Sector Classification for Crowd-Based Software Requirements, ***8.1.Kushagra Bhatia*** and Arpit Sharma, Electrical, DTU

68. Segmented Rotor Mutually Coupled Switched Reluctance Machine for Low Torque Ripple Applications, Siddharth Mehta, Md Ashfanoor Kabir, ***8.1.Prerit Pramod*** and Iqbal Husain, Electrical, DTU

69. Sensitivity analysis and multi-objective optimization of organic Rankine cycle integrated with vapor compression refrigeration system, Ashwni, Ahmad Faizan Sherwani, Deepak Tiwari and ***3.Anil Kumar***, Mechanical, DTU

70. Simulation analysis of mixing quality in T-junction micromixer with bend mixing channel, ***6.Ranjan Prakash***, ***3.Mohammad Zunaid*** and ***3.Samsher***, Mechanical, DTU

71. Smart Robot for Collection and Segregation of Garbage, ***8.Akanksha***, ***8.Muskan Gupta*** and ***3.Madan Mohan Tripathi***, Electrical, DTU

72. Software Defect Prediction using Dimensionality Reduction and Deep Learning, ***6.1.Massoud Massoudi***, ***8.Nameet Kumar Jain*** and ***8.Pranay Bansal***, CSE, DTU

73. Strategic framework for developing resilience in Agri-Food Supply Chains during COVID 19 pandemic, ***3.Pravin Kumar*** and Rajesh Kumar Singh, Mechanical, DTU

74. Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda, Rajesh Kumar Singh, ***3.Saurabh Agrawal, 6.Abhishek Sahu*** and Yigit Kazancoglu, DSM and Mechanical, DTU

75. Subthreshold Analytical Model of Asymmetric Gate Stack Triple Metal Gate All Around MOSFET (AGSTMGAAFET) for Improved Analog Applications, ***8.Arvind Ganesh***, ***8.Kshitij Goel***, ***8.Jaskeerat Singh Mayall*** and ***3.Sonam Rewari***, Electronics and Communication, DTU

76. Synthesis and fabrication of 2D Tungsten trioxide nanosheets based platform for impedimetric sensing of cardiac biomarker, ***6.Deepika Sandil***, Saurabh Srivastava, ***6.Ritika Khatri***, ***6.Kanika Sharma*** and ***3.Nitin K. Puri***, Physics, DTU

77. Text Based Hate-Speech Analysis, ***8.Janak Sachdeva***, ***8.Kushank Kumar Chaudhary***, ***8.Harshit Madaan*** and ***3.Priyanka Meel***, IT, DTU

78. The Evolution, progression and Impact of ISO 9000 in the manufacturing organizations around the world and India: An Analysis, DEEPA PRIYANSHU, RUBINA LIYAKAT KHAN, MAHERA MAQDADI and ***6.MONISHA GUPTA VASHISHTA,*** USME, DTU

79. Thermo-economic analysis of the combined solar based pre-compression supercritical CO2 cycle and organic Rankine cycle using ultra low GWP fluids, ***6.Yunis Khan*** and ***3.Radhey Shyam Mishra***, Mechanical, DTU

80. TRANSFORMING MANUFACTURING PROCESSES WITH THE HELP OF AI, ***8.1.Keshav Aggarwal***, ***8.Shaswat Garg*** and ***8.Prakash Budhaulia*** Mechanical, Production & Industrial Engineering, DTU

81. Using Gaussian based Bayesian Optimization in Recommender Systems, ***8.Garvit Rana***, ***8.Eklavya Bhatia***, ***8.Ejaz Ul Haq*** and ***3.Shailender Kumar***, CSE, DTU

82. Zinc micronutrient-loaded carboxymethyl tamarind kernel gum-based superabsorbent hydrogels: controlled release and kinetics studies for agricultural applications, ***6.Khatri Khushbu***, **3.Sudhir G. Warkar** and Nandkishore Thombare, Applied Chemistry, DTU

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

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

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

A Comprehensive Analysis of Deep Learning Techniques for Documentation Classification

Vinod Kumar
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
vinod_k@dtu.ac.in

Ayushdeep Dabas
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
ayushdeep1111@gmail.com

Ishika Jain
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
ishikajain.ij99@gmail.com

Kartik Pawar
Dept. of Computer Science and Engineering
Delhi Technological University
New Delhi, India
pawarkartik0@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395861

***Abstract*—The continuously increasing volume of documents in different fields has rendered document classification by manual labor infeasible. This has led to the genesis of automatic classification with the help of a myriad of techniques like Data Mining, Machine learning, and Deep Learning (N.L.P). This automation prevents human error and, more importantly, builds up speed, which is critical given the volubility of the data. Document classification has seen many developments in the past decade. This paper presents a comprehensive analysis of various state-of-the-art deep-learning algorithms for document classification and perform extensive experiments on these techniques. The performance of these algorithms is analyzed quantitatively using established metrics.**

***Index Terms*—Document Classification, Text Analysis, Natural Language Processing, Deep Learning**

I. Introduction

Text classification is one of the most fundamental tasks in NLP, which is further extended to classifying bigger-texts or documents. Document classification is the task of labeling the documents into several categories according to the content of the document. In this data-driven economy, new documents are produced every minute, adding to the already colossal mountain of data. These documents are a vital part of the workflow environment, providing intuitive insights to companies, which helps them to make the best decisions [1]. The document classification process can be performed manually or automatedly, allowing for easy maintenance, updating, and retrieval of multimedia. However, both types of classifications have their pros and cons. While manual classification gives superior control over the process and gives a much better verdict, it is a very arduous task. To overcome this problem, automated classification has become the need of the hour. It is much faster, as well as more cost-efficient. Its application can also be found in sentiment analysis [2], spam detection [3], question-answering [4]. In the past, many algorithms have been proposed to improve the accuracy of classification. One of the oldest techniques, proposed by J.J. Rocchio in 1971, called Rochhio Classification [5], used relevant feedback to query full-text databases [6]. Simple machine learning techniques like Support Vector Machines (S.V.M),Naïve Bayes, K-Nearest Neighbors (K.N.N), Decision Trees. These come under Supervised Learning. [7] Other techniques include Boosting and bagging, Logistic Regression. Unsupervised learning techniques include K-Means Clustering, Fuzzy C-Means. [8] The last decade has seen much more improvement in the process. Scientists and researchers have been able to develop much more complex algorithms yielding high accuracy. They include ReNN, MLP, CNN, RNN, Attention Networks, Trans, GNN.

However, just the use of the algorithms mentioned above does not solve the task. Any raw textual data presents its own challenges, like high dimensionality of feature space, un-organized text. To overcome them, we need a complete pipeline. ”Most text classification and document categorization systems can be deconstructed into the following four phases: Feature extraction, dimension reductions, classifier selection, and evaluations.” [9]. Text preprocessing also forms an integral part of the pipeline.

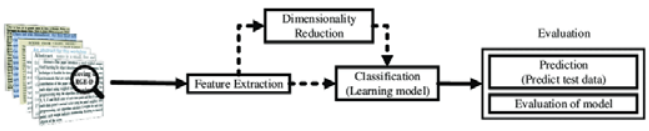


Fig. 1. Pipeline for text classification.

The complete process sees many variations induced by the different algorithms that these four phases have. Researchers have extensively performed experiments, mixing and matching various algorithms to get the best results.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DATASET DESCRIPTION TABLE

Dataset	Acronym	No. of Classes	Average sentence size	Total no. of Datapoints	Application
Yahoo Answer [10]	Yahoo	10	112	1,460,000	Topic Classification
DBpedia [11]	DBp	14	55	630,000	Topic Classification
20 Newsgroups [12]	20N	20	221	20,000	News Classification
SST2 [13]	SST2	5	19	11,855	Sentiment Analysis
IMDB [14]	IMDB	2	294	50,000	Sentiment Analysis
AG News [10]	AGN	4	45/7	127,600	News Classification
Yelp reviews [15]	Yelp	2	153	598,000	Sentiment Analysis
Amazon-2	Amz-2	2	91	4,000,000	Sentiment Analysis
Amazon-5	Amz-5	5	93	3,650,000	Sentiment Analysis
Movie Review	MR	2	20	10,662	Sentiment Analysis
Reuters 8	R8	8	66	7,674	News Classification
Reuters 52	R52	52	70	9,100	News Classification
Ohsumed	Ohsumed	23	136	7,400	Topic Classification
SNLI [16]	SNLI	3	-	570,152	Natural Language Inference
Reuters	R90	90	1	10,788	News Classification

II. RELATED WORK

In this section, we discuss numerous approaches to tackle document classification. Each algorithm has its pros and cons. Naive Bayes is probably the simplest and most broadly used model to classify documents. However, its assumption of independent features makes the algorithm a little unrealistic for practical purposes. While Rocchio classification is easy to implement and has a relevance feedback mechanism, it gives low accuracy because of its simple linear combination. It also often fails to classify multimodal classes and relationships. Logistic regression, on the one hand, is easy to implement and is computationally cheap. However, its inability to solve non-linear problems and to process dependent data makes it a terrible choice for advanced classification tasks. Decision Trees/Random Forests can overfit easily but are fast to train. Several other shallow learning algorithms also fall short in giving reliable results. These techniques suffer from multiple drawbacks, some of which include expensive computations, high variance and loss, rigid boundaries, slow training, and so on. Hence, over the last few decades, various deep learning techniques have been applied to improve the results. Feed-forward neural networks and recursive neural networks were the two initial deep learning models. Later on, researchers developed several RNN and CNN based approaches. RNN-Capsule [17] proposed an RNN based capsule model. For a given problem, for each category, one capsule is built having an attribute, a state, and three modules: representation module, probability module, and reconstruction module. The model can output the words that reflect the domain specificity of the dataset without any linguistic knowledge. Bao et al. [18] used a few-shot text classification. This model leverages the distributional signatures, mapped into attention scores using a pre-trained meta-learning framework. These attention scores then weigh lexical representations of the words.

Though CNN and RNN based models provide excellent results on document classification related tasks, these models offer poor interpretability due to the non-readability of any hidden data. The attention-based models improve upon the same by providing interpretability for text classification. Tan et al. [19] introduced a simple one-way attention mechanism affected through question context for generating answer embeddings along with CNN and RNN. The attention mechanism captures the complicated semantic relationships among the embeddings of questions and answers. Transformer-based models such as the BERT model [20] can parallelize computation without considering the sequential information making it suitable for large-sized datasets. Unlike deep learning models such as CNN, RNN, some researchers have used GNNs to work with arbitrarily structured graphs. GNN based models transform the task of text classification to the study of graph node classification. Peng et al. [21] converted the text into graph-of-words and learned multiple semantics with CNN models. Apart from the above approaches, there are some individual models as well, such as twin neural networks [22]. It used two distinct input vectors with equal weights to compute similar output vectors. Though deep learning models perform well on the text classification task, the need for extra hyperparameters increases the computational complexity of the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

network. Virtual adversarial training (VAT) [23] provided a robust optimization as it requires a small number of hyperparameters. Zhang et al. [24] applied Reinforcement Learning (RL) and learned the task structures to provide structured sentence representations.

In this paper, we compare some of the deep learning techniques that yield state-of-the-art results. We mention their advantages as well as their pitfalls. We also make a meticulous analysis of where each technique is best suited and how these may be further improved.

Table I thoroughly projects the datasets used in the experiments. We have stated the number of classes, average sentence size, the total number of data points, the primary application, and the acronym used in this paper. We compare the results using the metric, Accuracy.

III. METHODOLOGY

In this section we discuss some deep learning techniques and see how each of these techniques work in brief.

A. Attention based neural networks

The term attention pertaining to neural networks was first introduced by Bahdanau et al. [25] 2014. They proposed adding attention to fix the information problem that previous versions of neural networks, especially the Encoder-Decoder architectures, have caused. Such architectures pass only the last-hidden state, thereby causing only a fraction of the information to be sent at a time, which causes it to lose the context. Usage of attention helps the decoder-segment to attend the whole sequence and thus use the context of the entire sequence during the decoding step. [26] Building on the earlier approach, Yang et al. [27] proposed that better representations could be obtained by incorporating knowledge of document structure in the model architecture. As different sections have different relevance, some having more than others [27]; Finding these relevant sections and how they could be weaved together was their goal. The model used the fact that documents are built hierarchically (documents break down into sentences and sentences further break down into words). Working in a bottom-up manner, they first constructed sentence representations followed by their aggregation into a document. Because of the varying importance of words in their different contexts, the model contained two attention mechanism levels- one at the base vector level, i.e., word level, the other at the phrase level. This helped to determine the context and therefore rank the word's importance. The HAN consisted of a word sequence encoder, a word-level attention layer, a sentence encoder, and a sentence-level attention layer [27]. A GRU-based sequence encoder facilitated this to track the state of sequences. The document was converted into vector representation, which evolved from word vector to final document vector. After embedding the words and encoding them, the attention mechanism extracted the important words to aggregate them into sentence vectors. The same steps were repeated for the sentence vectors to yield v, the document vector.

The features for document classification:

$$p = softmax(W_c v + b_c) \tag{1}$$

The training loss was measured as:

$$L = -\sum_{d} \log p_{dj}, \tag{2}$$

where j is the label of document d.

The author has used the negative log likelihood of the correct labels as training loss.

B. XLNet Model

XLNet is a language model proposed by P. Zhou, Z. Qi [28]. XLNet is a pre-trained BERT type model. We can view XLNet as an extended version of BERT. It was able to outperform BERT in some NLP tasks, including Text Classification, Document Classification, Q/A. XLNet is the method of studying language representation by employing the generalized autoregressive pretraining technique. Its objective is to learn the language model. Pre-trained on a vast corpus using the permutation language modeling goal. XLNet uses Transformer XL as its feature extraction architecture, yielding better results than BERT, which can be credited to its additional repetition. This has the potential to give XLNet a deeper understanding of the language context. Since XLNet is a specific transformation learning method, we can first load the pre-trained XLNet model and then use the fine-tuning method to update the pre-trained model according to the required task. Here we have used library transformers deployed with Huggingface [29]. This library contains some pre-trained models for natural language processing like XLNet, GPT, BERT.

We need to set the document text into three types of attachments:

- Embedding tokens
- Word embedding mask
- Embedding segmentation

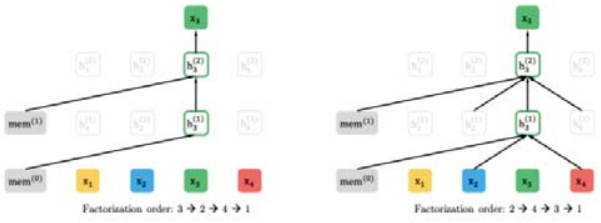


Fig. 2. Permutation Language Model, It has produced state-of-the-art results on several NLP tasks

When using transformational learning like XLNet, the process of training a new model with top-down data is called "fine-tuning". All we have done is just selected one of the pre-trained XLNet models and used our data to renew the model parameter to adjust our top-down NLP task. For the English NLP task, XLNet has two models: the base case model and the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

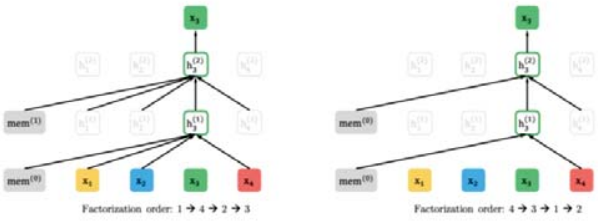


Fig. 3. Permutation Language Model (2)

large case model. The large case model is more comprehensive and gives more reliable performance than the base model, but it requires extra processing power and time.

C. Recurrent Neural Networks

RNN models [30] operate input text as a group of words and is expected to achieve the dependencies of words and "text structure" for classifying documents. Though vanilla-RNN models also don't perform very well and are usually inferior to feedforward networks. Among the various RNN models, the most famous architecture is Long Short-Term Memory (LSTM). The LSTM solves vanishing gradients and contains memory blocks in the hidden layer. Each memory block includes an input gate, an output gate and a forget gate. Work was done to improve RNN and Long Short-Term Memory models concerning document classification over getting more valuable knowledge such as tree structure in natural language and extensive words relationships in a corpus. The abstraction of LSTMs to tree-like network typologies has also been performed to explore rich semantic representations. S. Tai, R. Socher [31] developed a tree LSTM which is better than a chained LSTM for Natural language processing problems as the natural language has "syntactic properties that naturally combine words into phrases". Two studies were performed to find the correctness of tree-LSTM: classifying sentiments and determining the sentence similarity. X.Zhu, P.Sobihani [32] also extended the chained LSTM structure to tree structure by employing a memory gate to save the previous work done by child cells. It claimed that this new model presented a principled approach of looking at interactions at a distance over hierarchies. The Multi-Timescale LSTM neural network (MT-LSTM) [33] was also created to simulate long documents by obtaining relevant data at various time scales. MT-LSTM broke down the latent states of the conventional LSTM model in different collections. Every collection was initiated and renewed at distinct stages, and so, it was able to simulate large documents. RNN's can catch the local arrangement of word sequences, although they have difficulty remembering long-term dependencies. Many different and impressive models were released based on RNN. Liu et al. P. Liu, X. Qiu, and X. Huang [34] used multitasking training to prepare RNN to use marked and specified data from several similar tasks. Johnson and Rie [35] investigated "a text area embedding method using LSTM". Zhou et al. [30] integrated Bi-Directional LSTM

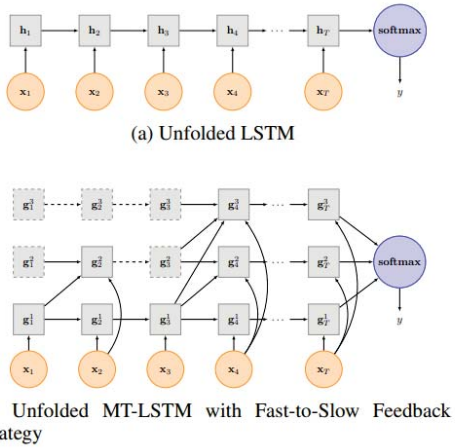


Fig. 4. The unfolded LSTM and unfolded MT-LSTM.

(BiLSTM) model with 2-dimensional "max-pooling to capture text objects".

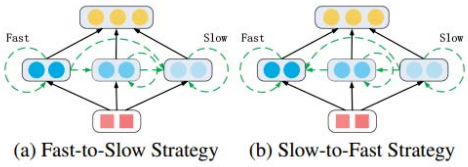


Fig. 5. Two feedback strategies of the model, dashed line indicate the feedback connection, the solid link indicates connection at contemporary time period.

When group Gk is activated at time period t,

$$i_t^k = \sigma(W_i^k x_t + \sum_{j\doteq 1}^{g} U_i^{j\to k} h_{t-1}^j + \sum_{j=1}^{g} V_i^{j\to k} c_{t-1}^j) \quad (3)$$

$$f_t^k = \sigma(W_f^k x_t + \sum_{j=1}^{g} U_f^{j\to k} h_{t-1}^j + \sum_{j=1}^{g} V_f^{j\to k} c_{t-1}^j) \quad (4)$$

$$o_t^k = \sigma(W_o^k x_t + \sum_{j=1}^{g} U_o^{j\to k} h_{t-1}^j + \sum_{j=1}^{g} V_o^{j\to k} c_t^j) \quad (5)$$

$$c_t^{\sim k} = \tanh(W_c^k x_t + \sum_{j=1}^{g} U_c^{j\to k} h_{t-1}^j) \quad (6)$$

$$c_t^k = f_t^k \odot c_{t-1}^k + i_t^k \odot c_t^{\sim k} \quad (7)$$

$$h_t^k = o_t^k \odot \tanh(c_t^k) \quad (8)$$

wherei_t^k, f_t^k and o_t^k are the vectors of input gates, forget gates, and output gates of group G_k at time period t respectively;c_t^k and h_t^k are the memory cell vector and hidden state vector of group Gk at time step t respectively. When group Gk is non-activated at time step t, its LSTM units remains unchanged.

$$c_t^k = c_{t-1}^k \quad (9)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

$$h_t^k = h_{t-1}^k \tag{10}$$

D. ALBERT

Z.Lan et al. [36] proposed the ALBERT (A lite BERT) architecture related to BERT. ALBERT employs a transformer encoder with GELU nonlinearities. Its models have much less parameter size than BERT models, reducing memory exhaustion and enhancing the training speed.
There are three main benefactions that ALBERT is a better design choice of BERT:

- Factorized Embedding parameterization
- Splitting of Cross-layer parameter
- Loss in Inter-sentence coherence.

The generated masked data for the Masked Language Model spots using n-gram masking [37], with the size of every n-gram mask is randomly picked and the probability formula having size "n" is represented as:

$$p(n) = \frac{1/n}{\sum_{x=1} 1/x} \tag{11}$$

The most significant size of n-gram can be chosen as 3, that is the Masked Language Modeling target can have up to a "3-gram" of full words, example: "India is great".

E. GNN based networks

Deep learning networks such as CNN, RNN work well on the regular structure but not for arbitrarily structured graphs. Many researchers have come across GNN based models for text classification on arbitrarily structured graphs [38] [39]. GNN-based models can learn the syntactic structure of the sentences, making them useful for text classification. Tong and Zhu [40] proposed the graph convolution network with attention. The model proposed consists of two parts: graph convolutional network and attention mechanism.

- **Graph convolution network:** This part finds the relationship between text and words, text and text, and learns the graph structure data.

The text passes through the middle hidden layer, and a new output matrix is generated to represent the text category. During training, the classification error was minimized using gradient descent. Meanwhile, the words are described as features and well synchronized with the training update.

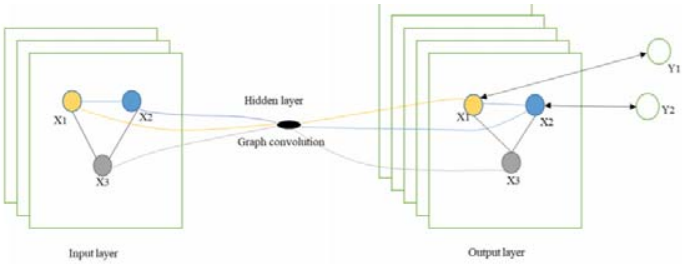


Fig. 6. Graph convolutional network

In Fig. 6, gray nodes represent words, colored nodes represent text, straight-line represents relationship between two nodes and the curve represents learning process showing updating of node features.
The update rule for a node is as follows:

$$X^{k+1} = \delta(AX^k\theta^k) \tag{12}$$

$$X_i^{k+1} = \delta(\sum_{j \in N_i} \frac{1}{d_{ij}} X^k\theta^k) \tag{13}$$

$\delta :$ *Relu activation function*
$X^k :$ *Input characteristics of kth layer*
$K :$ *Filter parameter matrix with training at kth layer k.*
$N_i :$ *Neighboring nodes of node i*
$d_{ij} :$ *Relationship between node i and node j*

TABLE II
COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES

Deep Learning Architecture	Novelty Introduced	Feature Extraction	Corpus	Validation Measure	Limitation
HANN	Hierarchical structure	Word embedding	Yelp, IMDB, Yahoo, and Amz	Accuracy	Works only for document-level
XLNet	Autoregressive language model, Permutation operation	Word embedding	IMDB, SST-2, Amz-2, Amz-5, Yelp	Accuracy	Computationally expensive*
BLSTM-2DCNN	Two-dimensional max pooling with bi-directional LSTM	Word embedding	Yelp, IMDB, and Amz-2, Amz-5	Accuracy	Doesn't work well for sentiment analysis
TextGCN	Graph Convolution Network	One hot encoding	20N, Ohsumed, MR, R52 and R8	Accuracy	Test nodes (without labels) are trained in GCN making it inherently transductive
ALBERT	Lowers memory consumption and increases the training speed of BERT	Word embedding	SST-2	Accuracy	Input scale how and what to add for different tasks is challenging

*Considered one of the best till now.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

- **Attention mechanism:** The attention mechanism in graph convolutional networks solves information loss between each node's attributes and the nodes due to average pooling operation.

$$Attention(N, D, V) = \ softmax(\frac{ND^T}{\sqrt{X_k}})V \quad (14)$$

N: Information matrix of nodes
D: Information matrix of edges
V: Parameter matrix

TextING [42] constructs separate graphs for each document and uses GNN to learn text-level word interactions to embed obscure words in new text properly. Several researchers proposed Graph attention networks (GATs) [43] for computing the hidden representation for each node. L.Hu [44] proposed the heterogeneous graph attention network with a dual-level attention mechanism. For a current node, the model learns the importance of neighboring nodes and various types of nodes. It addresses the semantic sparsity by capturing the relations for semi-supervised short text classification. To capture the correlation among the labels based on GATs, MAGNET [45] first learns the labels' dependencies. It uses a feature matrix and a correlation matrix to generate the classifiers. Texts in Event Prediction (EP) comprise a series of sequential sub-events, unlike other text classification tasks. Thus, to capture the relationship between various sub-events, SGNN [46] constructs an event graph and uses dense event connections for the EP task.

IV. RESULTS

Table II shows the comparison of different classification techniques. It compares the novelty, feature extraction or preprocessing, corpus, and the validation measure of the models. Lastly, it also mentions the limitations of each model. The performance of the models is compared using established metrics and then tabulated. Several metrics exist for classification tasks. Single-label Classification tasks can be evaluated with the help of a confusion matrix. Multi-label tasks are more complex and hence require more sophisticated metrics. Precision and Recall can be used here. These two metrics are inversely proportional. F1-score combines both of these metrics by taking their harmonic mean. Hence, F1-score gives a better measure of performance. Since there is not any universally accepted metric, we decided to use Classification Accuracy. [47].

Accuracy is the ratio of correct predictions to total inputs.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ Predictions\ made} \quad (15)$$

We also have more advanced metrics like Label-based Accuracy (L.B.A.) or hamming loss (H.L.) but are rarely used.
Table III, the results table, shows the approach and its paper along with the Dataset. The last column presents the Accuracy metric. We have highlighted the best accuracy achieved on all the datasets. Here, XLNet gives the best accuracy on many datasets, DBpedia, Yelp, Amazon, and AG Newsgroup. Furthermore, we can conclude that in comparison to other models, XLNet is relatively independent of the dataset taken. Meanwhile, the accuracy for other models varies significantly with the datasets. BLSTM-2DCNN performs the best on IMDB dataset. While TextGCN is able to achieve a high accuracy score on Reuters 8, its performance on other datasets is not that great. One of the lowest accuracy comparatively given by HANN. BERT gives good results on the IMDB dataset, falls just short of ALBERT and XLNet. ALBERT performs the best on SST2 dataset, with the help of additional data.

V. CONCLUSION

After performing extensive experiments and analyzing the results of the deep learning approaches, it can be observed that every approach excels in some conditions but also shows limitations in the others. The deep learning models enhance performance by improving the presentation learning method, model structure, and additional data and knowledge. This is in stark contrast to other techniques that improve their performance by

TABLE III
RESULTS TABLE

Approach	Paper	Datasets											
		Yahoo	SST-2	IMDB	Yelp	Amz-2	Amz-5	20NG	DBp	AGN	SNLI	R8	MR
HANN	Yang et al. [27]	75.8	-	49.2	71.0	-	63.6	-	-	-	-	-	-
XLNet	XLNet [28]	-	96.80	96.21	**98.45**	**97.60**	**67.74**	-	99.38	**95.51**	-	-	-
BLSTM-2DCNN	BLSTM-2DCNN [30]	-	89.50	**96.50**	-	-	-	-	-	-	-	-	**82.3**
TextGCN	TextGCN [41]	-	-	-	-	-	-	86.34	-	67.61	-	97.07	76.74
ALBERT	ALBERT [36]	-	**95.20**	-	-	-	-	-	-	-	-	-	-
BERT	BERT [20]	-	94.90	95.79	71.38	-	62.20	-	-	-	91.70	-	-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

improving the feature extraction scheme and classifier design [47]. In this paper, we compared four different techniques that belong to different areas. LSTM related techniques take considerable time to train, i.e., they are computationally one of the most expensive, but at the same time, they also give very high accuracy. Similarly, to increase the accuracy of XLNet and ALBERT, they have to be trained extensively. As mentioned earlier, every algorithm gives varied results on diverse datasets. Thus, there is a need to prepare a more robust and generic model that gives minor variance for most datasets while maintaining a good accuracy level. In addition to the above pitfalls, all of the techniques mentioned require labeled data for training. However, in the real world, labeled data is not always available. Hence, the models should be modified and applied for unsupervised classification. GCN, unlike other models, converts the linear text into graph nodes. Therefore, It works well even with a small set of training data. A complex model can be developed further by extending the models other than GCN for graph networks. So, as of now, no algorithm can be declared the best in general.

REFERENCES

[1] Benedict A. Rabut, Arnel C. Fajardo, and Ruji P. Medina. Multi-class document classification using improved word embeddings. In *Proceedings of the 2nd International Conference on Computing and Big Data*, ICCBD 2019, page 42–46, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[3] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A bayesian approach to filtering junk e-mail, 1998.

[4] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[5] S. T. Selvi, P. Karthikeyan, A. Vincent, V. Abinaya, G. Neeraja, and R. Deepika. Text categorization using rocchio algorithm and random forest algorithm. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 7–12, 2017.

[6] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[7] Yun Lin and Jie Wang. Research on text classification based on svm-knn. In *2014 IEEE 5th International Conference on Software Engineering and Service Science*, pages 842–844. IEEE, 2014.

[8] Yu Wanjun and Song Xiaoguang. Research on text categorization based on machine learning. In *2010 IEEE International Conference on Advanced Management Science(ICAMS 2010)*, volume 2, pages 253–255, 2010.

[9] Z. Jianqiang and G. Xiaolin. Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, 5:2870–2879, 2017.

[10] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016.

[11] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Christian Bizer. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6, 01 2014.

[12] empty. 20 newsgroups dataset, empty.

[13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[14] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). KDD '14, page 193–202, New York, NY, USA, 2014. Association for Computing Machinery.

[15] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[17] Yequan Wang, Aixin Sun, Jialong Han, Ying Liu, and Xiaoyan Zhu. *Sentiment Analysis by Capsules*, page 1165–1174. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018.

[18] Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot Text Classification with Distributional Signatures. *arXiv e-prints*, page arXiv:1908.06039, August 2019.

[19] Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. Improved representation learning for question answer matching. pages 464–473, 01 2016.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[21] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. *Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN*, page 1063–1072. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018.

[22] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[23] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *international conference on learning representations*, 2016.

[24] Tianyang Zhang, Minlie Huang, and Li Zhao. Learning structured representation for text classification via reinforcement learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6053–6060. AAAI Press, 2018.

[25] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[26] Matthias Aßenmacher. *Approaches in Natural Language Processing*, chapter Attention and Self-Attention for NLP. LMU Munich, 2020.

[27] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[28] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[30] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning based text classification: A comprehensive review, 2021.

[31] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics.

[32] Xiaodan Zhu, Parinaz Sobihani, and Hongyu Guo. Long short-term memory over recursive structures. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1604–1612, Lille, France, 07–09 Jul 2015. PMLR.

[33] Pengfei Liu, Xipeng Qiu, Xinchi Chen, Shiyu Wu, and Xuanjing Huang. Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2326–2335, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[34] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning, 2016.

[35] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings, 2016.

[36] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[37] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[38] mikael henaff, joan bruna, and yann lecun. Deep convolutional networks on graph-structured data. *CoRR*, 2015.

[39] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc.

[40] Genmei Tong and Min Zhu. Text classification based on graph convolutional network with attention. *Journal of Physics: Conference Series*, 1693:012038, dec 2020.

[41] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7370–7377. AAAI Press, 2019.

[42] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339, Online, July 2020. Association for Computational Linguistics.

[43] Petar Velickovic, Guillem Cucurull, A. Casanova, A. Romero, P. Liò, and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2018.

[44] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4821–4830, Hong Kong, China, November 2019. Association for Computational Linguistics.

[45] Ankit Pal, Muru Selvakumar, and Malaikannan Sankarasubbu. Magnet: Multi-label text classification using attention-based graph neural network. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 2020.

[46] Zhongyang Li, Xiao Ding, and Ting Liu. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4201–4207. AAAI Press, 2018.

[47] Qian Li, Hao Peng, Jianxin Li, Congyin Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:23:57 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

A Hybrid Machine Learning and Data Mining Based Approach to Network Intrusion Detection

Abhinav Singhal
Department of Information Technology
Delhi Technological University
Delhi, India
abhnav@gmail.com

Akash Maan
Department of Information Technology
Delhi Technological University
Delhi, India
maanakashdtu@gmail.com

Daksh Chaudhary
Department of Information Technology
Delhi Technological University
Delhi, India
dakshchaudhary07@gmail.com

Dinesh Vishwakarma
Department of Information Technology
Delhi Technological University
Delhi, India
dinesh@dtu.ac.in

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395918

***Abstract*—This paper outlines an approach to build an Intrusion detection system for a network interface device. This research work has developed a hybrid intrusion detection system which involves various machine learning techniques along with inference detection for a comparative analysis. It is explained in 2 phases: Training (Model Training and Inference Network Building) and Detection phase (Working phase). This aims to solve all the current real-life problem that exists in machine learning algorithms as machine learning techniques are stiff they have their respective classification region outside which they cease to work properly. This paper aims to provide the best working machine learning technique out of the many used. The machine learning techniques used in comparative analysis are Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) along with NSL-KDD dataset for testing and training of our Network Intrusion Detection Model. The accuracy recorded for Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN) and Support Vector Machines(SVM) respectively when tested independently are 98.088%, 82.971%, 95.75%, 81.971% and when tested with inference detection model are 98.554%, 66.687%, 97.605%, 93.914%. Therefore, it can be concluded that our inference detection model helps in improving certain factors which are not detected using conventional machine learning techniques.**

Keywords—Intrusion, Machine Learning, Inference Detection, K-Nearest Neighbors, Decision Tree, Support Vector Machine, Naïve Bayes.

I. Introduction

In 21st century, after the emergence of artificial intelligence and expert systems the new path ways for machine learning, deep learning, neural networks have emerged in various domains. As it has been observed, the most important thing for evolution is data.

In artificial intelligence too, data are the most essential tool which governs the performance of any modern-day systems and hence need to be protected. So, the data must be protected as "if something is valuable there will be long list of misusers attached to it." Here, the intrusion detection system (IDS) comes in play. As the intruders will try to access the valuable data in a network by entering the network through malicious ways, that are required to secure the devices on which data are being processed, stored or any kind of operations are taking place are to be secured. But even before that, it is vitally important to detect that the security breach exists. intrusion detection system helps in detecting such anomalous operations which might be a threat to our network.

Due to evolution of Network, now we are dealing with High Speed Networks even in our daily life i.e. optical fibre providing speed up to 1Gbps at normal household, 5G which is at our doorstep promises 1-2Gbps network link speed so, all the conventional network intrusion detection is becoming obsolete at the higher speed. Higher the data speed in a network, high speed intrusion detection system is to be needed along with high accuracy as we can't rely on slower speed NIDS (Network Intrusion Detection System). Due to factors listed above we have decided to design a new NIDS.

How to detect intrusion?

One of the oldest and still in use method is anomaly detection. As the name suggests, it detects the odd intrusion in a network i.e. something that is not behaving as per the general trend or accessing some sensitive data which are not supposed to be used by the user. An alert should be raised in such a case.

Nowadays due to increasing trend of machine learning and varieties of classification techniques available such as KNN, SVM, Naïve Bayes, etc accuracy in anomaly detection is peaking even above 99.5% which is trend setting for users even the slight improvement in accuracy can improve thousands of real life cases which are being ignored and path towards achieving perfect accuracy are opening.

II. Background

A. Classes of Anomalies

Anomalies can be classified as:

1. Point anomalies: An anomaly which is just in a clear sight no sign of any hiding attempt. E.g. High cash outflow at instance from credit card.
2. Contextual anomalies: An anomaly which is hiding in the plain sight it is too normal to be suspected. E.g., Low cash outflow at similar instance from credit card.
3. Collective anomalies: An anomaly which is in plain sight it is too normal to be suspected. E.g. Cash outflow from multiple credit cards.

B. Anomaly Detection Techniques

1) Simple Statistical Methods

Stats (these can be chosen with respect to method best suited for models and certain algorithms deemed right are to be used) are evaluated from the stored activities profile and every activity/process is evaluated and termed either as normal or harmful based on the degree of changes or dissimilarity from the optimum

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

ways. For example, your daily routine is studying, playing, eating if someday you decided to go on vacation so that will be termed as anomaly considering your usual behavior.

In simple words all the methods of anomaly detection are based on two factors:

1. Monitoring
   All the process/activities are monitored every time by the Intrusion Detection System(IDs) if any irregularities arise comparison mode is opted.
2. Comparison
   In this mode a dataset of all irregularities is stored i.e. all kind of symptoms (what is to be accessed by the process) these stats are called activities profiles. When there is a IDs call in case of any irregularities these signatures are compared.

2) Sandboxing

Monitoring process in the confined or restricted environment. Sandboxing is one of the most improper method Intrusion detection as it is not possible to first run the process in the restricted environment and then move on to the main system it is wastage of resource and time in general case if False Negative condition arises.

There are certain general issues which are very important regarding all Intrusion Detection system present:

1. Artificial Intelligence and Expert System are in desperate demands- There is a need for self-evolving models for Intrusion detection as there is no time to write thousands of lines of codes and compiling newer dataset with respect to the time.
2. Absence of Newer Datasets for model training- There are no new datasets for Intrusion and security related services so to train a machine learning model is going to be an uphill task.
3. Compilation of Dataset is extremely difficult in case of security related processes.
4. Consistent evolution of intruding systems is source of extreme misery.

III. Existing Work

In 2005, Wang Q et al.[1] new clustering algorithm for intrusion detection using KDD-99 shows the major path of using clustering for intrusion detection giving around 94% accuracy. In 2012, Chen et al. [2] introduce the community anomaly detection system (CADS) which consists relational pattern extraction and anomaly prediction in Collaborative information systems to detect intrusions. In 2018, Patgiri et al. [3] publish an investigation paper to evaluate machine learning algorithms such as Random Forest and Support Vector Machine(SVM) using NSL-KDD. In 2018, Shone at el. [4] and others show DL based method on KDD Cup'99 showing accuracy 97.85% and NSL-KDD showing accuracy 85.42%.

In 2019, Taher at el. [5] using Supervised Machine Learning Technique with Feature Selection using Artificial Neural Network(ANN) where input layer accepts 4 inputs and output layers give 2 output along with 2 hidden layers and Support Vector Machine(SVM) with Feature Extraction achieving accuracy 94.02% and 82.34% respectively which acts as a base of our proposed model. In same year, Raihan-Al-Masud et al. [6] and others proposed a NIDS model with ensemble i.e using 1 method as wrapper function and other as classification machine learning methods using NSL-KDD dataset. Later, Halimaa et al. [7], shows comparative analysis between various machine learning algorithms with respective results SVM showing 97.29%, Naïve Bayes 67.26%.

After progressive reading of all existing models, we observed that these models are overly reliant on the dataset and the outlier detection is totally based on the conditions outlined in the dataset no new inferences are created from quite sometimes.

By analyzing these existing models, we came to conclusions:

- No model is actually usable and effective in imitating real-life scenarios accurately.
- However, intrusions are not done in a straight and simple way. proper strategy is devised for particular step in intrusions. For example, there may not be a straight up ransomware attack in an organization, before certain attack there can be a strategic pattern like probe attack, rootkits and others in order to create a way for intrusion in a network.
- Some researchers have presented their analysis of the performance of machine learning techniques on NID. But none of them included a systematic comparative analysis of more than two techniques.
- To utilize the above patterns, we have created an inference network and inference detection models in order to mine and identify the attacks coming and block them in network.
- Substantial accuracy improvement is recorded when these methods are applied on current existing models. Accuracy and certain evaluated factors are described in results section of this paper.

IV. Proposed Methodology

Our hybrid intrusion detection model has two phases namely learning and detection. As the name suggests in learning phase, we would store the all types of pattern that an activity performs. Similarly, in detection phase, we would bring our model to use in order to calculate and observe the accuracy of the model.

There are various kinds of dependencies that exist publicly on network. Insiders in an organization can learn these dependencies and can use them to infer the value of one attribute by accessing another attribute. For example, let a and b are the two features and b are dependent on a, such as before writing b we need to read a, then there exists a probability that the value of b can be inferred if enough information about a is present. These features dependencies can be identified and exploited by an insider to access the sensitive data without deviating from their role profiles. Thus, this kind of an attack goes undetected from the M-NIDS approach described.

We are proposing a features selection using Chi-Square Test and setting up certain weights among features to set up a evaluating platform using Cramer's Rule which provides value between 0 and 1 even when 2 non categorical features are compared i.e. to show certain features are important than others as they have more impact on accuracy. Fig.1 shows architecture of our complete intrusion detection model.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

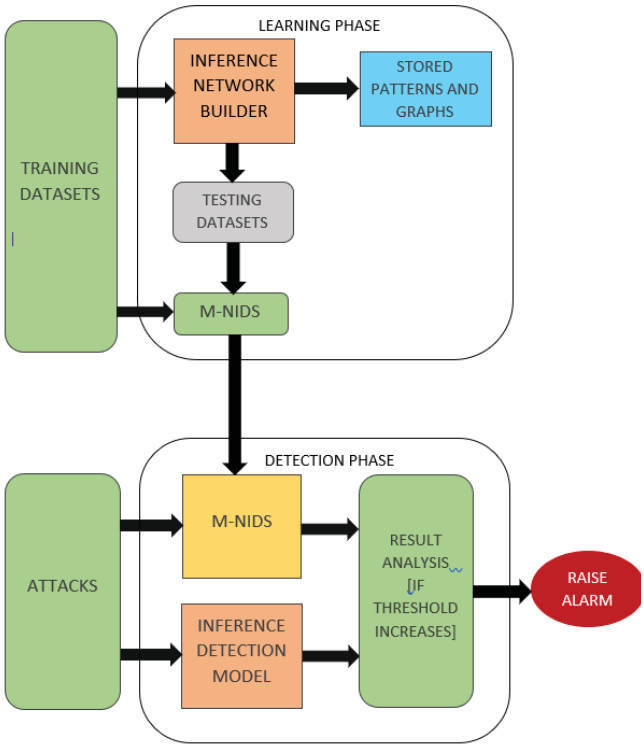


Fig. 1. Proposed model

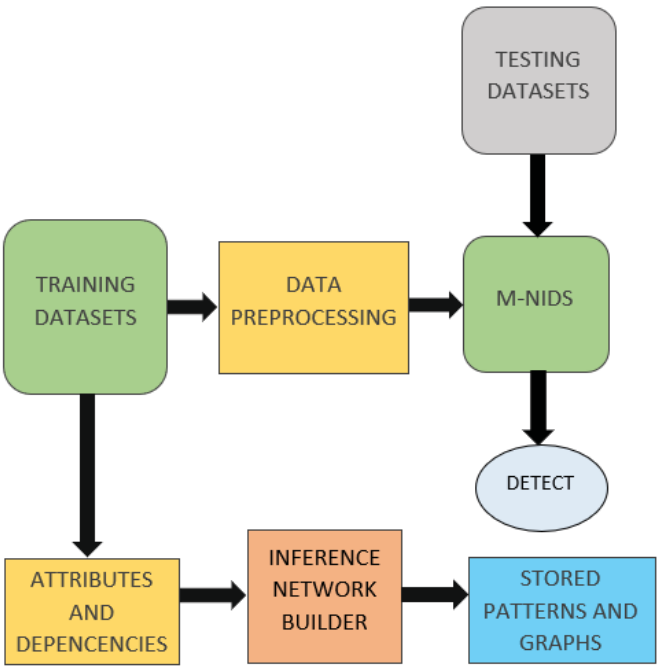


Fig. 2. Learning phase

A. *Learning Phase*

In this phase we use the dataset to train ML model for various attacks on network. Moreover, the transaction logs are also used to find out various dependencies that exist in the network, which can be exploited by a malicious user to attack the organization. The learning phase provides us with a framework which can then be used to detect malicious transactions and provide security. Fig. 2 shows the architecture of learning phase. It consists of two components that are explained in detail below.

1. Machine Learning Model for NIDS(M-NIDS)
2. Inference Network Builder using Data Mining

Data Preprocessing-Most machine learning models understand numbers better than words. In our dataset, four features (protocol, service, flag,label) had textual value. In this phase, each of these features assigned a numeric value using sklearn label encoder.

Finally each categorical column was converted to one-hot vectors.

B. *Detection Phase*

After constructing the inference graph and ML model in the learning phase, these are used to identify a malicious incoming network attacks in detection phase. Incoming activities are pre-processed and then checked parallelly against our previously learned models. Therefore, detection phase is divided into two parallel approaches.

1. Machine Learning Model for NIDS(M-NIDS)
2. Inference Detection Model

Results from M-NIDS and inference detection are then combined to get the degree of variation(anomaly) of an incoming activity and detect malicious activity.

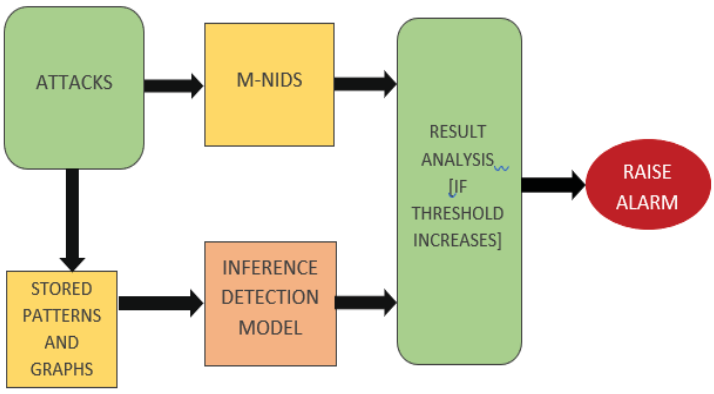


Fig. 3. Detection phase

C. *Machine Learning Model for NIDS(M-NIDS)*

i. Decision Tree – This algorithm uses a tree structure to analyse a choice made at each step. Each node's subtree signifies the scenerio if a particular decision is made. In this context, we have implemented Sklearn's DecisionTreeClassifier to filter hazardous intrusions out of the dataset.

ii. Naïve Bayes– It is a variation of bayes theorem, which uses a naïve assumption that different attributes of the dataset donot depend on each other.This assumption, which is not correct in many cases, is found to make specific systems highly effective in many problem statements. However, this assumption directly contradicts the idea of inference detection of features, which is based on mining the dependency of attributes on each other. Here, it was used to learn if a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

transaction is an attack, based on the features provided.

iii. K Nearest Neighbor (KNN)– This technique finds k records which are most similar to the record in question, among the whole dataset. If majority of these k records were hazardous, it is more likely that this record would also be hazardous. This technique is simple, yet efficient. The only drawback of this algorithm is that all the computation of distances between datapoints and prediction is done while testing as there is no training phase.

iv. Support Vector Machines (SVMs) – It is the most widely used technique used for classification problems. In this, the model tries to draw a boundary (hyperplane) to separate the hazardous and safe intrusions. It is considered the best classification model because it can even classify non linear data..

D. Inference Network Builder using Data Mining

This step focuses on inference detection in the network. We have to obtain dependencies which are given data items in the network. Dependencies refer to the relationship between different data points. We used sequential pattern mining to obtain these dependencies.

Let R(a) -> W(b) be a classification rule where a and b are the two data items and b are dependent on a. Since before writing b we need to read a, there exists a possibility that the value of b can be inferred if enough information about a is present. Thus, our model works on this concept.

There are various kinds of dependencies that exist publicly on network. Insiders in an organization can learn these dependencies and can use them to infer the value of one attribute by accessing another attribute. For example, let a and b are the two data items and b are dependent on a, such as before writing b we need to read a, then there exists a probability that the value of b can be inferred if enough information about a is present. These dependencies can be identified and exploited by an insider to access the sensitive data without deviating from their role profiles. Thus, this kind of an attack goes undetected from the M-NIDS approach described above.

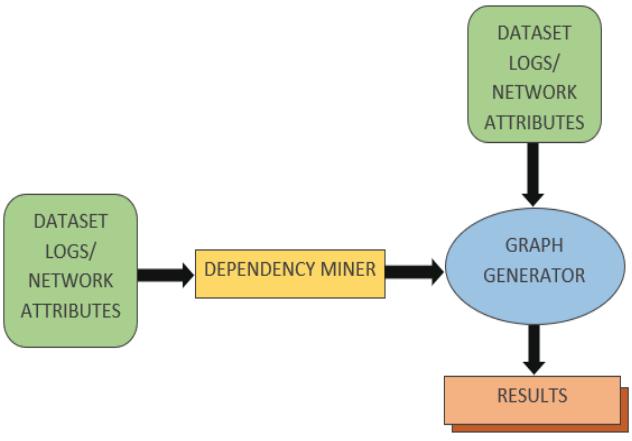


Fig. 4. Inference detection using data mining

Moreover, various inferences can be combined that may lead to inferring a sensitive data forming a complete inference channel. We construct an inference network that specifies all the possible inference channels to mitigate the inference attack. The technique employed uses frequent sequential pattern mining to generate data dependencies among data items. Dependencies are stored in form of classification rules. The tools used for the same are explained as follows:

1) The **Access Sequence** signifies the sequence in which the required transaction need to take place. Hence, a sequence of <(r1),(r2),(w2),(r3)> would mean these four operations will need to execute in the same sequence to avoid inconsistency.
2) **Inference Graph** It is graph in which each node represents a separate attribute from the dataset and the edge weights represent the correlation between the attributes.

During data mining we have mined several features for dataset suppose there is a feature set as follows.

$$F=<f_1, f_2, f_3, \ldots\ldots\ldots\ldots, f_n>$$

Now, we have certain features so inference graph is to drawn between features i.e. inference graph is drawn by first calculating the correlation among features. So, we will need some numerical values to make sure these features are equivalently compared for that we used modified Cramer's rule.

Constructing Inference Graph:

We construct the Inference Graph, in which the attributes of the network are represented as the vertices of the graph. Directed edges are formed using access rules mined above. For each access rule, a weighted edge is formed between all the attributes present in precedent and the antecedent of the rule, directed from the former to the latter.

To construct the inference network, we use dataset. Each pattern in the dataset is converted to read-write sequences using the sequence generator. These sequences are then passed through dependency generator. These read and write rules exhibit the dependencies between attributes and are used to construct the graph.

These edges so formed are assigned weights from the data correlation between the two connected vertices. Since one vertex is dependent on the other, highly correlated vertices increase the probability of inferring the value of the dependent attribute. Thus, weights of edges are directly dependent on the degree of correlation between the connected vertices. As we use corrected Cramer's V for correlation analysis, it is used as the weight of edges since the value lie for 0 to 1.

Correlation analysis

In order to find the weight on the edges of the graph we use correlation analysis. Correlation give a degree of statistical association between two variables. Given two highly correlated attributes with one dependent on another, an insider with knowledge of this dependency can infer the correct value of the dependent attribute with a high probability by accessing only the independent attribute.

We use corrected Cramer's V (reference) for correlation analysis. Since the value of Cramer's V lie between [0,1], we can directly use this as the edge weight. The edge weight gives us an intuitive probability of correct inference on that edge.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

Edge Weight W_{AB} is the weight assigned to the edge between features A and B. Then features A and B are two simultaneously distributed variables. Let p be the number of unique values of A and a $\in$ (1, p) and q be the number of unique values of B with b $\in$ (1, q) then,

$N_{a,b}$ = number of time values (A_a, B_b) occur together
N_a = number of time A has value A_a
N_b = number of times B has value B_b
N = total tuples

Chi square statistic is used to find the correlation of two variables as follows. This test is commonly known as Chi-Square Test for feature selection. The chi square is then given by:

$$\chi^2 = \Sigma_{a,b} (N_{a,b} - N_a.N_b/N)^2/(N_a.N_b/N) \quad (1)$$

$\varphi^2 = \chi^2/ N$, where φ is the phi coefficient in case of 2x2 contingency table.

Then the bias corrected weight is given by

$$W_{AB} = (\varphi^2 / \min (p'-1, q'- 1))^{½} \quad (2)$$

Where φ'^2, p', q' are corrected terms defined as

$$\varphi'^2 = \max (0, \varphi^2 - (p-1) (q-1) / (N-1)) \quad (3)$$
$$p' = p - (p-1)^2/ (N-1)$$
$$q' = q - (q-1)^2/ (N-1)$$

As Cramer's V is used to find the correlation between categorical data, continuous attributes in network are grouped into bins so that they can be used alongside other categorical variables. This is an intuitive approach since the value of continuous attributes are likely to be inferred in ranges rather than specific values. We use freedman-diaconis rule to divide the continuous data into groups.
Now, we have received the weight matrix which will look like as follows.

$$W=\{<W_{11}, W_{12}, W_{13}, \ldots\ldots\ldots\ldots, W_{1n}>$$
$$<W_{21}, W_{22}, W_{23}, \ldots\ldots\ldots\ldots, W_{2n}>$$
$$<\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots>$$
$$<W_{n1}, W_{n2}, W_{n3}, \ldots\ldots\ldots\ldots, W_{nn}>\}$$

E. Inference Detection Model

Various dependencies exist among the data. These dependencies may be due to functional dependencies or due to the structure of the organization like hierarchy in a company and respective remuneration. Insiders in an organization may have access to a part of or all the above information and over a period of time may learn these dependencies. Insiders can run multiple attacks and use these dependencies to infer sensitive data without accessing them. Moreover, two or more attacks can collude to infer data not accessible to them. Since collusion detection is not a trivial task we use inference detection to alleviate this issue. The aim of inference detection is to find all the inference channels a user can use to infer sensitive data.

As defined in the learning phase, we create an inference network that consists of all the dependencies that exist in the network. This network is used to detect for inference channels that can be used by malicious insiders to attack the network. However, we define the following terms.

1) Sensitive attributes are those attributes that store confidential information which when compromised may lead to attack. For example, salary of an employee, bank account number etc are sensitive attributes. If an insider is able to infer this information then it is a serious attack on the security of network. However, an employee in accounts department does have access to this kind of sensitive information. So, it not the access of account information that is sensitive but the link between name and the account number that poses threat to the security of network. Thus, for such an employee name can be viewed as a sensitive information. Thus, we define two types of attribute sets.

2) Access set: let AS_i be the access set for user role U_i , then AS_i is defined as the set of attributes accessed on the network by certain allowed individuals.
$AS_i = \{a_k|$ u access a_k where u e $U_i\}$

3) Sensitivity set: let SSi be the sensitivity set for user role Ui , then SSi is defined as the set of attributes sensitive with respect to that user on network.
$SS_i = \{s_k|$ s_k is sensitive to u where u e $U_i\}$

4) Inference path: let I_i be an inference path for network user U_i. Then I_i is defined on the inference graph as
$I_i = \{w_s, w_1.w_2...w_f\}$, w_k is a vertex
W_s is the starting vertex s.t.$w_s \in AS_i$, w_f is the terminal vertex s_t $w_f \in SS_i$.

Each attack's version of the inference network is stored. After any activity on network, access belief strength of the access set attributes is updated. Inferred belief strengths are then propagated along all the inference paths using the equation. If Inferred belief strength for any sensitive attribute comes as greater than a threshold, the activity is marked malicious.

An attack may not know about all the inference channels that exist, since some of the attributes may not be accessible to the attack. Moreover, an attack may not be using the accessed data to infer sensitive data, our models takes a strict approach by considering the possibility of inference and collusion by considering all the channels and attributes. This approach is combined by a somewhat lenient M-NIDS detect malicious activity.

When certain patterns are detected, in case the predicted attack is different from the actual attack, ratio of their possibility values is calculated. If this ratio is less than a certain user defined θ, we use the inference detection to check for malicious intention. If belief strength for any sensitive attribute is greater than threshold Ψ, is marked as malicious. Otherwise, the activity is non-malicious. If, however the possibility ratio is greater than θ, activity is straight away marked as malicious i.e. Ψ is greater than 0.9 than we say next activity is an attack where Ψ is a ratio always between 0 and 1.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

V. DATASET

NSL-KDD is a common data set and is a standard for present day internet traffic. It is an improved version of KDD dataset as it does not have redundant records. This data set is made up of 4 sub data sets i.e. KDDTest+, KDDTest-21, KDDTrain+, KDDTrain+_20Percent.

VI. RESULTS

For result calculation and validation of our model, we have used Jupyter Notebook along with Python 3.9 using KDD-NSL dataset. After testing the data using the four most commonly used machine learning techniques, with and without inference graph, the results obtained are shown in table 1.

In Table 1, Independent Accuracy Results are calculated directly with the conventional machine learning technique without any external algorithm used on our dataset. Whereas, Accuracy Results are calculated using our Inference Function where we are calculating the threshold Ψ if it exceeds 0.9 then we are labelling an activity as an attack even if independent result concludes it as normal activity. Due to the use of inference function we are able to increase accuracy in decision trees and SVM.

First, we applied each of the four techniques on the dataset and recorded the performance results in detecting network attacks. The results are shown in Fig. 5.

After implementing the inference function, the new accuracies are shown in Fig. 6.

TABLE I. PERFORMANCE COMPARISON OF THE ALGORITHMS

ALGORITHM	INDEPENDENT ACCURACY	ACCURACY WITH INFERENCE FUNCTION
KNN	95.755%	97.605
DECISION TREE	98.088%	98.554%
NAÏVE BAYES	82.971%	66.687%
SVM	81.263%	93.914%

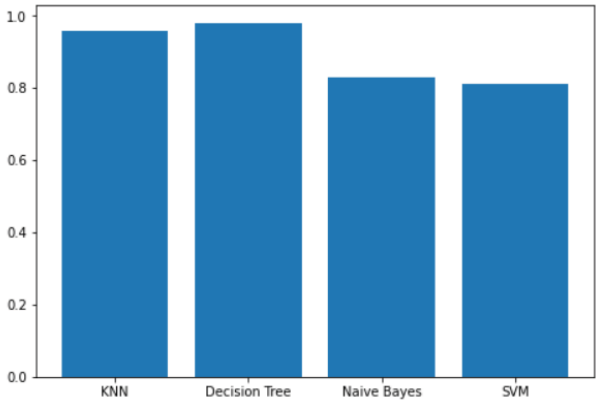


Fig. 5. Bar graph comparing performances of the four techniques used.

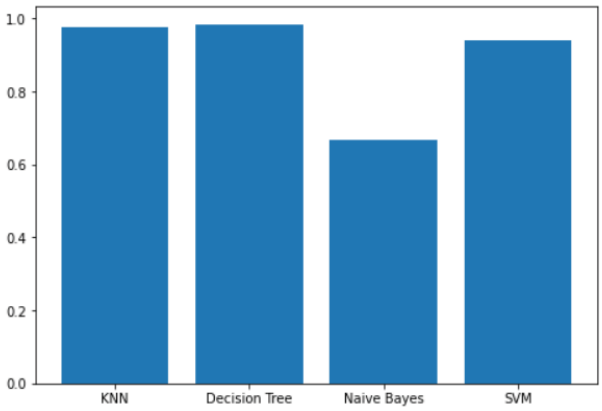


Fig. 6 Bar graph comparing the performance of the four techniques used along with the inference detection function

Finally, we plot a combined bar graph showing each algorithm's performance comparison with each other, in both the above-mentioned scenarios, as shown in Fig. 7.

We observed that:

- Decision tree gave the best result in detecting a network attack in both the environments we tested:
  i) When tested independently:
  Accuracy = 98.088%
  ii) With inference functionality:
  Accuracy = 98.554%
- K-Nearest Neighbours algorithm also performed exceptionally well, with a slightly less accuracy than Decision Trees.
  i) When tested independently:
  Accuracy = 95.755%
  ii) With inference functionality:
  Accuracy = 97.605%
- Naive Bayes technique gave the least accurate results in predicting network attacks. Also, it was the only technique in which accuracy further decreased on adding the inference functionality. The naive assumption in this algorithm might justify this result, as it assumes that individual attributes of a record have no impact on each other, which is contradictory to the idea of inference of attributes.
  i) When tested independently:
  Accuracy = 82.971%
  ii) With inference functionality:
  Accuracy = 66.687%
- Support Vector Machine technique (SVM) performed decently on the dataset but showed the highest improvement in accuracy (12.651%) when used along with inference detection.
  i) When tested independently:
  Accuracy = 81.263%
  ii) With inference functionality:
  Accuracy = 93.914%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

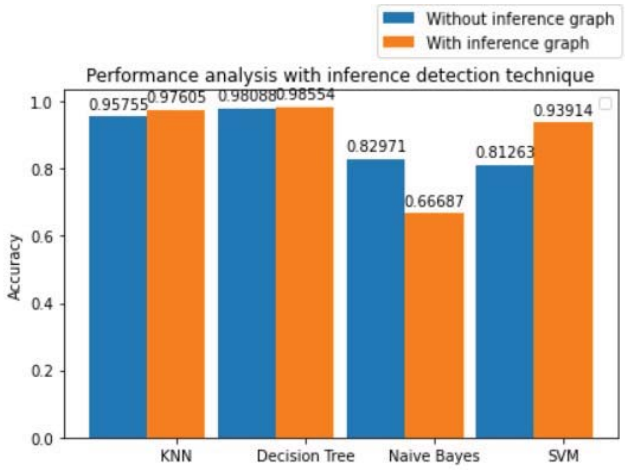


Fig. 7. Combined bar graph comparing accuracies by both the methodologies

VII. Conclusion

Hence, we conclude that three among the four machine learning techniques we considered, showed evident improvement in performance if accompanied by our inference detection model in detecting hazardous intrusions to a network, with Naïve Bayes as an exception.

It was also observed that SVM technique showed drastic improvement in accuracy, by using the inference function.

VIII. References

[1] Q. Wang and V. Megalooikonomou, "A clustering algorithm for intrusion detection," Mar. 2005, doi: 10.1117/12.603567.

[2] Y. Chen, S. Nyemba, and B. Malin, "Detecting Anomalous Insiders in Collaborative Information Systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 3, May 2012, doi: 10.1109/TDSC.2012.11.

[3] R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, "An Investigation on Intrusion Detection System Using Machine Learning," Nov. 2018, doi: 10.1109/SSCI.2018.8628676.

[4] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, Feb. 2018, doi: 10.1109/TETCI.2017.2772792.

[5] K. A. Taher, B. Mohammed Yasin Jisan, and Md. M. Rahman, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection," Jan. 2019, doi: 10.1109/ICREST.2019.8644161.

[6] Md. Raihan-Al-Masud and H. A. Mustafa, "Network Intrusion Detection System Using Voting Ensemble Machine Learning," Dec. 2019, doi: 10.1109/ICTP48844.2019.9041736.

[7] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," Apr. 2019, doi: 10.1109/ICOEI.2019.8862784.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:30:45 UTC from IEEE Xplore. Restrictions apply.

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)

A Novel Approach to Scene Graph Vectorization

Vinod Kumar[1], Deepanshu Aggarwal[2], Vinamra Bathwal[3], Saurabh Singh[4]
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
vinod_k@dtu.ac.in[1], deepanshu4929@gmail.com[2], vinamrabathwal99@gmail.com[3], saurabhs9913@gmail.com[4]

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397230

***Abstract*—In recent times due to the advancement in perceptual applications, focus in computer vision has been inclined towards tasks that require a significant level of semantic understanding of scenes. Combination of textual and visual information has resulted in a great improvement in performance on tasks like retrieval, captioning and visual q/a etc. In this regard, scene graphs have become a popular form of structural knowledge. But unlike Word embedding, general-purpose scene graph embedding has not been explored significantly. In this work, we propose a general-purpose scene graph embedding model that combines linguistic and graph processing techniques through a reconstruction based learning network to learn a low-dimensional data-driven vectorized embedding of scene graphs. Visualization of embedding of COCO dataset has shown to possess semantic separability and distance-based abstraction of scenes. When applied to a retrieval task and evaluated using Med-r and recall metric on COCO-stuff and VRD dataset, our model showed promising results.**

***Index Terms*—scene graph, embedding, knowledge representation, semantics**

I. Introduction

At times, there's a need to process the semantic information present in the scenes/images to enhance the quality of results provided by various computer vision algorithms employed on these images. Algorithms with the image as input to them, usually tend to focus on the absolute attributes(color, shape, size, IoU with the image, etc.) of the objects present in the image, i.e., on the conceptual information, and not on the relationship that these objects hold with each other,i.e., the contextual information (e.g. the man *plays with* dog). In such cases, we need to model/transform the images/scenes to a semantically rich representation. Semantics encapsulate both types of information: conceptual or absolute attributes of objects, as well as contextual, or their relationship with each other. Recently, scene graphs[1] have gained popularity to capture the semantics of a scene [1, 2, 3]. It resembles a graph data structure in its core form, as in the objects present in a scene are represented as nodes of the graph, while the relationships amongst them are represented as edges [4]. Fig. 1 shows an example of a scene graph corresponding to a scene. Although the scene graph captures the entire semantic information of a scene in a really precise and efficient manner, it still doesn't serve as a suitable input to most of the algorithms built upon to make use of the semantics because of its heterogeneous nature and the problem of transitive dependencies. At this point, there is a need to transform this graph representation to a rather linear representation such that it preserves the entire information of the graph and can very well serve as an input to most of the learning architectures.

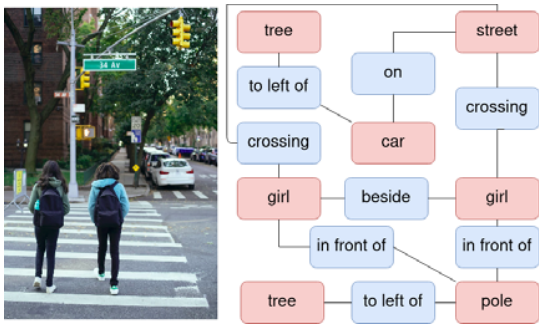


Fig. 1. **Scene Graph Example:** Figure contains a sample image from COCO-Stuff dataset and it's corresponding Scene Graph, The red boxes represent the objects present in the image(ex. girl, car, etc) and the blue boxes represent the visual relationship between objects(ex. in front of, to left of, etc).

In this paper, we present a novel architecture to convert a scene graph to its low-dimensional distributed vector representation that can further be applied easily on various downstream tasks. Previous works involving scene graphs have mostly focused on their application to a specific task such as image retrieval [1], image generation [5] whereas our approach is a general approach to obtain a vector representation of a scene graph. Our architecture comprises of three parts. First, the *linguistic embedding network* that builds upon the existing language model of Sentence-BERT[6] to capture the information represented by the labels, attributes of the objects and the relationships between them. Second, *graph embedding network* is constructed to work upon the structural information contained in the graph-like structure of the scene graph. Third, *representation network* is an autoencoder[7] based network used to learn the low-dimensional latent space representation of the scene graph using reconstruction based learning. The scene graph is fed into the linguistic embedding network and graph embedding network to obtain a higher dimensional representation. Finally, the obtained higher-order vector space is then converted into a lower-order vector space considering the global representation structure using representation network to obtain a lower-dimensional distributed vector embedding. In the further stages of our architecture, our model is sent for training on large scale datasets like Visual Relationship Dataset[8], MS-COCO[9] etc. The embedding obtained for the COCO dataset shows promising results when projected

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:19:06 UTC from IEEE Xplore. Restrictions apply.

onto a 2D vector space to visualize the clusters formed by the semantically similar scene graphs, while the dissimilar scene graphs have a lower inter-vector similarity. We also demonstrate the applicability of the embedding by using it on a retrieval task and quantitatively evaluating the results.

II. Related Work

Scene Graphs were instituted as a structured knowledge representation of semantic information [4, 5] in a format commonly used in [10, 11]. A scene graph consists of nodes, which are objects of the image, and visual relationships between these objects as edges. Scene graphs have found their application in a multitude of computer vision tasks like Image Generation [5, 12, 13], Image retrieval [1, 14], Image captioning [3, 15, 16], etc., and also have produced significant gains in many of the listed applications. Because of the applicability of scene graphs, many researchers have explored the task of scene graph generation from a given input image[17, 18] or textual description of image[19].

Many researches [5, 13, 20, 21, 22] have used intermediate layout prediction to create a learnt embedding. In a closely related paper[23] Schroeder et al. introduced additional triplet supervisory signals and data augmentation to improve upon the embedding and handle long-tail distribution of the visual genome dataset. In our work, we do not use layout prediction as a pretext for learning scene graph embedding. In another related work[24] Belilovsky et al. combined visual and scene graph features to generate a joint visual-scene graph embedding. On the similar lines of scene graph generation[25] Wan et al. generates new relationship triplets from existing sparse scene graphs using a convolution module for visual feature extraction and a hierarchical model for recombination of structural and visual embedding. Contrary to these approaches, we do not supply visual features to learn an embedding.

III. Method

Our goal is to convert the input, in the form of a scene graph representing the objects and the relationships between them in a scene, to a latent space representation. Our primary objectives are threefold:

- Develop a method to process the complex graph structure.
- Process the linguistic information about nodes and edges of the scene graph.
- Develop a method to convert the scene graph to a low dimensional latent space representation ensuring that the real context or meaning of the image is not lost.

Let us discuss the primary components of the proposed model and our algorithm in detail.

A. Linguistic Embedding Network

Linguistic embedding network converts a triplet of the form $< subject, predicate, object >$ to a fixed dimensional representation D_l. Textual description of each node needs to be processed to extract out meaning of the object and relationship among them. The network uses a Sentence-BERT network[6], which uses a Siamese and triplet network to extract the semantically correct meaning from the sentence and gives as output a vector representation which can be compared using cosine similarity. Sentence BERT is used because of its high computational efficiency over other sentence embedding methods and can efficiently handle semantically similar phrases written differently. The framework provides a dense vector representation of the phrases or sentences.

This network takes input in the form of triplet $< subject, predicate, object >$, which is then converted into a phrase by concatenating the subject, predicate, and object. The phrase obtained expresses a complete description of an action or relationship along with the entities involved. We will use Sentence-BERT pre-trained model fine-tuned for semantic textual similarity. The phrases are fed into the Sentence-BERT network to generate a 1-D embedding of size D_l for each triplet.

B. Graph Embedding Network

The graph embedding network takes input a raw graph in the form of nodes and edges without their feature vectors and generates a node embedding for each node of the graph based on the structures of the graph. The network converts the complex graph data structure to a vector-based embedding. It maps the vertices to a low dimensional space preserving the inherent property of the vertices and the topological information of the graph. It must preserve the first order as well as higher-order proximity[26] in the network.

- First order proximity is that two nodes are considered similar if they have an edge between them.
- Higher-Order proximity refers to the similarity of two vertices if they are present in the neighborhood. They do not have a direct edge between them but have common neighbors.

We use a Node2Vec[27] graph embedding technique, which is a walk based technique, which uses a skip-gram network for node embedding. It learns a vector representation for nodes in a graph by simulating biased walk based on a preserving neighborhood. The objective of the algorithm can be flexible,i.e., the algorithm can be modified according to various needs by trying to simulate random walks based on probabilities. Node2vec has proved to be less computationally expensive and words on diverse patterns of graphs. The flexibility node2vec offers in terms of identifying diverse graph patterns is suitable for us as scene graphs representing scenes can have a wide range of type and number of objects and relationships among them. The graph is fed into the node2vec network to generate a node embedding for each individual node. It outputs a vector of size $D_g \times n$ where n represents the number of nodes in the graph and D_g represents the fixed dimension of the individual node embedding.

C. Representation network

The representation network uses an autoencoder[7] architecture to convert the higher dimensional representation to a lower-dimensional representation. Autoencoder is an unsupervised learning technique to generate a low dimensional

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:19:06 UTC from IEEE Xplore. Restrictions apply.

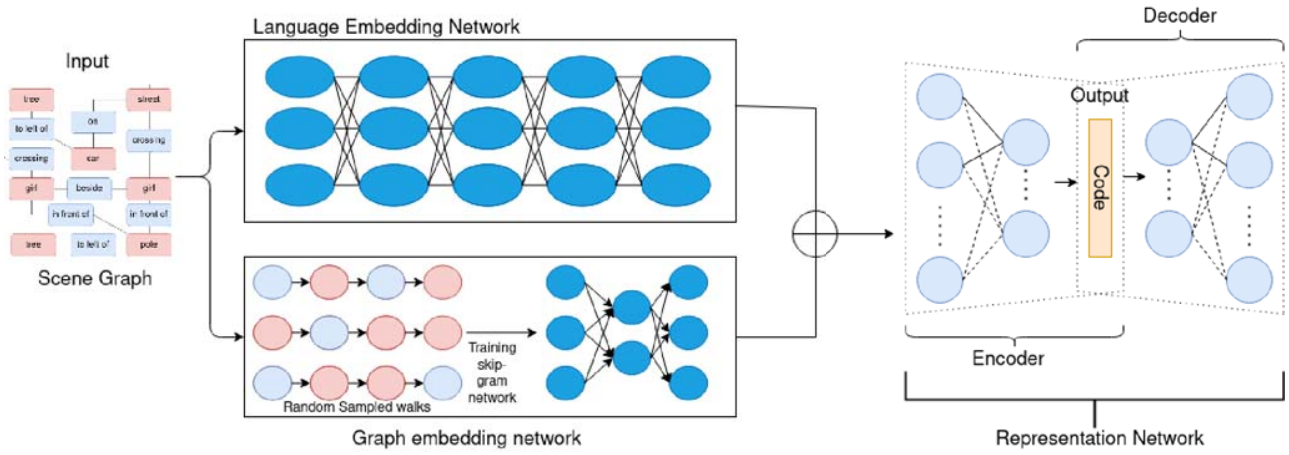


Fig. 2. **Architecture :** Here, we present a multi-step network that takes a scene graph as an input and generates a low dimensional embedding of the scene graph. The scene graph passes through a Language embedding network and a graph embedding network that encodes the linguistic and structural information respectively, then pools these together and passes it to the autoencoder based representation network for learning the low-dimensional representation.

representation code by having a bottleneck in the neural network. Autoencoders are able to learn linear as well as nonlinear relationships present in data. The low-dimensional embedding created must be able to encapsulate the information in the scene graph, while the separation between two scene graphs in the latent space being inversely proportional to their actual similarity.

We use an architecture inspired by [26], which uses two shared weight parallel autoencoders (see Fig. 3) to ensure that the embedding respects both the first order as well as higher-order proximity. This method was chosen to preserve the information of the scene graph and also respect the higher-order proximity with other representations in the latent space. The losses are calculated from the low-dimensional code as well as from the reconstructed output and back-propagated in the network to learn the representation.

D. Algorithm

Step 1: Preprocessing Convert the scene graph from a form where each node and edge belongs to more than one class to a phrase-based form where each node is represented by a triplet $< subject, predicate, object >$, and edges represent the involvement of a common subject or object in corresponding actions. The different nodes have edges between them if they have a common subject or object. This transformation is done for two reasons: converting the heterogeneous edges of the scene graph to a graph with homogeneous form makes it easier to process as all edges represents common subject or object, and to prevent the loss of context which happens when textual description of nodes and edges are processed individually.

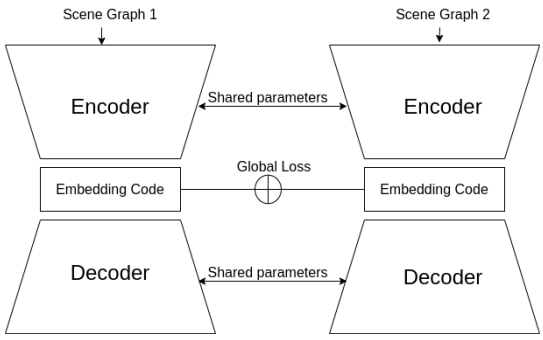


Fig. 3. **Shared weight Autoencoder :** This is a SDNE inspired shared weight architecture that aims to preserve higher order proximity. The loss computed with intermediate codes preserve the global structure of representations.

Step 2: Generating Linguistic Embedding Each node of the graph is then fed into the language embedding network to generate a linguistic embedding of nodes in form of 1-D vector X_l of size D_l.

Step 3: Generating Structural Embedding The scene graph is fed into the graph embedding network to generate the graph structure based node embedding X_g of size D_g.

Step 4: Generating Final Embedding The two representations are merged together to generate a vector X_t for each individual node. The vectors X_t of each node of the scene graph is pooled together and fed into the encoder part of the representation network to generate the low dimensional embedding of the scene graph as shown in Fig. 2.

This method ensures that semantically similar scene graphs are closer to each other in latent space and the information of the scene graph is well represented in the embedding.

IV. Experiment

A. Dataset

Visual Relationship Dataset Visual Relationship Dataset is a benchmark dataset for Visual relationship detection[8]. It consists of about 5K images and 37K relationships. The objects belong to 100 object categories, and relationships belong to 70 predicate categories. The predicates can be broadly classified into action, spatial, preposition, comparative, and verb. The VRD dataset provides real-world scene graphs

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:19:06 UTC from IEEE Xplore. Restrictions apply.

but is known to suffer from long-tailed distribution of classes, which does not occur quite frequently. The visual relationship dataset contains simple triplets of relationships. We convert these sets of triplets into a scene graph by connecting triplets that have a common subject or object.

COCO Dataset We use 2017 COCO stuff dataset which contains 40K train and 5K validation images[9]. We randomly selected 5K train images for training our model. To generate a scene graph from the COCO dataset, we use a technique similar to [5] where we generate relationships among objects based on their relative geometric positions. The six relations we consider are 'to left of','to right of','above','below','inside','surrounding'. Smaller objects appearing on the side of the image are ignored. We only consider objects which cover at least 5% of the image and use only those instances of data in which the number of objects is between 3 to 8.

B. Training

The model is trained to generate an embedding X_c of dimension D_f for the scene graph. The embedding generated should contain the information of the scene graph as well as it should respect other representations in the latent space. To ensure this, we use a combination of loss function for training our model. Before defining the loss function, let us get familiar with some terms.

Cosine Similarity Cosine similarity is a measure of similarity between two vectors. It refers to the cosine of the angle between the two vectors. It is used to compare vectors, especially embedding, in text analysis. Mathematically, Cosine Similarity $CS(X_i, X_j)$ between two vectors X_i and X_j, is defined as

$$CS(X_i, X_j) = \frac{X_i.X_j}{|X_i||X_j|} \tag{1}$$

IoU Similarity We use Intersection over Union (IoU) similarity to find similarity between two scene graphs. Each triplet $< Subject, Predicate, Object >$ is fed into the language embedding network to yield the embedding of each triplet of the scene graph. The representations generated is compared using cosine similarity, and then, Intersection over Union is computed as a measure of similarity. More similar scene graphs have more matching triplet; hence their IoU value is closer to 1. Similar to Cosine Similarity, IoU similarity also ranges between 0 to 1(inclusive). Mathematically, IoU similarity IS of two scene graphs S_i and S_j expressed as a set of triplets is defined as

$$IS(S_i, S_j) = \frac{S_i \cap S_j}{S_i \cup S_j} \tag{2}$$

We define two types of losses, Global loss and Local loss. The Global loss is defined as the mean-squared error(MSE) loss between the cosine similarity of the two embedding code generated in the paired shared weight autoencoder, and the IoU similarity between the respective scene graphs. The local loss is the reconstruction loss, which is defined as the MSE loss between the reconstructed output and the input of the learnable network. We use a parallel autoencoder architecture which allows us to train the model using a limited set of data, as for a dataset of n scene graphs we generate n^2 pairs of scene graphs to be fed into the network in pair. For two scene graphs, S_i and S_j fed into the network and let X_i and X_j be their respective representation code generated using parallel autoencoders. Mathematically, the global Loss L_g and local loss L_l is defined as

$$L_g = (CS(X_i, X_j) - IS(S_i, S_j))^2 \tag{3}$$

$$L_l = MSE(X_i - X_i') \tag{4}$$

The total loss L_t is computed as a combination of local loss L_l and global loss L_g as

$$L_t = \alpha L_l + (1 - \alpha) L_g \tag{5}$$

where α is a hyperparameter that determines the extent to which the embedding should respect the global affinity over local reconstruction.

C. Experimental Setup

The models were trained on Google Collaboratory with Tesla K80 GPU with 12 GB RAM. The language embedding output dimension D_l was set to 768, the graph embedding dimension D_g was set to 128. The final embedding dimension D_f was set to 128. The learning rate was kept 10^{-2} initially and was set to decay to $1/10^{th}$ of the value after every 10 epochs. We set the number of epochs to 50. Based on the analysis of [26] on first order proximity coefficient, hyperparameter α was set to 0.3 to balance the local reconstruction and global affinity. Each batch contained 64 pairs of scene graphs trained in parallel using the shared weight network.

V. Results

A. Visualizing results

To visualize the results on COCO dataset, we select 100 random scene graphs from our dataset. The embedding for these scene graphs is generated with our model. We use t-SNE[28] plot for visualization. Since directly obtaining context from the scene graph is difficult for humans, we use the image caption available in the COCO dataset as the description of each individual scenes. We identify many small clusters and extract image description for those clusters, as shown in Fig. 4. The plot of the embedding shows good separability with semantically similar scenes appearing together. For eg., the blue box in Fig. 4 represents scenes related to eating or selling of fruits and vegetables. The embedding is able to capture higher-level clusters such as indoor/outdoor scenes as blue, yellow, green and orange boxes represents scenes which are mostly indoor whereas magenta and cyan boxes represent the outdoor scenes. The separability can also be justified by the amount of variation in sentence description of the enclosing points for small and larger boxes. For instance, the yellow box contains scenes related to toys, cats and laptops whereas magenta box represents scenes related to zebras predominantly.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:19:06 UTC from IEEE Xplore. Restrictions apply.

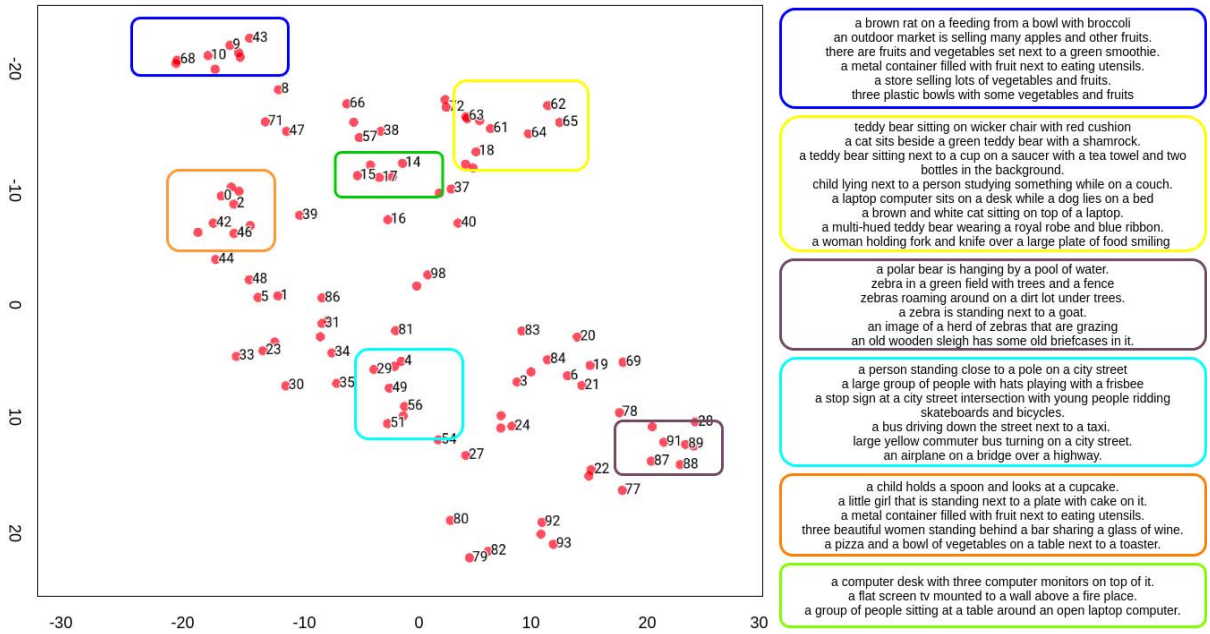


Fig. 4. **Embedding plots :** The figure shows 2-D t-SNE Embedding plots of test set Scene graphs from COCO-Stuff dataset, we identified some clustered data points and examined the textual descriptions of those scene graphs, a cluster is enclosed in a rectangular box with its descriptions on the right inside their respective colored boxes. Sentences in the same cluster demonstrate semantic correlation amongst them.

B. Retrieval Task Results

To evaluate the efficacy of our technique we use our scene graph embedding on a retrieval task and calculate two metrics, Med-r, that is the rank of the true image/the highest rank true image and recall at rank k for k= (1,5,10) .the image is classified as a true positive if the similarity of scene graph exceeds 80% according to the pairwise similarity heuristics.

The Med-r metric shows that through our model, the best matching scene graph is retrieved at the 8th rank for COCO stuff dataset and at 12th rank for VRD dataset. Considering that VRD consists of more relationship classes leading to more complex scene graphs, this was an expected trend. As shown in Table I recall at rank(1,5,10) for both COCO and VRD dataset proves that the embedding generated was able to represents the scene graph information.

TABLE I
RETRIEVAL TASK RESULTS

	COCO		VRD	
	Rand	**Our Model**	**Rand**	**Our Model**
Med-r	377	8	420	12
R@1	0	0.148	0	0.134
R@5	0.008	0.352	0.007	0.321
R@10	0.034	0.488	0.0027	0.453

VI. CONCLUSION

We propose a novel architecture to embed semantic information of the scene graph into a low dimensional embedding. We use a graph embedding network to embed the structural information and a linguistic embedding network to embed the language description of nodes and edges. The high dimensional representation is then converted to a low-dimensional latent space representation using reconstruction based learning. The embedding is able to embed the information of the scene graph, as well as the affinity of the embedding is maintained using a combination of global and local loss. The visualization of the model shows that it was able to achieve good embedding separability with the level of abstraction getting higher with distance. We applied our embedding on a retrieval task and calculated the metrics Med-r and Recall, which further proves that the embedding generated are able to effectively represent the scene graphs. In future, other graph embedding and linguistic embedding techniques can be applied, or loss functions can be improved for specific tasks like clustering.

REFERENCES

[1] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[2] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5213–5222, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:19:06 UTC from IEEE Xplore. Restrictions apply.

[3] Lizhao Gao, Bo Wang, and Wenmin Wang. Image captioning with scene-graph based semantic concepts. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pages 225–229, 2018.

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[5] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[7] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, 2012.

[8] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[10] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 423–429, 2004.

[11] Guodong Zhou, Min Zhang, DongHong Ji, and QiaoMing Zhu. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 728–736, 2007.

[12] Subarna Tripathi, Anahita Bhiwandiwalla, Alexei Bastidas, and Hanlin Tang. Heuristics for image generation from scene graphs. 2019.

[13] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. Interactive image generation using scene graphs. *arXiv preprint arXiv:1905.03743*, 2019.

[14] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–179, 2020.

[15] Chiranjib Sur. Tpsgtr: Neural-symbolic tensor product scene-graph-triplet representation for image captioning. *arXiv preprint arXiv:1911.10115*, 2019.

[16] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953*, 2019.

[17] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[18] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[19] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1261–1270, 2017.

[20] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael Bernstein, and Fei-Fei Li. Visual relationships as functions: Enabling few-shot scene graph prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[21] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9895–9904, 2019.

[22] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. Learning canonical representations for scene graph to image generation. In *European Conference on Computer Vision*, pages 210–227. Springer, 2020.

[23] Brigit Schroeder, Subarna Tripathi, and Hanlin Tang. Triplet-aware scene graph embeddings. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[24] Eugene Belilovsky, Matthew Blaschko, Jamie Kiros, Raquel Urtasun, and Richard Zemel. Joint embeddings of scene graphs and images. 2017.

[25] Hai Wan, Yonghao Luo, Bo Peng, and Wei-Shi Zheng. Representation learning for scene graph completion via jointly structural and visual embedding. In *IJCAI*, pages 949–956, 2018.

[26] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234, 2016.

[27] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:19:06 UTC from IEEE Xplore. Restrictions apply.

arXiv:2104.08518v1 [math.CV] 17 Apr 2021

A NOVEL CLASS OF STARLIKE FUNCTIONS

S. SIVAPRASAD KUMAR AND SHAGUN BANGA

Abstract. In the past several subclasses of starlike functions are defined involving real part and modulus of certain expressions of functions under study, combined by way of an inequality. In the similar fashion, we introduce a new class $\mathcal{S}_q^*(\alpha)$, consisting of normalized analytic univalent functions f in the open unit disk $\mathbb{D}$, satisfying

$$\operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right) \geq \left|1+\frac{zf''(z)}{f'(z)}-\frac{zf'(z)}{f(z)}-\alpha\right| \quad (0\leq\alpha<1).$$

Evidently, $\mathcal{S}_q^*(\alpha)\subset\mathcal{S}^*$, the class of starlike functions. We first establish $\mathcal{S}_q^*(\alpha)\subset\mathcal{S}^*(q_\alpha)$, the class of analytic functions f satisfying $zf'(z)/f(z)\prec q_\alpha(z)$, where q_α is an extremal function. Some necessary and sufficient conditions for functions belonging to these classes are obtained in addition to the inclusion and radius problems. Further, we estimate logarithmic coefficients, inverse coefficients and Fekete-Szegö functional bounds for functions in $\mathcal{S}^*(q_\alpha)$.

1. Introduction

Let $\mathcal{A}$ be the class of analytic functions $f(z)=z+a_2z^2+a_3z^3+\cdots$, defined in a unit disk $\mathbb{D}=\{z\in\mathbb{C}:|z|<1\}$ and the subclass $\mathcal{S}$ of $\mathcal{A}$ consists of univalent functions. A function f is said to be starlike if $k\mu\in f(\mathbb{D})$ for $\mu\in f(\mathbb{D})$ and $k\in[0,1]$. The class of starlike functions of order α $(0\leq\alpha<1)$ is defined by

$$\mathcal{S}^*(\alpha):=\left\{f\in\mathcal{S}:\operatorname{Re}\frac{zf'(z)}{f(z)}\geq\alpha\right\}.$$

For $\alpha=0$, $\mathcal{S}^*(0)=:\mathcal{S}^*$, the class of starlike functions. A function f is said to be convex if $k\mu_1+(1-k)\mu_2\in f(\mathbb{D})$ for μ_1, $\mu_2\in f(\mathbb{D})$ and $k\in[0,1]$. The class of convex functions of order α $(0\leq\alpha<1)$ is defined by

$$\mathcal{K}(\alpha):=\left\{f\in\mathcal{S}:1+\frac{zf''(z)}{f'(z)}\geq\alpha\right\}.$$

Note that $\mathcal{K}(0)=:\mathcal{K}$, the class of convex functions. Robertson [12] introduced the classes $\mathcal{S}^*(\alpha)$ and $\mathcal{K}(\alpha)$ and functions in either of them fail to be univalent for $\alpha<0$. Moreover, there is a two way bridge between these classes: $f\in\mathcal{K}(\alpha)$ if and only if $zf'(z)\in\mathcal{S}^*(\alpha)$. Another interesting class $\mathcal{M}(\beta)\subset\mathcal{A}$, studied by Uralegaddi [17] is characterized by $\operatorname{Re}(zf'(z)/f(z))<\beta$ $(\beta>1)$. Carathéodory class, denoted by $\mathcal{P}$, consists of the functions $p(z)=1+p_1z+p_2z^2+\cdots$, having positive real part in $\mathbb{D}$. Let $\mathcal{Q}(\alpha):=\{f\in\mathcal{A}:\operatorname{Re}(f(z)/z)\geq\alpha\}$ $(0\leq\alpha<1)$. Recall $\mathcal{S}^*(1/2)\subset\mathcal{Q}(1/2)$ from [2, p. 73] and the constant $1/2$ is the best possible. If there exist a Schwarz function $v(z)$ such that $f(z)=g(v(z))$ for two analytic functions f and g, then we say f is subordinate to g, written as $f\prec g$. Specifically, for the case when g is univalent, we say $f\prec g$ if and only if $f(\mathbb{D})\subset g(\mathbb{D})$ and $f(0)=g(0)$. In the recent past many new subclasses of $\mathcal{S}^*$ were introduced and studied extensively (see [1, 4, 5, 8, 11]). Motivated by these, we introduce another interesting subclass of $\mathcal{S}^*$:

2010 *Mathematics Subject Classification.* 30C45, 30C50.

Key words and phrases. starlike function, subordination, convolution, radius problem, coefficient estimate.

Definition 1.1. For $f \in \mathcal{S}$, we define below a class $\mathcal{S}_q^*(\alpha)$ if f satisfies:

$$\mathrm{Re}\left(\frac{zf'(z)}{f(z)}\right) \geq \left|1+\frac{zf''(z)}{f'(z)}-\frac{zf'(z)}{f(z)}-\alpha\right| \quad (0 \leq \alpha < 1). \tag{1.1}$$

We observe that the class $\mathcal{S}_q^*(\alpha) \subseteq \mathcal{S}^*$. The identity function $f(z) = z$ satisfies the inequality (1.1) for all $0 \leq \alpha < 1$, hence $\mathcal{S}_q^*(\alpha) \neq \emptyset$. For $\alpha = 0$, we set $\mathcal{S}_q^*(0) =: \mathcal{S}_q^*$. We also obtain $\mathcal{S}_q^*(\alpha)$ as a subclass of another fascinating class of starlike functions, which is introduced in the following section. The paper is structured as follows: In the next section, we introduce another class $\mathcal{S}^*(q_\alpha)$ related to $\mathcal{S}_q^*(\alpha)$ by means of a subordination. Recently, authors have defined new classes using the subordination concept (see [15, 16]). In the sequel we discuss many sufficient and necessary conditions for functions in $\mathcal{S}^*(q_\alpha)$. We also obtain certain inclusion and radius results pertaining to these classes. In the last section, we find the sharp estimate of logarithmic coefficients, Fekete-Szegö functional bounds and inverse coefficient bounds for functions in $\mathcal{S}^*(q_\alpha)$. For work in this direction (see [3, 6])

2. Results related to $\mathcal{S}_q^*(\alpha)$

We begin this section by showing the existence of some analytic function belonging to $\mathcal{S}_q^*(\alpha)$.

Theorem 2.1. *Let $f_\gamma(z) = z + \gamma z^2$. If $|\gamma| \leq r_0$, where r_0 is the smallest such $r < 1$ satisfying the equation:*

$$\begin{aligned}&\alpha^2(2r-1)^2(r-1)^3 - 33r^2 + 57r^3 - 48r^4 + 16r^5 + 2\alpha(r-1)^2r(2r-1)\\ &+ (2r-1)(r-1)(\alpha + r - 3\alpha r + 2\alpha r^2)(2-4r) + 9r = 1,\end{aligned}$$

then $f_\gamma \in \mathcal{S}_q^*(\alpha)$.

Proof. We know $f_\gamma \in \mathcal{S}$ whenever $|\gamma| \leq 1/2$, thus we choose $r_0 \in [0, 1/2]$. A simple calculation yields

$$\frac{zf_\gamma'(z)}{f_\gamma(z)} = \left(\frac{1+2\gamma z}{1+\gamma z}\right) \text{ and } 1 + \frac{zf_\gamma''(z)}{f_\gamma'(z)} - \frac{zf_\gamma'(z)}{f_\gamma(z)} = \frac{\gamma z}{(1+\gamma z)(1+2\gamma z)}.$$

Let $\gamma z = re^{i\theta}$, $0 \leq r \leq 1/2$ and $\theta \in [-\pi, \pi]$. Now, to prove the result, we show the following inequality holds:

$$\mathrm{Re}\left(\frac{1+2re^{i\theta}}{1+re^{i\theta}}\right) \geq \left|\frac{re^{i\theta}}{(1+re^{i\theta})(1+2re^{i\theta})} - \alpha\right|,$$

which is equivalent to show

$$\begin{aligned}&-\sqrt{\frac{\alpha^2 + r^2 - 6\alpha r^2 + 9\alpha^2 r^2 + 4\alpha^2 r^4 + 2\alpha(-1+3\alpha)r(1+2r^2)x + 4\alpha^2 r^2(2x^2-1)}{1 + 9r^2 + 4r^4 + 6(r+2r^3)x + 4r^2(2x^2-1)}}\\ &+ \frac{1 + 3rx + 2r^2}{1 + r^2 + 2rx} =: h(\alpha, r, x) \geq 0,\end{aligned}$$

where $x := \cos\theta$. The function h is increasing with respect to $x \in [-1, 1]$, thus it suffices to show

$$\begin{aligned}&\frac{2r-1}{r-1} - \sqrt{\frac{\alpha^2 + r^2 - 6\alpha r^2 + 13\alpha^2 r^2 + 4\alpha^2 r^4 - 2\alpha(-1+3\alpha)r(1+2r^2)}{1 + 13r^2 + 4r^4 - 6(r+2r^3)}}\\ &=: h(\alpha, r) = h(\alpha, r, -1) \geq 0.\end{aligned} \tag{2.1}$$

We observe that $h(\alpha, 0) > 0$ and $h(\alpha, 1/2) < 0$. Since h is a continuous function of r, thus there must exist $r_0 \in (0, 1/2)$, which is a smallest positive root of $h(\alpha, r) = 0$ by Intermediate value property. Consequently (2.1) holds for $0 < r \leq r_0$. Therefore, we conclude that $f_\gamma \in \mathcal{S}_q^*(\alpha)$ whenever $|\gamma| \leq r_0$. ∎

Theorem 2.2. *Let $f \in \mathcal{S}_q^*(\alpha)$. Then*

$$\frac{zf'(z)}{f(z)} \prec \frac{(1-2\alpha)z}{(1-z)(1-(1-z)^{1-2\alpha})} \quad (0 \leq \alpha < 1). \tag{2.2}$$

Proof. Let us define

$$p(z) = \frac{zf'(z)}{f(z)},$$

then equation (1.1) reduces to

$$\begin{aligned} \operatorname{Re} p(z) &\geq \left| \frac{zp'(z)}{p(z)} - \alpha \right| \\ &\geq \operatorname{Re}\left(\alpha - \frac{zp'(z)}{p(z)}\right), \end{aligned}$$

which yields

$$\operatorname{Re}\left(p(z) + \frac{zp'(z)}{p(z)}\right) \geq \alpha.$$

Using the subordination concept, we write the above inequality as follows:

$$p(z) + \frac{zp'(z)}{p(z)} \prec \frac{1+(1-2\alpha)z}{1-z}. \tag{2.3}$$

From [9, Theorem 3.3 d], we have:

$$q(z) + \frac{zq'(z)}{q(z)} = \frac{1+(1-2\alpha)z}{1-z}, \tag{2.4}$$

where q is the best dominant of the subordination (2.3). On solving equation (2.4), we obtain

$$q(z) = \left(\frac{z}{(1-z)^{2(1-\alpha)}}\right)\left(\int_0^z (1-t)^{-2(1-\alpha)} dt\right)^{-1}.$$

Hence the proof follows. ∎

Let us consider a function $f \in \mathcal{A}$, $c_n \geq 0$ and $b_n \in \mathbb{R}$ such that

$$\frac{z}{f'(z)} = \sum_{n=1}^{\infty} c_n z^n \text{ and } \frac{z}{f(z)} = 1 + \sum_{n=1}^{\infty} b_n z^n. \tag{2.5}$$

Theorem 2.3. *Let $f \in \mathcal{S}$ of the form (2.5) belongs to $\mathcal{S}_q^*(\alpha)$. Then*

$$\sum_{n=2}^{\infty} (n+\alpha-2)c_n \leq (1-\alpha)c_1.$$

Proof. A simple computation yields

$$z\left(\frac{z}{f(z)}\right)' = \frac{z}{f(z)} - \left(\frac{z}{f(z)}\right)^2 f'(z) \text{ and } z\left(\frac{z}{f'(z)}\right)' = \frac{z}{f'(z)} - \left(\frac{z}{f'(z)}\right)^2 f''(z).$$

Also, $f \in \mathcal{S}_q^*(\alpha)$ of the form (2.5) gives

$$\left|1 + \frac{zf''(z)}{f'(z)} - \frac{zf'(z)}{f(z)} - \alpha\right| \leq \operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right)$$

if and only if

$$\left|1-\frac{z\left(\frac{z}{f'(z)}\right)'}{\frac{z}{f'(z)}}+\frac{z\left(\frac{z}{f(z)}\right)'}{\frac{z}{f(z)}}-\alpha\right|\leq \operatorname{Re}\frac{\frac{z}{f(z)}-z\left(\frac{z}{f(z)}\right)'}{\frac{z}{f(z)}}$$

if and only if

$$\left|-1+\frac{\sum_{n=1}^{\infty}nc_nz^n}{\sum_{n=1}^{\infty}c_nz^n}-\frac{\sum_{n=1}^{\infty}nb_nz^n}{1+\sum_{n=1}^{\infty}b_nz^n}+\alpha\right|\leq \operatorname{Re}\left(1-\frac{\sum_{n=1}^{\infty}nb_nz^n}{1+\sum_{n=1}^{\infty}b_nz^n}\right).$$

Now if $z\in\mathbb{D}$ is real and tends to 1^- through reals, then from the above inequality we deduce

$$\frac{\sum_{n=1}^{\infty}nc_n}{\sum_{n=1}^{\infty}c_n}\leq 2-\alpha.$$

Therefore the results follows now. ∎

With the advent of Theorem 2.2, we define another interesting subclass of $\mathcal{S}^*$ as follows:

Definition 2.4. Let $\mathcal{S}^*(q_\alpha)$ denote the class of analytic functions $f\in\mathcal{S}$, satisfying

$$\frac{zf'(z)}{f(z)}\prec\frac{(1-2\alpha)z}{(1-z)(1-(1-z)^{1-2\alpha})}=:q_\alpha(z)\quad(z\in\mathbb{D},\ 0\leq\alpha<1). \tag{2.6}$$

We infer from Theorem 2.2, $\mathcal{S}_q^*(\alpha)\subset\mathcal{S}^*(q_\alpha)$, so is non-empty. This class generalizes many subclasses of $\mathcal{S}^*$, such as for $\alpha=0$, it reduces to $\mathcal{S}^*(1/2)$.

3. About $\mathcal{S}^*(q_\alpha)$

A function $f\in\mathcal{S}^*(q_\alpha)$ if and only if there exists an analytic function s, $s(z)\prec q_\alpha(z):=\frac{(1-2\alpha)z}{(1-z)(1-(1-z)^{1-2\alpha})}$ such that

$$f(z)=z\exp\int_0^z\frac{s(t)-1}{t}dt.$$

Let $g(z)=z+\sum_{n=2}^{\infty}g_nz^n$ and $h(z)=z+\sum_{n=2}^{\infty}h_nz^n$ be two analytic functions in $\mathbb{D}$, then convolution of g and h is defined as follows

$$(g*h)(z)=z+\sum_{n=2}^{\infty}g_nh_nz^n.$$

The following result gives the necessary and sufficient condition in terms of convolution for the functions in the class $\mathcal{S}^*(q_\alpha)$.

Theorem 3.1. *A function $f\in\mathcal{S}$ belongs to $\mathcal{S}^*(q_\alpha)$ if and only if it satisfies*

$$\frac{1}{z}\left(f(z)*\frac{z-\lambda z^2}{(1-z)^2}\right)\neq 0\quad(z\in\mathbb{D}), \tag{3.1}$$

where

$$\lambda:=\lambda(\alpha,\theta)=\frac{1-2\alpha}{(1-2\alpha)-(e^{-i\theta}-1)(1-(1-e^{i\theta})^{1-2\alpha})},\quad(\theta\in[-\pi,\pi]).$$

Proof. From Definition 2.4, we have $f\in\mathcal{S}^*(q_\alpha)$ if and only if it satisfies the subordination (2.6), equivalently

$$\frac{zf'(z)}{f(z)}\neq\frac{(1-2\alpha)e^{i\theta}}{(1-e^{i\theta})(1-(1-e^{i\theta})^{1-2\alpha})}\quad(z\in\mathbb{D},\theta\in[-\pi,\pi]),$$

further which yields

$$f'(z) \neq \left(\frac{(1-2\alpha)e^{i\theta}}{(1-e^{i\theta})(1-(1-e^{i\theta})^{1-2\alpha})}\right)\frac{f(z)}{z}.$$

Also, it can be expressed as

$$\frac{1}{z}\left(zf'(z) - \frac{(1-2\alpha)e^{i\theta}}{(1-e^{i\theta})(1-(1-e^{i\theta})^{1-2\alpha})}f(z)\right) \neq 0. \tag{3.2}$$

As we recall

$$zf'(z) = f(z) * \frac{z}{(1-z)^2} \text{ and } f(z) = f(z) * \frac{z}{1-z},$$

(3.2) becomes

$$\frac{1}{z}\left(f(z) * \frac{z - \frac{(1-2\alpha)e^{i\theta}}{(1-e^{i\theta})(1-(1-e^{i\theta})^{1-2\alpha})}(z-z^2)}{(1-z)^2}\right) \neq 0,$$

which upon simplification reduces to (3.1). Hence the result. ∎

Now, using the convolution properties, we have

$$\frac{1}{z}\left(f(z) * \frac{z-\lambda z^2}{(1-z)^2}\right) = \frac{1}{z}\left(zf'(z) - \lambda(zf'(z) - f(z))\right).$$

Substituting the above equation in (3.1) along with the Taylor series expansion of $f \in \mathcal{A}$, we deduce the following result:

Corollary 3.2. *A function $f \in \mathcal{S}$ belongs to $\mathcal{S}^*(q_\alpha)$ if and only if it satisfies*

$$\sum_{n=2}^{\infty} \frac{(e^{-i\theta}-1)(1-(1-e^{i\theta})^{1-2\alpha})n - (1-2\alpha)}{(1-2\alpha) - (e^{-i\theta}-1)(1-(1-e^{i\theta})^{1-2\alpha})} a_n z^{n-1} \neq 1.$$

Consequently, the next result follows.

Corollary 3.3. *Let $f \in \mathcal{S}$ satisfies*

$$\sum_{n=2}^{\infty} |\rho(\alpha,\theta)||a_n| := \sum_{n=2}^{\infty} \left|\frac{(e^{-i\theta}-1)(1-(1-e^{i\theta})^{1-2\alpha})n - (1-2\alpha)}{(1-2\alpha) - (e^{-i\theta}-1)(1-(1-e^{i\theta})^{1-2\alpha})}\right| |a_n| < 1, \tag{3.3}$$

then $f \in \mathcal{S}^*(q_\alpha)$.

Corollary 3.4. *The class $\mathcal{S}^*(q_\alpha)$ is convex.*

Proof. Let f_1, $f_2 \in \mathcal{S}^*(q_\alpha)$ having Taylor series expansion as

$$f_i(z) = z + \sum_{n=2}^{\infty} a_{i,n} z^n, \quad i = 1, 2$$

and $\mu \in [0,1]$. The result follows if we show

$$\mu f_1(z) + (1-\mu) f_2(z) = z + \sum_{n=2}^{\infty} (\mu a_{1,n} + (1-\mu)a_{2,n}) z^n \in \mathcal{S}_q^*(\alpha).$$

For this, let us contemplate the expression on the left hand side of the equality given in (3.3)

$$\begin{aligned}\sum_{n=2}^{\infty} |\rho(\alpha,\theta)|\,|\mu a_{1,n} + (1-\mu)a_{2,n}| &\leq \mu \sum_{n=2}^{\infty} |\rho(\alpha,\theta)|\,|a_{1,n}| + (1-\mu)\sum_{n=2}^{\infty} |\rho(\alpha,\theta)|\,|a_{2,n}| \\ &< \mu + (1-\mu) = 1.\end{aligned}$$

The last inequality follows from (3.3), concluding the proof. ∎

The next two results provide us with the sufficient conditions for functions to belong to $\mathcal{S}^*(q_\alpha)$.

Theorem 3.5. *If $f \in \mathcal{A}$ and*

$$\left|\frac{zf'(z)}{f(z)}\right| < \frac{|1-2\alpha|}{2(1+2^{1-2\alpha})} \quad (0 \leq \alpha < 1),$$

then $f \in \mathcal{S}^*(q_\alpha)$.

Proof. As $zf'(z)/f(z)$ lies in a disk with center 0 and radius $\frac{|1-2\alpha|}{2(1+2^{1-2\alpha})}$, then from Definition 2.4, it suffices to show

$$|q_\alpha(e^{i\theta})| \geq \frac{|1-2\alpha|}{2(1+2^{1-2\alpha})} \quad (0 < \theta < 2\pi), \tag{3.4}$$

where $q_\alpha(z)$ is given in (2.6). Consider

$$\begin{aligned}
|q_\alpha(e^{i\theta})| &= \frac{|1-2\alpha||e^{i\theta}|}{|1-e^{i\theta}||1-(1-e^{i\theta})^{1-2\alpha}|} \\
&\geq \frac{|1-2\alpha|}{|1-e^{i\theta}|(1+|1-e^{i\theta}|^{1-2\alpha})} \\
&= \frac{|1-2\alpha|}{\sqrt{2(1-\cos\theta)}\left(1+\left(\sqrt{2(1-\cos\theta)}\right)^{1-2\alpha}\right)} \\
&= \frac{|1-2\alpha|}{|2\sin(\theta/2)|+|2\sin(\theta/2)|^{2(1-\alpha)}}.
\end{aligned}$$

The function $|\sin(\theta/2)|$ attains maximum value 1 at $\theta = \pi$, which proves the inequality (3.4). ∎

Theorem 3.6. *Let $h(z) = z + az^n$ $(n \geq 2)$. Then $h \in \mathcal{S}^*(q_\alpha)$ $(\alpha \neq 1/2)$ if and only if*

$$|a| \leq \frac{3-2^{2(1-\alpha)}-2\alpha}{2n-2^{2(1-\alpha)}n+1-2\alpha} \quad (n \geq 2). \tag{3.5}$$

Proof. For function $h(z) = z + az^n$, we get

$$H(z) := \frac{zh'(z)}{h(z)} = \frac{1+naz^{n-1}}{1+az^{n-1}} \quad (n \geq 2).$$

The function H maps $\mathbb{D}$ onto a disk symmetric with respect to real axis. Therefore, using the concept of subordination, we deduce that $h \in \mathcal{S}^*(q_\alpha)$ if and only if

$$\frac{2\alpha-1}{2(1-2^{1-2\alpha})} \leq \frac{1-n|a|}{1-|a|},$$

which upon simplification yields (3.5). ∎

Here below, we obtain the bounds on $|f(z)|$ and $|f'(z)|$, respectively for functions in $\mathcal{S}^*(q_\alpha)$.

Theorem 3.7. *Let $f \in \mathcal{S}^*(q_\alpha)$. Then for $|z| = r$, we have*

$$\frac{r(1-|1-2\gamma|r)}{1+r} \leq |f(z)| \leq \frac{r(1+|1-2\gamma|r)}{1-r}, \tag{3.6}$$

where $\gamma = \gamma(\alpha)$, given in (4.1).

Proof. As $f \in \mathcal{S}^*(q_\alpha)$, from Theorem 4.5, we have

$$\frac{f(z)}{z} \prec \frac{1+(1-2\gamma)z}{1-z} \quad (0 \leq \gamma < 1),$$

from which, we infer that there exists $\upsilon(z)$ such that

$$\frac{f(z)}{z} = \frac{1+(1-2\gamma)\upsilon(z)}{1-\upsilon(z)}.$$

Let $\upsilon(z) = |\upsilon(z)|e^{i\theta}$, $-\pi \leq \theta \leq \pi$. Taking square of the modulus in the above equation, we deduce

$$\left|\frac{f(z)}{z}\right|^2 = \frac{1+(1-2\gamma)^2|\upsilon(z)|^2+2(1-2\gamma)|\upsilon(z)|\cos\theta}{1+|\upsilon(z)|^2-2|\upsilon(z)|\cos\theta}.$$

The left and the right inequalities follow respectively by minimizing and maximizing the above equation.

$$\frac{1-|1-2\gamma||\upsilon(z)|}{1+|\upsilon(z)|} \leq \left|\frac{f(z)}{z}\right| \leq \frac{1+|1-2\gamma||\upsilon(z)|}{1-|\upsilon(z)|}.$$

Using the fact $|\upsilon(z)| \leq |z|$, we arrive at the result. ∎

Theorem 3.8. *Let $f \in \mathcal{S}^*(q_\alpha)$. Then for $|z| = r$, we have*

$$|f'(z)| \leq \frac{|1-2\alpha|r(1+|1-2\gamma|r)}{(1-r)^2(1-(1+(-1)^{1-2\alpha}r)^{1-2\alpha})},$$

where $\gamma = \gamma(\alpha)$, given in (4.1).

Proof. As $f \in \mathcal{S}^*(q_\alpha)$, we have

$$\frac{zf'(z)}{f(z)} = \frac{(1-2\alpha)\upsilon(z)}{(1-\upsilon(z))(1-(1-\upsilon(z))^{1-2\alpha})},$$

for some Schwarz function $\upsilon(z)$. Taking modulus over above equation, we get

$$|f'(z)| = \left|\frac{f(z)}{z}\right|\left(\frac{|1-2\alpha||\upsilon(z)|}{|1-\upsilon(z)||1-(1-\upsilon(z))^{1-2\alpha}|}\right).$$

Now using the right part of the inequality (3.6) and applying triangle inequality, the above equation becomes

$$|f'(z)| \leq \left(\frac{1+|1-2\gamma||z|}{(1-|z|)}\right)\left(\frac{|1-2\alpha||\upsilon(z)|}{(1-|\upsilon(z)|)(1-(1+(-1)^{1-2\alpha}|\upsilon(z)|)^{1-2\alpha})}\right).$$

Now using the fact $|\upsilon(z)| \leq |z|$, the result follows. ∎

4. Inclusion Relation

We enlist below some inclusion relations of $\mathcal{S}_q^*(\alpha)$.

Theorem 4.1. *Let $0 \leq \alpha < 1$, then $\mathcal{S}_q^*(\alpha) \subseteq \mathcal{K}(\alpha)$.*

Proof. Let $f \in \mathcal{S}_q^*(\alpha)$, then inequality (1.1) yields

$$\operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right) \geq -\operatorname{Re}\left(1+\frac{zf''(z)}{f'(z)}-\frac{zf'(z)}{f(z)}-\alpha\right),$$

which upon a straightforward calculation yields

$$\operatorname{Re}\left(1+\frac{zf''(z)}{f'(z)}\right) \geq \alpha.$$

This completes the proof. ∎

The converse of the above result need not be true as we find a function $f(z) := \frac{(1-z)^{2\alpha-1}-1}{1-2\alpha}$, which clearly belongs to $\mathcal{K}(\alpha)$ but does not belong to $\mathcal{S}_q^*(\alpha)$ as it does not satisfy the inequality (1.1) at $z = 1/2 - i/4 \in \mathbb{D}$ for all α.

Silverman [14] introduced the class

$$\mathcal{G}_b := \left\{ f \in \mathcal{A} : \left| \frac{1 + zf''(z)/f'(z)}{zf'(z)/f(z)} - 1 \right| < b,\ 0 < b \leq 1 \right\}.$$

In the following result we show the relation between the classes $\mathcal{G}_b$ and $\mathcal{S}_q^*$.

Corollary 4.2. $\mathcal{S}_q^* \subseteq \mathcal{G}_1$.

Proof. Since $\mathcal{S}_q^* := \mathcal{S}_q^*(0)$, we have $\alpha = 0$. Let $f \in \mathcal{S}_q^*$, then inequality (1.1) becomes

$$\operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right) \geq \left| 1 + \frac{zf''(z)}{f'(z)} - \frac{zf'(z)}{f(z)} \right|,$$

which implies

$$\left|\frac{zf'(z)}{f(z)}\right| \geq \left| 1 + \frac{zf''(z)}{f'(z)} - \frac{zf'(z)}{f(z)} \right|,$$

therefore, we have

$$\left| \left(\frac{1 + zf''(z)/f'(z)}{zf'(z)/f(z)} \right) - 1 \right| \leq 1.$$

Hence the result. ∎

Mocanu [10] studied the class comprising of the linear combination of analytic representations of convex and starlike functions and defined it as $\rho-$ convex class as follows:

$$\operatorname{Re}(\rho(1 + zf''(z)/f'(z)) + (1-\rho)(zf'(z)/f(z))) \geq 0 \quad (\rho \in \mathbb{R}).$$

Corollary 4.3. *Let $f \in \mathcal{S}_q^*$, then f is -1-convex function.*

Proof. Let $f \in \mathcal{S}_q^*$, then we have

$$\operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right) \geq \operatorname{Re}\left(1 + \frac{zf''(z)}{f'(z)} - \frac{zf'(z)}{f(z)}\right),$$
$$\operatorname{Re}\left(2\frac{zf'(z)}{f(z)} - \left(1 + \frac{zf''(z)}{f'(z)}\right)\right) \geq 0.$$

This completes the proof. ∎

For $f \in \mathcal{S}^*(q_\alpha)$, we have the following inclusion results.

Theorem 4.4. *Let $0 \leq \gamma \leq \frac{2\alpha-1}{2(1-2^{1-2\alpha})}$, then $\mathcal{S}_q^*(\alpha) \subset \mathcal{S}^*(q_\alpha) \subset \mathcal{S}^*(\gamma)$.*

Proof. Since we know $\mathcal{S}_q^*(\alpha) \subset \mathcal{S}^*(q_\alpha)$, it is now left to show the latter inclusion. For this, let $f \in \mathcal{S}^*(q_\alpha)$. Now subordination (2.2) yields

$$\operatorname{Re}\frac{zf'(z)}{f(z)} > \frac{2\alpha - 1}{2(1 - 2^{1-2\alpha})} \geq \gamma.$$

Hence the result. ∎

Theorem 4.5. *Let $f \in \mathcal{S}^*(q_\alpha)$. Then we have*

$$\operatorname{Re}\frac{f(z)}{z} \geq \gamma(\alpha) := \frac{3 - 2\alpha - 2^{2(1-\alpha)}}{4 - 2\alpha - 3.2^{1-2\alpha}}, \tag{4.1}$$

which also implies $\mathcal{S}_q^*(\alpha) \subset \mathcal{S}^*(q_\alpha) \subset \mathcal{Q}(\gamma(\alpha))$.

Proof. For brevity, let us denote $\gamma(\alpha) =: \gamma$. We observe that $0 \leq \gamma \leq 1$ for $0 \leq \alpha < 1$. Let us define an analytic function p with $p(0) = 1$ as

$$p(z) = \frac{1}{1-\gamma}\left(\frac{f(z)}{z} - \gamma\right),$$

which upon simplification gives

$$\frac{zf'(z)}{f(z)} = 1 + \frac{(1-\gamma)zp'(z)}{(1-\gamma)p(z)+\gamma} =: \xi(p(z), zp'(z)),$$

where

$$\xi(a,b) := 1 + \frac{(1-\gamma)b}{(1-\gamma)a+\gamma}.$$

Let $\alpha' := \frac{2\alpha-1}{2(1-2^{1-2\alpha})}$ and $f \in \mathcal{S}^*(q_\alpha)$ then from Theorem 4.4, we have

$$\xi(p(z), zp'(z)) \subset \{w \in \mathbb{C} : \operatorname{Re} w > \alpha'\} =: \Omega_{\alpha'}.$$

Let $\lambda, \tau \in \mathbb{R}$ be such that $\lambda \leq \frac{-1}{2}(1+\tau^2)$. Then we have

$$\begin{aligned}
\operatorname{Re}\{\xi(i\tau,\lambda)\} &= \operatorname{Re}\left(1 + \frac{\lambda(1-\gamma)}{(1-\gamma)i\tau+\gamma}\right) \\
&= 1 + \frac{\lambda\gamma(1-\gamma)}{(1-\gamma)^2\tau^2+\gamma^2} \\
&\leq 1 - \left(\frac{\gamma(1-\gamma)}{2}\right)\left(\frac{(1+\tau^2)}{(1-\gamma)^2\tau^2+\gamma^2}\right) = 1 - \frac{\gamma(1-\gamma)}{2}h(\tau^2), \qquad (4.2)
\end{aligned}$$

where $h(\tau^2) := \dfrac{(1+\tau^2)}{(1-\gamma)^2\tau^2+\gamma^2}$, which is a decreasing function of τ^2 for all γ, we deduce the following

$$\frac{1}{(1-\gamma)^2} \leq h(\tau^2) \leq \frac{1}{\gamma^2}.$$

Using the above inequality in (4.2), we obtain

$$\operatorname{Re}\{\xi(i\tau,\lambda)\} \leq 1 - \frac{\gamma(1-\gamma)}{2(1-\gamma)^2} = \frac{2-3\gamma}{2(1-\gamma)} = \alpha'.$$

This clearly shows that $\operatorname{Re}\{\xi(i\tau,\lambda)\} \notin \Omega_{\alpha'}$. Now, from [9, Theorem 2.3i., p.35], we conclude $\operatorname{Re} p(z) > 0$ for $z \in \mathbb{D}$. Therefore we obtain $\mathcal{S}^*(q_\alpha) \subset \mathcal{Q}(\gamma(\alpha))$. In addition, the fact $\mathcal{S}_q^*(\alpha) \subset \mathcal{S}^*(q_\alpha)$ proves the result . ∎

5. Radius and Coefficient Problems

We commence this section by the following radius problems for $\mathcal{S}^*(q_\alpha)$.

Theorem 5.1. *Let $f \in \mathcal{S}^*(q_\alpha)$. Then the following holds:*

(i) *f is starlike of order γ in $|z| < \tilde{r}$ whenever $\frac{2\alpha-1}{2(1-2^{1-2\alpha})} < \gamma < 1$, where $\tilde{r}$ is the smallest such $r < 1$ satisfying the equation*

$$(2\alpha-1)r - \gamma(1+r)(1-(1+r)^{1-2\alpha}) = 0.$$

(ii) *For $\alpha < 1/2$, $f \in \mathcal{M}(\beta)$ in $|z| < r_0$ whenever $\beta > 1$, where r_0 is the smallest such $r < 1$ satisfying the equation*

$$(1-2\alpha)r - \beta(1-r)(1-(1-r)^{1-2\alpha}) = 0.$$

Proof. As $f \in \mathcal{S}^*(q_\alpha)$, then f satisfies the subordination (2.6).

(i) Thus we have

$$\operatorname{Re}\frac{zf'(z)}{f(z)} \geq \frac{(2\alpha-1)r}{(1+r)(1-(1+r)^{1-2\alpha})} =: \rho(r,\alpha), \quad |z| = r < 1.$$

We find $r \in [0,1)$ such that $\rho(r,\alpha) \geq \gamma$ for given α and γ. Let $\tilde{\rho}(r,\alpha,\gamma) := (2\alpha-1)r/((1+r)(1-(1+r)^{1-2\alpha})) - \gamma$. Now, we observe $\tilde{\rho}(0,\alpha,\gamma) = 1-\gamma > 0$ and $\tilde{\rho}(1,\alpha,\gamma) = \frac{2\alpha-1}{2(1-2^{1-2\alpha})} - \gamma < 0$ for all γ in the given range. So, there must exist $\tilde{r}$ such that $\tilde{\rho}(r,\alpha,\gamma) \geq 0$ for all $r \in (0,\tilde{r}]$, where $\tilde{r}$ is as described in the hypothesis. This completes the proof for part (i).

(ii) We have

$$\operatorname{Re}\frac{zf'(z)}{f(z)} \leq \frac{(1-2\alpha)r}{(1-r)(1-(1-r)^{1-2\alpha})} =: \mu(r,\alpha), \quad |z| = r < 1.$$

Now, to find $r \in [0,1]$ such that $\mu(r,\alpha) - \beta < 0$ for given α and β, let us assume $\tilde{\mu}(r,\alpha,\beta) := (1-2\alpha)r/((1-r)(1-(1-r)^{1-2\alpha})) - \beta$. Clearly, $\tilde{\mu}(0,\alpha,\beta) < 0$ and $\tilde{\mu}(1,\alpha,\beta) > 0$ for the given range of β. Thus, there must exist r_0 such that $\tilde{\mu}(r,\alpha\,\beta) \leq 0$ for all $r \in (0,r_0]$, where r_0 is as defined in the hypothesis. Therefore it concludes the proof for part (ii). ∎

Theorem 5.2. *Let $f \in \mathcal{S}^*(q_\alpha)$ and $\beta \geq 1$. Then for $|z| < \tilde{r}$, we have*

$$\left|\frac{zf'(z)}{f(z)} - 1\right| < \beta,$$

where $\gamma := \gamma(\alpha)$ given by (4.1) and $\tilde{r}$ is the smallest such $r < 1$ satisfying the equation

$$2(1-\gamma)r(1+r) - \beta(1-r^2)(1-|1-2\gamma|r) = 0.$$

Proof. As $f \in \mathcal{S}^*(q_\alpha)$, thus from (2.2) we deduce the following

$$\frac{f(z)}{z} = \frac{1+(1-2\gamma)\upsilon(z)}{1-\upsilon(z)}, \tag{5.1}$$

for some Schwarz function $\upsilon(z)$. Logarithmic differentiating the above equation yields

$$\frac{zf'(z)}{f(z)} - 1 = \frac{2(1-\gamma)z\upsilon'(z)}{(1+(1-2\gamma)\upsilon(z))(1-\upsilon(z))}.$$

Using the triangle inequality on the modulus of above equation, we get

$$\left|\frac{zf'(z)}{f(z)} - 1\right| \leq \frac{2(1-\gamma)|z||\upsilon'(z)|}{(1-|1-2\gamma||\upsilon(z)|)(1-|\upsilon(z)|)} \tag{5.2}$$

Upon applying the Schwarz-Pick inequality

$$|\upsilon'(z)| \leq \frac{1-|\upsilon(z)|^2}{1-|z|^2} \quad (z \in \mathbb{D})$$

and $|\upsilon(z)| \leq |z|$ in the equation (5.2), we deduce

$$\begin{aligned}\left|\frac{zf'(z)}{f(z)} - 1\right| &\leq \frac{2(1-\gamma)|z|(1+|z|)}{(1-|1-2\gamma||z|)(1-|z|^2)} \\ &= \frac{2(1-\gamma)r(1+r)}{(1-|1-2\gamma|r)(1-r^2)} =: u(r,\gamma).\end{aligned}$$

Now to show $u(r,\gamma)-\beta < 0$. For this consider, $\tilde{u}(r,\gamma,\beta) := 2(1-\gamma)r(1+r)-\beta(1-r^2)(1-|1-2\gamma|r)$. We observe $\tilde{u}(0,\gamma,\beta) = -\beta < 0$ and $\tilde{u}(1,\gamma,\beta) = 4(1-\gamma) > 0$. Thus, there must exist $\tilde{r}$ such that $\tilde{u}(r,\gamma,\beta) \leq 0$ for all $r \in [0,\tilde{r}]$, where $\tilde{r}$ is defined in the hypothesis. Hence the result. ∎

Remark 5.3. Since $\mathcal{S}_q^*(\alpha) \subset \mathcal{S}^*(q_\alpha)$, all the above results hold even for $\mathcal{S}_q^*(\alpha)$, however, they need not be sharp.

The logarithmic coefficients β_n for functions $f \in \mathcal{A}$ is given by:

$$\log \frac{f(z)}{z} = \sum_{n=1}^{\infty} 2\beta_n z^n \quad (z \in \mathbb{D}).$$

In the literature, the sharp bound $|\beta_n| \leq 1/n$ $(n \geq 1)$ for functions in $\mathcal{S}^*$ is already proved and equality holds for the Koebe function. Here we derive the bounds of $|\beta_n|$ for functions in the class $\mathcal{S}^*(q_\alpha)$.

Theorem 5.4. *Let $f \in \mathcal{S}^*(q_\alpha)$. Then, we have*

$$|\beta_n| \leq \frac{1-\alpha}{2n}.$$

The inequality is sharp.

Proof. For $f \in \mathcal{S}^*(q_\alpha)$, we have (2.6), or equivalently

$$z\left(\log \frac{f(z)}{z}\right)' \prec q_\alpha(z) - 1. \tag{5.3}$$

By subordination concept, we obtain

$$z\left(\log \frac{f(z)}{z}\right)' = q_\alpha(\upsilon(z)) - 1,$$

where $\upsilon(z)$ is a Schwarz function. Integrating the above equation, we deduce

$$\log \frac{f(z)}{z} = \int_0^z \frac{q_\alpha(\upsilon(t)) - 1}{t} dt,$$

which proves

$$\log \frac{f(z)}{z} \prec \int_0^z \frac{q_\alpha(t) - 1}{t} dt,$$

Now using (5.3), we have

$$\sum_{n=1}^{\infty} 2n\beta_n z^n \prec \sum_{n=1}^{\infty} B_n z^n := q_\alpha(z) - 1.$$

Applying Rogosinski's result [13], we obtain

$$2n|\beta_n| \leq |B_1| =: 1 - \alpha.$$

Hence the result. ∎

Theorem 5.5. *Let $f(z) = z + \sum_{n=2}^{\infty} a_n z^n \in \mathcal{S}^*(q_\alpha)$. Then*

(i) $|a_n| \leq 2(1-\gamma)$,

(ii) $|a_3 - \eta a_2^2| \leq \begin{cases} 2(1-\gamma)(1-2\eta(1-\gamma)) & \eta \leq 0 \\ 2(1-\gamma) & 0 \leq \eta \leq \frac{1}{1-\gamma} \\ 2(1-\gamma)(2\eta(1-\gamma)-1) & \eta \geq \frac{1}{1-\gamma}, \end{cases}$

where γ is given by (4.1). *The result is sharp.*

Proof. (i) As $f \in \mathcal{S}^*(q_\alpha)$, it satisfies equation (5.1). Let $p(z) = 1 + p_1 z + p_2 z^2 + \cdots \in \mathcal{P}$. Define $p(z) = (1 + \upsilon(z))/(1 - \upsilon(z))$. Substituting $\upsilon(z)$ in (5.1), we obtain

$$f(z) = ((1-\gamma)p(z) + \gamma)z = z + (1-\gamma)(p_1 z^2 + p_2 z^3 + p_3 z^4 + \cdots).$$

Comparing the coefficients on both sides of equation, we get

$$a_2 = (1-\gamma)p_1,\ a_3 = (1-\gamma)p_2,\ a_4 = (1-\gamma)p_3, \cdots a_n = (1-\gamma)p_{n-1}. \tag{5.4}$$

Since $|p_i| \leq 2$, thus the result follows at once.
(ii) From equation (5.4), we deduce

$$|a_3 - \eta a_2^2| = (1-\gamma)|p_2 - \mu(1-\gamma)p_1^2|.$$

Now the result follows directly from [7, Lemma 1, p.162]. ∎

Recall that inverse of a function $f(z) = z + \sum_{n=2}^{\infty} a_n z^n \in \mathcal{S}$ is given by $f^{-1}(f(z)) = z$, $z \in \mathbb{D}$ and $f(f^{-1}(w)) = w$ ($|w| < r_0$, $r_0 > 1/4$), for which

$$f^{-1}(w) = w - a_2 w^2 - (2a_2^2 - a_3)w^3 - (5a_3^2 - 5a_2 a_3 + a_4)w^4 + \cdots . \tag{5.5}$$

Corollary 5.6. *Let $f \in \mathcal{S}^*(q_\alpha)$ and the inverse be given by $f^{-1}(w) = w + \sum_{n=2}^{\infty} b_n w^n$. Then*

$$|b_2| \leq 2(1-\gamma) \text{ and } |b_3| \leq 2(1-\gamma)(4(1-\gamma) - 1),$$

where γ is given by (4.1). *The inequalities are sharp.*

Proof. Now, comparing the Taylor series of f^{-1} given in (5.5) and in the hypothesis, we get $|b_2| = |a_2|$ and $|b_3| = |2a_2^2 - a_3|$. Now the result follows from Theorem 5.5 (i) and (ii)($\eta = 2$), respectively. ∎

Acknowledgment. The second author is supported by a Research Fellowship from the Department of Science and Technology, New Delhi, Grant No. IF170272.

References

[1] M. K. Aouf, J. Dziok and J. Sokół, *On a subclass of strongly starlike functions*, Appl. Math. Lett. **24** (1), 27–32, 2011.
[2] P. L. Duren, *Univalent functions*, Grundlehren der Mathematischen Wissenschaften, 259, Springer-Verlag, New York, 1983.
[3] Y. C. Kim, J. H. Choi and T. Sugawa, *Coefficient bounds and convolution properties for certain classes of close-to-convex functions*, Proc. Japan Acad. Ser. A Math. Sci. **76** (6), 95–98, 2000.
[4] S. S. Kumar and S. Banga, *On a Special Type of Non-Ma-Minda Function*. arxiv preprint arXiv:2006.02111, 2020.
[5] S. S. Kumar, M. G. Khan, B. Ahmad and W. K. Mashwani, *A Class of Analytic Functions associated with Sine Hyperbolic Functions*. arXiv preprint arXiv:2011.04875, 2020.
[6] S. K. Lee, V. Ravichandran and S. Supramaniam, *Bounds for the second Hankel determinant of certain univalent functions*, J. Inequal. Appl. **2013**, 2013:281, 17 pp.
[7] W. C. Ma and D. Minda, *A unified treatment of some special classes of univalent functions*, in *Proceedings of the Conference on Complex Analysis (Tianjin, 1992)*, 157–169, Conf. Proc. Lecture Notes Anal., I, Int. Press, Cambridge, MA.
[8] H. Mahzoon, R. Kargar and J. Sokół, A new subclass of starlike functions, Turkish J. Math. **43** (5), 2354–2365, 2019.
[9] S. S. Miller and P. T. Mocanu, *Differential subordinations*, Monographs and Textbooks in Pure and Applied Mathematics, 225, Marcel Dekker, Inc., New York, 2000.
[10] P. T. Mocanu, Une propriété de convexité généralisée dans la théorie de la représentation conforme, Mathematica (Cluj) **11** (34), 127–133, 1969.
[11] R. K. Raina and J. Sokół, Some properties related to a certain class of starlike functions, C. R. Math. Acad. Sci. Paris **353** (11), 973–978, 2015.
[12] M. S. Robertson, *Certain classes of starlike functions*, Michigan Math. J. **32** (2), 135–140, 1985.
[13] W. Rogosinski, *On the coefficients of subordinate functions*, Proc. London Math. Soc. **48** (2), 48–82, 1943.
[14] H. Silverman, Convex and starlike criteria, Int. J. Math. Math. Sci. **22** (1), 75–79, 1999.

[15] H. M. Srivastava, M. Darus and R. W. Ibrahim, *Classes of analytic functions with fractional powers defined by means of a certain linear operator*, Integral Transforms Spec. Funct. **22** (1), 17–28, 2011.
[16] H. M. Srivastava, D. Răducanu and P. Zaprawa, *A certain subclass of analytic functions defined by means of differential subordination*, Filomat **30** (14), 3743–3757, 2016.
[17] B. A. Uralegaddi, M. D. Ganigi and S. M. Sarangi, *Univalent functions with positive coefficients*, Tamkang J. Math. **25** (3), 225–230, 1994.

Department of Applied Mathematics, Delhi Technological University, Delhi–110042, India
Email address: spkumar@dce.ac.in

Department of Applied Mathematics, Delhi Technological University, Delhi–110042, India
Email address: shagun05banga@gmail.com

Journal of Materiomics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Materiomics

journal homepage: www.journals.elsevier.com/journal-of-materiomics/

Opinion paper

A novel method of water remediation of organic pollutants and industrial wastes by solution- route processed CZTS nanocrystals

Pooja Semalti [a, b], Vikash Sharma [a], Shailesh Narain Sharma [a, b, *]

[a] CSIR-National Physical Laboratory (NPL), Dr. K.S. Krishnan Road, New Delhi, 110012, India
[b] Academy of Scientific and Innovative Research (AcSIR), Council of Scientific & Industrial Research (CSIR) - Human Resource Development Centre, (CSIR-HRDC) Campus, Ghaziabad, Uttar Pradesh, 201002, India

ARTICLE INFO

Article history:
Received 30 October 2020
Received in revised form
21 February 2021
Accepted 8 April 2021
Available online xxx

Keywords:
Photocatalysis
Copper zinc tin sulfide (CZTS)
Converging lens
Industrial waste
Hot injection
Organic pollutants
Nanoparticles
TOP
TOPO

ABSTRACT

CZTS (Cu_2ZnSnS_4), a P-type semiconductor with a direct bandgap (1.2–1.7eV), earth-abundant, non-toxic, and has a large absorption coefficient makes it extremely useful in optoelectronics and light-harvesting applications. In this work, CZTS is prepared by an ingenious, cost-effective colloidal route using the 'hot-injection' method with the usage of different ligands. The XRD and Raman spectroscopy shows the single-phase highly crystalline CZTS nanoparticles with kesterite structure. The TEM results show that the size of CZTS nanoparticles is about 2–5 nm and monodispersity is confirmed by DLS (Dynamic Light Scattering). FTIR confirms the presence of different ligands used in CZTS preparation. The Uv–vis absorption shows the direct bandgap of 1.5–1.7eV. The contact angle study shows the hydrophobic nature of as-synthesized CZTS nanoparticles which were further ligand exchanged with L-cysteine hydrochloride to make it hydrophilic to study the photocatalytic degradation activity of organic pollutants and industrial waste in the water. The photocatalysis experiments were performed under two conditions: (i) under bare sunlight (Intensity ~ 900 W/m^2) (ii) focussing the sample under the sunlight via converging lens (1800 W/m^2). The photocatalytic efficiencies were then compared and the best photocatalytic efficiency achieved under sunlight was 98.4% for organic pollutants and 75% for industrial waste *via* converging lens while the corresponding efficiencies with bare sunlight were 98.1% and 73% respectively. To the best of the author's knowledge, a rapid and highly efficient photocatalysis of CZTS NPs employing a converging lens for water-remediation without the usage of noble & transition-metals has been reported for the first time.

© 2021 The Chinese Ceramic Society. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Semiconductor nanomaterials owing to tuneable bandgap as a function of particle size possess various incredible properties such as light-harvesting applications including photovoltaics and photocatalysis, LEDs, optical properties based on morphology, large surface area to volume ratio which has required a terrific research effort in developing a range of new nanomaterials since the last decade [1]. Typically, oxides materials are preferred, due to their wide bandgap results from oxygen vacancies, for light-absorbing material [2]. In particular, a multinary metal-chalcopyrite quaternary nano-compound CZTS (Copper Zinc Tin sulfide- Cu_2ZnSnS_4 belongs to the group (I_2-II-IV-VI_4) of the bandgap of 1.4–1.5eV, having p-type conductivity and high absorption coefficient, shows a promising light-harvesting application, not only in photovoltaics but CZTS has also flagged in the area of photocatalysis. It shows excellent stability than any other conventional material such as CdS and n-TiO_2 [3–5]. Several reports have successfully shown the tunability of physical and optical properties by varying the nanocrystal size and various synthesis parameters. CZTS has come up with higher photocatalytic activity in its nano form rather than its bulk counterpart, which further is easily incorporated in counter electrodes and photocathodes for solar cells [6,7]. There are varieties of vacuum and non-vacuum approaches that have been explored a lot to synthesize nanoparticles and absorber layers. Vacuum based processes include the usage of either all elements or a few numbers of elements at a time in an inert atmosphere, followed by sulphurization to obtain the pure phase of CZTS. This method is a reliable one as the films produced are highly uniform

* Corresponding author. CSIR-National Physical Laboratory (NPL), Dr. K.S. Krishnan Road, New Delhi, 110012, India.
E-mail address: shailesh@nplindia.org (S.N. Sharma).
Peer review under responsibility of The Chinese Ceramic Society.

https://doi.org/10.1016/j.jmat.2021.04.005
2352-8478/© 2021 The Chinese Ceramic Society. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: P. Semalti, V. Sharma and S.N. Sharma, A novel method of water remediation of organic pollutants and industrial wastes by solution- route processed CZTS nanocrystals, Journal of Materiomics, https://doi.org/10.1016/j.jmat.2021.04.005

with commendable phase purity. However, there are concerns over this method especially for CZTS due to its expensive processing cost, the volatility of Sn and Zn as the Continuous re-evaporation makes it difficult for achieving phase accuracy with phase at the industrial level and this is why co-evaporation and co-sputtering remains a better alternative for larger yield and high efficiencies [8–10]. Contrariwise, a non-vacuum perspective is absolutely a different approach, involving a solution-based process such as sol-gel, hydrothermal, *etc.*, it is cost-effective to a great extent, ease in scalability for large-area applications with high potential [11]. The most effective advantage of this technique is the high tunability within the limited cost range. This method also allows us to obtain pure phase nanocrystals and command over the tedious film deposition once optimized accurately. Hence, it can be tagged as one of the best ways to synthesize the nanoparticles [12].

Nowadays, textile industries are one of the most growing chains world-wide; production is itself a tedious process with different stages of production. At each stage of production, from beginning to end, the textile industry emits a lot of pollutants and bi-products that include wastewater, air pollutants, solid waste, etc. [13–15]. The primary concern about these industries is the wastewater and a load of chemicals in it. A rough estimate shows, to produce 1 Kg of textile a 200 L of water is used [16]. The water left out after production is far behind the actual standards containing dyes, and other harmful chemicals in large amounts, leaving a great impact on the environment, and showing the need for regular assessment of hazard projection [17].

Extensive work literature has been reported for photocatalytic degradation of dyes under LED light of low intensity which results in good photocatalytic activity [18–21]. In this paper, the photocatalytic degradation of organic pollutants and textile effluents (consisting of various chemicals and dyes) using CZTS nanoparticles produced by the hot-injection technique is reported. The hot injection technique allows the production of nanoparticles with precise control of the size. Long-chain ligands are used to obtain controllability over nanoparticles' size, which also hinders the charge transfer property, However, the choice of capping ligands would play an important role in overcoming this situation. Miscibility with water is a primary problem with different capping ligands to execute photocatalysis [22–24]. The simple, eco-friendly, and cost-effective ligand exchange process is one solution to alter the hydrophobic nature of synthesized nanoparticles in the form of hydrophilic behavior. L-cysteine (hydrochloride) and NaOH are used for the post-synthesis ligand exchange process and do not alter the basic behavior of CZTS. This hydrophilic CZTS was further effectively used as a catalyst for photocatalytic degradation of organic pollutants and textile waste. The photocatalytic degradation effect was further studied under two conditions: (i) bare sunlight exposure (intensity ~ 900 W/m^2) and (ii) sunlight exposure using a converging lens (1800 W/m^2).

2. Experimental section

2.1. Material required

The precursor salts, Copper (II) chloride dihydrate (98%), Zinc (II) Chloride (95%), Tin (IV) Chloride (98%), S powder (99%) were used for Copper, Zinc, Tin and Sulfur source respectively. The capping agents being n-Butylamine, and Octadecene. All the chemicals are of AR grade, supplied by Sigma Aldrich, and used without further purification.

2.2. Instrumentation

The Powder X-Ray Diffraction (XRD) within 20°–80° range was measured by Bruker-AXS D8 advanced X-Ray Diffractometer (Cu-K_α, $\lambda = 0.154$ nm) with scanning speed 0.01 deg/s. The optical absorbance spectrum was measured using UV-3101 PC (Shimadzu UV–Vis–NIR double-beam spectrophotometer). TEM measurement was done by using JEM-200CX (JEOL model), with an acceleration potential of 200 kV. After each modified step, contact angles of the surface and surface energies were recorded on the Drop Shape Analysis System, model DSA10MK2 (Kruss GmbH, Germany). FTIR studies were done using a model FTIR spectrometer (Spectrum two; PerkinElmer). Dynamic light scattering (DLS) measurement was done by nano series zetasizer (Nano-s model) with laser 50 mW (532 nm). The data provided by the various characterization tools are entirely adequate to support the correlation between the results as done in this work.

2.3. Synthesis of CZTS nanoparticles and structures

Synthesis of CZTS compounds with different ligands was done. Trioctylphosphine Oxide (TOPO) acting as a solvent for metallic precursors while Trioctylphosphine (TOP) is a medium for complexing the anions. The size-selective, highly monodispersed CZTS NPs were produced using an optimized Hot-injection method (colloidal route) [25–27]. Here, the preparation was contemplated using the 'High-temperature solid-state-junction' mechanism where the desired reactants dispersed in a solvent and injected rapidly to the other dispersed reactants solvent at elevated temperatures. In this typical procedure for CZTS1, 0.2 g, 0.6 g, and 0.1 g of Copper, Zinc, and Tin precursors were added to a 2-neck round bottom flask with 1.5 g of TOPO, maintaining a Zinc-rich ratio which exhibits better photocatalytic and photovoltaic activity [28]. The temperature is maintained between 250 and 300 °C for an hour, the color change is seen from yellow to orange. The elemental S (0.4 g) as sulfur source with TOP (1 ml) is then rapidly injected into the flask and maintained at a temperature of 230 °C for 20 min. The pale color of brown was obtained after 2–3 min of injection**,** and finally, deep black color was seen, heated up to 250 °C. Stirring was done continuously throughout the experimental process. The synthesis was carried out under the argon atmosphere connected via the Schlenk line. The Round bottom 2-neck flask was allowed to heat via a heating mantle having a magnetic stirrer. A rubber septum sealed the flask to provide accurate temperature within the flask and hence a precisely controlled reaction was obtained [29]. The as-synthesized NPs have washed alternatively for 3–4 times with methanol and toluene to remove unbounded ligands and other impurities such as secondary elements formed during the reaction. Centrifugation was done after washing and supernatants were discarded and a dark black colored powder precipitate was obtained. Finally, the precipitate was dispersed into toluene, and to use it as a powder it was dried at NTP (Normal Temperature and Pressure). Since the NPs are highly stable**,** these NPs can be stored in centrifuge tubes at NTP for a long time. A similar process was used to make another CZTS NPs (CZTS2), to compare the photocatalytic studies, with n-butylamine and Octadecene as ligands instead of TOPO and TOP.

2.4. Ligand exchange for CZTS nanoparticles

The basic behavior of pristine CZTS nanoparticles is hydrophobic, hence not suggested for degradation of organic pollutants, which impedes the performance of NPs in aqueous solutions to the insulating layer on the surface of nanoparticles. So, the hydrophobicity can be changed to hydrophilicity using the proper choice of ligand and experimental process [24,30]. Since TOP and TOPO are the toxic and most hydrophobic ligands [31] while n-Butylamine and Octadecene are non-toxic but less hydrophobic ligands used in

CZTS, hence, L-cysteine (non-toxic/edible) was used to ultimately make the CZTS NPs hydrophilic.

2.4.1. Direct ligand exchange

Post synthesized CZTS1 was used for direct ligand exchange with n-Butylamine. CZTS NPs were stirred and sonicated in Butylamine for few minutes. The final product became hydrophilic, but the CZTS structure was destroyed completely which was confirmed by the XRD pattern in Fig. 1(C) while CZTS2 doesn't show any change and remains as it is. However, before ligand exchange, the water contact angle is comparatively lower than that of CZTS1, which justifies the use of n-Butylamine and Octadecene.

2.4.2. Indirect ligand exchange

Before the photocatalysis of organic pollutants, the NPs were transferred to aqueous media from the organic media via L-cysteine, 20 mg/ml L-cysteine in DI water, maintaining the PH of 10 by dropwise adding NaOH solution, stirred for 20 min 2–3 mg of concentrated CZTS-toluene solution was added dropwise and heated for 15 min at 60 C, while stirring continued for 24 h**.** A single layer of clear insoluble solution (i.e. toluene) remains at the upper aqueous phase**,** which was discarded. On the contrary, CZTS NPs moved to the lower aqueous phase and a pale-yellow color was obtained, which signifies the purified NPs, that were kept stored for photocatalysis. This process includes the ligand exchange of CZTS NPs, capped with different ligands (CZTS1 capped with TOPO, TOP, and CZTS2- capped with n-Butylamine, Octadecene), with L-cysteine and NaOH.

The hydrophobic tail is replaced by the L-cysteine, the hydrophilic tail gets adhered to the surface of CZTS (both). The collected CZTS was then monitored for photocatalytic degradation of organic pollutants (Methylene Blue) and textile waste (consisting of Dyes/chemicals collected from textile industrial wastewater outlets). Fig. 1(a) and (b) show the structure of CZTS and the synthesis setup for the same. Fig. 1(c) and (d) shows the ligand exchange process for post synthesized CZTS NPs.

3. Results and discussion

3.1. X-ray diffraction and Raman spectroscopy

The nanoparticles were characterized by the XRD for identification and phase analysis. The result manifests the kesterite phase before ligand exchange and wurtzite phase after ligand exchange, for CZTS1 (with TOP-TOPO ligand) and CZTS2 (with n-Butylamine and Octadecene ligand). The synthesized nanoparticles are found to be polycrystalline in nature [32–34]. The JCPDS 26–0575 illustrates the presence of characteristic peaks (112), (200), (220), (312), (008), and (332) as shown in Fig. 2(A). In CZTS1, the XRD pattern shows two peaks corresponding to Cu_2S [35] at 31.9° (200) and 45.86° (220) which is the secondary phase of CZTS, while no such secondary phases are seen in CZTS2. The CZTS NPs deteriorate completely and does not show any significant characteristic peaks after 'direct ligand exchange', as shown in Fig. 2(C) while after the 'indirect ligand exchange' shown in Fig. 2(B) with L-cysteine, both the CZTS has changed the phase from kesterite to wurtzite (with characteristic peaks (002), (101), (110), (112) and (104)) [35,36], which can be attributed to the change of state from solid to liquid and the stress/strain caused by ligand exchange [37,38]. CZTS2 doesn't show any sign of any secondary/binary by-products [39] but CZTS1 again has a spurious peak for Cu_2S at the same 2θ values [34] which is responsible for lower catalytic efficiency [40,41]. CZTS generally shows the kesterite phase, which is considered to be the most stable structure [41,42]. However, the metastable phase of CZTS (wurtzite) has attained a huge attention owing to its flexibility of stoichiometry and the ability to tune the Fermi energy over a wide range [42]. The wurtzite phase of CZTS is found to be more attractive for catalytic applications because of a higher carrier-concentration as inferred from the Hall-effect measurements and hence lower resistivity and higher electrical conductivity as compared to kesterite CZTS [43–46]. Also, the wurtzite CZTS exhibits higher solar energy conversion and catalytic efficiency compared to its kesterite phase [43–46]. Both CZTS NPs are highly crystalline even after the ligand exchange. Here, the crystallinity of CZTS2 has enhanced as sharp XRD peaks are seen, thus more electron-hole pair generation**;** hence more charge transfer takes place due to low recombination centers. The particle size has also

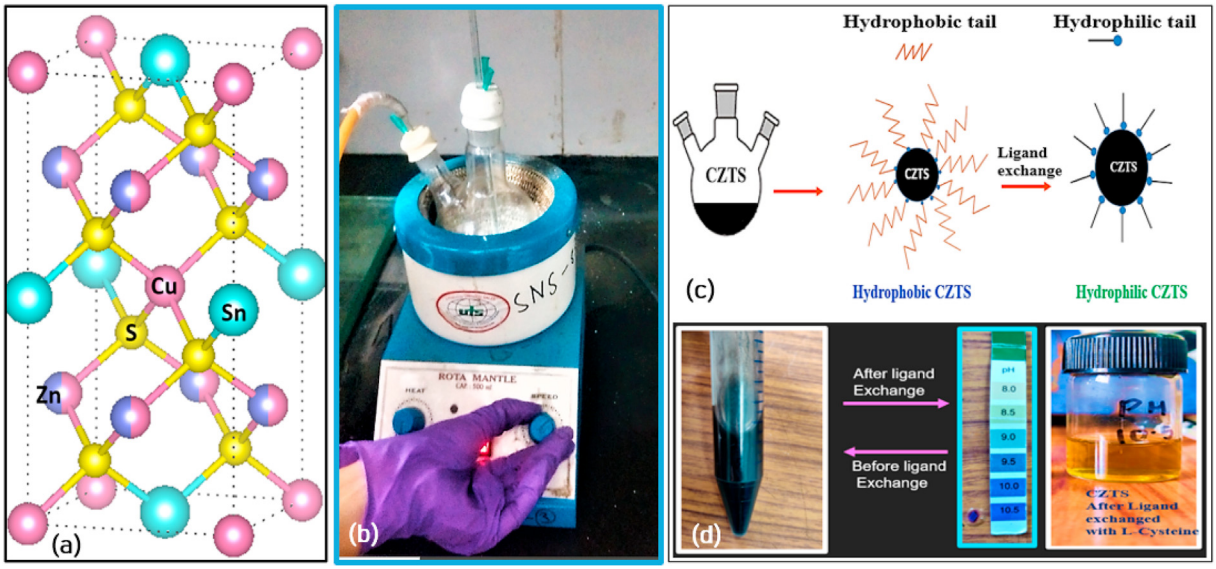


Fig. 1. (a) Structure of CZTS (b) synthesis setup for CZTS preparation (c) ligand exchange mechanism (from hydrophobic to hydrophilic); (d) before & after ligand exchange treatment.

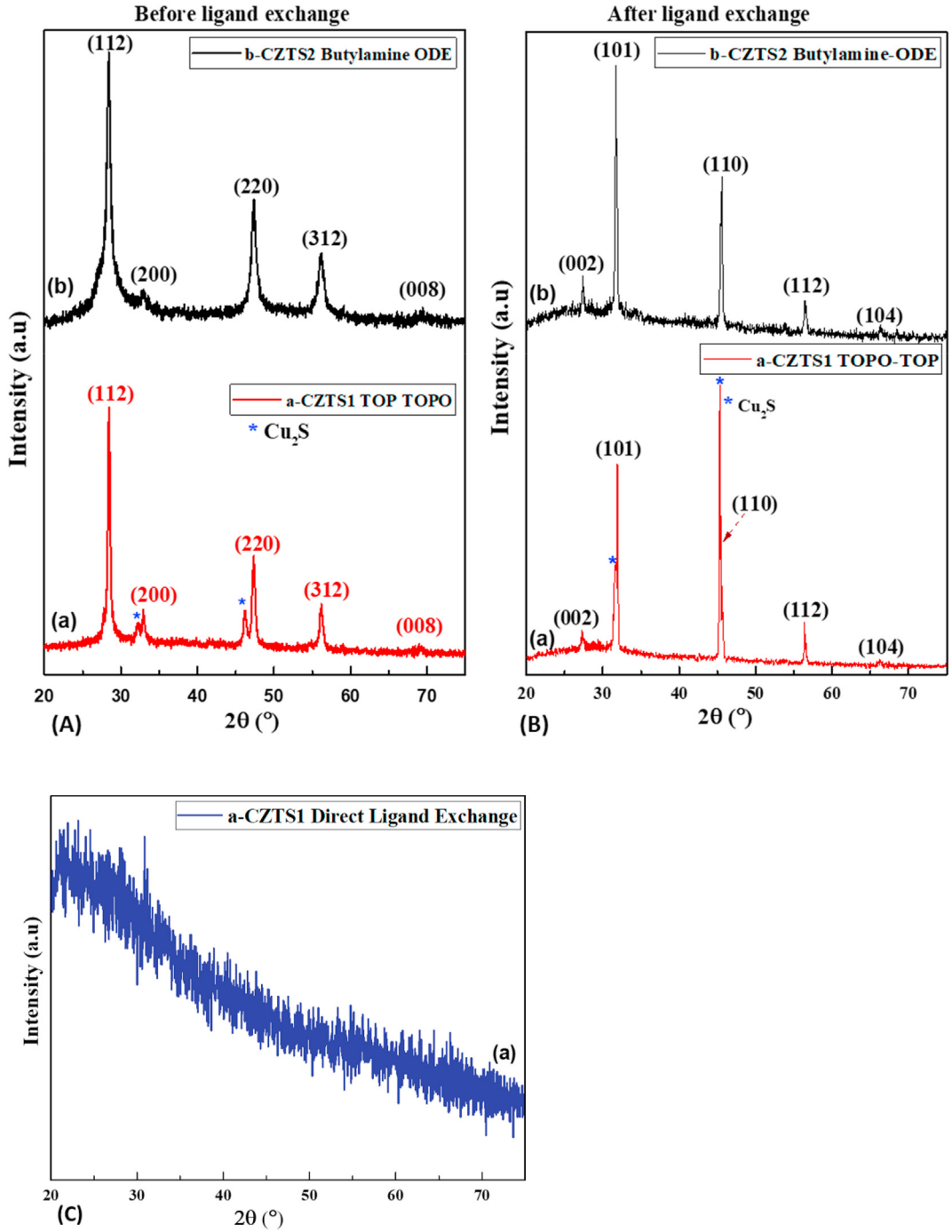


Fig. 2. (A) X-ray diffraction pattern a-CZTS1 and b-CZTS2; (B) Post synthesis Ligand exchange with L-Cysteine and (C) Direct Ligand exchange.

increased after the ligand exchange, therefore fewer grain boundaries hence, free radical generation and more effective catalytic efficiency in CZTS2 are expected. The average crystallite size using Debye Scherrer's method is ~16 nm and ~6 nm for CZTS1 and CZTS2, respectively before ligand exchange [43] while the crystallite size after ligand exchange is ~100 nm and 22 nm for CZTS1 and CZTS2, respectively.

Fig. 3 shows the Raman spectra of CZTS1 and CZTS2 nanocrystals. For CZTS, the intense characteristic Raman peak corresponding to A-Raman mode as reported by others mainly lies between 329 and 336 cm^{-1} and is accompanied by a relatively weaker peak at 285 cm^{-1} [47,48]. A slight Raman peakshift towards the low wavenumber is attributed to the presence of internal stress, which is very common in CZTS films formed without annealing under sulfur atmosphere, due to the shrinking of the unit cells' volume while cooling down the reaction [49,50]. The deconvoluted Raman spectrum (after baseline correction), as shown in Fig. 3(A, B) displays the characteristic peaks of CZTS appearing at ~ 285 cm^{-1} and 332 cm^{-1} [47,48,51–53]. Another weak peak at ~367/368 cm^{-1} corresponding to B Raman mode [54] also indicated the crystalline nature of CZTS. In addition to the above peaks, CZTS1 also exhibits weak signatures of impurity peak at ~475 cm^{-1} corresponding to

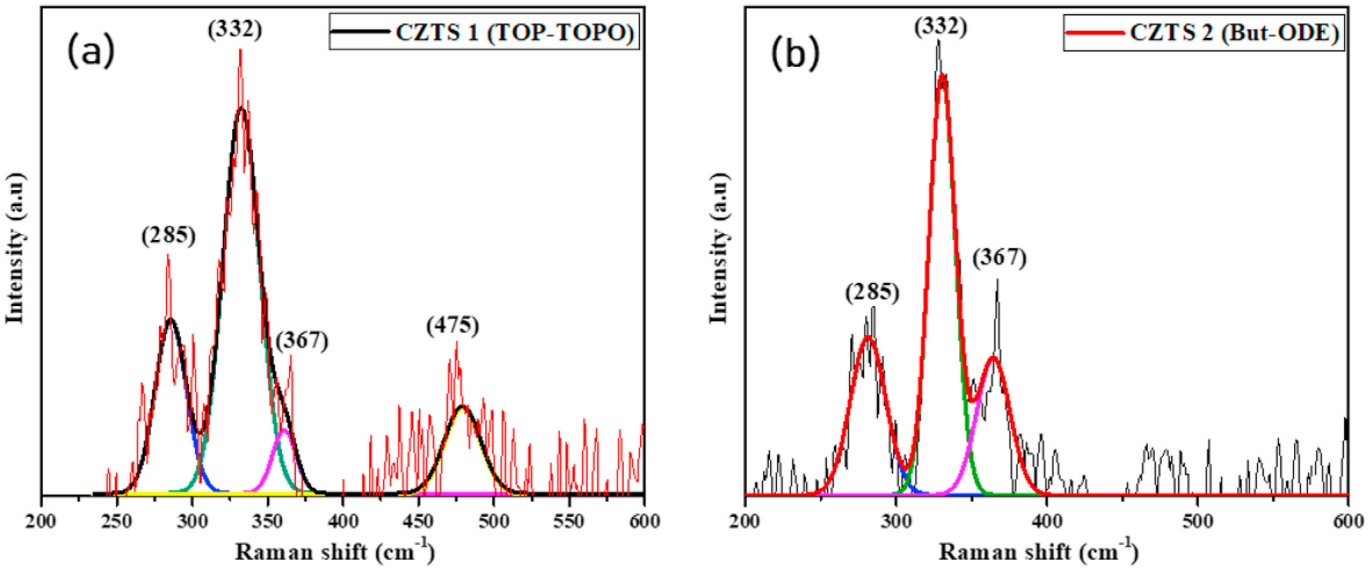


Fig. 3. The deconvoluted Raman spectra for (a) CZTS1 (b) CZTS2.

$Cu_{2-x}S$ [55]. However, no other characteristic peak of impurities like Cu_2SnS_3 (298 cm^{-1}), ZnS (351 cm^{-1}) and Cu_2SnS_4 (318 cm^{-1}) [55,56] could be observed. On the other hand, CZTS2 is devoid of any impurity binary/ternary peaks (absence of peaks corresponding to Cu_2SnS_3 (298 cm^{-1}), ZnS (351 cm^{-1}), Cu_2SnS_4 (318 cm^{-1}) and 475 cm^{-1}($Cu_{2-x}S$)) while the characteristic peaks at 285 cm^{-1}, 332 cm^{-1} and 368 cm^{-1} for pure CZTS are present. Thus, indicating phase purity of the as-synthesized nanocrystals [51–54]. It is conjectured that a Butylamine ligand owing to its simple, linear structure with short chain, renders efficient passivation of CZTS nanocrystals. On the other hand, the long chain of the bulky TOPO/TOP structure leads to a weak binding strength & a steric hindrance to the passivation of CZTS nanocrystals which leads to the presence of secondary phase Cu_2S in CZTS1. The weak signature of $Cu_{2-x}S$ in CZTS1 nanocrystals is seen, which is also due to the low miscibility of TOPO/TOP with sulfur and the low binding energy of TOPO/TOP with the CZTS compound. The deconvoluted Raman spectra also confirms the polycrystalline nature of CZTS1 and CZTS2 respectively as reported by others [57–59] and is in agreement with XRD results. From the above XRD and Raman studies, we can infer that both these studies have drawn reliable and similar results that commensurates well with each other.

The FWHM of the main Raman peak is related to the phono-confinement effect due to the nanometric dimensions of CZTS particles. The mean phonon lifetime can be calculated from the following Heisenberg's uncertainty principle [19]:

$$1/\tau = \Delta E/\hbar = 2\pi c\Gamma$$

Where ΔE represents uncertainty in phonon energy state, h is plank's constant, c is the velocity of light, and Γ is FWHM of intense Raman peak at 332 cm^{-1}, and τ is mean phonon lifetime. Using the above relation, τ value as determined for CZTS1 and CZTS2 comes to ~1.28 x 10^{-11} s and 2.58 x 10^{-11} s respectively, and FWHM for CZTS1 is 41.35 cm^{-1} & CZTS2 is 20.6 cm^{-1}. The lower values of FWHM or higher mean phonon lifetime indicate the better crystalline nature of CZTS2 as compared to CZTS1.

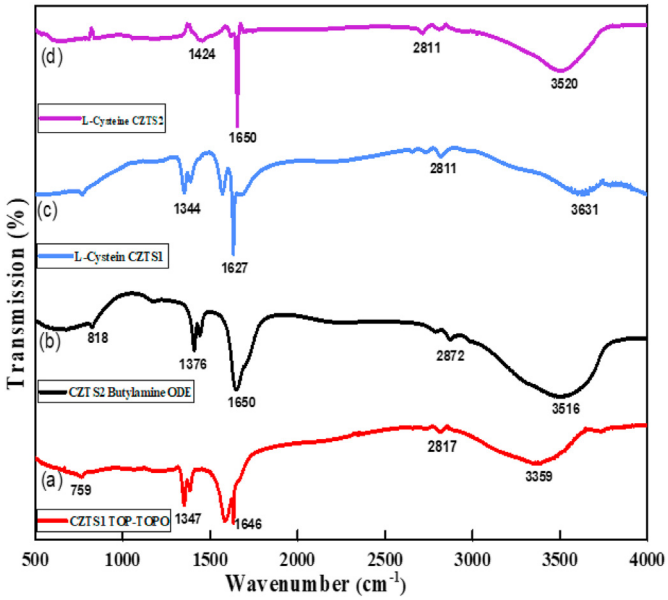


Fig. 4. FTIR Spectra for a-CZTS1, b-CZTS2, and c-CZTS1, d-CZTS2 after ligand exchange with L-cysteine hydrochloride.

3.2. FTIR (Fourier-Transform Infra-red spectroscopy)

Fourier-Transform Infra-red spectroscopy, Fig. 4 shows the FTIR spectra of different CZTS, and analysis of FTIR peaks corresponding to different groups with types of bonds attached in CZTS before and after ligand exchange is given in Table 1.

All the characteristic peaks are present with their corresponding material with different ligands. The P═O bond has been removed after ligand exchange with L-cysteine, shows the complete exchange of ligands, and hydrophilicity has increased after the ligand exchange, which is due to the amino group present in L-Cysteine.

Table 1
FTIR Peak values corresponding to different organic compounds.

Compound	FTIR Peaks
CZTS1 (TOP-TOPO) Before Ligand exchange	1347- P=O, 1646- C=C, 759-P–C, (2817, 3359)- C–H stretching
CZTS2 (Butylamine-ODE) Before Ligand exchange	1650- N–H bending, 3516- Amine (NH_2 stretching vibrational mode), 2872- C–H (symmetric stretching mode), 1376- CH3, 818- C=C bending
L-Cysteine hydrochloride (CZTS1) After Ligand exchange	(3631,1627)- NH_2, 1344- OH bending, 2811- S–H stretching
L-Cysteine hydrochloride (CZTS2) After Ligand exchange	(3454, 1650) –NH_2, 1424-OH bending, 2811- S–H stretching

The NH_2 stretching mode in CZTS2 (Fig. 4(d)) is more prominent as compared to CZTS1 (Fig. 4(c)) after ligand exchange with L-cysteine**,** which is also an amino group hence CZTS2 (Fig. 4(d)) is more hydrophilic and better for catalytic efficiency.

3.3. TEM (transmission electron microscopy)

The as-synthesized nanocrystals were characterized by transmission electron microscopy (TEM), and their corresponding micrographs are shown in Fig. 5. For CZTS1 nanocrystals, agglomerated nanorods (120–140 nm) with a fringe width of 0.23 nm shown in Fig. 5(a) and **(b),** respectively. The SAED pattern of CZTS1 (inset Fig. 5(b)) shows agglomerated crystallites of CZTS embedded in the crystalline matrix. Micrographs correspond to CZTS2 nanoparticles, owing to a nano spherical morphology of diameter 15–35 nm with a fringe width of 0.2–0.4 nm shown in Fig. 5(c) and (d), respectively. Fig. 5(c) inset shows the magnified image of nanospheres and (inset Fig. 5(d)) SAED images show a ring pattern that reveals fine grains, randomly-oriented CZTS crystallites.

3.4. Contact angle and surface energies

Here, the contact angle of CZTS1 and CZTS2 NPs with water before the ligand exchange was 168.18 and 106.18 , shown in Fig. 6(a) and 6**(b),** respectively. While the after-ligand-exchange process, water contact angle reduces to 41.33 and 18.31 (both are < 90 , therefore hydrophilic) as shown in Fig. 6(c) and **(d)** respectively.

The surface energy was also measured by Fowkes method [47], which is described as follows:

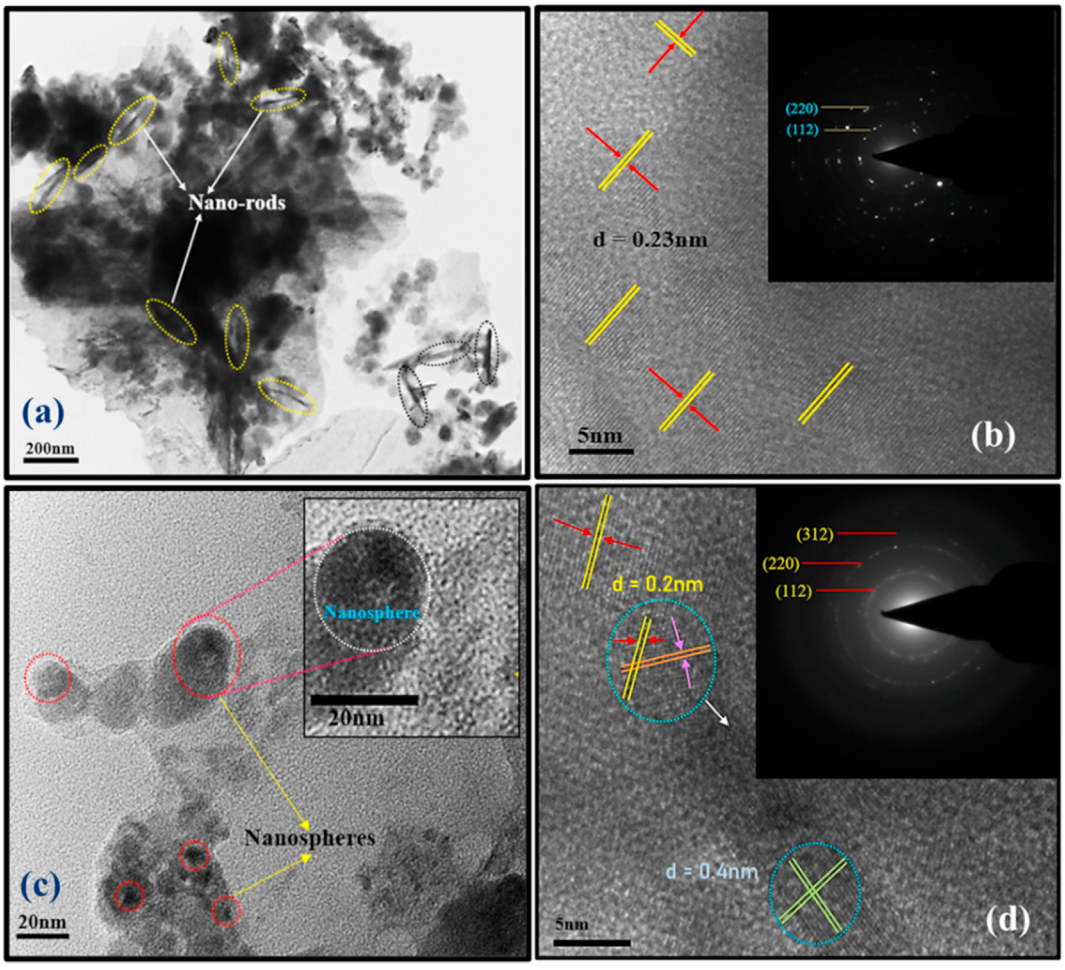


Fig. 5. TEM images of CZTS; (a) CZTS1, (c) CZTS2, (b) & (d) their corresponding HRTEM images & SAED patterns (insets).

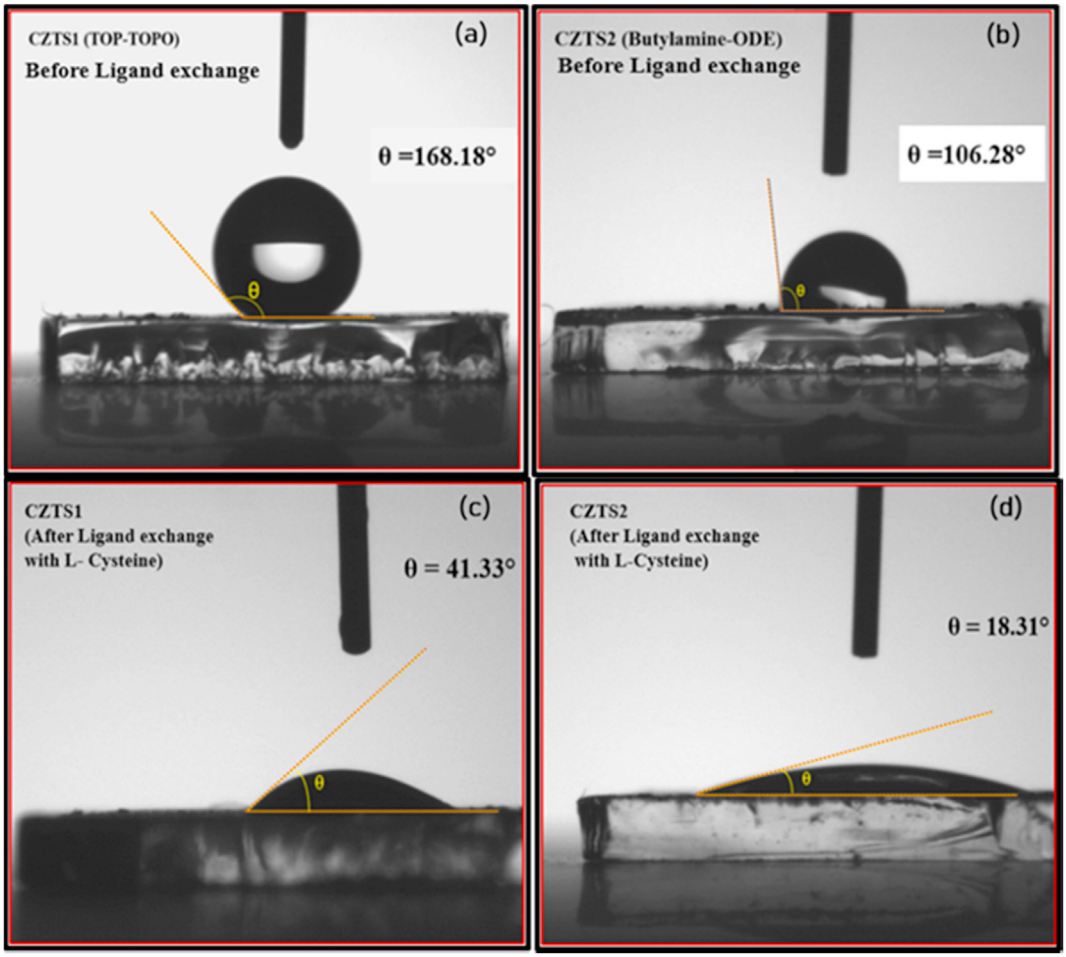


Fig. 6. Contact angle measurements for (a) CZTS1 and (b) CZTS2, before ligand exchange; (c) CZTS1 and (d) CZTS2, after ligand exchange.

This method is one of the most common methods for calculating surface energies for polar surfaces published in 1963 [60]. This theory estimates the interaction between liquid and solid in terms of dispersive and polar interactions**;** these interactions will sum up and gives the overall energy value.

$$\sigma_S = \sigma_S^D + \sigma_S^P$$

Here σ_s^D and σ_s^P are respectively the Dispersive and Polar components of the surface energy of solid.

For CZTS1, Surface Energy (Total) = 3.32 mN/m, (Disperse Part = 2.88, Polar Part = 0.45).

For CZTS2, Surface Energy (Total) = 153.73 mN/m, (Disperse Part = 116.21, Polar Part = 37.52).

This data shows that Butylamine in CZTS2 holds a stronger bond than TOP and TOPO in CZTS1. The higher surface energy of CZTS2 nanoparticles can be understood from the fact that CZTS2 has smaller nanometric dimensions than CZTS1**,** as inferred from TEM studies and hence higher surface area.

3.5. UV–vis

The bandgap determination by Tauc's equation was employed to utilize the obtained values of absorbance for direct band semiconductor CZTS nanoparticles [61]**,** which is ~1.6eV and 1.75eV before the ligand exchange and 1.50eV and 1.54eV after the ligand exchange for CZTS1 & CZTS2**,** respectively, shown in Fig. 7(a-d), which is consistent with the reported values (1.4–1.7eV) [62]. The bandgap lies in the visible range and can be used effectively in photovoltaics and other photonic applications [19,63]. It is important to realize the scope for tuning the physical properties of CZTS-related materials that surpass the previously established materials with their unique properties and present an optimal range of bandgap values ideal for charge separation and conversion [64].

3.6. Dynamic light scattering (DLS)

DLS is Used to determine the size distribution profile of small particles in a suspension & the temporal fluctuation gives an idea for the hydrodynamic size distribution**,** which includes the ligands, ions, surface charges, and whatever in the vicinity of the surface and therefore the difference is seen in TEM and DLS results. The correlation spectroscopy in DLS shows the relation between how small and large particles absorb and emit the irradiated light. Here, a small amount of CZTS NPs was dispersed into its solvent (toluene here), sonicated for 30 min, and immediately used for the analysis. In Fig. 8(a) and 8(b), there are sharp intense peaks for the size distribution profile before and after the ligand exchange**,** signifying the monodispersed size distribution of these particles. For CZTS1, the hydrodynamic size calculated by DLS is 831 nm and for CZTS2, the size is 177 nm before ligand exchange and 535 nm and 956 nm after the ligand exchange for CZTS1 and CZTS2**,** respectively. This also concludes the participation of long-chained TOP and TOPO in CZTS1 and small chained Butylamine in CZTS2, resultant in a size difference in these NPs**,** as shown in Fig. 8. The monodispersity of these NPs was also confirmed by the polydispersity index (PDI)

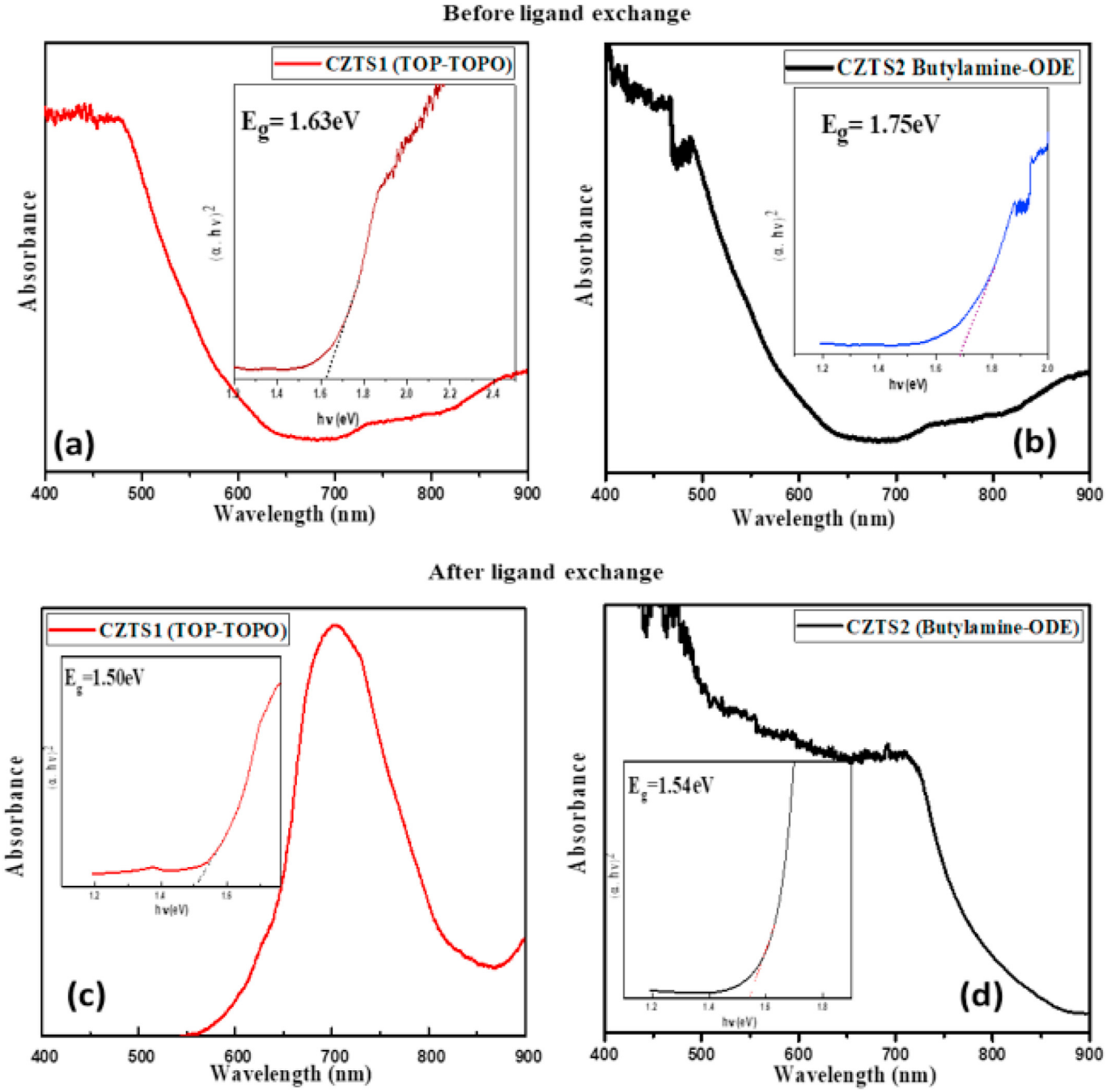


Fig. 7. (a) UV–Vis spectra for (a) CZTS1 (inset Tauc's Plot); (b) CZTS2 (inset Tauc's Plot) before ligand exchange and (c) CZTS1 (inset Tauc's Plot); (d) CZTS2 (inset Tauc's Plot) after ligand exchange.

value, which is less than 1 before and after the ligand exchange. Fig. 8(a, b) insets show the correlation coefficient for both NPs, CZTS1 shows large particles and hence there's a plateau, which refers to the slow decay of bigger particles as compared with CZTS2 because of Butylamine being a smaller ligand binds tightly on the surface of CZTS2 nanoparticles, and hence its relatively smaller size with high monodispersity as compared to CZTS1.

4. Photocatalytic studies

Photocatalytic degradation of pollutants: the constitutive process of photocatalysis follows a few progressive steps:

(1) Absorption of incident photons by the catalyst (CZTS); hence, band engineering occurs, i.e. electron (e), hole (h) pair generation with substantial potential for pollutant degradation.
(2) Charge separation, recombination inhibition, and charge migration to their corresponding surface sites.
(3) Surface reaction site construction for hydrogen (H_2) and oxygen (O_2) evolution as the main product.
(4) Redox reactions must co-occur [65–67], as shown by the mechanism in Fig. 9.

The catalyst maintains its photoactivity even after exposed for several months to the indoor light [68].

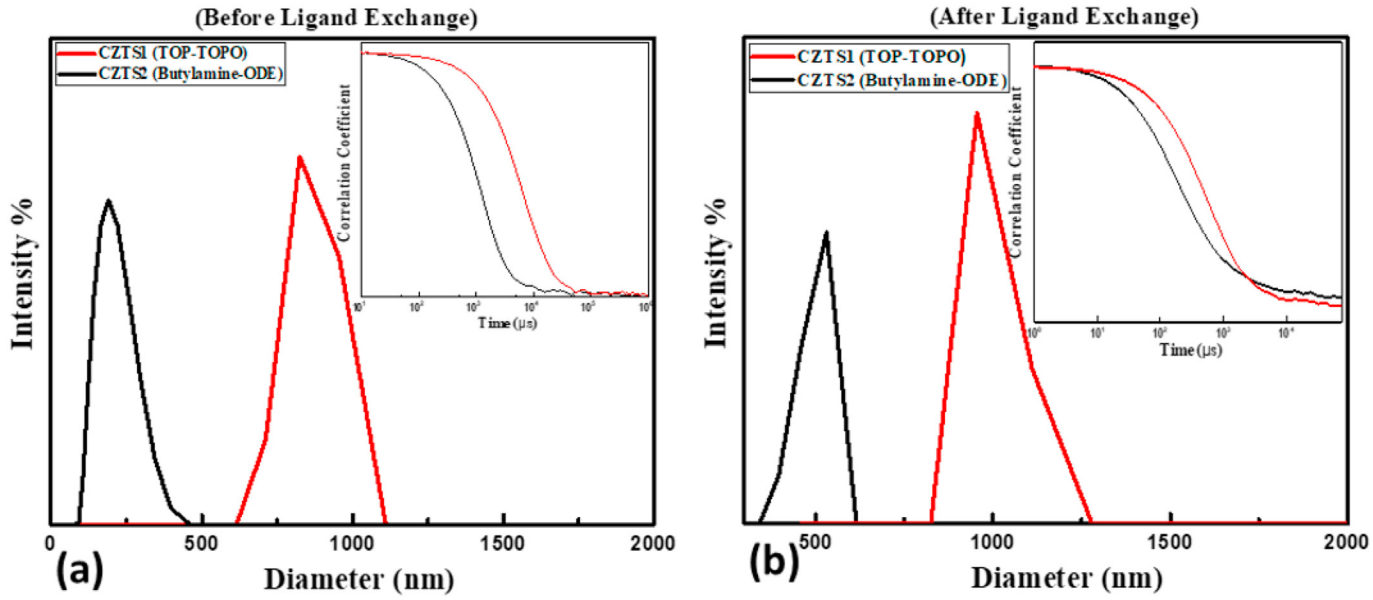


Fig. 8. Size distribution of CZTS1 and CZTS2 (a) Before and (b) After Ligand Exchange; with their corresponding Correlation coefficient vs. Time plot (Insets).

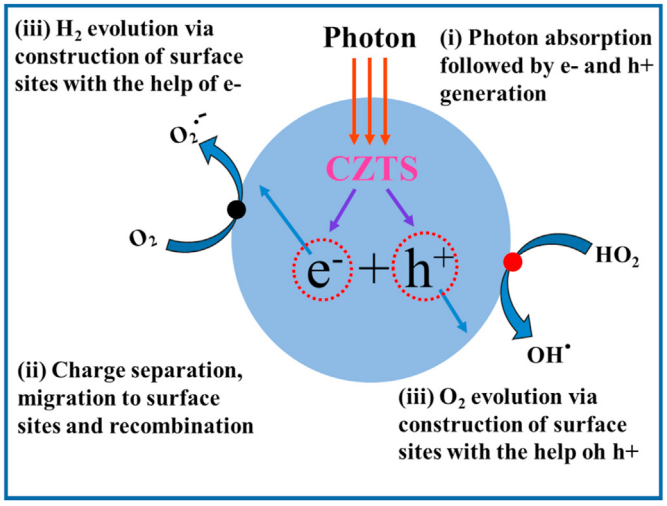


Fig. 9. Photocatalysis mechanism of CZTS.

The following equations are utilized here to calculate the degradation efficiency & other related parameters given by:

$$\textit{Catalytic Degradation Efficiency}\ (\%) = \frac{C_t - C_0}{C_0} \times 100\%$$

Rate constant K_{app} will be calculated from the equation [69]:

$$\ln \frac{C_t}{C_0} = -K_{app} t$$

Where C_0 = Initial absorbance/Concentration of Dye, C_t = Absorbance/Concentration of Dye with catalyst after time t.

For the potential photocatalytic evaluation of CZTS metal-chalcopyrite heterostructure, the degradation under visible light (Sunlight) (with and without lens) is tested as a model system. In this measurement, particularly, 200 μl CZTS NPs solution was suspended into a 10 ml aqueous methylene blue solution as an organic pollutant [70]. This prepared solution was continuously stirred in a dark room for 2 h to achieve the equilibrium adsorption. The solution was then kept under a visible light-source and simultaneously, the absorption spectra were monitored at regular time intervals. The method mentioned above was exactly followed for the photocatalytic degradation of textile waste as well. Table 2 highlights the comparison between previously reported data of CZTS prepared by different non-vacuum processes, along with their efficiencies, degradation rate constants, and the time taken for the complete degradation of the dyes or Organic pollutants. The absorption spectra, along with the concentration ratio with time for the different solutions, were carried out until the solution becomes colorless visibly, as shown in Fig. 10(a-o).

Fig. 10(a) shows the photodegradation of organic pollutants by CZTS1 in 160 min in direct sunlight and the corresponding rate constant and degradation efficiency shown in Fig. 10(b). Similarly, Fig. 10(c)–(d) shows the photodegradation of the same material but using a concentrator (converging) lens, and Fig. 10(e) shows the color degradation bar of the material. In this similar manner, Fig. 10(f)–(j) follows the same for CZTS2. Fig. 10(k)–(o) follows the same pattern by CZTS2 for the textile pollutants.

The reactions taking place here are as follows:

$$CZTS + h\nu \rightarrow CZTS(e^- + h^+)$$

Oxidation:

$$H_2O + (CZTS)h^+ \rightarrow CZTS + OH^- + H^+$$

$$OH^- + (CZTS)h^+ \rightarrow CZTS + OH^{\cdot}$$

Reduction:

$$(CZTS)e^- + O_2 \rightarrow CZTS + O_2^{\cdot -}$$

$$O_2^{\cdot -} + H^+ \rightarrow HO_2^{\cdot}$$

$$2HO_2^{\cdot} \rightarrow H_2O_2 + O_2$$

$$H_2O_2 + (CZTS)e^- \rightarrow OH^- + OH^{\cdot}$$

Table 2
Efficiencies and Degradation constants with the time taken by different CZTS with different pollutants (Comparison of previously reported data and the Present Work) [71–73].

Material with Crystal structure	Method	Band gap (eV)	Nano Structure	The rate constant (-Kapp) (min-1) & degradation percentage (%)	Degradation Time (min)	Ref.
Kesterite CZTS	One-pot	1.5	Nanocrystals	0.0317 and 91%	60min	[71]
CZTS-TiO2	Heating	1.46	Nanoparticles	27.6%	8hours	[72]
Kesterite CZTS	Hydro-thermal	1.53	Agglomerated Nanoparticles	0.018 & 50%	45 min	[19]
Kesterite CZTS	Hydro-thermal	1.72	Spherical	0.0144	120 min	[13]
Wurtzite CZTS	One-pot	1.54	Granular nanoparticles	81.79%	350 min	[73]
Wurtzite CZTS Au-CZTS Pd- CZTS	Hot Injection	1.40	Core-shell or Dimer morphology	CZTS- 44% Au-CZTS- 62% Pd- CZTS- 94%	90 min	[23]
CZTS1 Wurtzite	***Hot Injection***	***1.50***	***Nano-Rods***	***0.014 & 90.3% (Without Lens) 0.04 & 90.4% (With Lens) For Organic Pollutants***	***160 60***	***Present Work***
CZTS2 Wurtzite	***Hot Injection***	***1.54***	***Nano-Spheres***	***0.04 & 98.1% (Without Lens) 0.14 & 98.4% (With Lens) For Organic Pollutants***	***90 30***	***Present Work***
CZTS2 Wurtzite	***Hot Injection***	***1.54***	***Nano-Spheres***	***0.008 & 73% (Without Lens) 0.02&75% (With Lens) For Textile Effluents***	***150 50***	***Present Work***

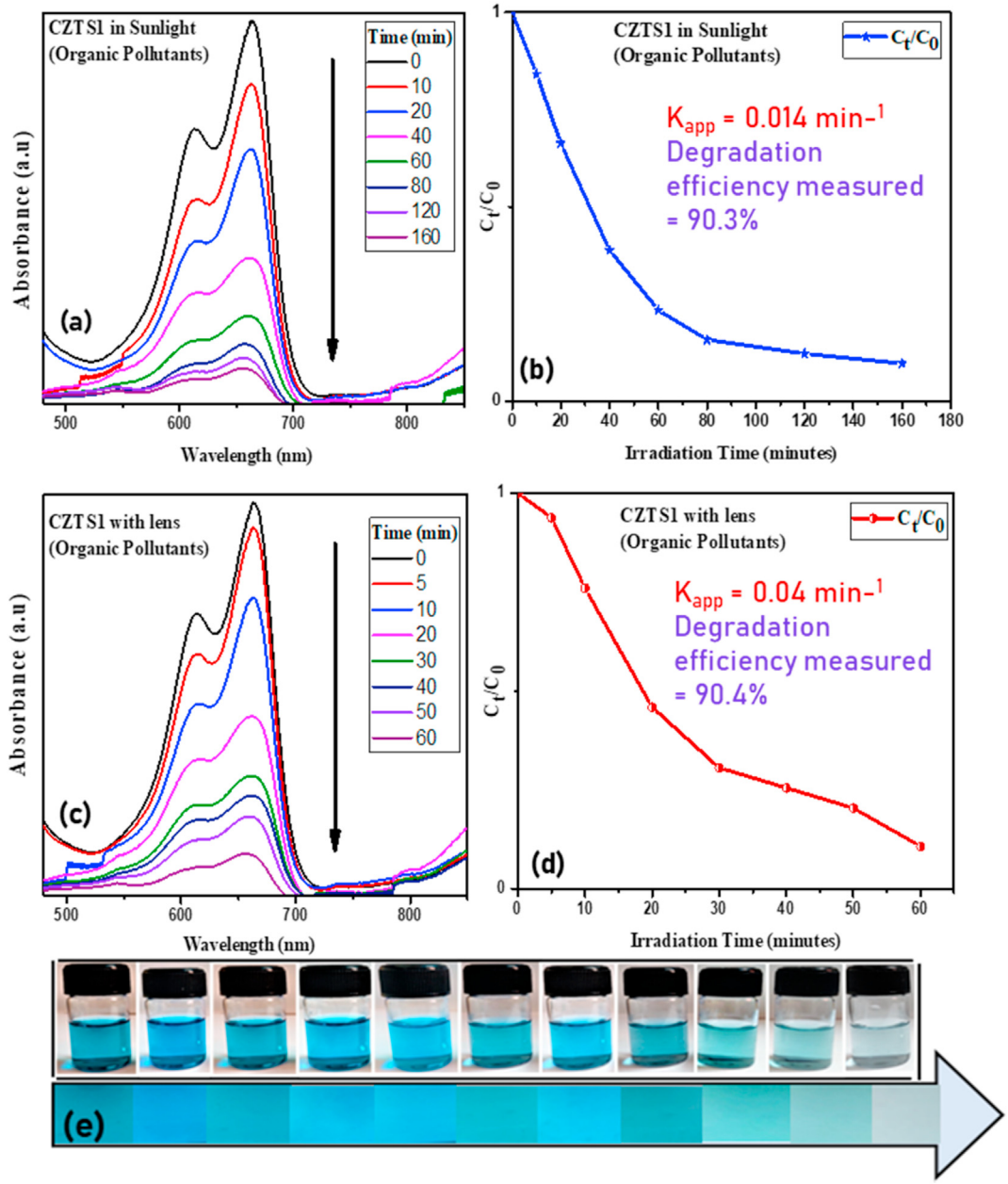


Fig. 10. Absorption spectra for CZTS1 (Organic Pollutants) (a) In Sunlight and (b) Its Concentration (C_t/C_0) Vs Time (c) With Lens and (d) Its Concentration (C_t/C_0) Vs Time (e) the corresponding color degradation bar;
Absorption spectra for CZTS2 (Organic Pollutants) (f) In Sunlight and (g) Its Concentration (C_t/C_0) Vs Time (h) With Lens and (i) Its Concentration (C_t/C_0) Vs Time (j) Color degradation bar.
Absorption spectra for CZTS2 (Textile waste); (k) In Sunlight and (l) Its Concentration (C_t/C_0) Vs Time (m) With Lens and (n) Its Concentration (C_t/C_0) Vs Time (o) the corresponding color degradation bar.

Organic Dye (MB) + $OH^{\cdot} \rightarrow$ Degradation products

Here Table 2 represents the overall comparison of the efficiencies and degradation constants with the time taken by different CZTS nanoparticles with different pollutants as reported by others along with our experimental data. As per the survey, for the work done by different groups in photocatalysis of CZTS nanoparticles, the minimum time reported by various groups is 45 min with 50% efficiency**,** while the maximum efficiency reported by them is 91% in 60 min of the degradation time. However, in our present work, the minimum degradation time is 30 min and 98% efficiency with the rate constant of 0.14 min^{-1} which is the fastest and most efficient reported till now without using Xenon Lamp or any other typical photocatalysis setup. Here, we haven't performed the photo-degradation of textile waste using CZTS1 because the percent degradation is slower than CZTS2 (in the case of Organic pollutants), and quite obviously the time-taken, efficiency**,** and the rate constant for textile waste using CZTS1 would be lower than CZTS2. However, CZTS1 shows comparable results with previously reported data by different groups.

From the above studies, we can infer that CZTS is a magnificent photocatalyst that gives excellent performance in degrading such

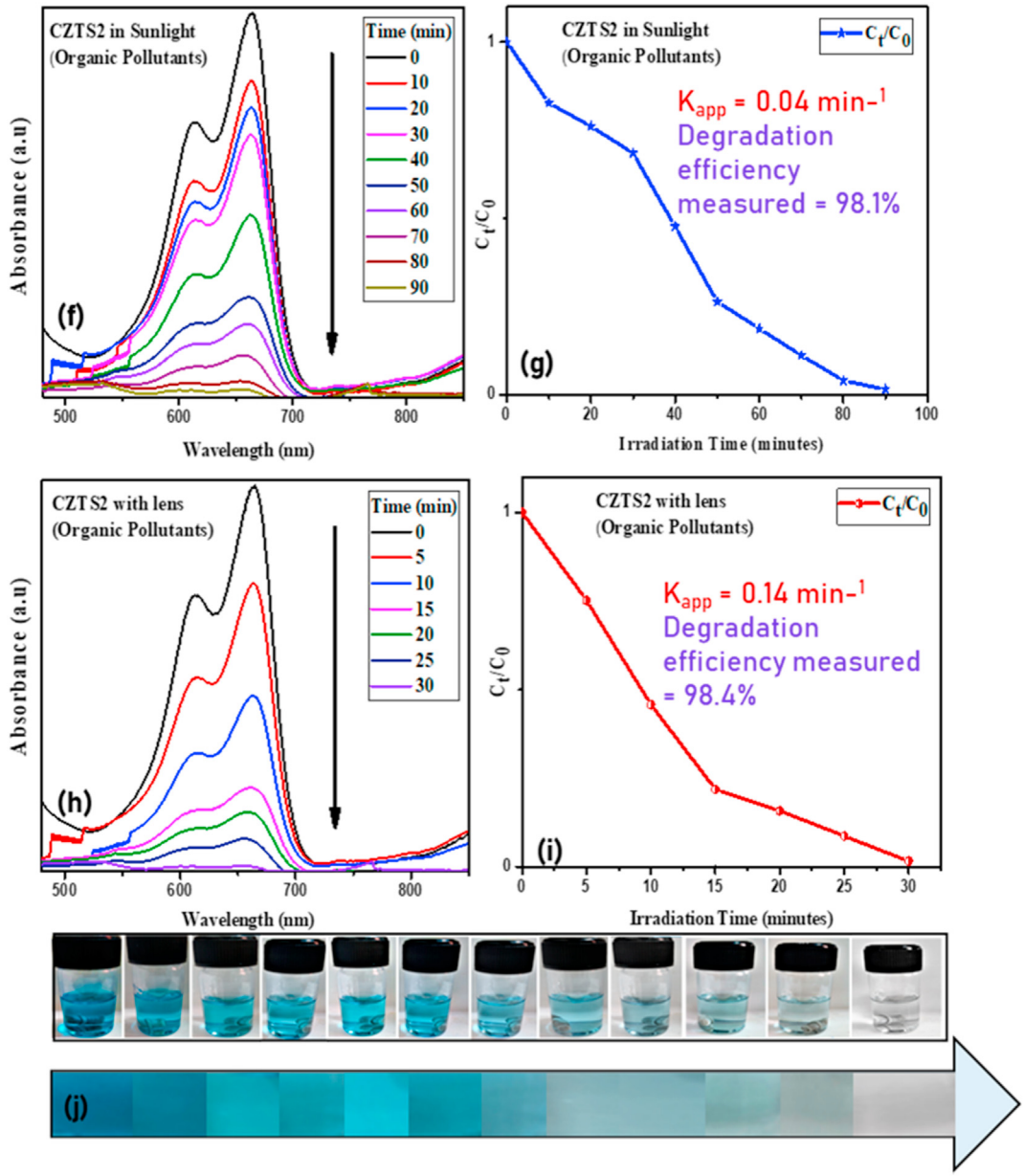


Fig. 10. (*continued*).

hazardous pollutants (such as metal salts e.g., sodium salts, chloride salts, sulfide salts, etc.). It is important to mention here that even with a high initial value of absorbance ~2.0 (CZTS + pollutant concentration) still, we were able to achieve excellent photocatalytic degradation efficiency up to 98.4% using our synthesized CZTS NP's for different pollutants without using any noble/transition metals. This is the highest efficiency recorded with such a large rate constant, which can be treated as a benchmark for other various light-harvesting and environmental applications.

5. Synthesis and growth mechanism of CZTS1 and CZTS2 NPs

The precursors of Zn, Cu, Sn are initially heated with TOPO at 250–300 °C, forming a metal complex. Afterward, elemental Sulfur, and TOP were injected into the metal complex resulting in the formation of $Cu_{2-x}S$ in the nucleation time, which is due to the reduction of Cu^{2+} to Cu^{1+} (by interaction with TOP-S complex). This formation of $Cu_{2-x}S$ (nanospheres) can be explained by its lowest solubility product constant, which remains stable below 100 °C. Thus, as the reaction proceeds, an increase in temperature results in high mobility; hence the exchange of metal ions, and diffusion of Sn^{4+} and Zn^{2+} cations take place at low energy which results in the CZTS compound. As the reaction proceeds, Zn cation infusion occurs, leading to the formation of rod-shaped CZTS NPs (CZTS1). The low binding energy of TOPO with CZTS compound also results in the growth of NPs along the c-axis of the kesterite crystal while in case of Butylamine (large binding energy with CZTS compound) holds strong attachment on the NPs, which in turn avoids the non-uniform growth in a particular direction and hence the NPs are spherical in the case of CZTS2. Here, the concentration of nuclei increases rapidly after precursor injection, which rises above the saturation concentration in minimum time. The NP grows rapidly with a burst of nucleation which forms many nuclei, lowers the concentration below the nucleation level, and allows the NPs to

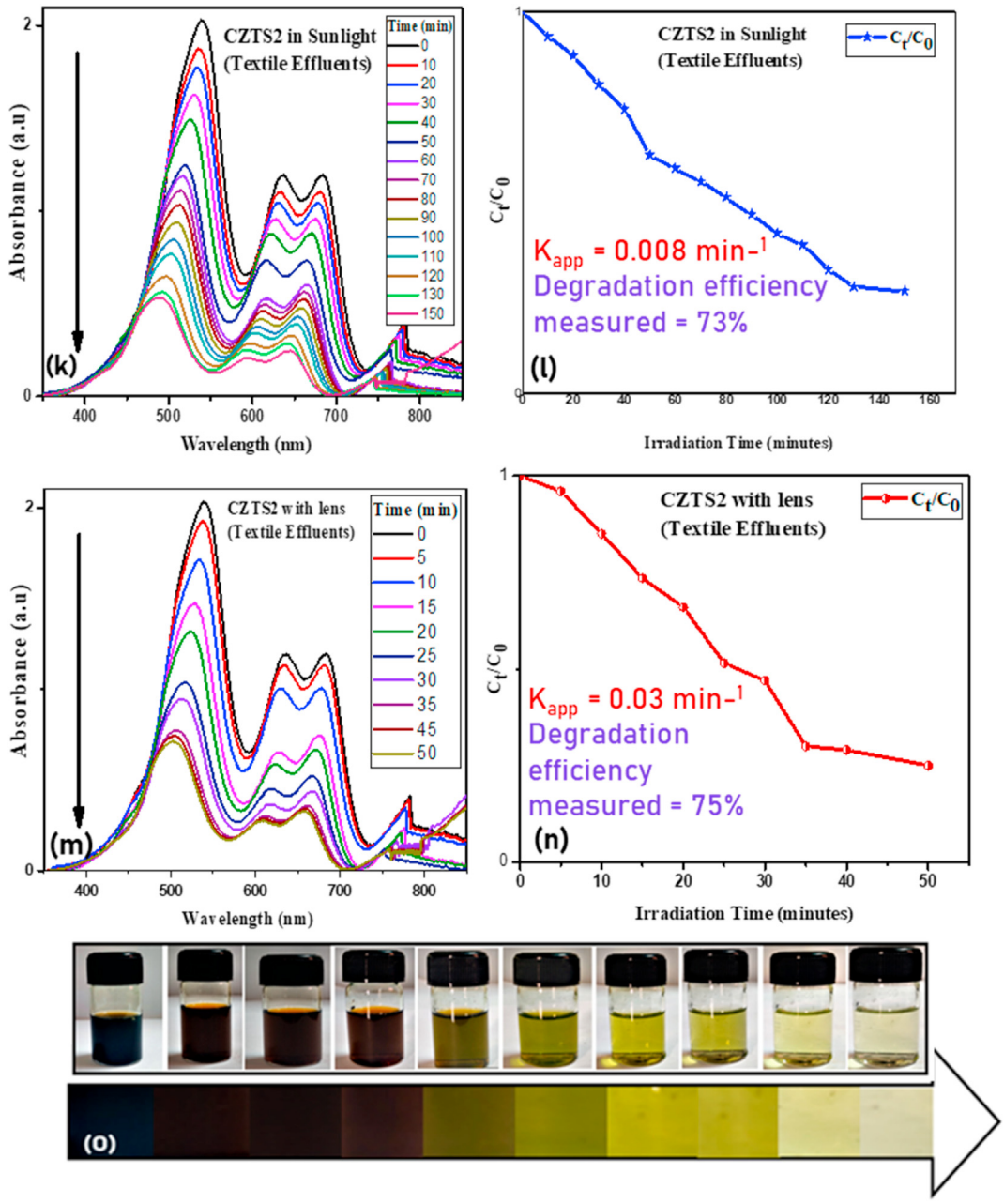


Fig. 10. (*continued*).

grow further. The nucleation takes place after some time with a constant concentration of monomers. As surface growth of NPs increases, the monomer concentration decreases eventually, and below critical level nucleation stops. Initially, the solution isn't in thermodynamic equilibrium but, allows the formation of large particles by taking small particles which reduces the surface energies and reaches thermodynamic equilibrium.

We can see Cu_2S peak in XRD and Raman spectra as well which is the binary phase of CZTS which can be attributed to the low binding energy of ligand (TOPO/TOP) to the CZTS compound which also result in the growth of NPs along c-axis, moreover, low miscibility of TOPO and TOP result in the formation of nanorods, while no such spurious peaks were found in CZTS2 because of the high binding energy of ligands (butylamine) with CZTS compound. Moreover, its high miscibility produces uniform spherical NPs, which again results in pure phase NPs of CZTS (here CZTS2).

Reactions Involved.

(1) **CZTS1**

$$CuCl_2 + ZnCl_2 + SnCl_2 + O{=}P(C_8H_{17})_3\ (TOPO) \rightarrow (Cu_XSn_YZn_W) = P(C_8H_{17})_3\ (Metal\text{-}Complex) \quad (1)$$

$$S + P(C_8H_{17})_3\ (TOP) \rightarrow S = P(C_8H_{17})_3\ (Complex) \quad (2)$$

$$Cu^{2+} \rightarrow Cu^{+}\ (Reduction\ process)$$

$$S = P(C_8H_{17})_3 + (Cu_XSn_YZn_W) = P(C_8H_{17})_3 \rightarrow Cu_{2-X}S + Zn^{2+} + Sn^{4+} + 2P(C_8H_{17})_3 \quad (3)$$

$$Cu_{2-x}S + Sn^{4+} \rightarrow Cu_2SnS_3 \quad (4)$$

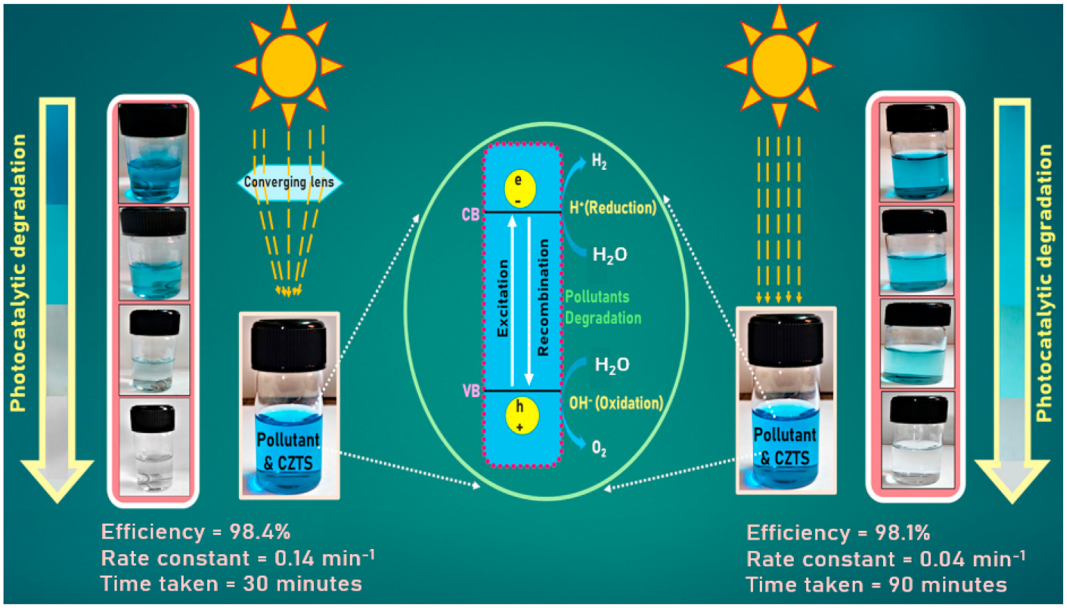


Fig. 11. A Graphical representation of photocatalytic efficiencies of CZTS2 (considered the best) NPs with and without Converging Lens under sunlight.

$$Cu_2SnS_3 + Zn^{2+} \rightarrow Cu_2ZnSnS_4 \quad (5)$$

(2) **CZTS2**

$$CuCl_2 + ZnCl_2 + SnCl_2 + C_4H_{11}N \text{ (n-Butylamine)} \rightarrow (Cu_XSn_YZn_W) = N(C_4H_9) \text{ (Metal Complex)} \quad (6)$$

$$S + H_2C{=}CH{-}C_{16}H_{33} \text{ (1-Octadecene)} \rightarrow S{=}CH{-}CH_2\text{-}(C_{16}H_{33}) \text{ (Complex)} \quad (7)$$

$$Cu^{2+} \rightarrow Cu^{+} \text{ (Reduction process)}$$

$$S{=}CH{-}CH_2\text{-}(C_{16}H_{33}) + (Cu_XSn_YZn_W) = N(C_4H_9) \rightarrow Cu_{2\text{-}X}S + Zn^{2+} + Sn^{4+} + C_4H_9{-}N{=}CH{-}CH_2{-}C_{16}H_{33} \quad (8)$$

$$Cu_{2\text{-}x}S + Sn^{4+} \rightarrow Cu_2SnS_3 \quad (9)$$

$$Cu_2SnS_3 + Zn^{2+} \rightarrow Cu_2ZnSnS_4 \quad (10)$$

5.1. Photocatalysis

The effect of morphology (Spherical and Rod-like) on the photocatalytic efficiency of the NPs can be understood from the fact that here, CZTS2 (nanospheres) shows greater efficiency than CZTS1 (nanorods) which can be attributed to their structures because the photocatalytic efficiencies follow the order of Sphere-like > Rods-like > Flakes-like structures [74,75]. CZTS can reduce metal salts such as sodium salts, chloride salts, sulfide salts, etc. Fig. 11 represents the overall study of CZTS NPs based photocatalysis with and without converging lens in Sunlight. As seen from the above table, upon employing lens has increased the rate constant for faster reaction and efficiency as well because the intensity in bare sunlight is (900 W/m^2). In contrast, upon employing a converging lens, the intensity is enhanced to (1800 W/m^2) which is almost equivalent to 2 Sun intensity. Moreover, the sunlight focussing was easily done by adjusting the clamp holding the lens.

6. Conclusion

The present work highlights the photocatalytic degradation of hazardous water pollutants especially the organic and industrial wastes using CZTS nanocrystals which were synthesized by the rapid hot-injection route (synthesis time ~ 45 min) using different ligands viz. TOPO/TOP and butylamine/n-octadecene. The as-synthesized CZTS nanocrystals were rendered hydrophilic upon ligand-exchange with L-cysteine for its application in photocatalysis. The photocatalysis experiments were performed under two conditions: (i) Under bare sunlight (Intensity ~ 900 W/m^2) (ii) focussing the sample under the sunlight via converging lens (1800 W/m^2). The photocatalytic efficiencies of the CZTS1 & CZTS2 nanocrystals prepared with different ligands (TOPO/TOP & Butyl amine/n-octadecene) were then compared and CZTS2's photocatalytic performance was found to be superior. CZTS2 exhibited the best photocatalytic efficiency ~98.4% (rate constant ~ 0.14 min^{-1}) under sunlight for organic pollutants and ~75% (rate constant ~ 0.02 min^{-1}) for industrial wastes via converging lens while the corresponding efficiencies with bare sunlight being 98.1% and 73% with rate constants as 0.04 & 0.008 min^{-1} respectively. The photocatalytic efficiency and rate constant values for CZTS1 under bare-sunlight and via converging lens were ~90.3%; 0.014 min.$^{-1}$ and ~90.4%; 0.04 min.$^{-1}$ for organic pollutants respectively. The superior phase purity as elucidated from the detailed structural and optical studies and thus enhanced photocatalytic performance of CZTS2 as compared to CZTS1 nanocrystals could be attributed to an efficient passivation rendered by butylamine ligand owing to its simple, short-chained linear structure as compared to the steric-hindrance to the passivation of CZTS1 nanocrystals imposed by the long chained, bulky TOPO/TOP ligand.

To the best of the author's knowledge, the present work is the first-hand information on the rapid and highly efficient photocatalysis of CZTS nanocrystals employing a converging lens. It is noteworthy to mention here that CZTS NPs are highly reproducible, with high crystallinity and purity. The photocatalytic experiments were found to be repeatable (tested for more than 6 times) with the least material usage (~200 μl of CZTS in 10 ml of water-pollutant), thus manifesting in the strong enhancement of visible light

absorption and efficient charge separation for an efficient photocatalytic device without resorting to the use of noble metals as co-catalysts. Furthermore, in a bid to increase the photocatalytic efficiency and the rate-constant, for the degradation of organic pollutants and industrial wastes, the employment of a converging lens can be done as a useful tool.

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.

Acknowledgment

The author (PS) sincerely acknowledges the Council of Scientific & Industrial Research (CSIR) and AcSIR for providing necessary facilities to carry out the experimental work and also thank CSIR for providing Junior Research Fellowship (#31/001(521)/2018-EMR-I).

References

[1] Han B, Wei W, Chang L, Cheng P, Hu YH. Efficient visible light photocatalytic CO2 reforming of CH4. ACS Catal 2016;6(2):494–7.
[2] Zhang K, Wang L, Kim JK, Ma M, Veerappan G, Lee C-L, et al. An order/disorder/water junction system for highly efficient co-catalyst-free photocatalytic hydrogen generation. Energy Environ Sci 2016;9(2):499–503.
[3] Wang L, Wang W, Sun S. A simple template-free synthesis of ultrathin Cu 2 ZnSnS 4 nanosheets for highly stable photocatalytic H 2 evolution. J Mater Chem 2012;22(14):6553–5.
[4] Wang X, Ruan Y, Feng S, Chen S, Su K. Ag clusters anchored conducting polyaniline as highly efficient cocatalyst for Cu2ZnSnS4 nanocrystals toward enhanced photocatalytic hydrogen generation. ACS Sustainable Chem Eng 2018;6(9):11424–32.
[5] Qian J, Chen Z, Sun H, Chen F, Xu X, Wu Z, et al. Enhanced photocatalytic H2 production on three-dimensional porous CeO2/carbon nanostructure. ACS Sustainable Chem Eng 2018;6(8):9691–8.
[6] Fan F-J, Wu L, Yu S H. Energetic I–III–VI 2 and I 2–II–IV–VI 4 nanocrystals: synthesis, photovoltaic and thermoelectric applications. Energy Environ Sci 2014;7(1):190–208.
[7] Thomas SR, Chen C-W, Date M, Wang Y-C, Tsai H-W, Wang ZM, et al. Recent developments in the synthesis of nanostructured chalcopyrite materials and their applications: a review. RSC Adv 2016;6(65):60643–56.
[8] Chawla V, Clemens B, editors. Effect of composition on high efficiency CZTSSe devices fabricated using co-sputtering of compound targets. 2012 38th IEEE Photovoltaic Specialists Conference. IEEE; 2012.
[9] Redinger A, Berg DM, Dale PJ, Siebentritt S. The consequences of kesterite equilibria for efficient solar cells. J Am Chem Soc 2011;133(10):3320–3.
[10] Li W, Tan JMR, Leow SW, Lie S, Magdassi S, Wong LH. Recent progress in solution-processed copper-chalcogenide thin-film solar cells. Energy Technol 2018;6(1):46–59.
[11] Chavda A, Patel M, Mukhopadhyay I, Ray A. Facile, noncyanide based etching of spray deposited Cu2ZnSnS4 thin films for secondary phase removal. ACS Sustainable Chem Eng 2016;4(4):2302–8.
[12] Thangaraju D, Karthikeyan R, Prakash N, Babu SM, Hayakawa Y. Growth and optical properties of Cu 2 ZnSnS 4 decorated reduced graphene oxide nanocomposites. Dalton Trans 2015;44(33):15031–41.
[13] Ansari MZ, Faraz M, Munjal S, Kumar V, Khare N. Highly dispersible and uniform size Cu2ZnSnS4 nanoparticles for photocatalytic application. Adv Powder Technol 2017;28(9):2402–9.
[14] Akpan UG, Hameed BH. Parameters affecting the photocatalytic degradation of dyes using TiO2-based photocatalysts: a review. J Hazard Mater 2009;170(2–3):520–9.
[15] Liu X, Shao Q, Wang Y, Zheng Y, Song H, Wang J, et al. One-pot in-situ microwave hydrothermally grown zeolitic imidazolate framework-8 on ZnIn-layered double oxides toward enhanced methylene blue photodegradation. Industrial & engineering chemistry research 2020.
[16] Ntuli F, Omoregbe DI, Kuipa PK, Muzenda E, Belaid M, editors. Characterization of effluent from textile wet finishing operations2009: world congress on engineering and computer science.
[17] Eisenberg DA, Yu M, Lam CW, Ogunseitan OA, Schoenung JM. Comparative alternative materials assessment to screen toxicity hazards in the life cycle of CIGS thin film photovoltaics. J Hazard Mater 2013;260:534–42.
[18] Coutts J, Hintze P, Meier A, Devor R, Surma J, Maloney P, et al., editors. Visible-light-responsive photocatalysis: Ag-doped TiO2 catalyst development and reactor design Testing2016: 46th international conference on environmental systems.
[19] Phaltane SA, Vanalakar S, Bhat T, Patil P, Sartale S, Kadam L. Photocatalytic degradation of methylene blue by hydrothermally synthesized CZTS nanoparticles. J Mater Sci Mater Electron 2017;28(11):8186–91.
[20] Casado C, Timmers R, Sergejevs A, Clarke C, Allsopp D, Bowen C, et al. Design and validation of a LED-based high intensity photocatalytic reactor for quantifying activity measurements. Chem Eng J 2017;327:1043–55.
[21] Yin S, Liu B, Zhang P, Morikawa T. Yamanaka K-i, Sato T. Photocatalytic oxidation of NO x under visible LED light irradiation over nitrogen-doped titania particles with iron or platinum loading. J Phys Chem C 2008;112(32):12425–31.
[22] Yu X, Liu J, Genc A, Ibanez M, Luo Z, Shavel A, et al. Cu2ZnSnS4–Ag2S nanoscale p–n Heterostructures as sensitizers for photoelectrochemical water splitting. Langmuir 2015;31(38):10555–61.
[23] Ganai A, Maiti PS, Houben L, Bar-Ziv R. Bar sadan M. Inside-out: the role of buried interfaces in hybrid Cu2ZnSnS4–noble metal photocatalysts. J Phys Chem C 2017;121(12):7062–8.
[24] Yang Y, Que W, Zhang X, Yin X, Xing Y, Que M, et al. High-quality Cu2ZnSnS4 and Cu2ZnSnSe4 nanocrystals hybrid with ZnO and NaYF4: Yb, Tm as efficient photocatalytic sensitizers. Appl Catal B Environ 2017;200:402–11.
[25] Chawla P, Singh S, Sharma SN. An insight into the mechanism of charge-transfer of hybrid polymer: ternary/quaternary chalcopyrite colloidal nanocrystals. Beilstein J Nanotechnol 2014;5(1):1235–44.
[26] Jain S, Chawla P, Sharma SN, Singh D, Vijayan N. Efficient colloidal route to pure phase kesterite Cu2ZnSnS4 (CZTS) nanocrystals with controlled shape and structure. Superlattice Microst 2018;119:59–71.
[27] Jain S, Verma S, Singh SP, Sharma SN. An electrochemical biosensor based on novel butylamine capped CZTS nanoparticles immobilized by uricase for uric acid detection. Biosens Bioelectron 2019;127:135–41.
[28] Yu X, Shavel A, An X, Luo Z, Ibanez M, Cabot A. Cu2ZnSnS4-Pt and Cu2ZnSnS4-Au heterostructured nanoparticles for photocatalytic water splitting and pollutant degradation. J Am Chem Soc 2014;136(26):9236–9.
[29] Babu KS, Reddy AR, Reddy KV. Controlling the size and optical properties of ZnO nanoparticles by capping with SiO2. Mater Res Bull 2014;49:537–43.
[30] Carrete A, Shavel A, Fontané X, Montserrat J, Fan J, Ibánez M, et al. Antimony-based ligand exchange to promote crystallization in spray-deposited Cu2ZnSnSe4 solar cells. J Am Chem Soc 2013;135(43):15982–5.
[31] Hoshino A, Fujioka K, Oku T, Suga M, Sasaki YF, Ohta T, et al. Physicochemical properties and cellular toxicity of nanocrystal quantum dots depend on their surface modification. Nano Lett 2004;4(11):2163–9.
[32] Yang Y, Kang X, Huang L, Pan D. Tuning the band gap of Cu2ZnSn (S, Se) 4 thin films via lithium alloying. ACS Appl Mater Interfaces 2016;8(8):5308–13.
[33] Chawla P, Jain S, Vashishtha P, Ahamed M, Sharma SN. Transition from CZTSe to CZTS via multicomponent CZTSSe: potential low cost photovoltaic absorbers. Superlattice Microst 2018;113:502–9.
[34] Roelofs KE, Guo Q, Subramoney S, Caspar JV. Investigation of local compositional uniformity in Cu 2 ZnSn (S, Se) 4 thin film solar cells prepared from nanoparticle inks. J Mater Chem 2014;2(33):13464–70.
[35] Srinivas B, Kumar BG, Muralidharan K. Stabilizer free copper sulphide nanostructures for rapid photocatalytic decomposition of rhodamine B. J Mol Catal Chem 2015;410:8–18.
[36] Méndez-López A, Morales-Acevedo A, Acosta-Silva Y, Ortega-López M. Synthesis and characterization of colloidal CZTS nanocrystals by a hot-injection method. J Nanomater 2016:2016.
[37] Carter CB, Norton MG. Solid-state phase transformations and reactions. Ceramic materials: springer. 2013. p. 457–75.
[38] Salje EK. Crystallography and structural phase transitions, an introduction. Acta Crystallogr A: Foundations of Crystallography 1991;47(5):453–69.
[39] Tiong VT, Bell J, Wang H. One-step synthesis of high quality kesterite Cu2ZnSnS4 nanocrystals–a hydrothermal approach. Beilstein J Nanotechnol 2014;5(1):438–46.
[40] Jung HR, Shin SW, Gurav K, Suryawanshi M, Hong CW, Yang HS, et al. Phase evolution of Cu2ZnSnS4 (CZTS) kesterite thin films during the sulfurization process. Ceram Int 2015;41(10):13006–11.
[41] Kumar M, Dubey A, Adhikari N, Venkatesan S, Qiao Q. Strategic review of secondary phases, defects and defect-complexes in kesterite CZTS–Se solar cells. Energy Environ Sci 2015;8(11):3134–59.
[42] Abdellaoui A, Ghaffour M, Bouslama M, Benalia S, Ouerdane A, Abidri B, et al. Structural phase transition, elastic properties and electronic properties of chalcopyrite CuAlX2 (X= S, Se, Te). J Alloys Compd 2009;487(1–2):206–13.
[43] Zhang X, Xu Y, Zhang J, Dong S, Shen L, Gupta A, et al. Synthesis of wurtzite Cu 2 ZnSnS 4 nanosheets with exposed high-energy (002) facets for fabrication of efficient Pt-free solar cell counter electrodes. Sci Rep 2018;8(1):1–13.
[44] Xu L-C, Wang R-Z, Liu L-M, Chen Y-P, Wei X-L, Yan H, et al. Wurtzite-type CuInSe 2 for high-performance solar cell absorber: ab initio exploration of the new phase structure. J Mater Chem 2012;22(40):21662–6.
[45] Kong J, Zhou Z-J, Li M, Zhou W-H, Yuan S-J, Yao R-Y, et al. Wurtzite copper-zinc-tin sulfide as a superior counter electrode material for dye-sensitized solar cells. Nanoscale research letters 2013;8(1):1–5.
[46] Li M, Zhou W-H, Guo J, Zhou Y-L, Hou Z-L, Jiao J, et al. Synthesis of pure metastable wurtzite CZTS nanocrystals by facile one-pot method. J Phys Chem C 2012;116(50):26507–16.
[47] Qiu L, Xu J, Cai W, Xie Z, Yang Y. Fabrication of Cu2ZnSnS4 thin films by microwave assisted sol-gel method. Superlattice Microst 2019;126:83–8.
[48] Sampath M, Sankarasubramanian K, Archana J, Hayakawa Y, Ramamurthi K, Sethuraman K. Structural, optical and photocatalytic properties of spray deposited Cu2ZnSnS4 thin films with various S/(Cu+ Zn+ Sn) ratio. Mater Sci

Semicond Process 2018;87:54–64.
[49] Sun L, He J, Kong H, Yue F, Yang P, Chu J. Structure, composition and optical properties of Cu2ZnSnS4 thin films deposited by Pulsed Laser Deposition method. Sol Energy Mater Sol Cell 2011;95(10):2907–13.
[50] Wang Y, Huang Y, Lee AY, Wang CF, Gong H. Influence of sintering temperature on screen printed Cu2ZnSnS4 (CZTS) films. J Alloys Compd 2012;539:237–41.
[51] Jenkins R, Snyder RL. Introduction to X-ray powder diffractometry1996.
[52] Rajesh G, Muthukumarasamy N, Subramanian E, Venkatraman M, Agilan S, Ragavendran V, et al. Solution-based synthesis of high yield CZTS (Cu2ZnSnS4) spherical quantum dots. Superlattice Microst 2015;77:305–12.
[53] Liu W, Guo B, Wu X, Zhang F, Mak C, Wong K. Facile hydrothermal synthesis of hydrotropic Cu 2 ZnSnS 4 nanocrystal quantum dots: band-gap engineering and phonon confinement effect. J Mater Chem 2013;1(9):3182–6.
[54] Fontané X, Izquierdo-Roca V, Calvo-Barrio L, Alvarez-Garcia J, Pérez-Rodríguez A, Morante J, et al. In-depth resolved Raman scattering analysis of secondary phases in Cu-poor CuInSe 2 based thin films. Appl Phys Lett 2009;95(12):121907.
[55] Currier JM, Holland JM, Drescher K, Foy D. Initial psychometric evaluation of the moral injury questionnaire—military version. Clin Psychol Psychother 2015;22(1):54–63.
[56] Su Z, Sun K, Han Z, Cui H, Liu F, Lai Y, et al. Fabrication of Cu 2 ZnSnS 4 solar cells with 5.1% efficiency via thermal decomposition and reaction using a non-toxic sol–gel route. J Mater Chem 2014;2(2):500–9.
[57] Irkhina A, Levcenko S, Hinrichs V, Plate P, Unold T. Metal acetate based synthesis of small-sized Cu 2 ZnSnS 4 nanocrystals: effect of injection temperature and synthesis time. RSC Adv 2017;7(19):11752–60.
[58] Fernandes P, Salomé P. Da Cunha A. Study of polycrystalline Cu2ZnSnS4 films by Raman scattering. J Alloys Compd 2011;509(28):7600–6.
[59] Durant BK, Parkinson B. Photovoltaic response of natural Kesterite crystals. Sol Energy Mater Sol Cell 2016;144:586–91.
[60] Fowkes FM. Attractive forces at interfaces. Ind Eng Chem 1964;56(12):40–52.
[61] Krishnan A, Ali KR, Vishnu G, Kannan P. Towards phase pure CZTS thin films by SILAR method with augmented Zn adsorption for photovoltaic applications. Materials for Renewable and Sustainable Energy 2019;8(3):16.
[62] Riha SC, Parkinson BA, Prieto AL. Solution-based synthesis and characterization of Cu2ZnSnS4 nanocrystals. J Am Chem Soc 2009;131(34):12054–5.
[63] Kush P, Deori K, Kumar A, Deka S. Efficient hydrogen/oxygen evolution and photocatalytic dye degradation and reduction of aqueous Cr (VI) by surfactant free hydrophilic Cu 2 ZnSnS 4 nanoparticles. J Mater Chem 2015;3(15):8098–106.
[64] Wei SH, Zunger A. Band offsets and optical bowings of chalcopyrites and Zn-based II-VI alloys. J Appl Phys 1995;78(6):3846–56.
[65] Li R, Li C. Photocatalytic water splitting on semiconductor-based photocatalysts. Advances in catalysis, vol. 60. Elsevier; 2017. p. 1–57.
[66] Deng F, Zou J-P, Zhao L-N, Zhou G, Luo X-B, Luo S-L. Nanomaterial-based photocatalytic hydrogen production. Nanomaterials for the removal of pollutants and resource reutilization. Elsevier; 2019. p. 59–82.
[67] Albonetti S, Perathoner S, Quadrelli EA. Horizons in sustainable industrial chemistry and catalysis. Elsevier; 2019.
[68] Zhu M, Zhai C, Qiu L, Lu C, Paton AS, Du Y, et al. New method to synthesize S-doped TiO2 with stable and highly efficient photocatalytic performance under indoor sunlight irradiation. ACS Sustainable Chem Eng 2015;3(12):3123–9.
[69] Mirzaeifard Z, Shariatinia Z, Jourshabani M. Rezaei Darvishi SM. ZnO photocatalyst revisited: effective photocatalytic degradation of emerging contaminants using S-doped ZnO nanoparticles under visible light radiation. Ind Eng Chem Res 2020.
[70] Dilsaver PS, Reichert MD, Hallmark BL, Thompson MJ, Vela J. Cu2ZnSnS4–Au heterostructures: toward greener chalcogenide-based photocatalysts. J Phys Chem C 2014;118(36):21226–34.
[71] Hou X, Li Y, Yan J-J, Wang C-W. Highly efficient photocatalysis of p-type Cu2ZnSnS4 under visible-light illumination. Mater Res Bull 2014;60:628–33.
[72] Covei M, Perniu D, Bogatu C, Duta A. CZTS-TiO2 thin film heterostructures for advanced photocatalytic wastewater treatment. Catal Today 2019;321:172–7.
[73] Gao Y, Long F, Wang J, Zhang J, Mo S, Zou Z. Understanding the growth mechanism of wurtzite Cu2ZnSnS4 nanocrystals and the photodegradation properties. Mater Des 2017;123:24–31.
[74] Gołąbiewska A, Kobylański M, Zaleska-Medynska A. Fundamentals of metal oxide-based photocatalysis. 2018.
[75] Clament Sagaya Selvam N. Vijaya JJ, Kennedy LJ. Effects of morphology and Zr doping on structural, optical, and photocatalytic properties of ZnO nanostructures. Ind Eng Chem Res 2012;51(50):16333–45.

Pooja Semalti received the B.Sc. (Hons) degree in Physics from Delhi University, Delhi, India in 2013, and the M.Sc. degree in Physics from Chaudhary Charan Singh University, Uttar Pradesh, India, in 2015. Her research interest is study on energy conservation and photonics devices. She is presently working as research scholar in CSIR-National Physical Laboratory, Delhi, India.

Dr Vikash Sharma has received his doctoral degree in Applied Physics in 2017 from Delhi Technological University, Delhi, India. Presently he is working in the CSIR-National Physical Laboratory as Senior Technical officer. He has 17 years of research experience and his research area is focused on Soft-lithography, Polymer based Microfluidics networks and novel electroactive carbon nanocomposite materials and their applications towards the fabrication of biochips (Lab-on-a-chip). He has 60 publications (Research papers in SCI journals: 15; Patents: 1; papers in International/National conferences proceedings: 44) to his credit. He has also worked for Calibration of Biomedical devices such as Defibrillator Analyzer.

Dr. Shailesh Narain Sharma (Senior Principal Scientist, CSIR-NPL and Professor-AcSIR) did his graduation and post-graduation studies from University of Delhi and Indian Institute of Technology (IIT) Delhi respectively. He did his Ph.D. from Indian Association for the Cultivation of Science (IACS), Kolkata in 1994. He went to University of Barcelona, Spain for his post-doctoral fellowship. In 1997, he has joined as a scientist in National Physical Laboratory (NPL). His Areas of Research Interest are Synthesis & characterization of Semiconductor Nanoparticles for its possible application in photovoltaics, light-emitting diodes & sensors. His current research activities are perovskites, InP Quantum Dots and TiO_2-based Hybrid Composites. He has two Indian patents, and about 130 research papers in international journals and ~50 National and International conference publications to his credit.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

A review on deep learning based pneumonia detection systems

Dr Shailender Kumar
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

Pankaj Singh
Delhi Technological University
Delhi, India
singh.pankaj88848@gmail.com

Mohit Ranjan
Delhi Technological University
Delhi, India
mohit.ranjan.910@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395868

***Abstract*—Pneumonia is a disease in which small air sacs of lungs called alveoli are affected. Pneumonia is a life-threatening disease caused by bacteria, fungi, or viruses. Millions of children die every year due to pneumonia. Earlier, only highly experienced radiologists examined pneumonia, but recent development in Image Processing and Deep Learning automated detection of diseases is possible. Early detection of pneumonia can save many lives. Various data augmentation techniques make the model achieve high accuracy even without having large datasets. Multiple types of research have used different techniques like CNNs, Transfer Learning, and Ensemble Learning. One of the current disadvantages of pneumonia detection is not having high-annotated data sets and the patient's medical history. The paper presents the overview of literature currently available on pneumonia classification by using chest x-ray. After reviewing the topic, the review investigates a comparison of various automated systems' drawbacks and limitations.**

***Index Terms*—CNN, Pneumonia, Deep Learning, Ensemble Learning, Transfer Learning**

I. Introduction

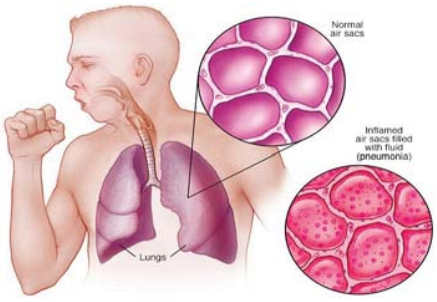


Fig. 1. Pneumonia in lungs
Source: https://www.mayoclinic.org

Pneumonia is a serve disease in which small air sacs of lungs called alveoli are affected. The leading causes of pneumonia can be either bacterial or viral infection. In this disease, the air sacs present in the lungs fill with fluids that lead to high fever, coughing, and breathing problems. Factors such as increased pollution and malnutrition increase the chances of having the disease. Pneumonia can be a very deadly disease for children and the elderly with a weak immune system. India has the second-highest death rate of children due to pneumonia. The UNICEF (United Nations International Children's Emergency Fund), in its report, found that 800,000 children died last year due to pneumonia. Infants under the age of two have a high mortality rate than others. It was also observed that more than 50% of children's deaths are from only five countries.

TABLE 1: PNEUMONIA CASES STATISTICS (2020)

Country	Child Deaths
Nigeria	162,000
India	127,000
Pakistan	58,000
Democratic Republic of Congo	40,000
Ethiopia	32,000

Source: United Nations International Children's Emergency Fund

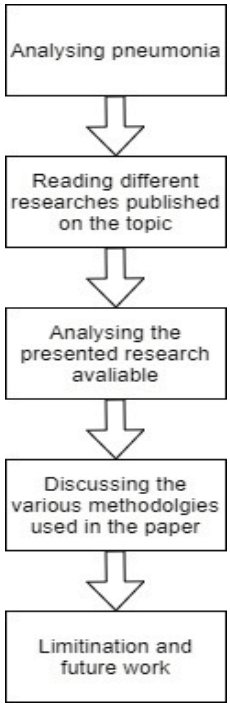
The table 1 shows the statistics of the pneumonia cases in different countries. All the countries mentioned in the table 1 are developing countries where expenditure on healthcare is very low. This is the official figure of the cases, due to poor healthcare system this number can be very big in actual. Most deaths due to pneumonia is in Nigeria with 162,000 deaths, the second most affected country is India with around 127,000 children death due to pneumonia.

One most common medical technique used to diagnose pneumonia is a chest X-ray. An expert radiologist examines the chest X-ray to find white spots present due to pneumonia fluid. The limitation of X-rays is the color scheme consisting of tints of white and black. It becomes challenging to determine if the spot is due to pneumonia fluid or not. Expert Radiologists need well-trained eyes to differentiate between the heterogeneous color distribution of air while flowing in the lungs. A human error in diagnosing will have a severe consequence on the patient. With the advancements in deep learning and machines with enormous computations, the new techniques and algorithms achieve the accuracy of the human eye accuracy. Neural Networks are mathematical models inspired by the human brain to store information that has been very successful in classifying and segmentation images. In the medical industry, deep learning (DL) can play an important part.

In the past, DL is used to detect tumors in the brain,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

breast cancer, and more recently, with the ongoing pandemic, researchers have used DL models to detect novel coronavirus (SARS-CoV-2) through X-rays and CT scans.
CNN (Convolutional Neural Network) is a Deep Neural Network that analyzes visual images and performs tasks like object classification, detection, and segmentation. Its input is an image that gives a class/classes that best describes the input image. In CNN, the input layer's information is an image, the hidden layers do the feature extraction, and the output layer gives the output to the system.
Many popular CNNs are available such as VGG16, VGG19, Resnet50, DenseNet201, EfficientNet, and many more. These are some of the best performing deep learning architectures. Many researchers use these models due to their high accuracy.
Transfer Learning is a method that stores knowledge of one problem and applies it to another related problem. The training of these models has a training dataset of the ImageNet dataset
Due to the less availability of medical data, data augmentation techniques are becoming very important. Data augmentation techniques help in avoiding the problem of overfitting by adding variance to the image datasets. Better data augmentation for medical imaging is an all-new research topic.
Zeshan Hussain[1] optimized the data augmentation methods; he achieved an accuracy of 88% compared to 68%. In the mixup [2] technique, a linear combination of images from the input dataset is the mix to add variation to the dataset images. Mixmatch[2] is mixing done on the amount of foreground label. The mixup and mix-match techniques outperform the other methods.
U-Net [3], SegNet [4], and CardiacNet [5] are some of the leading architectures for medical image analysis. These models are specifically designed for medical purposes, as medical data requires very fine-tuning. Also, the error in the medical examination can be critical. Much work is done in medical image segmentation and analysis, and a lot of research is still going on.
Rajpurkar et al. [6] proposed a CheXNeXt CNN model, to detect 14 different diseases, including pneumonia. Ke et al. [7] used the spatial distribution of Hue, Saturation, and Brightness in the X-ray image data set. Brightening the image improved the accuracy of the approach. The multi-purpose NSGANetV1[8] algorithm produces elements to measure all Pareto before simultaneously using genetic functions to reproduce and transform all construction components gradually. By carefully reducing the architecture and creating the emerging patterns allocated to the previous successful architecture using the Bayesian Network, operator distribution estimates during NSGANetV1 improve search efficiency. In many cases, the architecture we have found is beneficial by hand and by other architectures that are automatically designed for various data sets.
The conceptual flow of this paper is as follows.



The primary purpose behind writing this review paper is analysis and comparison of various methodologies related to the autonomous detection of pneumonia disease from X-Ray images of the chest. We will also compare transfer learning algorithms on different models like VGG16, ResNet, and DenseNet. This paper is designed to have a vast review of the topic, this will help the other researchers to analyse different methodologies. The merits and demerits will help researchers to apply these methodologies to different diseases classification and detection.
This paper consists of multiple sections: Section II is about related work; Section III is about findings of the review it has following sub-sections CNNs, transfer learning, different pre-trained models, ensemble learning, and Performance matrices; Section IV contains description of dataset available; Section V provides the conclusion and discusses future scope; In the last all References are mentioned.

II. RELATED WORK

A. "Transfer Learning with Convolutional Neural Network (CNN) for Pneumonia Detection using Chest X-rays"[9]

In this approach, four pre-trained Convolutional Neural Networks (CNNs): SqueezeNet, AlexNet, DenseNet201, and ResNet18 were chosen for the process of transfer learning. 5,247 images of viral, bacterial, and healthy chest X-rays were used to train the models after applying different data augmentation techniques.
CNN's perform pretty well upon training with a large dataset, as compared to a smaller one; therefore, transfer learning is used to apply the learnt knowledge on a relatively smaller dataset.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

Three different comparisons were made in the study: between pneumonia and non-pneumonia, between bacterial pneumonia, viral pneumonia, and non-pneumonia. In each of the comparisons, the dataset was divided according to the cases considered, and the results were summarized.
Before training the models, images were preprocessed and augmented. They were scaled down to 224*224 pixels for DenseNet201 and ResNet18, and to 227*227 pixels for SqueezeNet and AlexNet. After resizing and normalizing the images according to the model standards, the dataset was then augmented using three techniques: Rotation (Images were rotated by 45 degrees counterclockwise, and the area consisting of no pixels was filled with solid color), Scaling (Images were magnified by 10%) and Translation (Images were shifted horizontally by 10% and vertically by 10%).
In all the experiments, DenseNet201 outperformed all the other models.
Limitation: This research only uses few data augmentation techniques and due to small dataset, the models may be overfitted on the dataset. Also, For faster computation hyperparameters can be optimized.

B. "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks"[1]

In this approach, different data augmentation strategies were applied on a dataset consisting of mammograms, which had 1650 images consisting of mass cases and 1651 non-mass cases. Data augmentation becomes a necessity in medical imaging classification because of the scarcity of images required to train the CNNs for providing reasonable accuracy.
Although there has been much research on different augmentation strategies for natural imaging tasks, comparatively lesser work has been done in the case of medical imaging, therefore in this work, the authors have applied eight different augmentation techniques on the available dataset and then trained eight VGG-16 nets independently on the uniquely augmented sets. The results were then summarized, and the training and validation accuracy of mass/non-mass mammogram classification were compared.
For each image I in the dataset, the preprocessing included cropping the full-sized image to 1000*1000 pixels, and then dividing them among training and validation sets.One of the eight augmentation techniques were performed on each image in the training set, and then the VGG-16 nets were trained on the augmented sets.
Every image that contained mass was cropped around the mass lesion, whereas in normal images, a random 1000 * 1000 pixel crop of breast tissue was taken. Then, approximately 80% of the images were placed in the dataset for training, and remaining were placed in the set for validation.
Eight different augmentation techniques were as follows:

- **Flips:** A vertical and a horizontal flip was applied to the images. In case of natural images, a vertical flip would not have been of much use, but a vertical flip of the mass tissue image still results in a realistic image.
- **Gaussian Noise:** An array of numbers, A, was generated, in which every number has a value from a gaussian distribution that has mean=0 and variance lies in [0.1,0.9].
  A noisy image was obtained by following operation:
  $$I_0 = I + A$$
- **Jitter:** For every image I, a small amount of contrast ($\pm$1-4 intensity values) was added to it.
- **Scaling:** Each image I was scaled either horizontally or vertically; or in other words, an affine transformation was applied
  $$\begin{pmatrix} S_x & 0 \\ 0 & S_y \end{pmatrix}$$
  on every image I.
- **Powers:** Each image I was taken to a power. To calculate the power, p, the following equation was used, $p = n * r + 1$, where n is a random value taken from a Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$, while $r < 1$. Then, to generate the augmented image, I_a, we have, $I_a = sign(I) * (|I|^p)$. The sign and power are both taken elementwise.
- **Rotations:** The following affine transformation, $A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ where is between 10 and 175 degrees, was applied.
- **Gaussian Blur:** Each image I was blurred by a gaussian distribution function that had a variance between 0.1 and 0.9. The filter size is then generated internally by scikit-image, where the radius of the kernel is, $r = 4 * \sigma$.
- **Shears:** Finally, each image I was sheared, represented by the following affine transformation, $A = \begin{pmatrix} 1 & S \\ 0 & 1 \end{pmatrix}$.sdefines the amount that the image is sheared, and it is in the range of [0.1, 0.35].

TABLE 2: ACCURACIES OF DIFFERENT AUGMENTATION TECHNIQUES

Augmentation type	Training accuracy	Validation accuracy
Noise	0.625	0.660
Gaussian Filter	0.870	0.881
Jitter	0.832	0.813
Scale	0.887	0.874
Powers	0.661	0.737
Rotate	0.884	0.880
Shear	0.891	0.879
Flips	0.830	0.842

C. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning"[6]

Chest X-ray images of pneumonia disease often overlap with other diseases that cause variability in the result of pneumonia detection even with a radiologist's help. This

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

research paper proposed a novel model ChexNet, ChexNet model is described as consisting of 121 layer Dense Convolutional Network (DenseNet), in which Chest X-rays are fed and the output is in the form of probability of pneumonia and heatmap localizing the area of the image that has high probability of indication of pneumonia. This model is trained on Chest Xray 14 dataset consisting of 112,120 frontal view chest X-ray, in which each image is labeled with 14 different labels of thoracic pathologies according to radiology reports. Dataset is rearranged into training, validation and test sets with 98637, 1672 and 6351 images in each set respectively with no overlapping of images. Before feeding the images as input they were downscaled to 224*224 pixels. Normalization of the images was done on the basis of their mean and standard deviation. In DenseNet the fully connected final layer was replaced with one having a single output, and after that a sigmoid nonlinearity was applied. The network's weights were then initialized with those of a model that was pre-trained on ImageNet Dataset. This research paper also compared the F1-score of radiologist and the model proposed. The difference in F1 scores between them is 0.051(95% CI 0.005, 0.084). Four labels form radiologist and one from model were taken to perform the comparison, and in each one of them, the performance of CheXNet was significantly higher than that of the radiologists.

There are some limitations of this comparison, and these are:

- It provides conservative estimation of performance because only frontal view was presented to radiographs.
- Radiologist diagnostic performance decreased because in interpreting chest radiographs, they were not allowed to use patient history.

D. "A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images"[8]

This study aimed to propose an innovative approach based on transfer learning. Five different pre-trained models were combined, and their performance was analyzed. Due to the small size of the dataset consisting of chest X-rays, the models were highly likely to overfit. Therefore, to prevent it, some distortion in the image dataset was introduced to the dataset's images, resulting in significant improvement in the performance. Here, the noise also acted as a kind of dataset augmentation technique. All the models used in the study have input image size 224*224. First of all, the images resized to 224 × 224 × 3 pixels, then three different approaches: random resized crop, random flipping (horizontal), and augmenting images with varying intensities were used. Ensemble Classification was then used to combine the predictions of all those neural networks. Then a prediction vector was formed by combining the output of the networks. Then majority voting that combined the predictions from various other models was used for final prediction. Two approaches to majority voting prediction for classification were hard voting and soft voting. In hard voting, it sums all the predictions for every label, and then predicts class labels with the majority of votes, while in soft voting, it sums all the predicted probabilities for every label class and then predicts the class label that has the largest probability.
Limitation: One of the limitations found in this research is the unavailability of patients' medical history and smaller datasets. A large and varied dataset containing different images will give more accuracy.

III. FINDING OF THE REVIEW

A. Convolution Neural Network

A convolutional neural network(CNN) is a type of neural network to work on image data. The CNN architecture has many layers in it. There is a layer for input and output called the input layer and output layer, respectively. Apart from these, there are many intermediate layers such as the CLs(convolutional layers), PLs(Pooling layers), and FLs(fully connected layers). The input layer has the image as input. The image can be 2D or 3D depending whether the image is grey scale or color. As the name, CNN indicates that the convolution layer is an important and basic part of the CNN . The objective of convolution layer is to derive the peculiarities from the input image. It contains different filters(kernals) for additional feature extraction. The convolution operation is applied to it. Different stride value is used to have better result while passing through the convolution layer. This layer typically reduces the image size to preserve the size and information so padding is used in the convolution layer.
Pooling layer is used to reduces the complexity of the image by further extracting the domination features in the image. MaxPool and MinPool are 2 types of pooling operation.
FCL is the layer which takes images from CL or PL and process the results to used it for classification. It contains an activation function in this layer.
Output Layer is the layer at which does the final classification. For medical purpose the most common classifier used is the SoftMax classifier. Fig. 2 shows the overview

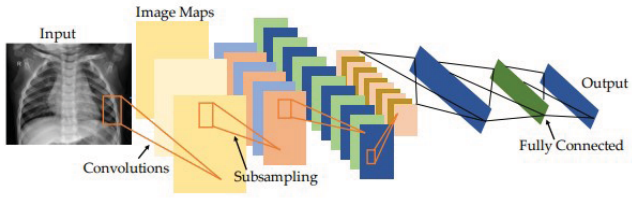


Fig. 2. Architectural overview of CNN Source:[9]

architecture of the CNN. The input to the CNN is the Chest XRay. The convolution layers extracts the features and image maps are created. The subsampling shown in the figure is the pooling layer where important features are derived. The resultant matrix from PL is passed to FCL, from which it is passed to the output layer.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

B. Transfer Learning

Transfer Learning is a DL technique that stores knowledge of one problem and applies it to another related problem. Transfer learning improves learning by transferring knowledge from a related similar job that has already been known. A model which trained on a dataset and is used to sove for another dataset is called pre-trained model. Some of the popular pretrained CNNs are VGG, InceptionNet, DenseNet, ResNet, MobileNet, and U-Net. These models have many versions of their own. The pre-trained models applied for biomedical image analysis are trained on the imagenet dataset, a vast dataset 14million images and a thousand classes. Due to such a vast dataset, the models have high accuracy in image processing. Using fine-tuning, these models are used on the medical dataset. Fig.3 shows

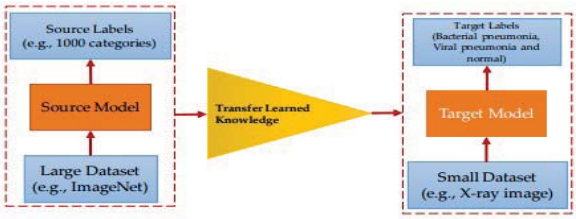


Fig. 3. Transfer Learning Source:[9]

how the pre-trained model source is trained on the ImageNet dataset (a huge dataset). The knowledge gained from that training is used to solve the desired problem. In this case, it is used to solve the pneumonia detection, so it is trained on the X-ray dataset.

C. Pre-trained models

1) VGG16: VGGs are object detection models, it has 2 versions VGG16(having 16 layers) and VGG19 (having 19 layers). VGG16 is one of the most popular pre-trained CNNs. Most of the recent research papers use VGG16 because of its simple design and very high accuracy. This model achieves 92accuracy in ImageNet Dataset. It is designed explicitly for Image Recognition. VGG16 has convolution layers of 3X3 configuration of filters with stride=1 and padding=" same." The advantages of using such a structure for this filter are:(1) We can use a combination of non-linear rectified units to make the decision function more distinctively accurate. (2) Drastic reduction in the number of training parameters or weights. We use the activation function, Rectified Linear Unit (ReLu), for these convolutional layers so that no negative value gets passed onto the next layers. It has pooling layers with Max Pool function, and the filter size is 2X2 with stride=2. This network is significantly deep, having about 140 million (approx.) parameters. In the end, the output layer uses a softmax classifier.

2) InceptionNet: InceptionNet has many versions. InceptionV1 is known as GoogleNet. It is a CNN which has 27 layers deep. InceptionNet has inception layers. The inception

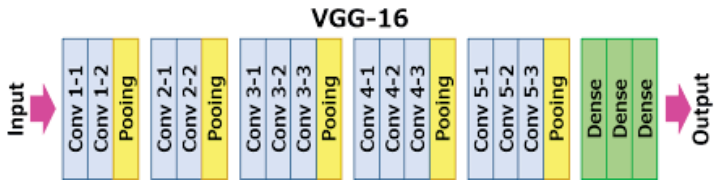


Fig. 4. VGG-16 Architecture

layer combines 3 Convolutional 3 layers of 1*1,3*3, and 5*5. The output of all the layers is incorporated into a single output vector forming the next stage's input.

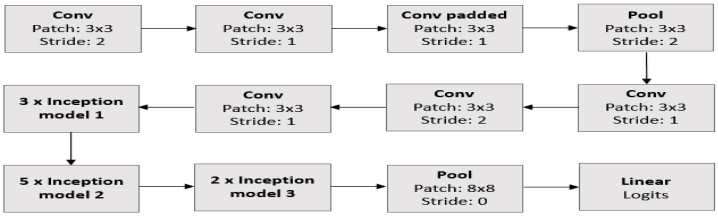


Fig. 5. InceptionNet Architecture

3) DenseNet: DenseNet is a Dense CNN; it has many versions, such as DenseNet169 and DenseNet201. In the DenseNet201 there are 201 layers and the input size of the images are 224*224. DenseNet201 is an improvement of ResNet that adds dense connections in layers. It has half the parameters compared to Resnet, but the accuracy on the ImageNet dataset is similar. Compared to traditional networks, the performance of DenseNet can increase by reinforcing feature propagation and reducing parameters, and encouraging feature reuse.

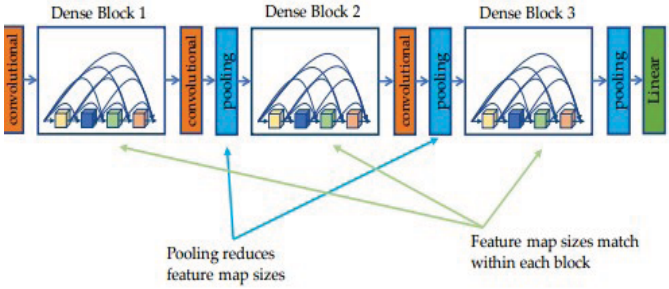


Fig. 6. DenseNet Architecture

4) ResNet18: ResNet originates from Residual Network (see figure given below). Resent was developed to solve two problems in CNNs: the vanishing gradient in which the gradient becomes 0 and no further update in weights takes place, and the second is the degradation problem. There are three different ResNet CNN are based on the parameters

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

such as input dimensions, activation function, no of layers, etc. ResNet is used in medical image segmentation and classification by using transfer learning techniques.

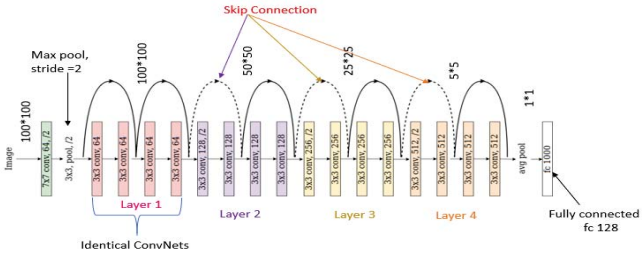


Fig. 7. ResNet18 Architecture

TABLE3: MODELS COMPARISON

Model	Layers (CL,PL,FL)	Para-meters	Size of Model (in MB)	Novelty	Limitations
AlexNet	5,2,3	61	233	Deep archi-tecture	Depth of the model is very less, unable to extract high features.
VGG16	13,5,2	138	527	good architecture for bench-marking on a particular task.	Takes very long time to train.
Inception Netv3	99,15,1	23.83	91	Factorising the n*n filters into 1*n and n*1, Factorising 7*7 convolution, Faster and efficient computation	Very complex architecture
DenseNet 201	200,5,1	20	80	Uses Dense Blocks, Compact-ness and non redundancy of features, Shows one of the best results in transfer learning.	Uses Dense Blocks, Compact-ness and non redundancy of features, Shows one of the best results in transfer learning.
Mobile Net	3CL,1PL, 7Bottle-necks,	3.4	14	Depth wise separable convolu-tions, very efficient mobile oriented model, Faster computation	Less accuracy when compared to VGG, densenet.

D. Ensemble Learning

It is a method in which many models are combined to improve the computational problem. Ensemble learning can be bragging and boosting.
The weighted Ensemble Learning models shows the most promising results. The weight given to each model can be the same, or it may be different. Assigning correct weightage to each model in ensemble learning is a topic of further research, but other statistical methods are used to set the weights. Once the weights are given, the result depends on the ratio of the weights assigned.
Many researchers verified that the ensemble learning model performs better than other models.

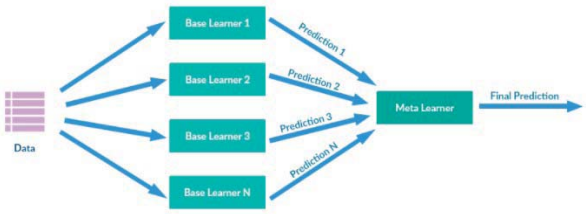


Fig. 8. Concept of Ensemble Learning

E. Performance metrices

Performance metrcices are the ways to evaluate, judge and compare the performance of the machine learning model. The most popular matrices is called the confusion matrix. It is a matrix that contains the predicted class and the actual class of the machine learning problem.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig. 9. Concept of Ensemble Learning

For pneumonia detection the dimensions of confusions matrix are:
False positive(FP) = Patient is normal but model predicted pneumonia. It is known as type I error.
True positive(TP) = Patient have pneumonia but model predicted pneumonia.
False negative(FN) = Patient have pneumonia but model predicted normal. It is known as type II error.
True negative(TN) = Patient is normal but model predicted normal.
Accuracy: It simply means how many predictions of model correct to all the prediction made by the model.
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
Precision: Precision means out of how many positives that model predicted are actually positive.
$Precision = \frac{TP}{TP+FP}$
Recall: Recall means out all of positives how many positive

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

model had predicted.
$Recall = \frac{TP}{TP+FN}$ Recall is more significant in medical problem because if the disease goes undetected it can cause harm.
F1-Score: F1-score contains both recall and precision. It is better to compare F1-score as recall and precision iversely depend on each other if we increase one, then the other one will decrease.
$F1 - Score = 2/(\frac{1}{Recall} + \frac{1}{Precision})$

IV. DATASETS

There are only a few datasets available of chest x-rays images of pneumonia.

A. Pneumonia DataSet Kaggle[11]

This dataset is available on Kaggle. It contains 5863 X Ray. It contains 1590 normal images and 4273 pneumonia images of chest Xray. The database has 3 folders in its train, test and val. Each of the 3 folders contains 2 subfolders of normal and pneumonia. These X Ray images of the chest are in anterior-posterior position because these are the most common for detecting the diseases. This dataset is collected from Guangzhou women and Children's Medical centre. All the images in this dataset are in JPEG format.

B. RSNA Pneumonia dataset[12]

This dataset was introduced in a competition on kaggle. The Radiological Society of North America(RSNA) is the organizer of the competition. This dataset has around 30287 XRay images out of which 9555 are pneumonia images and the rest 20672 images are the normal chest x ray images. MD.ai labelled the pneumonia cases in this dataset .The input size of the images are mostly 1024*1024 pixels. This dataset is collected by the NIH(National Institutes of Health). The dataset has csv files that have different labels 0 for normal, 1 for pneumonia.

C. NIH Chest X-ray Dataset[13]

The NIH clinical center holds this dataset which it made public in order to facilitate autonomous detection using xrays. It is the largest ChestX-ray dataset available. It contains 112,120 Chest X Rays of the front of chest. This dataset contains 14 different types of diseases. Size of this data is very large and it contains meta data in csv form describing details of patients such as age and sex. The different disease this dataset contains are: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural thickening, Cardiomegaly, Nodule, Mass and Hernia.The image sizes are different in this dataset. I

V. CONCLUSION AND FUTURE SCOPE

We conclude that chest radiography can plays an crucial role in testing and diagnosis of chest-related diseases. Various algorithms are available nowadays using different techniques. The paper presents the overview of literature pneumonia classification by using chest x-ray that is currently accessible. It reviews the topic and analyzes present algorithms in various aspects like usability, goodness, and complexity in computation associated with different techniques. Many datasets are available, but we observed that most of them are imbalanced, also only a few researchers are following balancing techniques.

In most of the cases, authors concluded results using very few images in dataset, most of them used under sampling. Therefore, we conclude that their results are unreliable, and suggest that their work should not be exploited on an industrial scale, because doing so will lead to danger to life of patients. Massive-scale data which is balanced and having a million pneumonia images should be used. Much research in using Machine Learning and Deep Learnings in the Medical industry is going on worldwide. Different versions of CNN's show promising results in detecting Pneumonia. We all know the shortage of expert radiologists, even when imaging machines are available. The automatic detection of Pneumonia will assist the expert radiologist, and this technique will be advantageous for developing countries which don't have proper medical facilities. If detection of Pneumonia is done timely, then it will save many lives. Studying and understanding the research work, we have found some limitations that are there.

One of the limitations of all the researches is that only a frontal view of the chest x-ray is available. All the datasets currently available only contain Posteroanterior (PA) of Anteroposterior (AP) X Rays, but research shows that about 15% of diagnosis requires a Lateral view. Hospitals and Clinics should try to provide more robust data, and it will help make the automation of the disease detection more accurate and efficient. One more limitation is not having access to the patient's medical history, and it decreases the performance of both radiologists and Deep learning algorithms. The deep learning methods do not give any reason as to how it concludes. With the use of Natural language with semantic segmentation, it should provide such an explanation. There is a need for an extensive and well-defined dataset and the patient's medical history for such a task.

Due to the less amount of medical data available, there are always very high chances that the algorithms try to overfit for the given problem. The problem with datasets is that they have few images and don't cover the variation of a specific disease, e.g., Pneumonia can be both bacterial and viral. A lot of research is going on how to optimize the hyperparameters and data augmentation for medical imaging. Several architectures like U-Net, NM-Net specially built for medical purposes, these models have shown promising results for various disease detection. We are in this paper combine many types of research work available on this topic. Different researchers have used other techniques and datasets.

TABLE4: COMPARISON OF DIFFERENT

Author	Approach	Dataset	Accuracy	Merits	Demerits
Pranav Rajpurkar [6]	121 Layer CNN with transfer learning [6]	Chest Xray14 dataset [13]	76.80	Heatmaps to visualize the area of disease in image, multiclass classification	Only frontal x-rays available, No patient history dataset available
Nada M.Elshennawy [14]	CNN, LSTMCNN, Resnet 152 V2, MobileNetV2 [14]	Pneumonia Dataset Kaggle [11]	99.22	Uses Novel LSTMCNN and CNN models, More epochs in training	Preexisting model's accuracy is more than custom CNN, Overfitting, no explanation of result

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

Author	Approach	Dataset	Accuracy	Merits	Demerits
Almubarok A. et al. [15]	mask RCNN +Adam + FPN, (deep ResNet) [15]	RSNA [12]	85.60	Uses a new mask RCNN model to detect pneumonia	Has a very low accuracy and recall value.
Khalid EL ASNAOUI [7]	Resnet50, DenseNet201, XCeption, MobileNetV2, VGG16, VGG19, InceptionV3, Inception-ResnetV2 [17]	Pneumonia Dataset Kaggle [11]	96.61	Uses advanced data preprocessing techniques such as intensity normalization and Contrast Limited Adaptive Histogram Equalization (CLAHE)	Confusion matrix shows comparatively high false negative ratio which is not desired in medical classification, Smaller dataset is used, No heatmaps for output result
Okeke Stephen [18]	CNN [18]	Pneumonia Dataset Kaggle [11]	95.31	Uses a custom CNN which is trained from scratch and has comparable accuracy to the pretrained state of are models, Sigmoid classifier is used.	The larger the size of the transformed images the lesser the validation accuracy, Prone to overfitting.
Tausif Ur Rahman [19]	AlexNet, ResNet18, DenseNet201, SqueezeNet [19]	Pneumonia Dataset Kaggle [11]	98.00	Activation maps for different models	No result justification, Limited dataset
Mohammad Farukh Hashmi [9]	Weighted Classifer of ResNet, DenseNet, InceptionNet, Xception, MobileNetV2 [9]	Pneumonia Dataset Kaggle [11]	98.43	Uses an ensemble model with weighted classifier a novel technique	Weighted classifier consumes more time, Results from the models requires radiologist to give justification
Deniz Yagmur Urey [20]	Resnet with Increased Contrast [20]	RSNA [12]	78.73	Uses different preprocessing on images for enhancing colors and brightness	Less accuracy when compared to other models
Vikash Chouhan [2]	Ensemble model of AlexNet, ResNet18, DenseNet121, InceptionNetV3, GoogLeNet [2]	Pneumonia Dataset Kaggle [11]	96.39	Uses Ensemble learning majority voting a novel technique	Scarcity of image data, model overfitting due to limited dataset, No explanation of results.
Dimpy Varshni [6]	DenseNet121, VGG16, VGG19 DenseNet169, XCeption, Resnet50 with SVM, KNN, Random Forest, Naive Bayes [16]	Chest Xray14 dataset [13]	80.02	Uses different classifiers Random Forest, SVM, KNN and Naïve Bayes with different pretrained models.	very large computations, Only frontal x rays available, No patient history dataset available

REFERENCES

[1] Zeshan Hussain1 , Francisco Gimenez2 , PhD, Darvin Yi2 , Daniel Rubin2 , MD, MS 1Stanford University, Department of Computer Science, Stanford, CA; 2Stanford University, Department of Radiology, Stanford, CA Differential Data Augmentation Techniques for Medical Imaging Classification Tasks.

[2] Zach Eaton-Rosen Felix Bragman Sebastien Ourselin M. Jorge Cardoso Improving Data Augmentation for Medical Image Segmentation.

[3] R. Olaf, F. Philipp, and Bomas, U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI Springer, New York, NY, USA, 2015..

[4] B. Vijay, K. Alex, and C. Roberto, "Segnet: Deep convolutional encoder-decoder architecture for image segmentation," 2015, http://arxiv.org/abs/1511.00561.

[5] M. Aliasghar, K. Rashed, R. Kawal, B. Jeremy, and B. Ulas, Cardiacnet, Segmentation of Left Atrium and Proximal Pulmonary Veins from MRI using Multi-View CNN, MICCAI, Springer, New York, NY, USA, 2017.

[6] Pranav Rajpurkar * 1 Jeremy Irvin * 1 Kaylie Zhu 1 Brandon Yang 1 Hershel Mehta 1 Tony Duan 1 Daisy Ding 1 Aarti Bagul 1 Robyn L. Ball 2 Curtis Langlotz 3 Katie Shpanskaya 3 Matthew P. Lungren 3 Andrew Y. Ng CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

[7] Ke, Q.; Zhang, J.; Wei, W.; Połap, D.; Wo ´zniak, M.; Ko´smider, L.; Damaševi ˇcius, R. A neuro-heuristic approach for recognition of lung diseases from X-ray images. Expert Syst. Appl. 2019, 126, 218–232.

[8] Vikash Chouhan 1 , Sanjay Kumar Singh 1 , Aditya Khamparia 1 , Deepak Gupta 2 , Prayag Tiwari 3 , Catarina Moreira 4 , Robertas Damaševi ˇcius 5,* and Victor Hugo C. de Albuquerque A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images.

[9] Tawsifur Rahman1, Muhammad E. H. Chowdhury2, Amith Khandakar2, Khandaker R. Islam3, Khandaker F. Islam2, Zaid B. Mahbub4, Muhammad A. Kadir1, Saad Kashem5 Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection using Chest X-ray.

[10] S.-H. Wang, S. Xie, X. Chen, D. S. Guttery, C. Tang, J. Sun, et al., "Alcoholism identification based on an AlexNet transfer learning model," Frontiers in psychiatry, vol. 10, 2019.

[11] Mooney, P. Chest X-ray Images (Pneumonia). Available online: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[12] Radiological Society of North America (RSNA). https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[13] NIH Chest X-ray dataset. https://nihcc.app.box.com/v/ChestXray-NIHCC

[14] Nada M. Elshennawy 1,† and Dina M. Ibrahim 1,2,,† Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images.

[15] In Proceedings of the International Conference of Artificial Intelligence and Information Technology (ICAIIT),Ouargla, Algeria, 4–6 March 2019; pp. 486–489.

[16] Dimpy Varshni Rahul Nijhawan Kartik Thakral Ankush Mittal Lucky Agarwal Pneumonia Detection Using CNN based Feature Extraction.

[17] Khalid EL ASNAOUI1, Youness CHAWKI2 , Ali IDRI1, 3 Automated Methods for Detection and Classification Pneumonia based on X-Ray Images Using Deep Learning.

[18] WASIF KHAN1, NAZAR ZAKI1, and LUQMAN ALI1 Intelligent Pneumonia Identification from Chest X-Rays: A Systematic Literature Review.

[19] Okeke Stephen1 Mangal Sain , 2 Uchenna Joseph Maduh , 3 and Do-Un Jeong An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare.

[20] Mohammad Farukh Hashmi 1,† , Satyarth Katiyar 2,†, Avinash G Keskar 3,† , Neeraj Dhanraj Bokde 4,† and Zong Woo Geem 5, Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning.

[21] Childhood pneumonia: Everything you need to know. https://www.unicef.org/stories/childhood-pneumonia-explained

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:37:11 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395767

A review on machine learning based student's academic performance prediction systems

Rahul Katarya
Computer Science Department
Delhi Technological University
New Delhi, India
rahuldtu@gmail.com

Jalaj Gaba
Computer Science Department
Delhi Technological University
New Delhi, India
jalajgaba12345@gmail.com

Aryan Garg
Computer Science Department
Delhi Technological University
New Delhi, India
aryangarg811@gmail.com

Varsha Verma
Computer Science Department
Delhi Technological University
New Delhi, India
varshaverma53@gmail.com

***Abstract*—Prediction of academic performance of students beforehand provides scope to universities to lower their dropout rate and help the students in improving their performance. In this field, research is being done to find out which algorithm is best to use and which features should be considered while predicting the academic performance of students. This kind of research work has been increasing over the years. This paper performs a survey on the techniques used in various research papers for academic performance prediction and also point out the limitations if any, in the methodology used.**

Keywords—Classification; Data Mining; Machine Learning; Prediction; Performance

I. INTRODUCTION

Education is the process of facilitating learning and acquiring skills, concepts, knowledge for the mental, physical, and spiritual development of an individual. An educated person is respected everywhere in the world and leads to the betterment of the society and prosperity of his nation. More importantly, it is necessary for his personal development.

Academics play a crucial part in education, and a student's academic performance in colleges and high school is an important criterion to judge a student's success. The academic performance comprises, in most cases, the grade point in each course studied and the cumulative grade point for the entire year (also known as CGPA). In the modern education system, it is the best and most suitable way known to judge one's caliber in academics. The grades that a student achieves greatly influence his chances of getting good colleges for higher studies and his job.

In the past few decades, there have been many attempts to get a prediction of a student's academic performance before the starting of his course, so that results can be predicted. This also becomes necessary as a student may like to know the areas where he lacks so that he can improve. Timely prediction of the future result may help an instructor know on which student to focus in particular and in what areas the improvement might be needed. An institute or the government would also like to know it to check the efficiency of its current education system and to make it better in the longer run.

Sachio Hirokava showed that academic performance depends on various factors such as demographic, behavioral, past results, habits, etc. [1] Sometimes unexpected factors like the place a student is living in might also produce a significant impact on it. There have been many kinds of research going on to tap these effects and the importance of different variables on a student's academic performance. These factors may not be the same globally, and each university and school may have individual differences when it comes to them.

In today's world, data has become very powerful, and Machine Learning has become very useful in harnessing the power of that data. Deep learning (DL) can be seen as an essential component of Machine learning which uses a computer model based on neural networks.[2]

Machine learning, along with deep learning techniques, has played a very pivotal part in the prediction of student's academic performance. There have been various Machine learning and deep learning methods like SVM, KNN, clustering, etc. that have been studied on various datasets belonging to different institutions to find hidden and unexpected patterns. Some new machine learning methods have also been developed and applied.

This paper has surveyed and analyzed the various work done in this field, and a comprehensive study is presented. The aim is to study the best variables used in prediction and the best

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:31:35 UTC from IEEE Xplore. Restrictions apply.

algorithms so further work can be done to improve the existing models and create a general system for prediction which can be globally applicable and is accurate.

This paper is divided as- Section II consists of the literature that we have reviewed along with some specific findings. Section III contains our contributions and Section IV contains the conclusion with the suggestion of future work followed by the references.

II. LITERATURE REVIEW

Students' CGPA is predicted by using ANN. The data used include the social and economic background of students. It also includes competitive exam results. Factors that impact students' performance are found out by statistical evaluations. The neural network has 11 input variables. Total Neuron layers- 2 hidden and 1 output [3]. One limitation is that attributes are not taken holistically. Other attributes like extra-curricular activities, online classes/courses can also be considered.

It was predicting at-risk students at an early stage using SGPA and CGPA to decrease the chance of student failure. It predicts the minimum required GPA for a particular student for each semester [4]. It helps in early identification, constant monitoring, and timely remedial action.

Students' scores are predicted using the Gartner Analytics Ascendancy model. The data used includes marks scored by students previously in a particular STEM course. It makes use of 8 Machine Learning algorithms in total. A student's academic performance is predicted in a particular course. Results are used to improve STEM education by developing appropriate tools [5]. The Linear regression algorithm predicted scores with the best accuracy, whereas the Naive Bayes classifier predicted with worst accuracy.

In [6], a novel framework MTLTR-APP is developed to predict students' academic performance. The multi-task learning framework is used. Behavioral data of students is used for prediction like sleep patterns.

Support Vector Machines are used to segregate those students who have low performance in academics. Students with the risk of failures are accurately identified. It uses SVM with a radial basis function kernel for training our model for each course [7]. It uses a dataset from 2 universities- University A (10 weeks courses), University B(12 weeks course).

Various Machine learning techniques are used to find out whether course grades of individual subjects or GPA should be used for predicting academic performance. By looking at the results it can be said that course grades should be used until the third term. After that, the GPA should be used [8]. This helps to determine which is to be used when.

Students' academic performance is predicted using a Multi-label ensemble model. It makes use of data-mining to improve students' academic performance. It predicts the performance of students in the next semester for different courses using different machine learning algorithms like SVM, RF, KNN, MLP to train the model. Label Powerset is used to further improve the prediction accuracy [9].

Predicting student's performance before the commencement of a course [10]. SVM, MIMLSVM, SISL-Circle, MIML Circle are used for the above purpose. More focus is put on traditional classroom teaching scenes, rather than online courses. MIML circle gave the best result.

New features like the number of days students have been absent are introduced, which led to an increase in accuracy of academic performance prediction by 10-15% compared to when such features were not used [11]. The dataset used contains 500 different records and 17 features. ANN performs better than other classification algorithms used. However, more psychological factors can be taken into account in this research, so as to get a more accurate analysis.

A hybrid classifier is used for predicting students' academic performance. It uses a predictive data-mining model to distinguish between fast and slow learners. Irrelevant inputs were eliminated using sensitivity analysis. It makes use of a hybrid classifier consisting of Fuzzy Artmap and Bayesian Artmap. Performance measures used include MCC, accuracy rate, TP, and FP [12].

The data that was analyzed belonged to students of public schools of brazil. Classification methods based on Gradient Boosting machines are used. 14 nominal and 3 numerical variables were considered. In particular, the variables - 'neighborhood' and 'school' had a major effect on student's performance. Attendance records also played a major role in predicting the performance [13].

The academic performance of students is predicted using several machine learning methods like gradient boosting classifiers and decision trees. Random forest and extreme gradient boosting classification techniques have also been used. [14]. The resultant accuracy achieved of the model is 95% and correlation heatmap is used to study relationships between various attributes. Here, multiclass classification is used instead of binary classification.

Prediction of students' academic performance using Ontological Modelling based on Fuzzy logic. It predicts academic performance in pervasive environments. It takes into account internal and external assessment and viva voce marks for prediction of the final grade. It is implemented in MATLAB [15]. Parameter Tuning can be applied to increase accuracy further.

The use of a new technique was developed by the researchers named MANFIS-S to predict student's performance[16]. MANFIS-S outperformed all other algorithms used for validation which were MANFIS, ANFIS, oneR, Random trees.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:31:35 UTC from IEEE Xplore. Restrictions apply.

Analysis of the factors which influence the students' academic performance prediction using Learning analytics. It uses data from 2 MOOCs on edX on Java programming [17]. The factors which influence this include previous grades, course duration, clickstream data, forum variables, etc.

Both classification and clustering algorithms are employed, and a hybrid model is developed for academic performance prediction. Therefore, techniques used include Support Vector Machine, Decision tree, clustering using K-means, Neural Networks, and Naive Bayes [18]. The results achieved by this hybrid model are better as compared to other models. However, more courses could have been analyzed and more complex algorithms could have been used.

Use of 2 datasets and 5 machine learning algorithms to predict the student's academic performance. Backpropagation and Long-Short Term Memory are used. In addition to this, BP, SVR, and Gradient Boosting Classifiers are used in classification. In conclusion, we see that SVR produced the best result. Even Backpropagation (BP) produced good results [19].

Three classification classifiers- KNN, Naive C4.5, and Bayes are used to create a student academic prediction model. Other factors, along with entrance examination scores like high school awards, admission area, etc. are also considered [20]. The result is that the KNN algorithm is better than C4.5 and Naive Bayes. One limitation could be that it is a binary classifier.

TABLE I. ANALYSIS OF VARIOUS RESEARCH PAPERS

S.No	Authors	Year	Dataset used	Methods used	Result and Remarks
1	V. L. Uskov, J. P. Bakken, A. Shah, A. Byerly	2019	Marks scored by students in assignments of STEM courses.	The Gartner Analytics Ascendancy Model.	Different accuracy- Linear Regression: 96.3%, SVM : 93.78%, RF: 93.41%, Decision tree:92.19%, KNN:69.99%, ANN : 67% and Naive Bayes:55.88% accuracy.
2	W. Nuankaew, J. Thongkam	2020	9500 students of Rajabhat Maha Sarakham University in Thailand	Naive Bayes, SMO, ANN, Random Forest, KNN, Partial decision trees, REPTree	Random Forest gave the best performance, with a precision of 94.70%
3	E. T. Lau, L. Sun, Q. Yang	2019	Chinese University- 1000 students	ANN	Accuracy: ANN -84.8% ,AUC-0.86. One limitation is the poor performance of ANN in classifying students according to gender.
4	R. Ravikumar, V. Akre, F. Aljanahi, A. Rajan,	2018	Dummy grade combinations for students.	Early warning system based on the GPA of students	Students who scored C grade or less than that from the 1st semester have a high chance of falling into risk.
5	Sachio Hirokawa	2018	University Data(480 students)	SVM	Using SVM, accuracy is 84.3% .The author can use more advanced machine learning techniques.
6	H. YAO, D. LIAN, Y. CAO, Y. Wu, T. ZHOU	2019	Data from one university during 2011/09/01 to 2015/06/30	Multi-task learning	The proposed method is very effective but there is a limitation in data collection.
7	S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, L. Porter	2019	Dataset used was obtained from 2 public universities in North America	SVM with radial basis function kernel.	The result improves on the addition of more data. It identifies at-risk students at the earliest point in time, which helps both students and teachers. It also helps in reducing the chance of student failure. The model has not considered courses related to other fields.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:31:35 UTC from IEEE Xplore. Restrictions apply.

8	A. E. Tatar, D. Düstegör	2020	357 students who were admitted to the CCSIT at IAU from Fall 2011 to Fall 2013	Naïve Bayes, logistic regression, random forest	Better accuracy is achieved using the 'course grade' variable if the prediction takes place before the third term. Otherwise in all cases using 'Grade Point Average' yields the best accuracy. The methodology could be applied on more datasets.
9	E. Admasu, A. Teklay	2019	Dataset was collected from 3 high schools of Ethiopia	Multi-label Ensemble model	SVM performs best in Hamming loss(22.4%), micro F1(83.3%), and macro F1(82.4%), Random Forest performs best in subset accuracy(34.4%), MLP performs best in Jaccard Similarity(72.3%). Partitioning using RAkEL gives better results as compared to a fast-greedy approach.
10	Yuling MA, Chaoran CUI , Jun YU, Jie GUO, Gongping YANG, Y. YIN	2019	Dataset of 1,020 university students from various computer-related courses	SVM, MIMLSVM, SISL-Circle, MIML Circle	MIML-Circle gave the best result. Model needs to be trained with a dataset with more training samples.
11	Sana, Isma Farrah Siddiqui, Qasim Ali Arain	2019	Kiteboard 360 e-learning system(500 records, 17 features)	Naive Bayes classifier, ANN, and Decision Tree	increase in accuracy of academic performance prediction by 10-15% by using new features
12	Roshani Ade	2019	Educational data set	Hybrid classifier consisting of Fuzzy Artmap and Bayesian Artmap.	Hybrid classifier is more accurate. The accuracy was 92% using Hybrid classifier whereas Fuzzy Artmap with 84% accuracy and Bayesian Artmap with 87% accuracy.
13	E. Fernandes, M. Holandaa, M. Victorinoa, V. Borgesa, R. Carvalhoa, G. V. Ervena	2018	School students of Brazil	Classification models that are based on gradient boosting machine	The variables ' neighborhood', 'school', and 'attendance record' influenced the result greatly.
14	Abhinav Jain, Shano Solanki	2019	UCI	Decision tree, gradient boosting classifier, random forest classifier, and extreme gradient boosting classifier	Random forest classifier Gave the best result. The resultant accuracy achieved of the model is 95%.
15	Ramanathan Veeraiyan, Sivakumar Ramakrishnan	2020	CIA marks, external marks, viva voce marks	Ontological Modeling based on Fuzzy logic	The final grade is directly proportional to the internal assessment and also external assessment whereas viva voce marks slightly influence the final grade.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:31:35 UTC from IEEE Xplore. Restrictions apply.

16	Le Hoang Son · Hamido Fujita	2018	Dataset of students from Vietnam National University and KDD cup datasets as well as the dataset from UCI.	MANFIS-S	MANFIS-S outperformed all the other algorithms used for validation which were MANFIS, ANFIS, OneR, and Random tree.
17	P. M. Moreno-Marcos, T.C. Pong, P. J. M. Merino, C. D. Kloos	2019	MOOC data is collected from edX.	Learning Analytics	A total of 6 models are made. MOOC data is considered. Forum variables were not useful in prediction whereas exercise variables were most useful in prediction. Clickstream data also gave useful predictions in the absence of exercises.
18	B. K. Francis, S. S. Babu	2019	Students' data from Kerala	SVM, Decision tree, K-means, Neural Networks, and Naive Bayes	The hybrid model has an accuracy of 0.7547.
19	B. Sekeroglu, Kamil Dimililer,Kubra Tuncal	2019	Student Performance Dataset and Student's Academic Performance Dataset	Backpropagation, SVR, and Long-Short Term Memory	SVR produced the highest R^2 and EV values
20	Xinyu Du	2019	Students' data from a Chinese University	KNN, Naive C4.5, and Bayes	KNN algorithm is better than C4.5 and Naive Bayes. There is a limitation in data collection due to which attributes may not be optimal

III. FINDINGS OF THE REVIEW

We have studied and analyzed the above research papers and have presented it in a structured format along with the summaries for future references. In our study, we have found out that the variables used for analysis in almost all researches do not include all the relevant factors. Thus, we need a new holistic approach so that we can capture those factors which may not be intuitive but would play a significant role in a student's life. Thus, better surveys are needed to be developed, and new datasets are needed to be generated for a more accurate analysis. We can include factors such as the amount of social media usage, sleeping habits, etc. A thorough analysis of student life is needed.

IV. CONCLUSION AND FUTURE WORK

The purpose of this paper is to find out the methods and techniques which are employed currently for predicting the academic performance of students and also observe the results achieved by them. The machine learning techniques and data mining has been used extensively in this area and provide future scope also to work with huge amounts of data. In the past few decades, numerous studies and researches have already been done on this subject, we have attempted to compare them and put the relevant studies in one place. We conclude that a lot of research can still take place to make academic performance prediction more accurate and globally applicable. Moreover, the practical application of such work is also required and it could be very beneficial for the betterment of our education system.

Many factors that are not usually intuitive but have a direct impact on student's performance need to be considered. More research is also needed in the area of analyzing a student's life from different perspectives and identifying more variables that are left out in the research that we have studied. There also needs to be a system so that an educational institute can directly tap onto the weak points of an individual using the predictions and can give him a more holistic treatment in those specific areas only. There could also be a classification of students based on similar characteristics. A student should also be able to know the areas improving in which would lead to the maximum growth in his performance. So that he could work optimally and more efficiently.

REFERENCES

[1] Sachio Hirokawa. 2018. Key attributes for predicting student academic performance. In Proceedings of the 10th International Conference on Education Technology and Computers (ICETC '18). Association for Computing Machinery, New York, NY, USA, 308–313. DOI:https://doi.org/10.1145/3290511.3290576

[2] W. Nuankaew and J. Thongkam, "Improving Student Academic Performance Prediction Models using Feature Selection," *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, Thailand, 2020, pp. 392-395, doi: 10.1109/ECTI-CON49241.2020.9158286.

[3] Lau, E.T., Sun, L. & Yang, Q. Modelling, prediction, and classification of student academic performance using artificial neural networks. *SN Appl. Sci.* **1,** 982 (2019). https://doi.org/10.1007/s42452-019-0884-7

[4] R. Ravikumar, F. Aljanahi, A. Rajan, and V. Akre, "Early Alert System for Detection of At-Risk Students,"

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:31:35 UTC from IEEE Xplore. Restrictions apply.

2018 Fifth HCT Information Technology Trends (ITT), Dubai, United Arab Emirates, 2018, pp. 138-142, doi: 10.1109/CTIT.2018.8649508.

[5] V. L. Uskov, J. P. Bakken, A. Byerly, and A. Shah, "Machine Learning-based Predictive Analytics of Student Academic Performance in STEM Education," *2019 IEEE Global Engineering Education Conference (EDUCON)*, Dubai, United Arab Emirates, 2019, pp. 1370-1376, doi: 10.1109/EDUCON.2019.8725237.

[6] Huaxiu Yao, Defu Lian, Yi Cao, Yifan Wu, and Tao Zhou. 2019. Predicting Academic Performance for College Students: A Campus Behavior Perspective. ACM Trans. Intell. Syst. Technol. 10, 3, Article 24 (May 2019), 21 pages. DOI:https://doi.org/10.1145/3299087

[7] Soohyun Nam Liao, Daniel Zingaro, Kevin Thai, Christine Alvarado, William G. Griswold, and Leo Porter. 2019. A Robust Machine Learning Technique to Predict Low-performing Students. ACM Trans. Comput. Educ. 19, 3, Article 18 (June 2019), 19 pages. DOI:https://doi.org/10.1145/3277569

[8] Tatar, A.E.; Düştegör, D. Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average? *Appl. Sci.* **2020**, *10*, 4967.

[9] Yekun, E., and Teklay, A., 2020. *Student Performance Prediction With Optimum Multilabel Ensemble Model*. [online] arXiv.org. Available at: <https://arxiv.org/abs/1909.07444> [Accessed 14 September 2020].

[10] Ma, Y., Cui, C., Yu, J., *et al.* Multi-task MIML learning for pre-course student performance prediction. *Front. Comput. Sci.* **14,** 145313 (2020). https://doi.org/10.1007/s11704-019-9062-8

[11] S. B., I. Farrah Siddiqui, y Q. Ali Arain, Analyzing Students' Academic Performance through Educational Data Mining, *3C Tecnología*, pp. 402-421, May 2019.

[12] Ade, R., 2020. *Students Performance Prediction Using Hybrid Classifier Technique In Incremental Learning | International Journal Of Business Intelligence And Data Mining*. [online] Inderscienceonline.com. Available at: <https://www.inderscienceonline.com/doi/pdf/10.1504/IJBIDM.2019.101255> [Accessed 14 September 2020].

[13] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., and Erven, G., 2020. *Educational Data Mining: Predictive Analysis Of Academic Performance Of Public School Students In The Capital Of Brazil*.

[14] A. Jain and S. Solanki, "An Efficient Approach for Multiclass Student Performance Prediction based upon Machine Learning," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, pp. 1457-1462, doi: 10.1109/ICCES45898.2019.9002038.

[15] Veeraiyan, R., and Ramakrishnan, S., 2020. [online] Ijaema.com. Available at: <http://ijaema.com/gallery/271-january-3311.pdf> [Accessed 14 September 2020].

[16] Son, L.H., Fujita, H. Neural-fuzzy with representative sets for prediction of student performance. *Appl Intell* **49,** 172–187 (2019). https://doi.org/10.1007/s10489-018-1262-7

[17] P. M. Moreno-Marcos, T. Pong, P. J. Muñoz-Merino, and C. Delgado Kloos, "Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics," in IEEE Access, vol. 8, pp. 5264-5282, 2020, doi: 10.1109/ACCESS.2019.2963503.

[18] Francis, B.K., Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J Med Syst* **43,** 162 (2019). https://doi.org/10.1007/s10916-019-1295-4

[19] Boran Sekeroglu, Kamil Dimililer, and Kubra Tuncal. 2019. Student Performance Prediction and Classification Using Machine Learning Algorithms. In Proceedings of the 2019 8th International Conference on Educational and Information Technology (ICEIT 2019). Association for Computing Machinery, New York, NY, USA, 7–11. DOI:https://doi.org/10.1145/3318396.3318419

[20] Du, X. (2019). A Prediction Model for High School Students' Academic Performance in College Based on Machine Learning.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:31:35 UTC from IEEE Xplore. Restrictions apply.

A Stochastic Model for Characterizing Fluctuations in Chemical Sensing

Abhishek Grover and Brejesh Lall

***Abstract*—The chemical sensing process can be understood as the physical process consisting of binding and unbinding reactions on the sensing surface. The random nature of the reactions lead to fluctuations in the response of a sensor. The response of the sensor is modeled in the form of a stochastic differential equation (SDE). The model is developed by formulating the binding and unbinding reactions in the form of Bernoulli trials. The transition probability density function (pdf) of the number of binded analyte molecules X_t, at a given time instant t, has been evaluated using the Bernoulli trial representation. The obtained pdf is used to derive the Fokker Planck equation (FPE) for the stochastic process. The FPE has a direct correspondence with the desired SDE model. A maximum aposteriori (MAP) estimation method has been proposed for evaluating the parameters. The SDE model for the fluctuations in the sensor signal has been evaluated. The frequency spectrum of the fluctuations is evaluated using the SDE. The proposed model is used to analyze noise due to the ligand-receptor kinetics in the receiver in molecular communication. The utility of the model in agent identification applications has been demonstrated.**

***Index Terms*—Stochastic differential equation, ligand-receptor kinetics, diffusion process, parameter estimation, spectrum estimation.**

I. Introduction

NOISE is defined as undesired variations in the output of the system. Clearly, the identity of noise depends on the distinction between desirable and undesirable. The notion of desirability is based on the system and signal model. The output signal features in correlation with the signal model are classified as desired output, and the uncorrelated term is classified as noise. Thus, there is always a scope for increasing the information content by framing a clear distinction between the signal's desired and undesired variations. In biological and chemical sensing, fluctuations are due to the continuous binding and unbinding of analyte molecules on the sensing surface. These fluctuations can be classified both as noise and a signal based on the application. In molecular communication, the fluctuations due to the ligand-receptor kinetics in the receiver is a noise [1]. A physical and stochastic model has been proposed for ligand-receptor kinetics [2]. It has been shown that the knowledge of the frequency spectrum of the receiver system can be used to improve the bit error rate [3]. Thus a stochastic model of fluctuations may help in designing an optimal receiver in molecular communication.

In some applications, fluctuations can be used as a signal. Using fluctuations in the signal for making an output decision is called fluctuation sensing. In chemical sensing, the technique is known as Fluctuation Enhanced Sensing (FES) [4]. FES is usually employed by connecting a differential amplifier across the sensor and using the output signal's Power spectral density (PSD) for final evaluation [5]. There are some concerns related to the implementation of the technique. The fluctuations are a low magnitude signal. The inherent noise in the differential amplifier would be detrimental to the signal. Another demerit is that standard PSD algorithms use the stationarity assumption, which is not applicable since the fluctuation signal is non-stationary. The previous works do not take into account the signal model for generations of fluctuations. In the absence of a defined signal model, the notion of fluctuations is meaningless. The previous attempts [6], [7] to design the stochastic model for biological sensing cannot be extended to other applications. The previous works' designed model have been used to find an optimal sensor design but does not answer the problem of estimating non-stationary spectra of fluctuations in the sensor response. A parameter estimation method [8] for a stochastic differential equation (SDE) model had been proposed, but this method assumes a specific form of SDE model. It assumes the drift and diffusion coefficients of the model based on the specific application of affinity biosensor considered in the referenced work. It should be noted that building a model and identifying its parameters are critical problems in process modeling, and utmost care is required for solving these problems in a particular application. This study aims to perform this task for the fluctuation sensing method, which could prove useful in increasing the accuracy and robustness in chemical and biological sensing applications.

This study proposes a stochastic differential equation model for the fluctuations in chemical sensing by modeling the binding and unbinding reactions at the sensor surface. The binding and unbinding reactions occur at the receptors or the assumed active sites on a sensor surface. The occurrence of a binding/unbinding reaction at an active site is assumed independent of binding/unbinding reaction at another active site. A binding reaction at an active site is represented as a Bernoulli trial. Similarly, an unbinding reaction is represented as another independent Bernoulli trial. A large number of such trials lead to independent Binomial trials. The Gaussian approximation is employed to

Manuscript received October 6, 2020; revised February 1, 2021; accepted March 6, 2021. Date of publication March 9, 2021; date of current version April 5, 2021. This work was supported by Department of Telecom, Government of India under project RP03521. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pia Addabbo. *(Corresponding author: Abhishek Grover.)*

The authors are with the Bharti School of Telecommunication Technology and Management, Indian Institute of Technology Delhi, New Delhi 110016, India (e-mail: abhishekgrover13@gmail.com; brejesh@ee.iitd.ac.in).

Digital Object Identifier 10.1109/TSP.2021.3064977

1053-587X © 2021 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 June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

determine the conditional pdf for the number of binded molecules on a sensor surface in a given time interval. The knowledge of the conditional pdf allows us to classify the stochastic process as a diffusion process. The conditional pdf is used to evaluate the drift and diffusion coefficients of the diffusion process. The time derivative of the conditional pdf is used to determine the forward Kolmogorov equation. The obtained forward equation allows us to determine the SDE for the process. The maximum aposteriori (MAP) estimates of the parameters in SDE have been evaluated. The proposed SDE model has been used to define and derive the model for fluctuations in the signal. The dynamical model of the variations is used to find the PSD. The proposed algorithm has been used to analyse the noise in the receiver's chemical transduction process in molecular communication. The algorithm has also been tested on dataset related to agent identification in chemical sensing. The contributions of this work are as follows:

- A SDE model has been proposed for the binding and unbinding reactions on the sensing surface.
- The notion of fluctuations in the sensor signal has been defined based on the proposed SDE model.
- The frequency spectrum of the fluctuations have been evaluated.
- The application of the proposed stochastic model has been presented in the case of noise analysis due to ligand-receptor kinetics in molecular communication and in fluctuation enhanced sensing in chemical sensors.

The remainder of this article is organized as follows: Section II presents the previous works related to the methods studied in this paper. Section III presents the stochastic model for chemcial sensing. Section IV presents the parameter estimation method for the model proposed in Section III. Section V presents the spectrum estimation method for fluctuations in the signal. Section VI presents the numerical results. Finally, Section VII concludes this article.

II. Related Work

The fluctuations can be modeled by learning the dynamics equations of the number of molecules binded on the sensor surface. The dynamics model is represented in the form of a SDE [9] [10]. SDE models have been used in population models [11], economics [12], chemical reactions [13], communications [14], and pharmacokinetic models [15]. Building models using SDE involves modeling the physical phenomenon as a Markov process and then identifying the first and second order statistics [16]. Various likelihood estimation-based methods such as exact likelihood, pseudo-likelihood, and approximate likelihood have been suggested for SDEs [17]. In some cases the likelihood based methods might not be appropriate since they do not take into account the system's physical characteristics. In such cases, a maximum aposteriori (MAP) estimation [18] is more robust since it considers the apriori distribution for the parameters. Evaluating apriori distribution is a challenge since the knowledge of physical models and physical constants might not be available. Various studies have reported the utility of PSD [19] and higher-order statistics [20] of the fluctuations in the sensor signal. Spectrum estimation techniques for non-stationary signal [21] should ideally be utilised to measure PSD in this application. The techniques proposed for estimating noise PSD in circuit theory could prove useful. The techniques can be divided into time-domain techniques and frequency domain techniques. Frequency domain techniques involves evaluating the time varying autocorrelation function [22] or finding the transfer function of the system [23]. Time-domain techniques are based on evaluating the frequency spectrum based on the dynamics model of the signal. Monte-Carlo based techniques [24] and the SDE model based techniques [25] are the two time domain methods. Monte-Carlo based techniques are computationally intensive, and it is difficult to implement the algorithms on a low-cost sensing unit. Therefore in this study, the SDE based technique [25] has been used for spectrum estimation.

Stochastic models for chemical sensing find applications in receiver design for molecular communication, chemical analyte identification, and modeling biosensors. In the literature related to molecular communication, various methods have been suggested to increase the receiver's accuracy [26]. The information related to diffusion and reception noise has been utilised in designing the receiver system [27]. The noise due to the transduction process has not been used in previous studies due to the absence of a reliable stochastic model. The fluctuation enhanced sensing system has been primarily used in chemical [28], [29] and biological [30], [31] sensing systems for agent identification applications. The technique has been found useful in finding defects in semiconductor devices [32]. Field Effect Transistor (FET) based biosensors utilizing this technique have been proposed [33]. A diffusion process model has been proposed to understand the scaling laws of affinity-based biosensors [34]. A stochastic model for nanowire-based gas sensors has been proposed to find the optimal sensor design [6]. The following section presents the stochastic model proposed in this study.

III. Stochastic Model for Chemical Sensing

The sensing process can be understood as the interaction between analyte molecules and the sensor surface. The analyte molecules in the ambient environment of the sensor diffuse into the sensor surface and cause changes in its electrical characteristics. The reaction between the analyte molecule and sensing surface is always a reversible reaction. This property leads to continuous binding and unbinding of analyte molecules on the sensor surface. The continuous binding and unbinding leads to variations in the sensor output. The pattern of the continuous binding/unbinding is a characteristic of the sensing surface and the analyte molecule. This characteristic feature is used for agent identification in chemical sensing applications. This section presents the SDE model for the total number of binded molecules based on the random binding/unbinding process.

The modeling strategy is based on the Langmuir's adsorption theory [35] [36]. The sensing surface is assumed to have n active sites where the analyte molecules can get binded. Binding and unbinding reactions are constantly occuring on the sensor surface. It has been assumed that the probability of a reaction occuring is uniform throughout the sensor surface. A binding

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

reaction adds a binded molecule and an unbinding reaction removes an analyte molecule from the sensor surface. Let X_t be the number of analyte molecules in the binded state at a given time instant t. It is assumed that the rate of binding reaction is proportional to the total number of available active sites $n - X_t$ at time instant t

$$\frac{dX_t}{dt} \propto (n - X_t), \tag{1}$$

and the rate of unbinding reaction is proportional to the number of molecules binded to the sensor surface

$$\frac{dX_t}{dt} \propto -X_t. \tag{2}$$

Assume a finitely small interval δt in which partial pressure or the concentration of ambient analyte is constant. It is assumed that the number of binded and unbinded molecules does not change drastically in this small interval δt. On the sensing surface, a large number of binding and unbinding reactions are simultaneously occuring at the active sites. Let $p_1(\delta t)$ be defined as the probability of a binding reaction in δt interval and $p_2(\delta t)$ be the probability of an unbinding reaction in δt interval. The following equations can be deduced using the relation in (1) and (2)

$$\begin{aligned} p_1(\delta t) &= b_1(n - X_t)\delta t, \\ p_2(\delta t) &= d_1(X_t)\delta t, \end{aligned} \tag{3}$$

where b_1 and d_1 are coefficient of proportionality which may vary with time. We want to estimate the number of molecules X_t in the binded state after a time interval t. The interval t is divided into N equal intervals $(t_1, t_2, \ldots, t_N)$ of size δt. Let δX_i be the change in number of binded molecules in the interval $\delta t = t_i - t_{i-1}$ where $i = 1, \ldots, n$. The change in number of binded analyte molecules is equal to the difference between the number of binding reactions and the number of unbinding reactions in δt interval. This difference is evaluated as shown in equation (4)

$$\delta X_i = \sum_{j=1}^{n_1} \delta X_{ij} - \sum_{k=1}^{n_2} \delta X_{ik}, \tag{4}$$

where j, k are indices, n_1 denotes the number of binding reactions in δt interval and n_2 denotes the number of unbinding reactions in δt interval. The variable δX_{ij} is a Bernoulli random variable indicating the occurrence of a binding reaction at an active site. The variable takes a value 1 with $P(\delta X_{ij} = 1) = p_1$, where $p_1 = b_1(n - X_{t_{i-1}})\delta t$ as defined in equation (3). The variable takes a value 0 with $P(\delta X_{ij} = 0) = q_1 = 1 - p_1$. The variable δX_{ik} is a Bernoulli random variable indicating the occurrence of an unbinding reaction at an active site. The variable takes a value 1 with $P(\delta X_{ik} = 1) = p_2$, where $p_2 = d_1(X_{t_{i-1}})\delta t$ as defined in equation (3). The variable takes a value 0 with $P(\delta X_{ik} = 0) = q_2 = 1 - p_2$. The total number of binded molecules at the end of the t_i^{th} partition is

$$X_{t_i} = X_{t_{i-1}} + \delta X_i. \tag{5}$$

It has been assumed that there would not be any large changes in the number of binded analyte molecules in the δt interval. This also implies that the number of molecules in the vicinity of the sensor surface will remain constant in δ_t interval. Fig. 1 shows a schematic for visualizing the dynamics at the sensor surface. The term n_1 in (4) denotes the maximum number of binding reactions that may happen in δ_t interval. At the instant t_{i-1}, the maximum number of analyte molecules in the close proximity to the sensor will be approximately equal to $X_{t_{i-1}}$ (number of binded molecules at instant t_{i-1}). These are the molecules which have a high probability of binding with the surface. Therefore, a good approximation for n_1 is $X_{t_{i-1}}$. The Bernoulli trials consists of evaluating which of these molecules will eventually take part in a binding reaction. The term n_2 in (4) denotes the maximum number of unbinding reactions that may happen in δ_t interval. Since large changes are not permissible, the maximum number of vacant active sites at the end of δ_t interval should be approximately equal to $n - X_{t_{i-1}}$. Hence, the maximum number of unbinding reactions (n_2) can be approximated as $n - X_{t_{i-1}}$. By approximating the upper limits n_1 and n_2, we are able to completely characterize the Binomial trials that may occur on the sensor surface. Let T_1 denotes the binomial trial ($T_1 = \sum_{j=1}^{n_1} \delta X_{ij}$) corresponding to the first term in equation (4) and T_2 denotes the binomial trial ($T_2 = \sum_{k=1}^{n_2} \delta X_{ik}$) corresponding to the second term in equation (4). The distribution for the random variables T_1 and T_2 are as follows

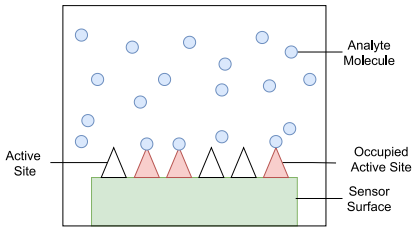


Fig. 1. A schematic for depicting dynamics of the analyte molecules at the sensor surface. The binded molecules are closest to the sensor surface.

$$\begin{aligned} T_1 &\sim B(X_{t_{i-1}}, p_1), \\ T_2 &\sim B((n - X_{t_{i-1}}), p_2), \end{aligned} \tag{6}$$

where $B(m_B, v_B)$ denotes a Binomial distribution with m_B number of trials and v_B as the success probability. It has been assumed that T_1 and T_2 are independent random variables. This assumption is valid in most of the cases as the binding and unbinding reactions at different active sites can be assumed to be independent. In any chemosensor, it is always ensured that the sensor is working within its saturation limits. This ensures that both the quantities $X_{t_{i-1}}$ and $(n - X_{t_{i-1}})$ are large. Hence Gaussian approximation can be used for binomial random variables as shown in equation (7)

$$\begin{aligned} T_1 &\sim N(X_{t_{i-1}}p_1, X_{t_{i-1}}p_1q_1)), \\ T_2 &\sim N((n - X_{t_{i-1}})p_2, (n - X_{t_{i-1}})p_2q_2), \end{aligned} \tag{7}$$

where $N(m_G, v_G)$ denotes a Gaussian distribution with mean m_G and variance v_G. Observing equations (4) and (7) and substituting the values of p_1, q_1, p_2, q_2, the probability density

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

of δX_i is given as

$$\begin{aligned} &p(\delta X_i/X_{t_{i-1}}) \sim N(\mu_{T_1} - \mu_{T_2}, \sigma_{T_1} + \sigma_{T_2}),\\ &p(\delta X_i/X_{t_{i-1}}) \sim N(X_{t_{i-1}}(n - X_{t_{i-1}})(b_1 - d_1)\delta t,\\ &X_{t_{i-1}}(n - X_{t_{i-1}})(b_1 + d_1)\delta t - X_{t_{i-1}}(n - X_{t_{i-1}})\\ &(b_1^2(n - X_{t_{i-1}}) + d_1^2 X_{t_{i-1}})\delta_t^2), \end{aligned} \tag{8}$$

where $\mu_{T_1}, \mu_{T_2}, \sigma_{T_1}$ and σ_{T_2} denote the mean and variance for the random variables T_1 and T_2. Using equation (8) and (5), the transition density $p(X_{t_i}/X_{t_{i-1}})$ is given as

$$\begin{aligned} &p(X_{t_i}/X_{t_{i-1}}) \sim N(m, v)\\ &m = X_{t_{i-1}} + X_{t_{i-1}}(n - X_{t_{i-1}})(b_1 - d_1)\delta t,\\ &v = X_{t_{i-1}}(n - X_{t_{i-1}})(b_1 + d_1)\delta t\\ &\quad -X_{t_{i-1}}(n - X_{t_{i-1}})(b_1^2(n - X_{t_{i-1}}) + d_1^2 X_{t_{i-1}})\delta t^2. \end{aligned} \tag{9}$$

The above density function representations are for the discrete samples from a continuous stochastic process X_t. The next step is to characterize the continuous process X_t based on the obtained density function in (9). It has been assumed δt is small and there cannot be any large changes in the distribution of adsorbed molecules on the sensor surface. Therefore δX_i can be assumed to lie in the range $\delta X_i \in (-\varepsilon, +\varepsilon)$ where ε is a positive finitely small value. Mathematically, this is equivalent to

$$P[|X_{t+\delta t} - X_t| > \varepsilon/X_t] = o(\delta t). \tag{10}$$

The above equation states that the sample path of the process X_t is continuous. As observed in equation (9), the transition density $p(X_{t_i}/X_{t_{i-1}})$ at instant t_i depends only on the state at previous time instant t_{i-1}. Thus the process X_t is a Markov process of order 1. The expectation of the random variable $(X_{t+\delta t} - X_t)/X_t$ is given as

$$\begin{aligned} &E[(X_{t+\delta t} - X_t)/X_t] = E[X_{t+\delta t}/X_t] - X_t,\\ &= \int_{-\infty}^{+\infty} (x - X_t)p(X_{t+\delta t}/X_t)dx\\ &= \int_{|X_{t+\delta t} - X_t| < \epsilon} (x - X_t)p(X_{t+\delta t}/X_t)dx + o(\delta_t). \end{aligned} \tag{11}$$

The last equation in (11) is based on the findings in (10). Substituting $E[X_{t+\delta_t}/X_t]$ using (9)

$$E[(X_{t+\delta t} - X_t)/X_t] = X_t(n - X_t)(b_1 - d_1)\delta t. \tag{12}$$

The drift coefficient $\mu(X_t)$ of the process X_t is evaluated as

$$\mu(X_t) = \lim_{\delta t \to 0} \frac{E[(X_{t+\delta t} - X_t)/X_t]}{\delta t} = X_t(n - X_t)(b_1 - d_1). \tag{13}$$

The variance of the random variable $(X_{t+\delta t} - X_t)/X_t$ is given as

$$\begin{aligned} &E[(X_{t+\delta t} - X_t)^2/X_t]\\ &= \int_{-\infty}^{+\infty} (x - X_t)^2 p(X_{t+\delta t}/X_t)dx,\\ &= \int_{|X_{t+\delta t} - X_t| < \varepsilon} (x - X_t)^2 p(X_{t+\delta t}/X_t)dx + o(\delta t), \end{aligned} \tag{14}$$

For $\delta t \to 0$, the integral in equation (14) is the variance of the random variable $X_{t+\delta t}/X_t$. The diffusion coefficient $\sigma(X_t)$ of the process X_t is evaluated as

$$\begin{aligned} \sigma(X_t) &= \lim_{\delta t \to 0} \frac{E[(X_{t+\delta t} - X_t)^2/X_t]}{\delta t}\\ &= \lim_{\delta t \to 0} \frac{X_t(n - X_t)(b_1 + d_1)\delta t}{\delta t}\\ &\quad - \frac{X_{t_{i-1}}(n - X_{t_{i-1}})(b_1^2(n - X_{t_{i-1}}) + d_1^2 X_{t_{i-1}})\delta t^2}{\delta_t},\\ &= X_t(n - X_t)(b_1 + d_1) \end{aligned} \tag{15}$$

where the variance has been evaluated using the density function in (9). A Markov process with continuous sample path and having finite drift and diffusion coefficients is called a diffusion process [37]. The following theorem summarizes the above findings about the process X_t which represents the number of binded analyte molecules on the sensor surface at time instant t.

Theorem: For a stochastic process X_t, with a transition density $p(X_{t+\delta t}/X_t)$ given in (9), the following points hold:

I) X_t is a Markov process of order 1. The distribution of the present state is dependent only on the previous state.

II) $lim_{\delta t \to 0} \int_{|X_{t+\delta t} - X_t| > \varepsilon} p(X_{t+\delta t}/X_t)dx = 0$,
Using the fact that there can be only small changes in δt interval.

III) $lim_{\delta t \to 0} \frac{1}{\delta t} \int_{|X_{t+\delta t} - X_t| < \varepsilon} (x - X_t)p(X_{t+\delta t}/X_t)dx = X_t(n - X_t)(b_1 - d_1) = \mu(X_t)$,
where $\mu(X_t)$ is called the drift coefficient of process X_t.

IV) $lim_{\delta t \to 0} \frac{1}{\delta t} \int_{|X_{t+\delta t} - X_t| < \varepsilon} (x - X_t)^2 p(X_{t+\delta t}/X_t)dx = X_t(n - X_t)(b_1 + d_1) = \sigma(X_t)$,
where $\sigma(X_t)$ is called the diffusion coefficient of the process X_t.

The above theorem indicates that the stochastic process with transition density given in equation (9) can be modeled as a diffusion process. This indication enables us to represent the stochastic process of number of binded molecules in the form of a stochastic differential equation. The SDE is identified by evaluating the differential of the stochastic process with respect to time. The following lemma 1 is used to identify the form of the SDE.

Lemma 1: A stochastic process X_t defined using a stochastic differential equation

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t,$$

satisfies a forward Kolmogorov equation, also known as the Fokker-Planck equation given by

$$\begin{aligned} \frac{\partial p(x,t;y,\tau)}{\partial t} &= -\frac{\partial (p(x,t;y,\tau)a(x,t))}{\partial x}\\ &\quad + \frac{1}{2}\frac{\partial^2 (p(x,t;y,\tau)b^2(x,t))}{\partial x^2}, \end{aligned}$$

where $p(x,t;y,\tau)$ denotes the transition density for the transition from the state value y at time τ to a state value x at time t. The parameters $a(X_t,t)$ and $b(X_t,t)$ are drift and diffusion coefficients of the process X_t respectively [16]. The

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

real-valued stochastic process W_t denotes the Wiener process. A Wiener process $(W_t, t \geq 0)$ satisfies $W_0 = 0$, has stationary and independent increments such that $W_t - W_s \sim N(0, t-s)$ for all $0 \leq s \leq t$, and is continuous for all t.

Lemma 1 establishes the connection between the forward Kolmogorov equation satisfied by the transition density $p(x,t;y,\tau)$ of a process X_t and its corresponding SDE notation. Thus identifying the forward differential equation for a given transition density will be useful in identifying the form of the SDE for the stochastic process X_t. In the following steps, we will first derive the forward Kolmogorov equation for the stochastic process characterized in the Theorem and then we will use Lemma 1 to identify the SDE. Let $G(x)$ be a non-negative arbitrary real function which is twice differentiable. The function $G(x)$ and its derivatives are assumed to have a compact support such that for $x_1, x_2 \in \mathbf{R}$ $(x_1 < x_2)$

$$\begin{aligned} &G(x) = 0, \qquad x < x_1 \quad \text{and} \quad x > x_2, \\ &G(x_1) = G(x_2) = G^{'}(x_1) = G^{'}(x_2) = 0, \\ &G^{''}(x_1) = G^{''}(x_2) = 0, \end{aligned} \tag{16}$$

where $G^{'}(x)$ and $G^{''}(x)$ denotes the first and second order derivatives of the function respectively. The time derivative of a transition density is

$$\frac{\partial p(x,t;y,\tau)}{\partial t} = \lim_{\delta t \to 0} \frac{p(x,t+\delta t;y,\tau) - p(x,t;y,\tau)}{\delta t}. \tag{17}$$

Multiplying with $G(x)$ both sides and integrating with respect to x

$$\begin{aligned} &\int_{-\infty}^{+\infty} \frac{\partial p(x,t;y,\tau)}{\partial t} G(x) dx \\ &= \int_{-\infty}^{+\infty} \left(\lim_{\delta t \to 0} \frac{p(x,t+\delta t;y,\tau) - p(x,t;y,\tau)}{\delta t} \right) G(x) dx. \end{aligned} \tag{18}$$

The Chapman-Kolmogorov equation for the transition density $p(x,t+\delta t;y,\tau)$ is shown in equation (19)

$$p(x,t+\delta t;y,\tau) = \int_{-\infty}^{+\infty} p(x,t+\delta t;z,t) p(z,t;y,\tau) dz. \tag{19}$$

Substituting equation (19) in the right hand side (RHS) of (18) leads to

$$\begin{aligned} &\lim_{\delta t \to 0} \int_{-\infty}^{+\infty} \frac{1}{\delta t} \int_{-\infty}^{+\infty} (p(x,t+\delta t;z,t) p(z,t;y,\tau) dz \\ &\quad - p(x,t;y,\tau)) G(x) dx \\ &= \lim_{\delta t \to 0} \frac{1}{\delta t} \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,t+\delta t;z,t) p(z,t;y,\tau) G(x) dz dx \right. \\ &\quad \left. - \int_{-\infty}^{+\infty} p(x,t;y,\tau) G(x) dx \right). \end{aligned} \tag{20}$$

Interchanging the variables x and z in the first integral of the above equation

$$\begin{aligned} &= \lim_{\delta t \to 0} \frac{1}{\delta t} \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(z,t+\delta t;x,t) p(x,t;y,\tau) G(z) dz dx \right. \\ &\quad \left. - \int_{-\infty}^{+\infty} p(x,t;y,\tau) G(x) dx \right) \\ &= \lim_{\delta t \to 0} \frac{1}{\delta t} \left(\int_{-\infty}^{+\infty} p(x,t;y,\tau) \left(\int_{-\infty}^{+\infty} p(z,t+\delta t;x,t) G(z) dz \right. \right. \\ &\quad \left. \left. - G(x) \right) dx \right). \end{aligned} \tag{21}$$

Expanding $G(z)$ using Taylor series expansion about x

$$\begin{aligned} &= \lim_{\delta t \to 0} \frac{1}{\delta t} \left(\int_{-\infty}^{+\infty} p(x,t;y,\tau) \left(\int_{-\infty}^{+\infty} (p(z,t+\delta t;x,t)(G(x) \right. \right. \\ &\quad + (z-x) G^{'}(x) + \frac{1}{2}(z-x)^2 G^{''}(x) + o(z-x)^2) dz \\ &\quad \left. \left. - G(x) \right) dx \right). \end{aligned} \tag{22}$$

As observed in point II of Theorem, $z - x < \varepsilon$. Therefore all the higher order terms have been included in $o(z-x)$ in equation (22). Equation (22) can be reduced using the findings mentioned in Theorem. The specific substitutions are

$$\begin{aligned} &\int_{-\infty}^{+\infty} p(z,t+\delta t;x,t) dz = 1, \\ &\int_{-\infty}^{+\infty} o(z-x)^2 p(z,t+\delta t;x,t) dz = 0, \\ &\lim_{\delta t \to 0} \frac{1}{\delta t} \left(\int_{-\infty}^{+\infty} (z-x) p(z,t+\delta t;x,t) dz G^{'}(x) = \right) G^{'}(x) \mu(x), \\ &\lim_{\delta t \to 0} \frac{1}{\delta t} \left(\int_{-\infty}^{+\infty} (z-x)^2 p(z,t+\delta t;x,t) dz \right) G^{''}(x) \\ &= G^{''}(x) \sigma(x). \end{aligned} \tag{23}$$

In general μ and σ can be dependent on both t and x, but in this study, as mentioned in the Theorem, both coefficients are explicitly dependent only on x. In equation (22), taking the limit inside the first integral and substituting the observations in (23)

$$\begin{aligned} &\int_{-\infty}^{+\infty} \frac{\partial p(x,t;y,\tau)}{\partial t} G(x) dx = \int_{-\infty}^{+\infty} p(x,t;y,\tau) G^{'}(x) \mu(x) dx \\ &\quad + \frac{1}{2} \int_{-\infty}^{+\infty} p(x,t;y,\tau) G^{''}(x) \sigma(x) dx. \end{aligned} \tag{24}$$

There are two integrals on RHS of equation (24). The first term is integrated once using integration by parts and the second term is integrated twice using integration by parts. The following equation is obtained:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

$$\int_{-\infty}^{+\infty} \frac{\partial p(x,t;y,\tau)}{\partial t} G(x)dx = \int_{-\infty}^{+\infty} \left(-\frac{\partial}{\partial x}(p(x,t;y,\tau)\mu(x)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(p(x,t;y,\tau)\sigma(x)) \right) G(x)dx. \quad (25)$$

Arranging the terms, substituting values of μ and σ and observing that equation (25) should be valid for arbitrary $G(x)$, the following partial differential equation can be deduced.

$$\frac{\partial p(x,t;y,\tau)}{\partial t} = -\frac{\partial}{\partial x}(p(x,t;y,\tau)X_t(n-X_t)(b_1-d_1)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(p(x,t;y,\tau)X_t(n-X_t)(b_1+d_1)). \quad (26)$$

Using Lemma 1 the SDE for the binding-unbinding process is

$$dX_t = X_t(n-X_t)(b_1-d_1)dt + \sqrt{X_t(n-X_t)(b_1+d_1)}dW_t. \quad (27)$$

The SDE in equation (27) represents the dynamics model for the number of binded molecules on sensor surface.

IV. Parameter Estimation for Binding-Unbinding Kinetics

The parameters b_1, d_1 and n can be determined from the available physical parameters for forward and reverse reaction rate constants. The deterministic equation for surface kinetics [3] is

$$\frac{dX_t}{dt} = k_f c_R(t)(n-X_t) - k_b X_t, \quad (28)$$

where X_t is the number of binded molecules on the sensor surface, n is the total number of active sites on the sensing surface, k_f is the forward reaction empirical rate constant and k_b is the backward reaction empirical rate constant. The forward reaction corresponds to the binding reaction at the sensor surface and the reverse reaction corresponds to the unbinding reaction at the sensor surface. The function $c_R(t)$ is the concentration of analyte molecules in the receptor space of the sensor. Differential equations have been used in many applications to determine the deterministic models for expected values of specific states in markov processes [38]. A similar methodology has been used in this study. The following equations compare the deterministic differential equation of surface kinetics with the $E[(\delta X_i/\delta t)/X_{t_{i-1}}]$ obtained in equation (8)

$$E[(\delta X_i/\delta t)/X_{t_{i-1}}] = X_{t_{i-1}}(n-X_{t_{i-1}})(b_1-d_1),$$
$$b_1 = \frac{k_f c_R(t)}{X_{t_{i-1}}},$$
$$d_1 = \frac{k_b}{n-X_{t_{i-1}}}. \quad (29)$$

By comparing the first equation in (29) with the ordinary differential equation (ODE) in (28), the last two equations in (29) can be obtained. The above substitutions (29) are implemented to ensure that the expected values of the modeled process are similar to the values obtained using the surface kinetics theory. A similar deterministic equation can be obtained using the langmuir theory introduced in Section III. Thus the knowledge of the constants k_f, k_b and n helps in identifying the parameters of the model proposed in this study.

If the physical parameters for rate constants are not available then the parameters are determined using a calibration routine. A training dataset has to be first generated by deploying sensors in factory settings. The ambient environment should be similar to the practical scenarios. Specific studies have been conducted [39], [40] for developing such factory calibration routines for sensors. The obtained training data is then used to calibrate sensors by identifying the three parameters. Let Y_M denote the set of M samples of training data $\{y_1, y_2, \ldots, y_M\}$. We want to estimate the probability distribution $p(\boldsymbol{\theta}/Y_M)$ for the parameters $\boldsymbol{\theta} = (b_1, d_1, n)$ given Y_M. Using Baye's rule

$$p(\boldsymbol{\theta}/Y_M) = \frac{p(Y_M/\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(Y_M)} \propto p(Y_M/\boldsymbol{\theta})p(\boldsymbol{\theta}). \quad (30)$$

Finding $\boldsymbol{\theta}$ which maximizes the term $p(Y_M/\boldsymbol{\theta})p(\boldsymbol{\theta})$ is known as maximum aposteriori estimation. To maximize the aposteriori term, the probability distribution $p(Y_M/\boldsymbol{\theta})$ and the apriori information $p(\boldsymbol{\theta})$ needs to be estimated. The likelihood function $p(Y_M/\boldsymbol{\theta})$ is evaluated as follows

$$\begin{aligned} &p(y_1, y_2, \ldots, y_M/\boldsymbol{\theta}) \\ &= p(y_n/y_1, y_2, \ldots, y_{M-1}, \boldsymbol{\theta})p(y_1, y_2, \ldots, y_{M-1}/\boldsymbol{\theta}) \\ &= \left(\prod_{k=1}^{M} p(y_k/y_{k-1}) \right) p(y_0, \boldsymbol{\theta}). \end{aligned} \quad (31)$$

The training data Y_M is assumed to be generated by the binding-unbinding stochastic process explained in the previous section. If the time interval between two samples y_k and y_{k-1} is small, then the distribution $p(y_k/y_{k-1})$ can be evaluated using equation (9). Assuming apriori distribution is Gaussian ($p(\boldsymbol{\theta}) \sim N(\boldsymbol{\mu_\theta}, \boldsymbol{\sigma_\theta})$) and $p(y_0/\boldsymbol{\theta})$ is a constant

$$p(\boldsymbol{\theta}/Y_M) \propto \left(\prod_{k=1}^{M} \frac{exp(-\frac{(y_k - y_{k/k-1})^2}{2R_{k/k-1}^2})}{\sqrt{2\pi}R_{k/k-1}} \right) \left(\frac{exp(-\frac{1}{2}\boldsymbol{\mu_\theta}^T \boldsymbol{\sigma_\theta}^{-1} \boldsymbol{\mu_\theta})}{|\boldsymbol{\sigma_\theta}|^{1/2}(2\pi)^{3/2}} \right) p(y_0/\boldsymbol{\theta})$$
$$y_{k/k-1} = y_{k-1} + y_{k-1}(n-y_{k-1})(b_1-d_1)$$
$$R_{k/k-1} = y_{k-1}(n-y_{k-1})(b_1+d_1). \quad (32)$$

The next step is to evaluate $\boldsymbol{\mu_\theta}$ and $\boldsymbol{\sigma_\theta}$ using training data. An electrical equivalent circuit has been proposed [35] for Langmuir theory. This electrical equivalent circuit has been used in this study to identify the physical constants. The sample mean and sample variance values of b_1, d_1 and n are taken as the mean ($\boldsymbol{\mu_\theta}$) and variance ($\boldsymbol{\sigma_\theta}$) respectively of the apriori distribution. It should be noted that there are some constraints associated with the parameters. The parameter n has to be greater than the maximum value of sensor response. This is to ensure a positive diffusion term of the stochastic process proposed in equation (27). The MAP estimate $\boldsymbol{\theta_{MAP}}$ is then obtained by minimizing

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

the negative of log likelihood function as shown in equation (33)

$$\boldsymbol{\theta_{MAP}} = \underset{\boldsymbol{\theta}}{argmin}\{-ln(p(\boldsymbol{\theta}/Y_M, y_0))\}. \tag{33}$$

Several points need to be noted regarding the parameter estimation process. In the above estimation process Y_k denotes the number of binded molecules at given instant k. In many cases sensor response is in the form of an electrically measurable parameter $Y_k^{'}$. $Y_k^{'}$ is related to Y_k by the transduction function $Y_k^{'} = g(Y_k)$. The function g is usually available from physical principles [41], [42]. If the function g is not available then an iterative method as suggested in Section VI- C is used to find the parameters. The proposed method is a robust way to estimate parameters if no information is available from physical models.

V. Spectrum Estimation Using SDE

The fluctuations in sensor values are non-stationary and depend on the ambient analyte concentration. The frequency information in the fluctuations is evaluated using the SDE model proposed in equation (27). The following lemma 2 is used to find the stochastic differential for the function of a process X_t.

Lemma 2: Let x_t be the unique solution of a stochastic differential equation

$dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

where W_t is the Wiener process with $E[dW_t] = 0$ and $E[dW_t^2] = dt$. Let $\psi(X_t)$ be a scalar-valued real function, continuously differentiable in t and having continuous second mixed partial derivatives with respect to x. Then the stochastic differential of ψ is

$$d\psi = (\psi_t + \psi_X a(X_t, t) + \frac{1}{2} b^2(X_t, t)\psi_{XX})dt + \psi_X b(X_t, t)dW_t$$

where $\psi_t = \frac{\partial \psi}{\partial t}, \psi_X = \frac{\partial \psi}{\partial x}, \psi_{XX} = \frac{\partial^2 \psi}{\partial x^2}$ [9].

The first step in spectrum estimation involves applying the lamperti transformation [43] on the SDE in equation (27). This is implemented to make the diffusion term independent of the state X_t. It has been found that the transformation increases the stability of the process [17]. It also helps in implementing a first order approximation of the fluctuations in the process. This approximation allows us to derive a linear model for the fluctuations in observed values. The obtained linear model is then used to estimate the PSD of the fluctuations in the adsorption-desorption process. The lamperti transformation is implemented as follows: Let $Z_t = \psi(X_t)$ be the transformed process. Using lemma 2 and substituting the drift $\mu(X_t)$ and diffusion coefficient $\sigma(X_t)$ of the process X_t defined in the Theorem, dZ_t is

$$dZ_t = (\psi_t + \psi_X \mu(X_t) + \frac{1}{2}\psi_{XX}\sigma(X_t))dt + \psi_X \sqrt{\sigma(X_t)} dW_t. \tag{34}$$

The function ψ is selected so that the transformed process has a state independent diffusion term. Therefore the following transformation is implemented

$$\psi(X_t) = \int \frac{1}{\sqrt{X(n-X)}} dX \Big|_{X=X_t} = \sin^{-1}\left(\frac{X_t - n/2}{n/2}\right). \tag{35}$$

The following relations can be deduced using the above transformation

$$X_t = \frac{n}{2}(1 + \sin(Z_t)),$$
$$\psi_t = 0, \psi_X = \frac{1}{\sqrt{X_t(n - X_t)}} = \frac{2}{n\cos(Z_t)},$$
$$\psi_{XX} = \frac{2X_t - n}{X_t^{3/2}(n - X_t)^{3/2}} = \frac{4sin(Z_t)}{n^2 cos^3(Z_t)}. \tag{36}$$

The expressions for $\mu(X_t)$ and $\sigma(X_t)$ in terms of Z_t are as follows

$$\mu(X_t) = (b_1 - d_1)\frac{n^2 cos^2(Z_t)}{4},$$
$$\sigma(X_t) = (b_1 + d_1)\frac{n^2 cos^2(Z_t)}{4}. \tag{37}$$

Substituting the expressions in (36) and (37) in the SDE (34) for the transformed process

$$dZ_t = (\frac{n}{2}(b_1 - d_1)cos(Z_t) + \frac{1}{2}(b_1 + d_1)\tan(Z_t))dt + \sqrt{b_1 + d_1} dW_t. \tag{38}$$

Let Z_{t_k} be the observed sample value at instant t_k after transformation and $H(Z_t) = \frac{n}{2}(b_1 - d_1)cos(Z_t) + \frac{1}{2}(b_1 + d_1)tan(Z_t)$ be the drift coefficient of the transformed process. Expanding $H(Z_t)$ around $Z_t = Z_{t_k}$ using Taylor series in equation (38)

$$dZ_t = (H(Z_{t_k}) + H'(Z_{t_k})(Z_t - Z_{t_k}))dt + o(Z_t - Z_{t_k}) + \sqrt{b_1 + d_1} dW_t,$$
$$dZ_t - H(Z_{t_k})dt = H'(Z_{t_k})(Z_t - Z_{t_k})dt + o(Z_t - Z_{t_k}) + \sqrt{b_1 + d_1} dW_t. \tag{39}$$

The higher order terms in the above equation have been included in $o(Z_t - Z_{t_k})$ which is in accordance with the assumption of small variations between consecutive samples. There are several things to be noted at this point. Every SDE has a drift component and a diffusion component. The drift component $H(Z_t)dt$ in equation (38) can be understood as the deterministic component of the differential equation and the diffusion component $\sqrt{b_1 + d_1} dW_t$ can be understood as the random component. The fluctuations in the signal are defined as the variations around the deterministic component of the signal. The left hand side (LHS) of the equation (39) represents the difference between the variation in the signal and its drift component. Thus the LHS essentially represents the fluctuations in the binding-unbinding process. Equation (39) represents the model for signal variations about the currently observed value and it allows us to define the fluctuations in the observed signal. The best estimate of Z_t between the observations is the observed value Z_{t_k}. Hence the term $H(Z_{t_k})dt$ of SDE in equation (39) can be denoted by dZ_{t_k}

$$d(Z_t - Z_{t_k}) = H'(Z_{t_k})(Z_t - Z_{t_k})dt + \sqrt{b_1 + d_1} dW_t. \tag{40}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

Let $R_t = Z_t - Z_{t_k}$ be the fluctuation around the recently observed value. The SDE for the fluctuation signal R_t can be derived as

$$dR_t = H'(Z_{t_k})R_t dt + \sqrt{b_1 + d_1}dW_t,$$
$$dR_t = BR_t dt + CdW_t, \tag{41}$$

where $B = H'(Z_{t_k})$ and $C = \sqrt{b_1 + d_1}$. The process R_t denotes the fluctuations in the transformed process Z_t. It is important to understand its relation with the original process X_t. Using equation (35), R_t can be expressed as

$$R_t = \sin^{-1}\left(\frac{X_t - n/2}{n/2}\right) - \sin^{-1}\left(\frac{X_{t_k} - n/2}{n/2}\right)$$
$$= \sin^{-1}\left(\frac{X_t - n/2}{n/2}\sqrt{1 - \left(\frac{X_{t_k} - n/2}{n/2}\right)^2}\right.$$
$$\left. - \frac{X_{t_k} - n/2}{n/2}\sqrt{1 - \left(\frac{X_t - n/2}{n/2}\right)^2}\right) \tag{42}$$

Since $X_t \in (0, n)$, $\frac{X_t - n/2}{n/2} \in (-1, 1)$ and $|X_t - X_{t_k}| < \epsilon$ from the theorem. Based on these relations the first square root term $\sqrt{1 - \left(\frac{X_t - n/2}{n/2}\right)^2}$ in the above expression can be approximated as $\sqrt{1 - \left(\frac{X_{t_k} - n/2}{n/2}\right)^2}$. Thus R_t can be expressed as

$$R_t \approx \sin^{-1}\left(\left(\frac{2}{n}\right)\sqrt{1 - \left(\frac{X_{t_k} - n/2}{n/2}\right)^2}(X_t - X_{t_k})\right) \tag{43}$$

Observing that the term $\left(\frac{2}{n}\right)\sqrt{1 - \left(\frac{X_{t_k} - n/2}{n/2}\right)^2}(X_t - X_{t_k}) \ll 1$

$$R_t \approx \left(\frac{2}{n}\right)\sqrt{1 - \left(\frac{X_{t_k} - n/2}{n/2}\right)^2}(X_t - X_{t_k}) \tag{44}$$

The term $\left(\frac{2}{n}\right)\sqrt{1 - \left(\frac{X_{t_k} - n/2}{n/2}\right)^2}$ is a constant in the interval (t_k, t_{k+1}). Hence R_t is directly proportional to the fluctuations $X_t - X_{t_k}$ in the original process X_t. The spectrum of R_t is the desired information in this study since it quantifies the frequency of the fluctuations in sensor values. Equation (41) gives the linear and time varying model for R_t. The spectrum is evaluated by decomposing the process of PSD estimation into the process of solving three ODEs. The Fourier transform $G(t, w)$ and the square of its modulus $I(t, w)$ of a segment of R_t of duration t is

$$G(t, w) = \int_0^t R(\tau)e^{-jw\tau}d\tau,$$
$$I(t, w) = G(t, w)G^*(t, w). \tag{45}$$

Taking derivative and then the expectation operator on both sides

$$\frac{d}{dt}(ESD) = E[R(t)G^*(t, w)]e^{-jwt},$$
$$+E[R^*(t)G(t, w)]e^{jwt}. \tag{46}$$

where $E[I(t, w)]$ denotes the energy spectral density (ESD). The LHS in equation (46) represents the instantaneous PSD. The average value of the instantaneous PSD for the complete measurement window is taken as the PSD value at a specific frequency w. The instantaneous PSD can be estimated if the value of $E[R(t)G^*(t, w)]$ is available. Let $e(t) = R(t)G^*(t, w)$ where R(t) is a real signal. $E[e(t)]$ is evaluated as

$$de(t) = dR(t)G^*(t, w) + R(t)R^*(t)e^{jwt}dt,$$
$$de(t) = (BR(t)dt + CdW_t)G^*(t, w) + R(t)R^*(t)e^{jwt}dt. \tag{47}$$

Taking expectation operator E both sides and using $E[dW_t] = 0$

$$\frac{dE[e(t)]}{dt} = BE[e(t)] + E[R^2(t)]e^{jwt}. \tag{48}$$

The ODE in equation (48) can be solved by finding values of $E[R^2(t)]$. Equation (41) is used to find the ODE for $E[R^2(t)]$. The differential equation for $R^2(t)$ can be evaluated from equation (41) using Lemma 2

$$d(R^2(t)) = (2BR^2(t) + C^2)dt + 2R(t)CdW_t. \tag{49}$$

Taking expectation on both sides of equation (49) and using $E[dW_t] = 0$

$$dE[R^2(t)] = 2BE[R^2(t)]dt + C^2dt. \tag{50}$$

The ODE in equation (50) is first solved to find $E[R^2(t)]$. Then $E[R^2(t)]$ is used to solve ODE in equation (48). The values of $E[R(t)G^*(t, w)]$ is substituted in equation (46) to find the instantaneous PSD. The above method can be used in frequency analysis for the reception process in molecular communication. The average PSD values in the frequency range 0–500 rad/s are used to form feature sets in various agent identification applications.

VI. Numerical Results

This section presents the efficacy of using the proposed stochastic model in the applications involving binding-unbinding process at the sensor surface. Subsection A presents an experimental setup to analyze the voltage response of a chemical sensor based on the adsorption of ions. Subsection B presents the simulation analysis of noise due to ligand-receptor kinetics in the receiver system used in molecular communication. Subsection C presents the utility of the proposed algorithm on a dataset related to agent identification in chemical sensing.

A. An Experimental Setup to Analyze the Response of an Adsorption Based Chemical Sensor

An experimental setup has been developed to present the response of an adsorption based chemical sensor. The setup has also been used to demonstrate the response of a receiver in molecular communication. Fig. 2 shows the experimental setup. The setup consists of a peristaltic pump, a pH glass electrode, an electronic interface circuit and a logic analyzer. Molecular communication based setups utilising pH levels for communication have been presented in previous studies as well [44], [45]. The

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

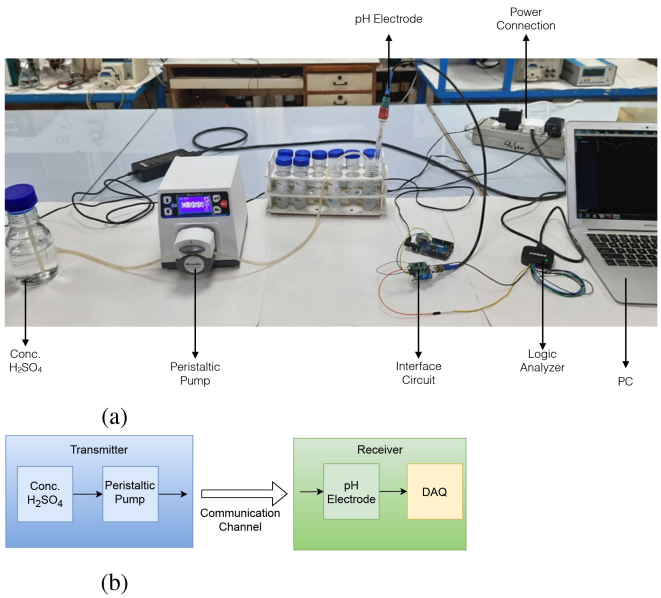


Fig. 2. Experimental setup to analyze response of a chemical sensor (a) Experimental setup (b) Block Diagram.

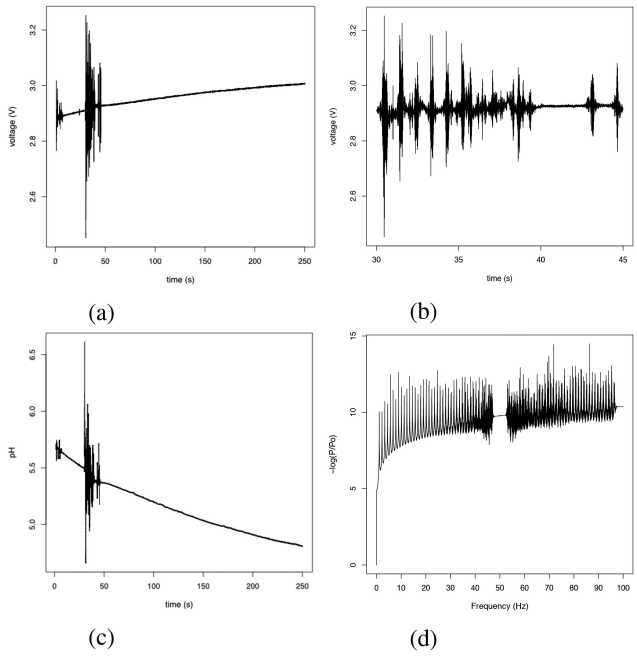


Fig. 3. (a) Voltage response of the electrode for a period of 250 s (b) Voltage response for the interval 30s–45 s (c) pH values for a period of 250 s (d) Power spectrum for the volatge waveform in (a).

electronic interface circuit has been used to add a DC bias to the voltage response from the pH electrode. The logic analyzer has been used as a data acquisition system (DAQ). The sampling rate is 15.626 KHz. Before analyzing, the obtained time series is passed through a Finite Impulse Response (FIR) low pass filter with cut-off frequency of 520 Hz to remove undesired high frequencies and a FIR comb filter to remove harmonics of 50 Hz power line frequency. The peristaltic pump has a conc. $H_2\ SO_4$ with a pH value of 2.4 on one side and a container with a pH electrode dipped in 20 ml RO (Reverse Osmosis) water on the other side. The conc. $H_2\ SO_4$ and the peristaltic pump represent the transmitter in the communication system and the container with the pH glass electrode represents the receiver. The operation of a pH glass electrode is based on the adsorption of H^+ ions. The adsorption of H^+ ions on the glass membrane leads to the development of a potential across its surface. The difference between the developed potential and the potential of the reference electrode is then used to estimate the pH of the solution [46]. The pH of the solution gives an estimate of the concentration of H^+ ions in the solution. Thus the voltage response of a pH electrode is based on the binding-unbinding kinetics of H^+ ions on the glass membrane surface. The pH electrode has been calibrated using buffer solutions with pH 4, 7, and 10. The experiment consists of transferring conc. $H_2\ SO_4$ to the container with RO water for a predetermined time. This predetermined time corresponds to the 'ON' time for 'High' pulse in the communication system. The 'ON' time is to be detected correctly using the response of a pH electrode. The experiment consists of first recording electrode response in RO water for a period of 30 s and then switching on the peristaltic pump for a fixed 'ON' interval.

Fig. 3(a) shows the voltage response for a specific experiment in which 'ON' time interval is 15 s. Fig. 3(b) shows the voltage response for the 'ON' time interval. As seen in figure 3(b), the magnitude of fluctuations increases when the solution is being transferred. In figure 3(a), the baseline of the voltage response has increased at the end of the experiment. This indicates increased H^+ ion concentration on the receiver side. Fig. 3(c) shows the pH values for the entire duration of the experiment. The 'ON' time interval can be identified using the pH response. The variations in the voltage response corresponds to the fluctuations. These variations are due to the continuous binding and unbinding of H^+ ions on the glass membrane surface. Fig. 3(d) shows the power spectrum of the the time series in figure 3(a). The power spectrum is evaluated using the Fast Fourier Transform (FFT). In figure (d), the y-axis represents $-log(P/Po)$ where P is the FFT sequence power at a specific frequency and Po is the FFT sequence power at 0 Hz. It can be seen that the signal mainly consists of direct current (DC) signal. Components with frequencies less than 40 Hz are also present. The power in these components is very small and it decreases with increase in frequency. These components represent the spectrum of fluctuations. The above analysis shows the existence of the fluctuations in the response of adsorption based sensors. The following subsections present the utility of parameter estimation and spectrum estimation presented in this study. Subsection B presents the case when the adsorption and desorption rates are known in a simulation scenario and subsection C presents the case when the adsorption and desorption rates are required to be evaluated from experimental data.

B. Simulation Analysis of Noise in the Receiver System Used in Molecular Communication

In a molecular communication system [47], communication between a transmitter and receiver is implemented using

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

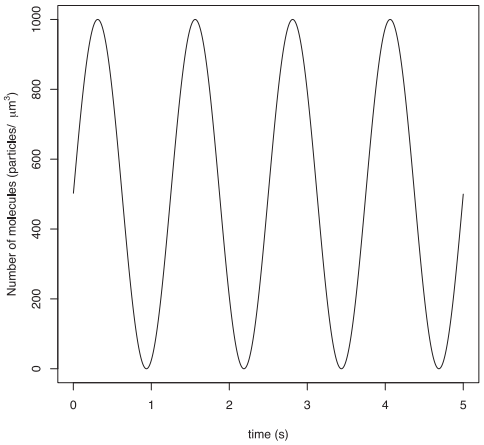


Fig. 4. Number of ambient analyte molecules ($c_R(t)$) in the receptor space of the receiver.

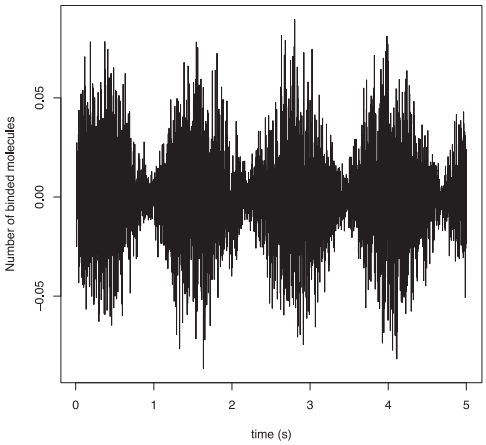


Fig. 6. Fluctuations (R_t) in the number of binded molecules on the sensor surface.

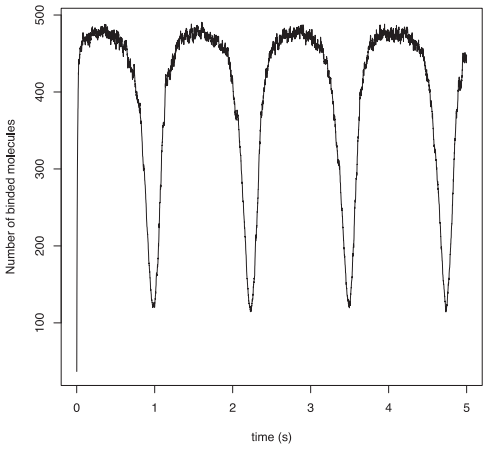


Fig. 5. Number of binded molecules (X_t) on the sensor surface.

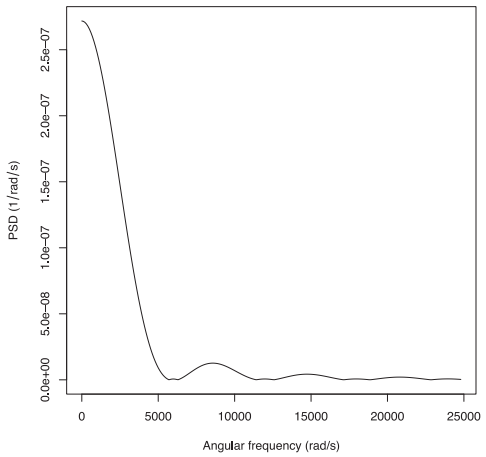


Fig. 7. Power spectral density of the fluctuations in the number of binded molecules on the sensor surface.

diffusing molecules. The information is encoded by varying the number, type or temporal pattern of molecular emission at the transmitter. The receiver system consists of a chemical receptor which detects the molecules sent by the transmitter. In simulation analysis $n = 500$, $k_f = 0.2 \quad \mu\text{m}^3/s$ and $k_b = 10 \quad 1/s$. The function $c_R(t)$ is assumed as $c_R(t) = Asin(2\pi f_a t) + A$, where $A = 500 \quad particles/\mu\text{m}^3$ and $f_a = 0.8 \quad Hz$. The MAP estimation is not required in this case since the parameters are available. The parameters n, b_1 and d_1 of the SDE in (27) are evaluated using the method proposed in Section IV

$$n = 500,$$
$$b_{1_i} = \frac{k_f c_R(t_i)}{X_i},$$
$$d_{1_i} = \frac{k_b}{n - X_i}, \tag{51}$$

where X_i is the value observed at the instant t_i. The simulation time step is 1 ms and the total simulation time interval is 5 s. Fig. 4 shows the waveform for $c_R(t)$. Fig. 5 shows the waveform for the number of binded molecule (X_t) which is simulated using the proposed SDE model. It can be seen that the number of binded molecules correctly follows the trend in increase or decrease of the ambient molecule concentration. The waveform does not goes to zero level when the concentration drops. There would always be some molecules which would remain binded to the receiver. This phenomenon depends on dynamic response of the receiver. If the reverse reaction rate is higher then all the molecules will unbind when the ambient concentration drops and the receiver response would accurately follow the ambient concentration waveform. Fig. 6 shows the noise waveform (R_t) simulated using the SDE model proposed in equation (41). It is to be noted that the magnitude of noise is larger when the number of ambient analyte molecule concentration is large. This is intuitive as an increase in number of ambient analyte molecule concentration leads to increased binding/unbinding reactions. Fig. 7 shows the power spectral density for the noise waveform shown in figure 6. PSD is evaluated using the method proposed in section V. The PSD for fluctuations is similar to a low pass signal. The cutoff frequency is around 2000 rad/s and is evaluated graphically using the obtained spectrum. The exact shape of the waveform in figure 5, its maximum and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

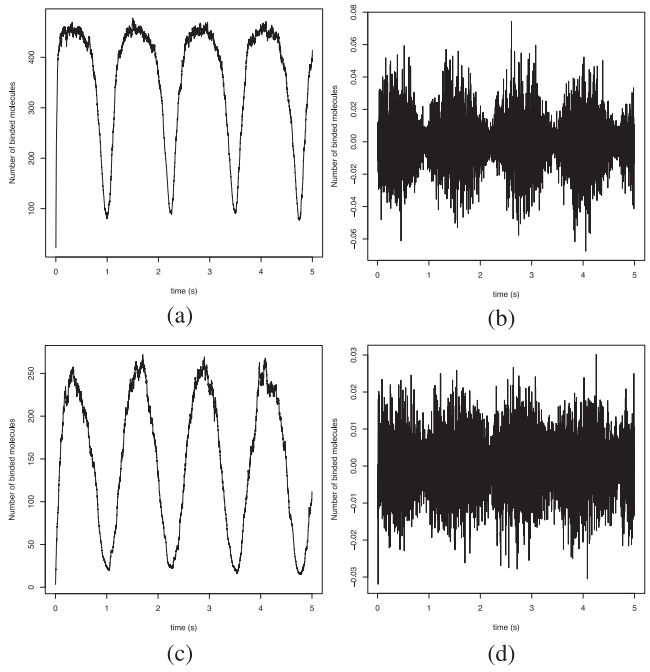


Fig. 8. Comparison of the number of binded molecule waveform and the fluctuation waveform for various values of forward and reverse reaction constants.(a), (b)$k_f = 0.1, k_b = 10$. (c), (d)$k_f = 0.01, k_b = 10$.

minimum values depends on the design specification. The forward and reverse reaction rate constants are the major design specifications in the receiver. Fig. 8 shows the number of binding molecule waveforms for various values of forward and reverse reaction rate constants. Fig. 8(a) and (b) are simulation results with parameters $k_f = 0.1$ and $k_b = 10$. The reverse reaction constant is similar to the simulation result shown in figure 5 but the forward reaction constant has been reduced by a factor of 10. This design choice decreases the forward reaction rate constant in comparison to the reverse reaction rate constant. Thus the number of binding reaction decreases. It can be seen in figure 8(a) that this receiver system follows the sinusoidal waveform more accurately at the peaks. The binding reactions are low and the unbinding reactions are high. When the ambient analyte molecule concentration increases, a low forward reaction constant ensures that the increase in number binded molecules is not drastic and a high reverse reaction constant ensures quick removal of molecules from the sensing surface. The magnitude of noise in figure 8(b) is less than that in fig. 6. A lower binding reaction rate leads to low number of binding reactions which in turn leads to low number of fluctuations. These observations are also visible in the third design choice ($k_f = 0.01, k_b = 10$) implemented in this study. Fig. 8 (c) and (d) show the simulation results. The response of the receiver at the peaks is in accordance with the input waveform. The magnitude of noise is lower than the waveforms shown in figures 6 and 8(b). It should be noted that these design choices depend on the specific application. In some applications where the specific molecular species are highly reactive, a lower reaction rate might be preferable. As can be seen in figures 5 and 8, a lower forward reaction rate decreases

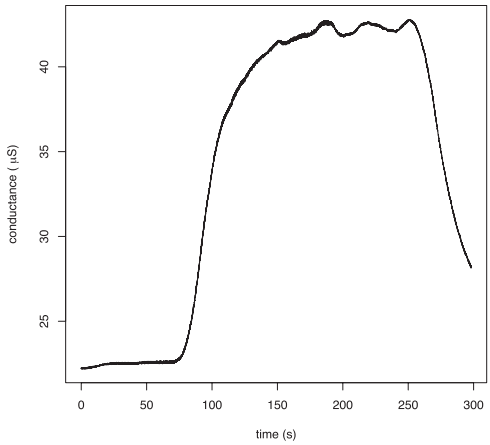


Fig. 9. Sensor conductance values for the complete duration of a specific experiment in the dataset.

the maximum number of analyte molecules that may bind to the surface. This lowers the detection range for the receiver. A design trade-off between the required detection range and the maximum tolerable noise levels needs to be implemented to design an optimal receiver system.

C. Fluctuation Enhanced Sensing in Chemical Sensors

This subsection presents an agent identification application using metal oxide semiconductor based gas sensors. The classification of presence/absence of analyte molecule has been implemented using the pattern in frequency spectrum of fluctuations in a single sensor. J. Fonollosa *et al.* [39] have devised an experimental setup for recording readings from an array of 8 sensors. The experiment consists of introducing a mixture of gases in a wind tunnel. In all experiments, the mixture consists of either ethylene and carbon monoxide or ethylene and methane. Each experiment consists of releasing varying volume of gases in the wind tunnel. The aim of the study was to detect the presence of ethylene in the mixture. The classification in the study [39] was implemented using the pattern from all 8 sensors. The dataset consists of 180 csv files with each file containing time series response from 8 sensors during a specific experiment. Each experiment consists of releasing gas into the measurement chamber for a period of 180 s after initial 60 s. At the end of 240 s, the gas mixture is removed from the chamber. Fig. 9 shows the sensor conductance values in one of the experiments. The initial 60 s data represents the baseline value for the sensors. The next 180 s data represents the adsorption process in the sensors and then the steady state when the adsorption and desorption rates are similar. The last 60 s denotes the desorption process corresponding to the gas removal from the sensing chamber. The observed variable in this case is the sensor conductance. The relationship between conductance of thin film and number of adsorbed gas molecules is given by [41]

$$\sigma_x = \sigma_o(1 + jC_o), \tag{52}$$

where σ_o is film conductance in air, C_o is the adsorbed target gas concentration in the thin film and j is the sensitivity coefficient.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

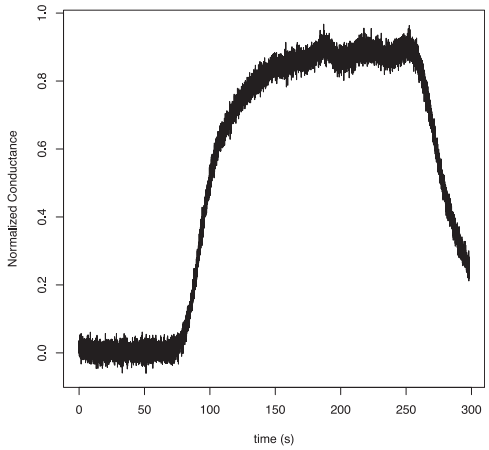


Fig. 10. Normalized conductance Y_k^s simulated using the SDE in (27). The parameters for the SDE have been learned using the specific case in figure 9.

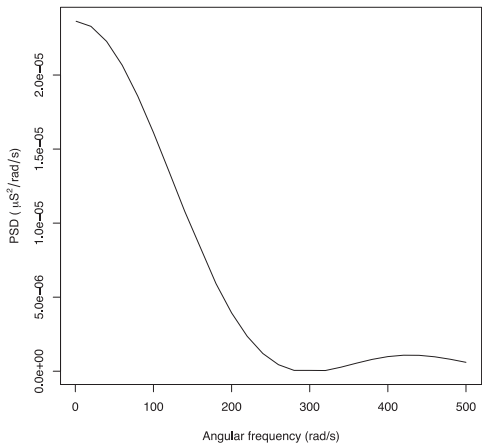


Fig. 11. Power spectral density of sensor fluctuations.

The total conductance of sensor film (σ_k) at instant k is evaluated as

$$\sigma_k = \int_0^L \sigma_x dx = \int_0^L \sigma_o(1 + jC_o)dx$$
$$= \sigma_o L + \sigma_o j C_{o_k} \tag{53}$$

where L is the total length of the sensing film and C_{o_k} is the total number of binded analyte molecules at the instant k. In the dataset, the values reported are the conductance σ_k for each instant k. The baseline value $\sigma_o L$ is evaluated for each experiment. The parameter j is not known. Therefore, an iterative method is used to estimate the parameters. A linear model of the following form is assumed for the transduction function

$$Y_k' = hY_k + v_k \qquad v_k \sim N(0, r), \tag{54}$$

where $Y_k^{'} = \frac{\sigma_k}{\sigma_o L} - 1$ is the normalized conductance value, $Y_k = jC_{o_k}$, h is the transduction coefficient and r is the noise variance. The normalized conductance Y_k' is used as the observation values for this use case. In the first iteration h is assumed to be unity and v_k is assumed to be identically zero. The parameters θ are estimated based on the evaluated Y_k values. The obtained parameters are used to simulate the time series Y_k using SDE in (27). The simulated time series is denoted as Y_k^s. Then a linear regression is implemented between Y_k' and Y_k^s. The parameter r is evaluated as the error variance in regression. A low noise variance r would correspond to a good linear approximation. The iteration is repeated if the euclidean norm between the previous and the new transduction coefficient is greater than a threshold. The model parameters are evaluated for each sensor.

Fig. 10 shows the simulated normalized sensor values for the experimental case shown in figure 9. The fluctuations between the observations represent the variations due to the adsorption-desorption process. The frequency spectrum of the fluctuations is used for classification application.

Fig. 11 shows the PSD of the fluctuations. These findings are similar to the $1/f$ noise reported in various studies [20], [48]. The PSD values at specific frequencies is used to form a pattern corresponding to each experimental file. The dataset was

TABLE I
CLASSIFICATION ACCURACY (%) USING THE PROPOSED FREQUENCY DOMAIN AND TIME DOMAIN FEATURES AS INPUT FOR THE CLASSIFICATION ALGORITHM

Sensor	Time Domain (T) f=100	Frequency Domain (F) f=25	T+F f=10
1	82.2	81.6	81.9
2	75.8	75.4	76.6
3	75.4	75.3	76.1
4	76.5	76.3	77.8
5	90.8	89.1	89.4
6	91.3	89.5	89.9
7	93.8	89.7	91.1
8	94.3	90.4	91.1

divided into 30 batches with each batch containing 6. csv files. Each batch contains all specific levels of ethylene (low to high) used in the experiment. Then for a specific experiment, a single file is included in the training set from each batch. The rest files are included in the test set. Thus in each experiment the training set consists of 30 samples and test set consists of 150 samples. The classification for presence/absence of ethylene is performed using KNN algorithm. The classification is implemented by designing a binary classifier. It should be noted that the technique can also be used to classify multiple agents by designing multiple classifiers. The number of classifiers would be equal to the number of agents to be detected. Each classifier would be used to detect a specific agent. In the KNN algorithm, the nearest 3 neighbours are used to make a decision. The above experiment is repeated 60 times. Various experiments have been conducted. Table I shows the average classification accuracy using one sensor. The first column presents the specific case when the feature set consists of only time domain samples. The length f of the feature set is 100. The second column presents the case when the feature set consists of only frequency domain values. The length f of the feature set is 25. The frequency spectrum was evaluated using the method proposed in section V. The third column presents the classification accuracy results with feature set formed using the combination of time domain and frequency domain features. The length f is 10 in this case with equal contribution of features from the time and frequency domain. In the table I, it can be observed that the accuracy achieved using frequency domain features is approximately similar to

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

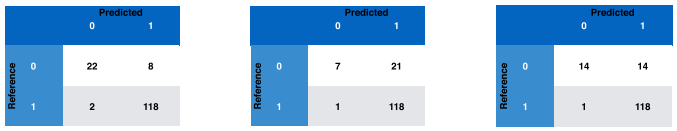


Fig. 12. Confusion matrices for the classification experiment with features: (a) Time domain features from all 8 sensors (b) Frequency domain features from all sensors (c) Time domain and frequency domain features from all sensors .

the accuracy achieved using time domain features. With the exception of sensor 7 and sensor 8, the obtained accuracy is almost same in both cases. This suggests that frequency domain features hold discrimination power equivalent to the time domain features. The improvements in accuracy in the third column are very small. More studies with sophisticated sensors need to be conducted to present the simultaneous application of features from both domains.The second experiment consists of evaluating classification accuracy using features from multiple sensors. Three cases were studied. In the first case, average sensor response values from all 8 sensors was used to perform classification accuracy. The accuracy obtained is 93.1%. Fig. 12(a) shows the confusion matrix for a specific experiment in this case. In the figure 12, the index 0 corresponds to the absence of ethylene and the index 1 corresponds to the presence of ethylene. In the second case, average frequency response values from all 8 sensors is used. The accuracy obtained is 86.5%. Fig. 12(b) shows the confusion matrix for this case. In the third case, both the average time domain and frequency domain values are used. The classification accuracy is 87.2%. Fig. 12(c) shows the confusion matrix for this case. Time domain features are useful but the prospects of utilizing frequency domain features cannot be neglected. In practical applications, sensor responses will be effected by noise. In such cases frequency domain features would prove more useful. It seems that more robust frequency domain features can be developed if sensor arrays consist of sensors selective to a single gas. Sensors labelled 4 and 6 are sensitive to carbon monoxide. Sensors labelled 5 and 7 are sensitive to methane. All other sensors were sensitive towards other gases. In spite of this fact, the accuracy from all individual sensors is good. This indicates that the fluctuations in the adsorption-desorption process are characteristic of a gas and can be used to increase the accuracy of sensors.

VII. Conclusion

A stochastic differential equation based model has been proposed for the chemical sensing process. The parameters of the SDE are usually evaluated using the known physical constants such as the binding and unbinding reaction rate constants. If the parameters are not available then the MAP estimation method can be used. The proposed SDE model is used to define the model for fluctuations in the signal. The obtained model is used to evaluate the instantaneous PSD using a time domain frequency estimation method. The forward and reverse reaction rate constants are primary parameters in determining the dynamics response of a chemical sensor. A choice of a lower forward reaction rate constant may help in diminishing noise levels but it also reduces the dynamic range of the sensor. It has been shown that the proposed model can be used in fluctuation sensing method for classification applications. The future work can be in the direction of using this model for optimal receiver design in molecular communication. It can be used for analyzing signal to noise ratio in biological and chemical sensors. Sophisticated techniques need to be designed for using the model in increasing the accuracy of chemical analyte quantification applications.

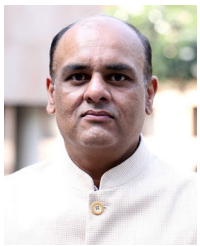
References

[1] M. Pierobon and I. F. Akyildiz, "Diffusion-based noise analysis for molecular communication in nanonetworks," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2532–2547, Jun. 2011.
[2] M. Pierobon and I. F. Akyildiz, "Noise analysis in ligand-binding reception for molecular communication in nanonetworks," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4168–4182, Sep. 2011.
[3] H. ShahMohammadian, G. G. Messier, and S. Magierowski, "Modelling the reception process in diffusion-based molecular communication channels," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2013, pp. 782–786.
[4] L. Lentka, J. Smulko, R. Ionescu, C. Granqvist, and L. Kish, "Determination of gas mixture components using fluctuation enhanced sensing and the LS-SVM regression algorithm," *Metrol. Meas. Syst.*, vol. 22, no. 3, pp. 341–350, 2015.
[5] S. Sonkusaic, Y. Kim, and A. Agarwal, "A CMOS front-end IC for fluctuation enhanced sensing," in *Proc. SENSORS*, p. 4, 2005, doi: 10.1109/ICSENS.2005.1597924}.
[6] G. Tulzer *et al.*, "Kinetic parameter estimation and fluctuation analysis of CO at SnO_2 single nanowires," *Nanotechnology*, vol. 24, no. 31, 2013, Art. no. 315501.
[7] G. Tulzer and C. Heitzinger, "Fluctuations due to association and dissociation processes at nanowire-biosensor surfaces and their optimal design," *Nanotechnology*, vol. 26, no. 2, 2014, Art. no. 025502.
[8] M. Shamaiah, X. Shen, and H. Vikalo, "Estimating parameters of sampled diffusion processes in affinity biosensors," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 3228–3239, Jun. 2012.
[9] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Chelmsford, MA, USA: Courier Corporation, 2007.
[10] V. Y. Kontorovich and V. Z. Lyandres, "Stochastic differential equations: An approach to the generation of continuous non-Gaussian processes," *IEEE Trans. Signal Process.*, vol. 43, no. 10, pp. 2372–2385, Oct. 1995.
[11] L. J. Allen and E. J. Allen, "A comparison of three different stochastic population models with regard to persistence time," *Theoretical Population Biol.*, vol. 64, no. 4, pp. 439–449, 2003.
[12] M. Baxter, A. Rennie, and A. J. Rennie, *Financial Calculus: An Introduction to Derivative Pricing*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[13] D. J. Higham, "Modeling and simulating chemical reactions," *SIAM Rev.*, vol. 50, no. 2, pp. 347–368, 2008.
[14] S.-P. Sheng, M. Liu, and R. Saigal, "Data-driven channel modeling using spectrum measurement," *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1794–1805, Sep. 2015.
[15] G. Zhao, X. Qian, B.-J. Yoon, F. J. Alexander, and E. R. Dougherty, "Model-based robust filtering and experimental design for stochastic differential equation systems," *IEEE Trans. Signal Process.*, vol. 68, pp. 3849–3859, Jun. 2020.
[16] E. Allen, *Modeling With Itô Stochastic Differential Equations*. Dordrecht, The Netherlands: Springer, 2007.
[17] S. M. Iacus, *Simulation and Inference for Stochastic Differential Equations: With R Examples*. New York, NY, USA: Springer-Verlag, 2008.
[18] L. Boubchir and B. Boashash, "Wavelet denoising based on the MAP estimation using the BKF prior with application to images and EEG signals," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 1880–1894, Apr. 2013.
[19] C. Kwan, G. Schmera, J. M. Smulko, L. B. Kish, P. Heszler, and C.-G. Granqvist, "Advanced agent identification with fluctuation-enhanced sensing," *IEEE Sensors J.*, vol. 8, no. 6, pp. 706–713, Jun. 2008.
[20] J. Smulko, C.-G. Granqvist, and L. B. Kish, "On the statistical analysis of noise in chemical sensors and its application for sensing," *Fluctuation Noise Lett.*, vol. 1, no. 3, pp. L147-L153, 2001.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

[21] J. Huillery, F. Millioz, and N. Martin, "Gaussian noise time-varying power spectrum estimation with minimal statistics," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5892–5906, Nov. 2014.
[22] A. Napolitano, "Discrete-time estimation of second-order statistics of generalized almost-cyclostationary processes," *IEEE Trans. Signal Process.*, vol. 57, no. 5, pp. 1670–1688, May 2009.
[23] L. Toth, G. Efthivoulidis, and Y. P. Tsividis, "Noise analysis of externally linear systems," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 47, no. 12, pp. 1365–1377, Dec. 2000.
[24] P. Bolcato and R. Poujois, "A new approach for noise simulation in transient analysis," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, 1992, pp. 887–890.
[25] V. Vasudevan, "A time-domain technique for computation of noise-spectral density in linear and nonlinear time-varying circuits," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 51, no. 2, pp. 422–433, Feb. 2004.
[26] D. Kilinc and O. B. Akan, "Receiver design for molecular communication," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 12, pp. 705–714, Dec. 2013.
[27] L.-S. Meng, P.-C. Yeh, K.-C. Chen, and I. F. Akyildiz, "On receiver design for diffusion-based molecular communication," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 6032–6044, Nov. 2014.
[28] G. Schmera, C. Kwan, P. Ajayan, R. Vajtai, and L. B. Kish, "Fluctuation-enhanced sensing: Status and perspectives," *IEEE Sensors J.*, vol. 8, no. 6, pp. 714–719, Jun. 2008.
[29] B. Ayhan *et al.*, "Fluctuation enhanced sensing (FES) with a nanostructured, semiconducting metal oxide film for gas detection and classification," *Sensors Actuators B Chem.*, vol. 188, pp. 651–660, 2013.
[30] G. Tulzer and C. Heitzinger, "Noise and fluctuations in nanowire biosensors," *IFAC-PapersOnLine*, vol. 48, no. 1, pp. 761–765, 2015.
[31] L. B. Kish *et al.*, "Fluctuation-enhanced sensing for biological agent detection and identification," *IEEE Trans. Nanotechnol.*, vol. 10, no. 6, pp. 1238–1242, Nov. 2011.
[32] M. J. Deen, "Low-frequency noise in semiconductor devices-state-of-the-art and future perspectives plenary paper," in *Proc. Int. Conf. Noise Fluctuations*, 2017, pp. 1–4.
[33] M. Deen, M. Shinwari, J. Ranuárez, and D. Landheer, "Noise considerations in field-effect biosensors," *J. Appl. Phys.*, vol. 100, no. 7, 2006, Art. no. 074703.
[34] S. Das, H. Vikalo, and A. Hassibi, "On scaling laws of biosensors: A stochastic approach," *J. Appl. Phys.*, vol. 105, no. 10, 2009, Art. no. 102021.
[35] A. D'Amico, C. Di Natale, C. Falconi, G. Pennazza, M. Santonico, and I. Lundstrom, "Equivalent electric circuits for chemical sensors in the Langmuir regime," *Sensors Actuators B Chem.*, vol. 238, pp. 214–220, 2017.
[36] S. Gomri, J.-L. Seguin, and K. Aguir, "Modeling on oxygen chemisorption-induced noise in metallic oxide gas sensors," *Sensors Actuators B Chem.*, vol. 107, no. 2, pp. 722–729, 2005.
[37] O. C. Ibe, *Elements of Random Walk and Diffusion Processes*. Hoboken, NJ, USA: Wiley, 2013.
[38] T. G. Kurtz, "Solutions of ordinary differential equations as limits of pure jump Markov processes," *J. Appl. Probability*, vol. 7, no. 1, pp. 49–58, 1970.
[39] J. Fonollosa, I. Rodríguez-Luján, M. Trincavelli, A. Vergara, and R. Huerta, "Chemical discrimination in turbulent gas mixtures with MOX sensors validated by gas chromatography-mass spectrometry," *Sensors*, vol. 14, no. 10, pp. 19336–19353, 2014.
[40] A. Vergara, J. Fonollosa, J. Mahiques, M. Trincavelli, N. Rulkov, and R. Huerta, "On the performance of gas sensor arrays in open sampling systems using inhibitory support vector machines," *Sensors Actuators B Chem.*, vol. 185, pp. 462–477, 2013.
[41] G. Sakai, N. Matsunaga, K. Shimanoe, and N. Yamazoe, "Theory of gas-diffusion controlled sensitivity for thin film semiconductor gas sensor," *Sensors Actuators B Chem.*, vol. 80, no. 2, pp. 125–131, 2001.
[42] A. Ghosh and S. Majumder, "Modeling the sensing characteristics of chemi-resistive thin film semi-conducting gas sensors," *Phys. Chem. Chem. Phys.*, vol. 19, no. 34, pp. 23 431–23 443, 2017.
[43] J. K. Møller and H. Madsen, "From state dependent diffusion to constant diffusion in stochastic differential equations by the lamperti transform," *DTU Informatics*, 2010.
[44] L. Grebenstein *et al.*, "A molecular communication testbed based on proton pumping bacteria: Methods and data," *IEEE Trans. Mol., Biol. Multi-Scale Commun.*, vol. 5, no. 1, pp. 56–62, Oct. 2019.
[45] N. Farsad, D. Pan, and A. Goldsmith, "A novel experimental platform for in-vessel multi-chemical molecular communications," in *Proc. IEEE Global Commun. Conf.*, 2017, pp. 1–6.
[46] D. J. Graham, B. Jaselskis, and C. E. Moore, "Development of the glass electrode and the pH response," *J. Chem. Educ.*, vol. 90, no. 3, pp. 345–351, 2013.
[47] T. Nakano, M. J. Moore, F. Wei, A. V. Vasilakos, and J. Shuai, "Molecular communication and networking: Opportunities and challenges," *IEEE Trans. Nanobiosci.*, vol. 11, no. 2, pp. 135–148, Jun. 2012.
[48] V. M. Aroutiounian *et al.*, "Noise spectroscopy of gas sensors," *IEEE Sensors J.*, vol. 8, no. 6, pp. 786–790, Jun. 2008.

Abhishek Grover received the B.Tech degree in electronics and communication engineering from Delhi Technological University, New Delhi, India, in 2015 and the M.Sc. (Research) degree in 2018 from the Bharti School of Telecommunication Technology and Management, Indian Institute of Technology Delhi, Delhi, New Delhi, India, where he is currently working toward the Ph.D. degree. His research interests include instrumentation and signal processing.



Brejesh Lall received the B.Tech. degree in electronics and communication engineering, the master's degree in signal processing from the Delhi College of Engineering, New Delhi, India, and the Ph.D. degree in signal processing from the Indian Institute of Technology (IIT) Delhi, New Delhi, India, in 1999. He is currently a Professor with the Department of Electrical Engineering, IIT Delhi.

In 1997, he joined Aricent. In 2005, he joined IIT Delhi as an Assistant Professor. From 2010 to 2018, he was an Associate Professor. He is actively working in the area of image processing, signal processing, and communication. His research interests include object representation, tracking and classification, odometry, depth map generation, representation and rendering, vector sensor based underwater acoustic communications, and performance issues in molecular communications.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:14:28 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395978

A Survey on Analysis of Fake News Detection Techniques

Shailender Kumar
Computer Science Department
Delhi Technological University
New Delhi, India.
shailenderkumar@dce.ac.in

Subodh Kumar
Computer Engineering
Delhi Technological University
New Delhi, India.
subodh9696kumar@gmail.com

Pooja Yadav
Computer Engineering
Delhi Technological University
New Delhi, India.
pooja796760@gmail.com

Meghna Bagri
Computer Engineering
Delhi Technological University
New Delhi, India.
meghnabagri98@gmail.com

Abstract- Social media is one of the most available news sources these days for many folks worldwide due to their low value, quick access, and fast spread. However, this comes with some confusing signs and significant risks of exposure to 'false stories' written to mislead readers. Such information can affect the public's voice and allow evil groups to control the outcome of public events, such as elections. Fake and misleading news can have a real impact on those who find themselves as targets. This paper focuses on analysis of 2017 to 2021 papers and analysis of different fake news detection techniques. This survey gives a far-reaching review about the recent and past examinations on false news detection using different ML algorithms.

Keywords- Fake news detection, machine learning, deep learning, social media, ensemble techniques, preprocessor.

I. INTRODUCTION

Once in a while, we all have come across the term what's App University. This is not just the term which is specifying the access of information through social media site but also concern with the reliability of information through sources like social media and various other platforms of information and entertainment. The widespread and accessibility of the internet gives information to the masses at a click, but it also made it hard to verify the information. With the question of reliability and verification, the question of fake news also comes.

The information and news regarding the spread of global pandemic covid-19 like self-verification of being infected, it spread based on temperature, about vaccination; the speech of political figures during public addressing and the unverified statement of them regarding the military invasion, about developing and doing public goods; false and misleading images of people for malign or praise them; manipulation of videos and audios are some of the cases and example of fake news.

Table1. DEFINING SOME TERMS RELATED TO FAKE NEWS

Word	**Meaning**
Rumor	"*An moving information whose truth status was not at the time of posting"*
Fake News	"*A news article that is definitely false"*
Hoax	"*A intentionally created falsehood made to pose as truth"*
Clickbait	*"An intentional strategy which is intended to attract attention and encourage visitors to collect revenue"*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:07:27 UTC from IEEE Xplore. Restrictions apply.

Misinformation	*"Fake or not so correct information which is spread unintended"*

Fake news is not a very new thing, but its extensive penetration is very new and this is highly supported by various media sources. Facebook, whatsApp, blogs, Twitter, YouTube, and sadly but the news channels are also included in the list of spreading fake news. This can't be said that these channels and online platforms are unaware of this. Not long ago Facebook owner Mark Zuckerberg was called by Senate and questions him about its minimum control over the spread of fake and hate news through its platform and even after this interrogation, the Fb is still used in the same way.

A view of an individual becomes information for others and based on those biased and unverified information others build their surroundings. The increase in information based on this approach made a society running with false ideas. This falsification of information is hardly verified by an individual as they busy themselves in their individual and virtual world. But the society based on the false and biased idea is a bomb which tickles every time to burst whenever a new idea intervenes and become a threat to the dominance of the exiting idea which is neither good for an individual or a society.

In the last few years, we have heard and read about the statement of ministers in national science congress about saying Darwin's theory of evolution a myth, saying the first plastic surgery done millions of years ago with some verse on Ganesha, the first airplane of Ravana with much advancement than today's plane, and the utility and usefulness of cow's waste are an example of fake news which are building people's perception and making them doubt of sciences. Similarly, during the time of demonetization, the news channels were spreading false news about the chip in new notes of rupees 500 and 2000. Also, with the help of social media one's view about the other community is used to create division at political, religious, caste, and nationality levels.

These examples show how fake news is controlling and making our scientific, religious, social, and other beliefs and relations. Although the main beneficiaries of this fake news are the party in power for their political gain, others are also gaining momentary.

II. RELATED WORK

In 2013, Aditi Gupta, Henmark Lamba and Anupam Joshi achieved more than 90% correct result in identifying false images from twitter of Hurricane Sandy which impacted the United States. Meanwhile, they did a character-ization analysis to analyze the impact patterns of the fake pictures by analyzing mor than 10,000 images on Twitter. During this time, they worked on Niave Bayes and Decision tree model. After applying these two ML algorithm they arrives at good result having accuracy of 97% by Decision Tree. [1]

In 2017, Elena Kochkina, Arkaitz Zubiaga & Maria Liakata worked on classification of rumor stance on social media platform with the help of sequential classifiers. In this they use Twitter as their social media platform and describe tweets into 4 categories: 1.Support, 2.Deny, 3.query and 4. comment on an earlier post. They used four sequential classifier-hawkes processes, Long Short Term Memory(LSTM), linear CRF and tree CRF on 8 data sets and all data are related to breaking news. They discover sequential classifiers that use the recitation property in social media interaction outperform non sequential classifiers also LSTM works better than other sequential classifiers. [2]

In 2018, Kalina Bontch., Ahmet Aker, Maria Liakata work on rumour detection using NLP and data mining Methods. They defines false news that circulates on social media into two types: long standing rumors and new emerging rumor generate during recent events. They develops a rumor classification system that consist of 4 parts: 1.Detection of rumor, 2.Tracking of rumor, 3. Stances of rumor and 4.Veracity of rumor. And use this system on the PHEME dataset which is publicly available for rumors and non rumors. [3]

In 2018, C.M.M Kotteti, Na Li and Lijun Qian work on increasing the detection of fake news with data imputation. To improve performance they used a novel data preprocessing method to fill the missing value in the raw dataset. With the help of data modeling, they applied missing values for numerical and hierarchical attributes. For hierarchies they select the most frequent value in columns and are numeric for the average value of the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:07:27 UTC from IEEE Xplore. Restrictions apply.

column. He did 3 things to cover the missing values. 1. Removed columns with missing values, 2. Missed values with empty text and 3. Used data impersonation techniques to apply missing values and found that multilayer perceptron (MLP) classes improved accuracy by 16%. [4]

In 2018, Supanya Aphiwongsophn and Prabhas Chongstitvatana purpose the ML algorithm to identify fake news. In this paper three popular methods are used: 1.Naive Bayes 2. Support Vector Machine and 3. Neural Network. They used normalization method for cleaning data so that it works better with correct data. In this paper they found that Naive Bayes has an accuracy of 96.08% and the other two complex techniques has an accuracy of 99.90%. [5]

In 2018, a. Jain and A. Kasbe work on detecting fake news and they proposed a method so that we can implement this method on Facebook. He used Naive Bayes for forecasting. They used a dataset from Github with 11000 articles divided into (index, text, title and label). Apart from politics, this data contains news related to science and business. For implementation they used both the title and text for their primary source and also added some references by n-gram And then he compared the results and find that Naïve Bayes (on text with n-grams) gives the accuracy of 0.931 and they also showed some ways to improve this model. [6]

In 2019, Deepayan Bhowmik**,** Oluwaseun Ajao and Shahrzad Zargari proposed a model that identify false news tweets from twitter post using combination of (CNN) and (RNN) models. For the dataset they collected 5,800 tweets centered on five rumor stories: Charlie Hebgo, Sydney Siege, Germanwing Crash, Ottawa Shooting, and Ferguson Shooting. Their proposed work on hybrid of CNN & RNN intuitionally identifies important feature related with false news stories without any prior knowledge of news and achieve more than 80% accuracy. [7]

In 2019, Varshil Mehta and Wenlin Han work on performance evaluation of fake news detection methods. They divide the dataset of fake news into 2 categories. The first is news and the second is the social context model and they divide news into 2 categories of visual (picture, video) and linguistic (text, title) based. They compared performance between traditional ML methods (Naïve Bayes , Random forest) and the latest deep learning methods (LSTM DROP, LSTM-CNN). The purpose of this paper is to provide a basis so that people can choose between these two approaches. They found that the hybrid CNN - RNN model gives better performance/ results. [8]

In 2019, J. C.S. Reece, A. Correia, F. Murai, A. Veloso and F. Benevuto works on searching on a wide variety of features from news articles, posts and stories that can help predict fake news with greater accuracy. He showed the importance of these new features for the evaluation of fake news. Some of those features are bias, reliability / trustworthiness, engagement, domain location, and temporal patterns. They used a dataset containing 2282 buzzfeed (news articles). They used KNN, Naïve Bayes, Random Forest, Support Vector Machine and XGBoost algorithm for evaluation and to discuss the opportunities and challenges of this approach and they found out that XGBoost work better than all with the accuracy of 0.86. [9]

In 2020, Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad classified fake news articles using machine learning models and ensemble techniques (Logistic Regression, Random Forest, Perez- LSVM). In this paper various textual properties are used to differentiate fake new from real news. The experiment was conducted on 4 publicly available dataset which is of different domians and also calculated the performance by performance metrics. The maximum accuracy is 99% achieved by random forest and Perez-LSVM on ISOT Fake News Dataset. [10]

In 2020, C. Yuan, Q. Ma., W. Zhou proposed a model structure-aware multi-head attention network (SMAN) based approach to detect fake news. This method is based on the reliability of both publishers and users. The datasets used for this approach were real-world datasets. This approach can be used for early detection of fake news that optimizes the detection process with the help of asymmetric graphs between publisher and users. They use this model on 3 different dataset (Twitter 15, Twitter 16, Weibo) and find that this model gives very high accuracy. [11]

In 2021, S.M. Shifath, Md. S. Islam and Md. F. Khan proposed a transformer-based approach for detecting COVID-19 fake news. They performed experiments on

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:07:27 UTC from IEEE Xplore. Restrictions apply.

traditional language models and CNN. The dataset is social media posts related to COVID-19 and labels indicating whether the posts are fake or real. They also experimented with transformer-based models and tested different hyper parameters. The highest accuracy is 0.979 which is shown by RoBERTa. [12]

III. MACHINE LEARNING METHODS

Multiple supervised classification algorithms used to evaluate accuracy:

Accuracy - Accuracy is the display of most natural illustration and is only part of the preconceived notion of complete ideas. One might think that, if we could possibly achieve higher clarity, our model is correct. Indeed, precision is a rare measure but as long as you have metrics data where positive and negative estimates are almost the same.

Accuracy □ (TN +TP) / (TP+FP+TN+FN).

Precision - It is a measure of positively anticipated ideas in the perfectly expected positive outlook. Higher accuracy indicates a lower deceptive value.

Precision □ TP / (FP+TP)

Recall or Sensitivity- It is the, measure of the number of positive ideas that are expected of all ideas in the original class - yes.

Recall □ TP / (TP+FN)

F1 score -F1 Score is standard with Precision and Recall weights. As a result, the score is looking for false negatives and false positive both. Naturally it's not as straight forward as precision, but F1 score is often more important than accuracy, like in the event that it is still distributed in separate categories. Exactness works best when counterfeit and illegal pros and cons have costs to compare.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Where TP, TN, FP and FN indicate genuine positive, genuine negative, false positive and false negative, individually.

IV. RELATIVE STUDY OF TECHNIQUES IN FAKE NEWS DETECTION

Author & Year	Dataset	Technique	Accuracy
A. Gupta & H. lamba [1]	Twitter feeds (total 1,782,526 tweets)	Naïve Bayes Decision Tree	0.91 0.97
Arkaitz Zubiaga & Elena Kochkina [2]	Supporting tweets on twitter	SVM Linear CRF Tree CRF	0.657 0.603 0.552
M. kotteti, X. Dong, Na Li & L.Qian [4]	LIAR dataset containing (12,836 records)	SVC LinearSVC-CS Decision Tree MLP Classifier Gradient boosting	0.245 0.195 0.394 0.457 0.442
Supanya Aphiwon. & Prabhas Chongsti. [5]	Twitter feeds	Naive Bayes Neural Network SVM	0.96 0.99 0.99
A. Jain & A. kasbe [6]	From Github (11,000 articles)	NB (on title) NB (on text) NB(on title with n-grams) NB (on text with n-grams)	0.806 0.912 0.807 0.931
Oluwaseu Ajao, & Shahrzad Zargari [7]	5800 tweets on five rumoured stories	LSTM LSTM- DROP LSTM-CNN	0.82 0.73 0.80
Wenlin Han & Varshil Mehta [8]	Collect data in (multi-media, text, audio, hyperlink	Naïve Random forest LSTM LSTM DROP LSTM-CNN	0.67 0.56 0.82 0.73 0.80
J. C.S. Reece, A. Correia, F.	2282 Buzzfeed	KNN NB	0.80 0.72

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:07:27 UTC from IEEE Xplore. Restrictions apply.

Murai, & A. Veloso [9]	news article	RF SVM XGB	0.85 0.79 0.86
Iftikhar Ahmad, Md Yousaf & Suhail Yousaf [10]	ISOT Fake News Dataset	Logistic Regression Random Forest Perez-LSVM	0.97 0.99 0.99
C. Yuan, Q. Ma, W. zhou & J. Han [11]	Twitter 15 Twitter 16 Weibo	SMAN SMAN SMAN	0.929 0.935 0.956
Md. Faiyaz Khan and Md. Saiful Islam [12]	Social media posts related to COVID-19	Bi-LSTM 1D-CNN BERT ROBERTa	0.928 0.926 0.971 0.979

V. DISCUSSION AND FUTURE WORK

This section contains the limitations of the existing work on identifying fake news and the idea for future work.

As there is a lot of work already done on identifying fake news. Some researchers used machine learning techniques, while others used deep learning techniques and also there are some researchers who used artificial intelligence but yet no one got 100% accuracy on identifying fake news.

In the paper by Wenlin Hain and Varshil Mehta [8], he studied ML and Deep learning methods. And he gave the idea for future work to choose between ML and deep learning methods based on the balance between lightweight and accuracy.

But there are not many researches where we ensemble different machine learning methods for identifying fake news.

Our proposed method consists of combination of different ensemble techniques. The effect of ensembling different techniques will play an important role in the novelty of our research work.

Also, the preprocessor plays an important role in extracting the information from raw dataset and there are different algorithms for data preprocessing.

So, we can also study how these algorithms can affect our result, and also we can combine both of the above mentioned work(ensemble techniques of different algorithms for data preprocessing) for future work.

In the paper [4] they filled the missing value with imputation techniques, but they performed this only on 1 dataset, so in the future we can work on this technics with different types of datasets and compare the results.

VI. CONCLUSION

In this survey paper, we studied the work done till now in the field of fake news detection with different techniques and approaches. This paper also illustrated the tools available to simplify the analysis process and to compare existing methods and techniques.

Our research will be targeted towards applying better optimized ensemble technique and to also include some powerful word embedding techniques. We also intend to produce consistent results with the help of ensemble techniques with the dataset which contains data from different domains. Ensemble techniques can make better predictions and achieve better performance while reducing the dispersion.

VII. REFERENCES

[1] A. Gupta, H. Lamba, P. Kumaraguru and A. Joshi, "Faking sandy, identifying fake images of hurricane sandy on twitter," in *In proceedings of the 22nd international conference on World Wide Web. ACM*, 2013.

[2] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik and I. Augenstein, "Discourse-Aware Rumour stance classification in social media using sequential classifier.," *Information Processing & Management,* vol. 54, pp. 273-290, 2017.

[3] A. Aker, A. Zubiaga, K. Bontcheva, M. LIakata and R. oricter, "Detection and Resolution of rumours in social media: A survey," *ACM comput.,* vol. 51, p. 36, Feb 2018.

[4] C. M. M. Kotteti, X. Dong, N. Li and L. Qian, "Fake News Detection Enhancement with Data Imputation," in *IEEE 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing*, 2018.

[5] S. Aphiwongsophon and P. Chongstitvatana, "Detecting fake news with Machine Learning methods," in *15th international conference on Electronics, computer, Telecommunication and Information Technology*, 2018.

[6] A. Jain and A. Kasbe, "Fake News Detection," in *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences*, 2018.

[7] O. Ajao, D. Bhowmik and S. Zargari, "Fake news Identification on Twitter with combination of CNN and RNN models," in

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:07:27 UTC from IEEE Xplore. Restrictions apply.

International conference on social media & society, Copenhagen,Denmark SMSociety, 2018.

[8] W. Han and V. Mehta, " Han, Wenlin; Mehta, Varshil (2019). [IEEE 2019 IEEE International Conference on Industrial Internet (ICII) - Orlando, FL, USA (2019.1Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation.," in *IEEE 2019 IEEE International Conference on Industrial Internet (ICII)*, Orlando, FL, USA, 2019.

[9] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto and E. Cambria, "Supervised Learning for Fake News Detection," in *IEEE Intelligent Systems*, 2019.

[10] I. Ahmad, M. Yousaf, S. Yousaf and M. O. Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods," vol. 2020, p. 11, 2020.

[11] C. Yuan, Q. Ma, W. Zhou, J. Han and S. Hu, "Early Detection of Fake News by Utiliz ing the Credibility of News, Publishers, and Users Based on Weakly Supervised Learning," *arXiv:2012.04233v2 [cs.CL],* p. 11, 2020.

[12] S. Shifath, M. F. Khan and M. S. Islam, "A transformer based approach for fighting COVID-19 fake news," *arXiv:2101.12027v1 [cs.CL],* p. 9, 2021.

[13] S. I. Manzoor, Nikita and D. J. Singla, “Fake News Detection using Machine Learning Approaches: A systematic Review,” in *Third International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019.

[14] P. Ksieniewicz, P. Zyblewski, M. Choras, R. Kozik, A. Gielczyk and M. Wozniak, "Fake News Detection from Data Streams," in *IEEE 2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom, 2020.

[15] B. Ghanem, S. P. Ponzetto, P. Russo and F. Rangel, "FakeFlow: Fake News Detection by Modeling the Flow of Affective Information," *arXiv:2101.09810v1 [cs.CL],* 2021.

[16] N. Vo and K. Lee, "Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News," *arXiv:2010.03159v1 [cs.CL],* p. 15, 2020.

[17] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM. Comput. Surv. 1,* p. 37, 2020.

[18] S. Kwon, M. Cha, K. Jung, W. Chen and Y. Wang, "Prominent features of rumor propagation in online social media.," in *IEEE 13th International Conference on Data Mining (ICDM)*, 2013.

[19] S. H. Sharma, "(PDF) A Survey on Decision Tree Algorithms of Classification.," 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:07:27 UTC from IEEE Xplore. Restrictions apply.

2021 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021

A Unified Audio Analysis Framework For Movie Genre Classification Using Movie Trailers

Aditya Sharma
Biometric Research Laboratory,
Department of Information Technology
Delhi Technological University,
New Delhi, India
adityas2000@yahoo.com

Mayank Jindal
Biometric Research Laboratory,
Department of Information Technology
Delhi Technological University,
New Delhi, India
mayankjindal2017@gmail.com

Ayush Mittal
Biometric Research Laboratory,
Department of Information Technology
Delhi Technological University,
New Delhi, India
ayushheritage18@gmail.com

Dinesh Kumar Vishwakarma
Biometric Research Laboratory,
Department of Information Technology
Delhi Technological University,
New Delhi, India
dinesh@dtu.ac.in

2021 International Conference on Emerging Smart Computing and Informatics (ESCI) | 978-1-7281-8519-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ESCI50559.2021.9396892

***Abstract*— The audio content of the movie trailers carries various prominent characteristics that can be exploited to predict the genres of the movie. None of the previous approaches have focused on extracting the audio features from the trailers using a clustering-based unsupervised learning approach followed by distance-based supervised learning. Hence, in this paper, we propose a novel framework for movie genre classification using audio features of movie trailers. Movie trailers belonging to the five most generic and popular genres (i.e., Action, Romance, Horror, Science Fiction, and Comedy) are considered in the work. The proposed AFAnet architecture is trained on a total of 78 features including 68 audio features and 10 distance-based features extracted after the k-means clustering on the audio chunks. The cross-dataset validation is performed using the standard LMTD dataset to validate the performance of our proposed framework. The results obtained depict that our model has performed excellently and show the robustness of the proposed approach.**

Keywords*— Movie genre classification, Audio Segmentation, K-means clustering, Artificial Neural Network

I. INTRODUCTION

Audio content in movie trailers contains a lot of useful information for predicting the genre of the movie. Audio content comprises various back-ground and foreground sounds such as any environmental noises from vehicles, devices, animals, alarms, etc. along with dialogues and music. Hence, audio content provides many prominent features to predict movie genres. These audio features provide useful mapping with the movie genres. For example, sounds such as screaming, crying, demonic laugh, etc. maps with the 'horror' genre. Sounds of blasts, vehicular noise, etc., map with the 'action' genre.

In this study, we propose a novel framework for movie genre classification tasks utilizing only audio features from movie trailers of length varying from 120 sec to 300 sec. The audio is further chopped into several non-overlapping disjoint chunks of 5 seconds each**.** An unsupervised learning approach is presented to cluster the chunks based on standard audio characteristics such as chroma deviation, energy, energy entropy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, MFCC, chroma vector, zero-crossing rate, etc.

The presented approach is carried out in multiple phases for classifying the final genre. In phase 1, dataset preparation is done. The movie trailers belonging to the five most generic genres (Action, Romance, Horror, Science Fiction, and Comedy) are downloaded from the IMDb website. The trailers are preprocessed in phase 2, where the audio is extracted from the video trailers. The audio is split into small chunks. In phase 3, the chunks are clustered based on various audio characteristics using k-means clustering. Features are extracted based on the distance of a chunk with all the formed clusters. In phase 4, we design and train the AFAnet architecture using the audio features extracted in the previous stage and other audio features. Finally, in phase 5, the proposed architecture is evaluated using the AU (PRC) metric.

The major contributions of the paper are as follows:

- The proposed approach of multi-label movie genre classification using only audio features from the trailers is novel as far as we know in the domain of movie genre classification. None of the past literature has focused on extracting the features in the way proposed in this work.
- We propose a novel ASM and AFAnet architecture to predict movie genres using audio-based features
- The proposed architecture is evaluated using the AU(PRC) metric via cross-dataset validation on the standard LMTD-9 dataset. The results obtained depict that our model has performed excellently and show the robustness of our proposed approach.

The remaining portion of the paper is organized as: In Section II, the related past work on movie genre classification is described. In Section III, the proposed framework is explained comprehensively. Section IV demonstrates the model performance and cross-dataset validation results based on AU(PRC) metric. A conclusion derived from the work is presented in Section 0.

II. RELATED WORK

This segment describes the previous related work done in the movie genre classification domain. Several approaches were proposed by the researchers some based on low-level

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:41:37 UTC from IEEE Xplore. Restrictions apply.

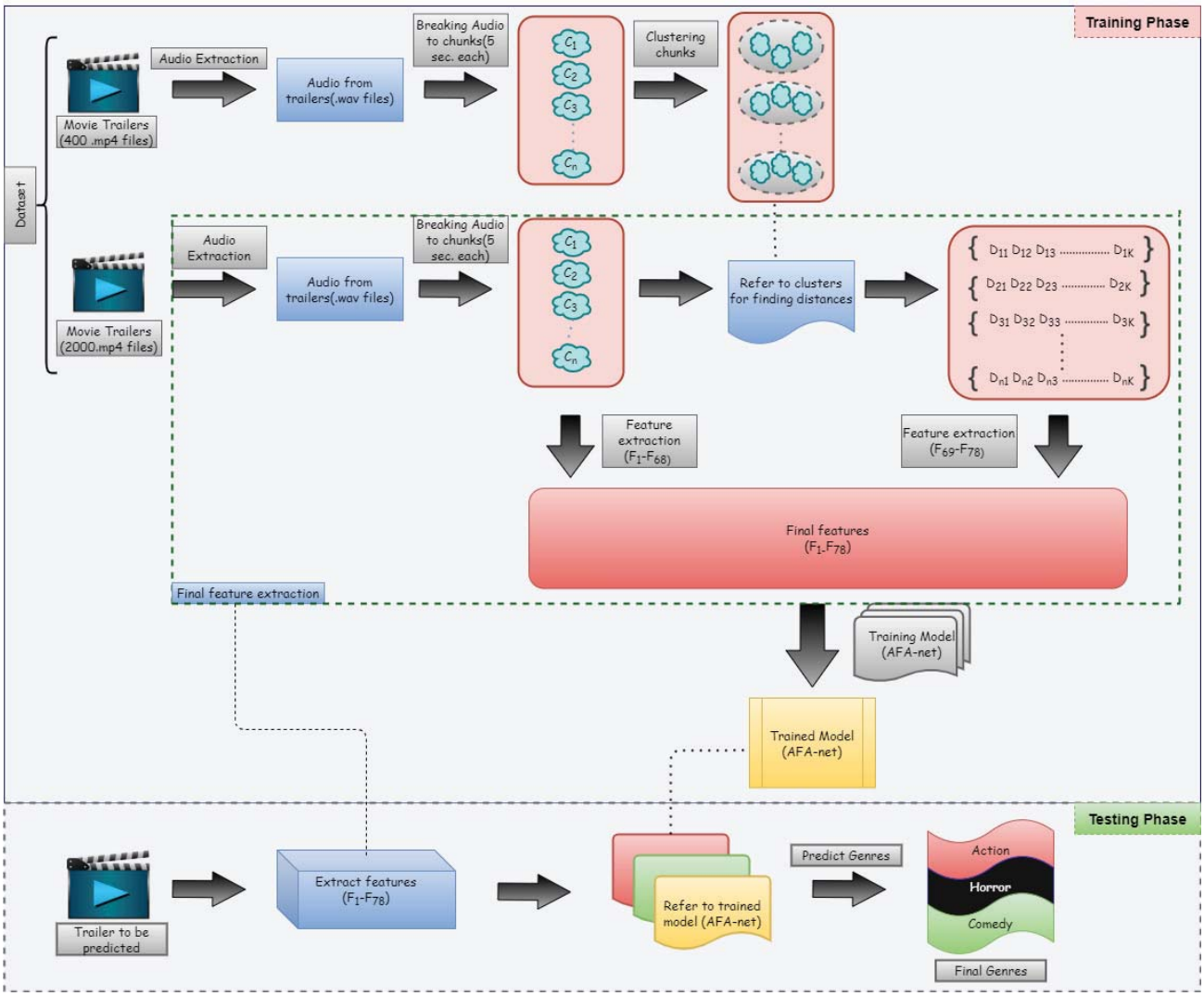


Fig. 1. Pipeline of Proposed Framework

features including color variances, motion, intensity, audio power, homological persistence, plot, signal, shot length, etc., and high-level features including visual, audio, and textual features from movie posters, movie trailers, and metadata.

Amidst the past work employing low-level features include, Rasheed et al. [1] who extracted various features such as motion, color variance, shot length, etc. from the movie previews, categorizing each for a specific genre as action or non-action genre based on only computable visual cues including average movie shot length and visual disturbance feature. Sugano et al. [2] proposed to use audio-visual-based MPEG compressed data from movies taking some low to mid-level features including motion intensity, shot length, and audio class into account to predict shot genre. Yuan et al. [3] proposed to extract spatial and temporal features from a video containing representative information by proposing a hierarchical ontology scheme on genres. Ertugrul et al. [4] proposed plot summary-based movie genre classification using Bi-LSTM networks. The plot summary is divided into sentences, and the results are obtained by considering the posterior probabilities assigned to each sentence. Álvarez et al. [5] utilized film characteristics, low-level descriptors, and visual information to perform film classification. Doshi et al. [6] applied features such as homological persistence through topological data analysis of movie plots to perform the detection of overlapping movie genres. Kos et al. [7] utilized low-level features, including color and edges from movie posters to perform multi-label movie genre classification. Jin et al. [8] proposed a multimodal approach from video features utilizing multiple low-level MPEG-7 audio-visual descriptors. Low to mid-level descriptors alone are not enough to classify the movie genre. Thus, several studies utilized high-level features from video content to perform movie genre classification task in a better way.

Several recent works focus on high-level features from movie posters, trailers, and meta-data (plots). High-level features describe a lot of information about the genre of the movie. Among these works, Zhou et al. [9] employed scene categorization by extracting state of art scene descriptors from movie trailers. The work gives a better understanding of high-level semantic features; utilizing a bag of visual words as model descriptors from the movie trailer. Simoes et al. [10] proposed CNN-MoTion encapsulating deep CNN-based architecture for detecting the relevant and higher-order complex features to predict movie genres. Wehrmann et al. [11] proposed an ultra-deep convolution network with residual connections by utilizing temporal facts from image-based features from the frames to predict movie genres. Chu

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:41:37 UTC from IEEE Xplore. Restrictions apply.

et al. [12] utilized movie posters for movie genre classification task utilizing a deep convolution network to describe visual appearances and object-based information. Yadav et al. [13] proposed an emotion-based approach by utilizing emotions from frames gathered from movie trailers to predict movie genres. Huang et al. [14] utilized 277 features, all based on audio and visual video from movie trailers to propose a self-adaptive harmony search algorithm, later on, used stacked SVM to perform the classification task. Fish et al. [15] performed fine-grained clustering on video clips extracted from movie trailers to perform the classification task. Mangolin et al. [16] proposed a multimodal approach utilizing audio, video frames, posters, subtitles, synopsis as features, and various model architectures, fusing them all to predict movie genres.

TABLE I. PARAMETERS OF AFANET

Layer Name : Type	Description	Input Shape	Output Shape	Parameters #
dense_1 : Dense	Dense Layer with 256 neurons	(78,)	(256,)	22784
dense_2 : Dense	Dense Layer with 256 neurons	(256,)	(256,)	65792
dropout_1 : Dropout	Dropout Layer with rate = 0.4	(256,)	(256,)	0
dense_3 : Dense	Dense Layer with 64 neurons	(256,)	(64,)	16448
dense_4 : Dense	Dense Layer with 32 neurons	(64,)	(32,)	2080
dropout_2 : Dropout	Dropout Layer with rate = 0.2	(32,)	(32,)	0
dense_5 : Dense	Dense Layer with 5 neurons	(32,)	(5,)	165

The main motive behind our work is to perform movie genre classification based on audio as it depicts a lot of information about a movie relating to its genres. Also, most of the previous studies lack proper audio analysis for movie genre classification. Thus, our work primarily focuses on the audio features for classifying the movie genres.

III. PROPOSED METHODOLOGY

The proposed approach is divided into two broad models 1) **ASM** (Audio Segmentation Model) is based upon unsupervised learning technique. 2) **AFAnet (**Audio Features based Artificial Neural Network) is a supervised learning architecture comprising of an artificial neural network that finally classifies trailers into genres. The complete methodology can be summarized as per Fig. 1. The detail of approaches is described in below sections:

A. ASM (Audio Segmentation Model)

This study is based upon the idea that the audio of the trailer consists of a collection of some set of sounds as discussed in Section I. Considering an action genre trailer, it mostly contains high amplitude shots whereas a romance genre trailer includes low amplitude voices. So, using these distinctions in the type of audio, it is possible to generate a strong mapping between these types of sounds and the final movie genre. Note that a trailer consists of more than one type of these characteristics; hence it is better to break the audio into some non-overlapping disjoint chunks of small interval (5 seconds). To proceed with this approach, it is required to create a collection of a different set of sounds to categorize the audio chunk. Instead of using some other dataset to get collection as in [17], a set of movie trailers is chosen and segmented into categories as described in Section III.A.c).

a) Features of Audio

To analyze the audio data, various representations of audios can be formed. To get a better view of the characteristics of audio data, a large set of features are considered. A total of 34 features namely zero-crossing rate, energy, the entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, MFCCS (13 features), chroma vector (12 features), chroma deviation are used to define audio chunk as in [18]. Each of these 34 features consists of waveform representation. The waveform can be represented in the form of the actual y coordinates or the change of values (delta representation). Both representations are used to sum a total of 68 features. For extracting features from every 5-sec chunk, a window of 50 msec with an overlap of 25 msec is taken. For a chunk of 5 sec, the extracted feature can be represented as an array of 200 values. The data seems to be vast as for every chunk, there are 200×68 features.

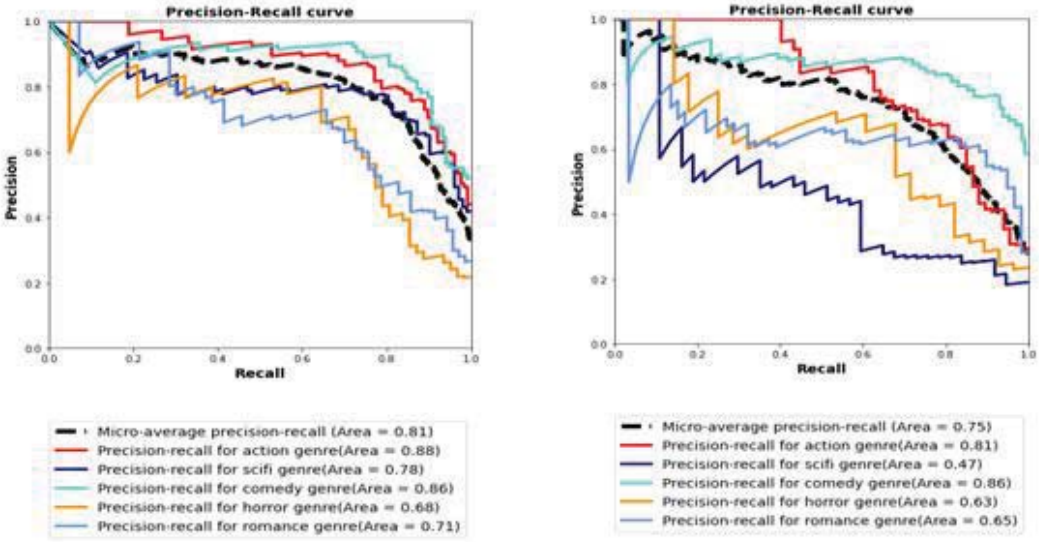


Fig. 2. Precision Recall Curve on 1) Our Dataset 2) LMTD-9 Dataset

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:41:37 UTC from IEEE Xplore. Restrictions apply.

For every chunk i, 68 features are mathematically represented in Eq. (1).

$$features = \{F_i^1\ ,\ F_i^2\ ,\ F_i^3\ , \ldots \ldots\ F_i^{68}\} \quad (1)$$

To reduce the dimensionality, for every feature in a chunk mean of a vector of size 200 is taken. Hence, a single chunk is represented in form of 68 values, by taking a mean of 200 values described above.

b) Preprocessing

Audio is extracted from the movie trailers and split into disjoint small segments of 5 seconds. The first and the last segments of every trailer is ignored as it mostly consists of silent audio in most of the trailers, thus, they do not depict any useful information. Upon extensive manual experimentation, it is found that the 5 seconds chunks are useful and may be classified into a single category. Hence, chunks of the duration of 5 seconds are considered. Every chunk is represented using the features discussed in Section III.A.a). Hence, the final dataset available for k-means clustering has a shape of (8000×68). As 500 trailers are taken for the analysis and on average each trailer has a length of 1.25 minutes so, around 8000 chunks are formed each having 68 features. These features are the same as mentioned in previous sections.

c) Audio Segmentation

The audio chunks generated are unlabeled, and consist of various types of sounds like screaming voices, laughing voices, etc. To make these chunks labeled, the segmentation technique of k-means clustering is used to cluster the data into some clusters based on the features and to identify centroid such that each chunk is as near as possible from one of the centroids, extracted as described in Section III.A.a). To choose the number of segments, extensive experiments are performed as discussed in Section IV. It is concluded that for 10 clusters, the data seems to be segmented well. Hence, finally, audio data is clustered into 10 clusters. The 10 centroids obtained after clustering can be defined as in Eq. (2) and Eq. (3).

$$j^{th}\ centroid\ is\ expressed\ as,\ \{G_j^1\ , G_j^2\ , G_j^3\ , \ldots . . G_j^{68}\} \quad (2)$$

Where j^{th} centroid and i^{th}value,G_j^i is defined in Eq. (3),

$$G_j^i = mean\big(F_{S_1}^i, F_{S_2}^i \ldots .\quad\big) \quad (3)$$

$$for\ every\ (chunk\ S_c) \in cluster$$

B. AFA-net (Audio Features-based Artificial Neural Network)

This study focuses to incorporate the type of sounds that trailer audio contains, and using these properties to make a mapping between movie genre and the type of sound trailer contain. In Section III.A.c) various sound segments have been created, now using those sounds as a separate feature, the final genre of the movie trailer can be predicted. There are a total of 10 centroids of the audio as described previously, the inverse of the distance of a particular chunk from all 10 centroids is taken as an independent feature. For chunk i, the distance from $j\hat{}th$ centroid can be calculated as in Eq. (4).

$$D_i^j = \sqrt[2]{\left(F_i^1 - G_j^1\right)^2 + \left(F_i^2 - G_j^2\right)^2 + \ldots\ldots\ldots + \left(F_i^{68} - G_j^{68}\right)^2}\ for\ 1 \le j \le 10 \quad (4)$$

Each trailer can be represented as described in Fig. 1. Every trailer is represented as a set of 5-sec chunks. For each chunk, apart from 68 features, 10 more features can be added as discussed. Hence an audio chunk in this model consists of 68 (actual audio features as in Section III.A.a)) + 10 (distance-based features) = 78 features. The distance-based features are defined in Eq. (5).

$$F_i^{69} = \left(D_i^1\right)^{-1},\ F_i^{70} = \left(D_i^2\right)^{-1} \quad F_i^{78} = \left(D_i^{10}\right)^{-1} \quad (5)$$

$$New\ set\ of\ features\ for\ i\ th\ chunk = \{\ F_i^1\ , F_i^2\ , F_i^3\ , \ldots \ldots\ F_i^{78}\ \}$$

Due to the variable length of a movie trailer, the number of chunks for each trailer varies. To make the shape of all data-points uniform, the arithmetic average is taken for every feature over all chunks as in Eq. (6) Each movie trailer can be expressed as a 78 sized vector of shape 1×78 as in Eq. (6).

$$New\ set\ of\ features\ for\ j\ th\ trailer\ , Ffinal^j$$
$$Ffinal^j = \{mean\ (\ F_i^1 \ldots), mean\ (\ F_i^2 \ldots)\ , \ldots\ mean\ (\ F_i^{78} \ldots)\ \}$$
$$where\ i\ \in chunks\ of\ trailer\ j \quad (6)$$

Next, to predict the multi-label genre of the trailers, an artificial neural network is proposed. To train a neural network, the data is normalized to lie in the range 0 to 1. The architecture as described in Table I is built to predict the genres. Later, the data with input shape (78,) is fed into the model. The architecture starts with first a dense layer with 256 neurons. Thus, the shape of output is (256,) after the first layer. Next data is fed into one more dense layer along with a drop out of 0.4 to avoid overfitting. This layer comprises 256 neurons gives an output of shape (256,) which is fed into two stacked dense layers. They consist of 64 and 32 neurons in order. Again, one more dropout layer is applied. The output shape after this layer is (32,). The last layer of the model is again a dense layer. It consists of 5 neurons to get a probability of the 5 genres considered in this work.

IV. EXPERIMENTS

All the experiments are performed on a GPU workstation with a 128 GB DDR4 Ram, a 64-bit machine with NVIDIA Titan RTX GPU (24 GB). For the dataset, a total of 2400 trailers are downloaded from IMDB, out of which 400 trailers are used in ASM architecture for clusters formation and 2000 trailers are used for training/validating purposes in the AFA-net as described in Section III.B. These 2000 trailers (different from ones used in ASM) are used to train and validate the AFAnet model. The trailers are split into 85:15 ratios for training and testing. The model is trained with a learning rate of 0.001 to make the model learn the probability of genres based on features as described in Section III.B. The saturation was achieved at 200 epochs. Model is validated on the above dataset and cross-tested on the standard LMTD-9 [10], [19] dataset for better evaluation. Evaluation scores are measured based on AU (PRC) which stands for (Area under the precision-recall curve).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:41:37 UTC from IEEE Xplore. Restrictions apply.

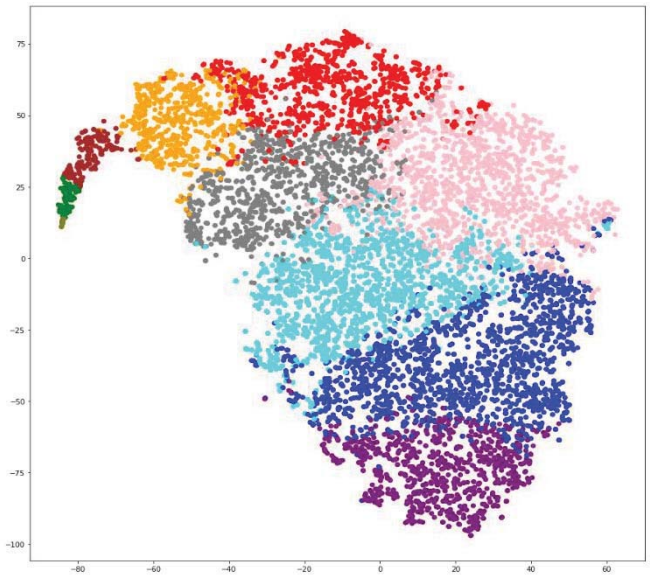


Fig. 3. t-SNE for 10 clusters

A. Variation According to Number of Clusters

Extensive experiments are performed to decide upon the number of clusters. The architecture defined in Section III.B. is trained on the different number of clusters and AU (PRC) observed is depicted in Table II. Upon testing with the different number of clusters, the best results were depicted at k= 10. Apart from this, analysis using t-SNE [20] is also made. t-SNE for 10 clusters is segmented disjointly as depicted in Fig. 2. Also, for 5 clusters AU (PRC) value is the least, and for other numbers of clusters also, AU (PRC) value lag as compared to the results observed with 10 clusters.

TABLE II. AU(PRC) VALUES WITH DIFFERENT NUMBER OF CLUSTERS

Clusters # \ Dataset	Our Dataset	LMTD-9
k = 05	0.69	0.67
k = 08	0.77	0.72
k = 10	**0.81**	**0.75**
k = 12	0.75	0.73
k = 15	0.78	0.73
k = 20	0.74	0.69

B. Classification results

Extensive experiments are performed to test the model on our dataset and to validate the approach on the standard LMTD-9 dataset. It is depicted in Table III that our model performed well in classifying genres. The performance of the model on our trailers downloaded from IMDB and on the LMTD-9 dataset is very well. Cross-validation on the LMTD-9 dataset shows that our proposed approach is robust, as it is performing well on unseen data too. Among all five genres, the action genre is performing best, as compared to other genres on both the dataset. The model performance is also evaluated using AU (PRC) as it is a stricter measure to evaluate multi-label classification problems. This approach gets an AU (PRC) of 81% on our trailers, and 75% on the LMTD-9 dataset as shown in Fig. 3.

TABLE I. CLASSIFICATION RESULTS OF AFAnet

Genres \ Dataset	Score	Our Dataset	LMTD-9
Action	P	0.83	0.84
	R	0.84	0.79
	F1	0.83	0.81
Science Fiction	P	0.81	0.71
	R	0.82	0.68
	F1	0.82	0.69
Comedy	P	0.85	0.78
	R	0.85	0.77
	F1	0.85	0.78
Horror	P	0.77	0.70
	R	0.80	0.81
	F1	0.78	0.74
Romance	P	0.81	0.75
	R	0.74	0.82
	F1	0.77	0.75

V. CONCLUSION

This work focuses on the audio extracted from the movie trailers to classify movie genres; this approach eased the genre classification process by considering audio from movie trailers instead of watching a whole movie of 2-3 hours. Audio content depicts a lot of information relating to the genre of the movie. Thus, we propose a novel and robust framework for movie genre classification. For validation purposes, we cross-tested our model on the standard LMTD-9 validation set as well. This work can also be used for classifying movies from different cinemas including Bollywood, Hollywood, South Korea, Tollywood, Cinema of Japan, etc. Also, this architecture can be extended to produce a multimodal framework in later stages of research. Apart from this, other features including video frames, subtitles can also be extracted from movie trailers. Plot, description, and movie posters can also be considered while classifying the movie genres. We also plan further to include multi-model approaches and more movie genres in the study.

REFERENCES

[1] Z. Rasheed and M. Shah, "Movie genre classification by exploiting audio-visual features of previews," in Proceedings - International Conference on Pattern Recognition, 2002, vol. 16, no. 2.

[2] M. Sugano, R. Isaksson, Y. Nakajima, and H. Yanagihara, "Shot genre classification using compressed audio-visual features," in IEEE International Conference on Image Processing, 2003, vol. 2.

[3] X. Yuan, W. Lai, T. Mei, X. S. Hua, X. Q. Wu, and S. Li, "Automatic video genre categorization using hierarchical SVM," 2006.

[4] A. M. Ertugrul and P. Karagoz, "Movie Genre Classification from Plot Summaries Using Bidirectional LSTM," in Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018, 2018, vol. 2018-January.

[5] F. Álvarez, F. Sánchez, G. Hernández-Peñaloza, D. Jiménez, J. M. Menéndez, and G. Cisneros, "On the influence of low-level visual features in film classification," PLoS One, vol. 14, no. 2, 2019.

[6] P. Doshi and W. Zadrozny, "Movie genre detection using topological data analysis," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11171 LNAI.

[7] M. Ivasic-Kos, M. Pobar, and L. Mikec, "Movie posters classification into genres based on low-level features," 2014.

[8] S. Jin, T. M. Bae, J. Choo, and Y. Ro, "Video genre classification using multimodal features," 2004, pp. 307–318.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:41:37 UTC from IEEE Xplore. Restrictions apply.

[9] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, “Movie genre classification via scene categorization,” 2010.

[10] G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz, “Movie genre classification with Convolutional Neural Networks,” in Proceedings of the International Joint Conference on Neural Networks, 2016, vol. 2016-October.

[11] J. Wehrmann and R. C. Barros, “Movie genre classification: A multi-label approach based on convolutions through time,” Appl. Soft Comput. J., vol. 61, 2017.

[12] W. T. Chu and H. J. Guo, “Movie genre classification based on poster images with deep neural networks,” 2017.

[13] A. Yadav and D. K. Vishwakarma, “A unified framework of deep networks for genre classification using movie trailer,” Appl. Soft Comput. J., vol. 96, 2020.

[14] Y. F. Huang and S. H. Wang, “Movie genre classification using SVM with audio and video features,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, vol. 7669 LNCS.

[15] E. Fish, A. Gilbert, and J. Weinbren, “Rethinking movie genre classification with fine-grained semantic clustering,” arXiv Prepr. arXiv2012.02639, 2020.

[16] R. B. Mangolin et al., “A multimodal approach for multi-label movie genre classification,” Multimed. Tools Appl., 2020.

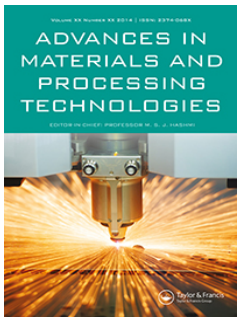
[17] J. F. Gemmeke et al., “Audio Set: An ontology and human-labeled dataset for audio events,” 2017.

[18] T. Giannakopoulos, “PyAudioAnalysis: An open-source python library for audio signal analysis,” PLoS One, vol. 10, no. 12, 2015.

[19] J. Wehrmann, R. C. Barros, G. S. Simoes, T. S. Paula, and D. D. Ruiz, “(Deep) Learning from Frames,” 2017.

[20] L. Van Der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, 2008.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:41:37 UTC from IEEE Xplore. Restrictions apply.



Taylor & Francis
Taylor & Francis Group

Advances in Materials and Processing Technologies

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tmpt20

Advances in applications of Non-Destructive Testing (NDT): A review

Mridul Gupta, Muhsin Ahmad Khan, Ravi Butola & Ranganath M. Singari

To cite this article: Mridul Gupta, Muhsin Ahmad Khan, Ravi Butola & Ranganath M. Singari (2021): Advances in applications of Non-Destructive Testing (NDT): A review, Advances in Materials and Processing Technologies, DOI: 10.1080/2374068X.2021.1909332

To link to this article: https://doi.org/10.1080/2374068X.2021.1909332

Published online: 04 Apr 2021.

Submit your article to this journal

Article views: 162

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tmpt20

ADVANCES IN MATERIALS AND PROCESSING TECHNOLOGIES
https://doi.org/10.1080/2374068X.2021.1909332

Taylor & Francis
Taylor & Francis Group

Check for updates

Advances in applications of Non-Destructive Testing (NDT): A review

Mridul Gupta, Muhsin Ahmad Khan, Ravi Butola and Ranganath M. Singari

Department of Mechanical Engineering, Delhi Technological University, Delhi, India

ABSTRACT

Manufacturing processes such as casting, rolling, forging, extrusion, material removal processes, etc. are some of the common techniques used today in manufacturing industries. However, these processes must be carried out with utmost precision and precaution. Even the slightest negligence can give rise to anomalies which can be classified as defects or discontinuities depending upon their repairability. Non-Destructive Testing or NDT is amode of testing which has been adopted by industries for decades to test mass manufactured products for anomalies.

This paper discusses various NDT methods, such as Visual Testing, Magnetic Particle Inspection, Penetrant Testing, Ultrasonic Testing, Radiographic Testing, Acoustic Emission and Eddy Current Testing. The paper also discusses how technical advancements have broadened the scope of NDT even in industries that may not be manufacturing oriented, indicating that NDT is not limited to detection of anomalies. Even though these developments have made NDT more favourable, this paper highlights the fact that these techniques are mostly manual and heavily dependent upon inspectors' knowledge and experience, leaving room for errors. This paper also discusses the key challenges and future scope in the field of NDT.

ARTICLE HISTORY
Accepted 24 March 2021

KEYWORDS
Non-destructive testing (NDT); manufacturing defects; eddy current testing; ultrasonic testing; acoustic emission

1. Introduction

One of the manufacturing industry's key objectives is to ensure product quality, that is, to make sure the product can perform its intended function for a long time. For a product to achieve this criterion, it is of utmost importance that the product is free of any defect and discontinuity. A defect is an imperfection in a product that can restrict its life by impeding its performance, eventually causing it to fail, even during operation under ordinary conditions. A discontinuity is similar to a defect, the only difference being that a discontinuity can be rectified whereas a defect cannot. These anomalies can arise during manufacturing stages or even during the operation of the product. These anomalies must be detected, studied and if possible, be rectified. Non-Destructive Testing or NDTconsists of a range of techniques that enable the detection and evaluation of said anomalies so that the product's condition is not exacerbated and can be rectified if possible. One of the most significant advantages of NDT is that they can be performed at all product life stages, even during their operation. This provides an accurate idea about the location and

CONTACT Mridul Gupta mridulgupta_2k18me131@dtu.ac.in
© 2021 Informa UK Limited, trading as Taylor & Francis Group

the properties (like dimensions, rate of propagation, etc.) and the severity of the defect or discontinuity.

The principles employed in various techniques of NDT are also employed in different fields. It is an implication that NDT is not only limited to manufacturing. For example, principles of NDT techniques are prevalent in medical diagnostics. X-Radiography used for detecting tumours, fractures, etc., and Remote Visual Testing used in procedures like endoscopy are some examples of NDT in the medical field. However today, major applications of NDT in manufacturing lie in automobiles, oil and gas, aerospace and so on. There are other fields that use the principles of NDT like geological exploration, architecture, etc. This paper aims at co-relating the principles of various NDT techniques with advances in applications in fields which may or may not be manufacturing-oriented.

2. Literature review

According to Cawley [1] eddy current, ultrasonic, magnetic particle, radiography and penetrant testing are the 'big five' techniques that dominate the NDT market. Thus, research is being done continuously to reduce the preparation required, increase the

Table 1. Capabilities and limitations of some of the key NDT Techniques.

NDT Technique	Capabilities	Limitations
Visual Testing	Suitable for surface inspection and monitoring manufacturing stages of a product.	1. Only restricted to surface flaws. 2. Highly dependent upon the inspector's experience.
Penetrant Testing	Suitable for surface inspection of mass-manufactured products.	1. limited to surface anomalies. 2. Too many variables affecting the test. 3. Messy process.
Magnetic Particle Inspection	Capable of detecting anomalies on or close to the surface in ferromagnetic materials.	1. Limited to anomalies in the vicinity of the surface. 2. Can be only used for testing ferromagnetic materials. 3. Testing of complex shapes may not be possible. 4. Removal of any coatings is required.
Radiographic Inspection	Capable of detecting surface and subsurface anomalies.	1. Safety hazards and waste disposal issues. 2. Time-consuming. 3. Expensive. 4. Dependent upon the orientation of anomalies.
Ultrasonic Testing	Capable of detecting surface and subsurface anomalies	1. Anomalies smaller than grain structure can go undetected. 2. Largely manual, therefore highly dependent upon the inspector's skill and experience. 3. Misinterpretations of signals can occur.
Eddy Current Testing	Capable of detecting flaws on or close to the surface in materials that are electrically conductive with or without coating (e.g. paint).	1. Limited to the testing of electrically conductive materials. 2. Assessment of subsurface anomalies in ferromagnetic materials is limited by this technique. 3. Theory of eddy currents, in general, is complicated.
Acoustic Emission	Capable of detecting surface and subsurface flaws with information about the propagation of the anomaly.	1. Structure under testing will attenuate stress waves. 2. Extrinsic sounds can cause misinterpretation.

speed of inspection and to inspect without shutting down the operation. How the process of using previous cases of data-interpretation to interpret new data can be helpful in solving the issue of NDT data interpretation is shown by Jarmulak et al. [2]. The article shows how Case-Based Reasoning (CBR) can be a successful alternative to expert systems and statistical classifiers. Zhijun [3] states that common defects and damages in composites can be effectively checked by ultrasonic inspection as compared to other methods of NDT. The author also introduces the development of studying and the applications of the ultrasonic method. According to G.S. Park and others [4], in a Magnetic Flux Leakage (MFL) type NDT system, the defect signals depend on the modification of the magnetic flux leakage in the area of defect. Thus, the authors describe a method to design a magnetic system which maximises the MFL in an NDT system.

Tschelisnig [5] presents the principles of hyperbolic triangulation and the basis of propagation of waves. The author also explains the AE inspection systems by taking an example from TUV Vienna's 32 channel software and equipment. The results gained after the tests performed on the TUV Vienna are discussed in integrity analysis and leakage tests. Crostack et al. [6] state that the ability to measure internal and induced stresses or to differentiate between structures of materials can be provided by Non-destructive testing. This, in turn, provides the data for the calculation of potential lifetime and reliability. They also state that the materials may be better understood by observing fast-moving processes like crack propagation when they are under impact load. Carlomagnoet et al. [7] examine the application of infrared thermography in the field of architectural restoration by taking three samples made of support of tuff, marble or brick and covering them with a layer of plaster. This layer of plaster is given inclusions to simulate cracks and detachments in frescoes. To detect the artificially induced flaws, various techniques, such as lateral heating thermography, pulse thermography, pulse phase thermography and modulated thermography are used. Bollas et al. [8] discuss the monitoring of defects in the wheelsets of locomotives using the Acoustic Emission (AE) monitoring method of NDT. The authors present the results of AE measurements performed on the wheelsets of the locomotives with and without defects, with varying speed and directions. The conclusions drawn show that the AE monitoring method can be an efficient method for monitoring the wheelsets of the trains.

Opportunities for the inspection of rock bolt in applications such as reinforcement of coal mine roofs are limited to destructive techniques like pull-out testonly. Thus, Beard et al. [9] tried to develop an NDT technique for monitoring the condition of rock bolts. Drinkwater et al. [10] reviewed the most relevant works on the use of ultrasonic arrays for non-destructive evaluation applications, in addition to which the author also discusses its future direction. The author also reviews the most relevant papers on the application of ultrasonic arrays in medical and sonar fields. On the other hand, Lim et al. [11] explain how one NDT technique is not sufficient for evaluation. Every technique has certain limitations which restrict their efficacy under different situations. The authors do so by differentiating the results between UT and ultra-response method for evaluation of defects in a concrete structure. A disadvantage of NDT is that most evaluations are heavily dependent upon the inspector's knowledge and experience. This is most relevant to magnetic particle inspection and penetrant testing, which also happen to be the most used inspection techniques. Luk et al. [12] research on some human factors which may

affect the accuracy of an evaluation and provide some suggestions in order to tackle these issues.

Tasca et al. [13] explain how the evaluation of complex parts can be done using NDT. Although x-ray provides accurate data about a defect and is mostly p`referred over other techniques, sometimes UT can provide better results. The authors also provide better insight as to how UT can provide more information other than just the nature of the defect or discontinuity but also how it occurred. Cau et al. [14] worked on the detection of defects in pipes using UT. They evaluated the problem with the help of appropriate Artificial Neural Network models. Numerical Techniques like FEM have been used to evaluate problems in pipes by studying the echoes produced by ultrasonic waves. Eddy Current testing has many variables that affect evaluations made by this method. Martin et al. [15] provide a summary on these variables as well as recent advancements in this mode of NDT. The authors also explain how development in electronics and data evaluation systems have made Eddy current testing more effective. One of the advantages of NDT is that it can be done simultaneously while operating a machine. For example, Pohl et al. [16] illustrated that Eddy current testing could be done on rail tracks simultaneously to trains travelling on them. The authors also emphasised other applications of NDT, such as the use of ultrasonic testing on railroad gauges and wheels.

Singh et al. [17] illustrate the application of magnetic particle separation techniques used in the Indian mining industry. The authors provide an insight as to how this process has made inspection simple and cost-effective as well as discuss the disadvantages of this technique. Advancements in NDT have resulted in better evaluation of defects and discontinuities. Bracun et al. [18] proposed a laser-based technique for 3-D modelling of defects which use advanced algorithms to produce 3-D images with the help of electronic visual aids, such as cameras. Most of the principles employed in NDT techniques are employed in other fields as well. In fact, some of these techniques can be incorporated into improvising existing technology. One such example has been illustrated by Benenson et al. [19] where they emphasise the issues with self-navigating cars. A combination of techniques such as ultrasonic waves and thermography can assess surrounding environments more accurately, thus increasing reliability. An advantage of NDT is that it can be conducted at any stage of production. Lattice defects which are microscopic in nature, arise during material manufacture and affect the overall physical integrity of any component. Dobmann et al. [20] show that NDT, if incorporated at every stage of material production, helps in making sure that these properties are uniform throughout.

One of the major advances in NDT is that these methods are no longer limited to detection but also play a major role in preservation. X-radiography is normally preferred for this application because of its ability to scatter, which can be used to study the deterioration of concrete. Kamal and Boulfiza [21] studied this by observing penetration in alkalis using X-radiography technique. Pulsed eddy current techniques are capable of providing detailed information about a specimen. However, a major drawback lies in its sensitivity to external factors which may cause deviations in results. Tian et al. [22] provide a way of reducing the problem of lift-off in ECT by using normalisation, also making it suitable for measuring the thickness of metal with various coatings. Shrivastava et al. [23] discuss the role NDT can play in advances in medical research. They also provide information regarding the current applications of NDT and discuss the

advantages of AE techniques in the biomedical field. NDT techniques today are being majorly used for evaluating mega structures like bridges. Sahama et al. [24] performed NDT on different parts of a bridge which include hangar sockets, welded joints and various concrete parts by using techniques such as PT, RT, MPT, etc.

One of the disadvantages of NDT techniques that depend upon magnetic behaviour is that they are not effective for metals with low magnetic permeability. One such material is Aluminium and its alloys. Hu et al. [25] used magnetic sensors and proposed a way for detecting anomalies in thin-plate aluminium alloys. Development in computation and electronics have provided NDT engineers with a brilliant and tricky opportunity for the development of this field. Bieder et al. [26] provide a solution to this problem by proposing quantification or classification of NDT results in terms of common parameters like type, location and size of the anomaly. NDT techniques enable measurement of those properties of materials which cannot be measured by other testing methods. For example, Hasar [27] measured microwave reflection from cement and mortar specimens which help in determining the electrical properties of the specimen. This is useful information which is used in the preservation of structures. Product quality over the years has improved largely due to automation of industries like CNC machines, die casting, rolling, etc. However, for further improvement and a greater level of satisfaction, Wolter et al. [28] suggest incorporation of NDT techniques during manufacturing stages. This can enable detection of any irregularities in machine settings or irregularities in the material.

Composites are a relatively new introduction in the field of material technology. Like metals, composites are also prone to defects like voids, cracks, etc. Banarjee et al. [29] discuss inspection of jute–polypropylene composite using a new technique called frequency-modulated thermal wave imaging (FMTWI). He et al. [30] discusses applications of pulsed eddy current testing, a relatively new form of eddy current testing, in the detection of subsurface defects using a rectangular PEC sensor. Automation has provided high rates of production. Bergmann et al. [31] discuss the requirement of rapid testing techniques that can keep up with the rate of production and provide the same sense of satisfaction ensuring customers quality products. They also discuss the applications and limitations of X-Ray computer tomography. FRP reinforced concrete is becoming an increasingly popular material for components, such as beams, slabs, etc. One of the problems with this material is that a proper surface examination cannot be done until the FRP layer is removed completely. A beam of this material may not even show signs of any anomalies when tested. Büyüköztürk et al. [32] provide a laser-acoustic NDT method for evaluating bonding in FRP reinforced concrete products.

Concrete has a unique self-healing ability. When concrete is exposed to the environment, it reacts with water to form grains. Tittleboom et al. [33] provide an NDT method that utilises AE for analysing self-healing capability of concrete. Evaluation of water pipelines is not only limited to studying the extent of the damage but also the type of flow. Open-type pipelines are usually placed under the surface; therefore, VT is not an option. Suzuki et al. [34] conducted experiments which showed that AE could be used to evaluate fluid flow conditions in these pipelines. The microstructure of metals is crucial as it determines many properties of a metal. It is usually determined by destructive techniques, such as hardness tests and other metallurgical techniques. Gür [35] discusses Magnetic Barkhausen Noise method, which is a form of NDT technique that has a high potential for classifying microstructures of steel. Ribeiro et al. [36] provide an

NDT method for the evaluation of non-ferromagnetic materials by providing an algorithm to model variation in ECT voltages. Usually, this method would involve the use of a probe for detecting readings. However, these readings are complex.

Mazal et al. [37] studied AE in cyclically loaded AlMg alloys and studied the changes in these AE through X-Ray diffractions. Modern NDT methods have been enhanced by advancements in technology. VT is the oldest mode of NDT. However, it was not considered to be as reliable as other modes of testing. Zhu et al. [38] discuss and review developments in optical technology used in NDT. Applications of NDT are prevalent in aerospace engineering. Sato et al. [39] discussed how planar defects in graphite are detected using UT technique which is used for quality assurance in rockets. Residual stresses in materials post-manufacturing or post-operation of a product can hinder its performance. Xu et al. [40] discuss how ultrasonic critical refracted longitudinal waves can be used to detect and evaluate residual stresses non-destructively and confirmed the accuracy of UT LCR wave method experimentally.

Mori et al. [41] present a new method for detection of defects in a concrete structure. This method is based on the dynamic response generated by concrete structures having flaws and are subjected to impact loading. Experimental and numerical results presented in the paper suggest that this method is effective as compared to the similar methods like impact-echo method. Pohl et al. [42] discussed that the ultrasonic inspection of the disc and the wheel rim of trains should be done without dismantling the wheel. While the inspection of track surface of railroad with a train speed of approximately 70 km/h should be carried out using eddy current method of NDT. N.P. Avdelid et al. [43] presents the applications of transient thermal NDT in assessing the aircraft composites. Pulsed thermography is used to monitor all the features in real-time. But, in the cases involving repair of composites, thermal modelling is also used in addition to pulsed-phase thermography to provide supplementary results. Maierhofer et al. [44] show that the impulse thermography is a good method for the detection of voids in concrete covers up to 10 cm and more. A computer program is also developed for investigation of the influence of various environmental conditions, parameters of the material and geometry on the thermal behaviour.

Mao et al. [45] propose a technique for determination of the location and the extent of defects and cracks which occur in oil pipelines using the time-frequency analysis of Hilbert-Huang. A signal processing technique developed by Hilbert–Huang which is based on direct extraction of the energy associated with the intrinsic time scales in the signal is used to process the signals from the oil pipelines with and without defects. Schajer et al. [46] present an NDT method for estimation of grain direction, dry density and moisture density of orthotropic materials like wood. This method is developed using the measurements of complex attenuation of microwaves which are transmitted via the material. These attenuations possess tensor characteristics like that of stress, strain, etc. Davis et al. [47] explain the usage of NDT methods in examining the tunnel lining grouting programs' efficiency with a focus on the impulse radar and impulse response methods. Kohl et al. [48] introduce the various scanning systems which are developed for on-site applications of Non-Destructive methods by presenting the results obtained from calculations which were performed at the laboratory at BAM as well as on-site at various bridges.

Shah et al. [49] present the results obtained from non-destructive ultrasonic testing of concrete specimens which are in the form of a cube. This testing is done using different frequency transducers. Orbán et al. [50] give an overview of the available methods of NDT, minor destructive and monitoring along with their applicability and efficiency in the testing of masonry arch railway bridges. Kornmeier et al. [51] discuss the appropriateness of various methods of non-destructive analysis to characterise Ci/SiC components like satellite nozzles. For characterisation of 3-D fibre orientation and integrity, matrix homogeneity, dimensions and distribution of micropores and synchrotron radiation using tomography at a fibre scale was used. A comparative study is given by Bayraktar et al. [52] for the developments in non-destructive controls of composite materials and manufacturing engineering applications. A review about the advantages, performance and usage at multiple scales of XR-CT (X-Rays Computed Tomography) medical scanner is also given.

Monitoring of seven-wire steel strands 15.7 mm in diameter is done by Chaki et al. [53] by using the guided ultrasonic wave procedure. For the optimisation of the measurement, the configuration was done by accounting the geometrical and mechanical properties of the prestressed strands followed by choice of mode of guided ultrasonic procedure at a suitable frequency. Capability and performance of shearography, ultrasonic, X-ray CT and thermography methods for the wind turbine blades inspection and delamination have been analysed by Amenabar et al. [54]. Recent advances in NDT with a focus on ultrasonic and electromagnetic techniques are reviewed along with ideas about the risk-based inspection by Abeele et al. [55]. Fatigue failure in offshore structures is predicted by using fracture mechanics-based damage models. A survey of NDT techniques for testing of building structures along with their application is done HOŁA et al. [56]. The major focus is given on the acoustic emissions method of NDT due to a lot of development in recent years and the ability to acquire important information on the tested structure using software and algorithms.

3. Methods of non-destructive testing

3.1. Visual Testing (VT)

VT is the oldest mode of NDT. This method involves a surface examination by visual aids, mainly the eye. The human eye is the most important visual aid that is required for conducting VT inspection. Light is an important requirement for conducting VT. The amount of light entering the eye determines the quality of the image. Several factors such as surface condition, brightness, temperature, size and shape of the object can affect the amount of light that is being reflected into the eye. Human factors such as physiological and environmental factors also affect the results of VT inspection.

There are two types of VT methods, Direct VT (or DVT) and Remote VT (or RVT). Direct VT is carried out when the eye can be placed within 25 inches and at a minimum angle of 30 degrees to the surface of the specimen that is to be evaluated. Remote VT (as shown in Figure 1.) is done for sites which do not have proper accessibility and no illumination. In this mode, light is supplied by an external source via cable along with a camera which transmits images or a live feed on to a monitor.

3.2. Penetrant Testing

PT is an application of capillary action. When a special liquid, called a penetrant, is applied on the surface of a component, usually by an aerosol, it fills any discontinuity that may be present by virtue of capillary action.

There are mainly two types of penetrants, fluorescent and visible dye penetrants. However, due to the better contrast ratio of fluorescent dyes, they are typically used for PT. The component to be inspected is immersed in the penetrant (as shown in Figure 2 (a)) for some time and is then washed, dried and then observed under black light (fluorescent dyes). The set-up for PT can be manual and/or even automatic. However, for simplicity in field inspection, dye penetrants (which do not require complicated setups) are used. PT done with dye requires an additional liquid called developer (as shown in Figure 2(b)). The developer draws out the dye which may be entrapped in a discontinuity, making it visible. Factors such as wettability, capillary action, dry concentrate and viscosity determine the choice of the penetrant to be used. Before any PT is done, there are certain prerequisites that should be noted: 1) temperature, if the surface is too cold a longer period of immersion is required and if it is too high the penetrant may evaporate, 2) environmental considerations which include flammability, fumes, etc., 3) lighting and 4) surface condition such as coatings or rust should be addressed as it can affect the action of the penetrant.

3.3. Magnetic Particle Inspection

MPI (Figure 3) is a test based upon the behaviour of ferromagnetic materials when exposed to a magnetic field. Due to their high magnetic permeability, the domains align themselves in the direction of the applied magnetic force creating continuous lines. The underlying principle of MPI is studying the continuity of these lines of force; if the material is uniform, then the lines are continuous. Whereas the lines distort in the case of a discontinuity because they tend to follow a path where they experience the least reluctance (the ability of a domain to resist alignment in the direction of applied magnetic force). In order to do so, these lines bend around the discontinuity creating a 'distorted' field. Sometimes if discontinuities are large enough, these lines tend to exit

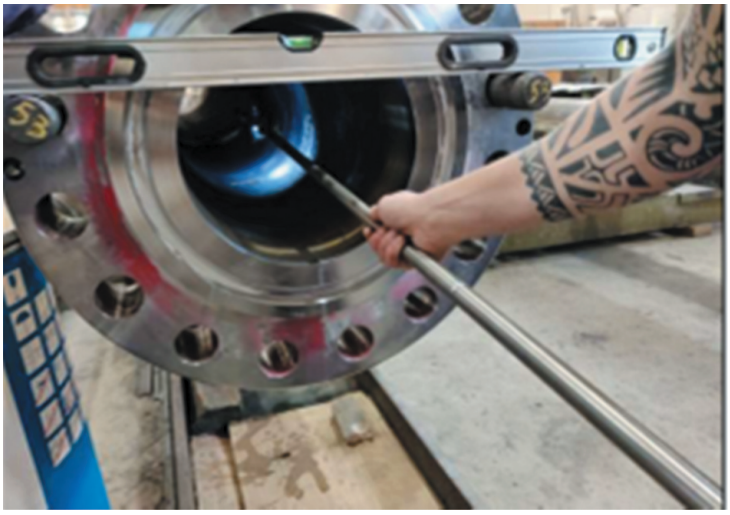


Figure 1. RVT of a pipe using borescopes with lighting source.

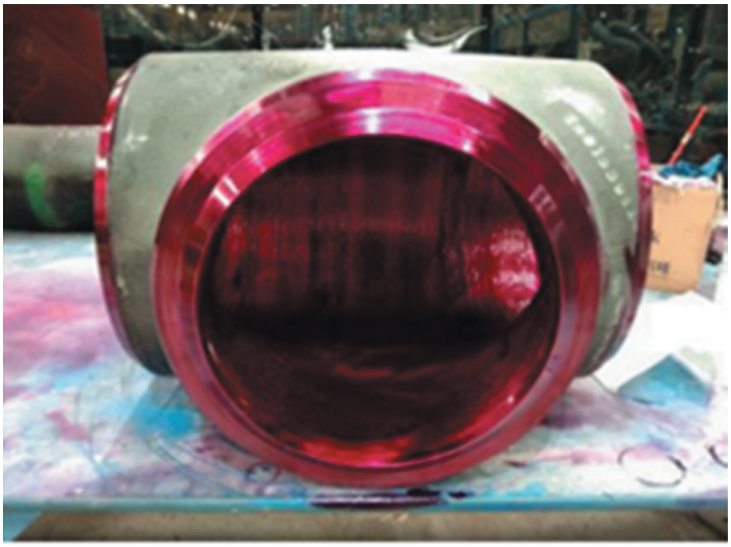


Figure 2. (a). Application of penetrant in PT.

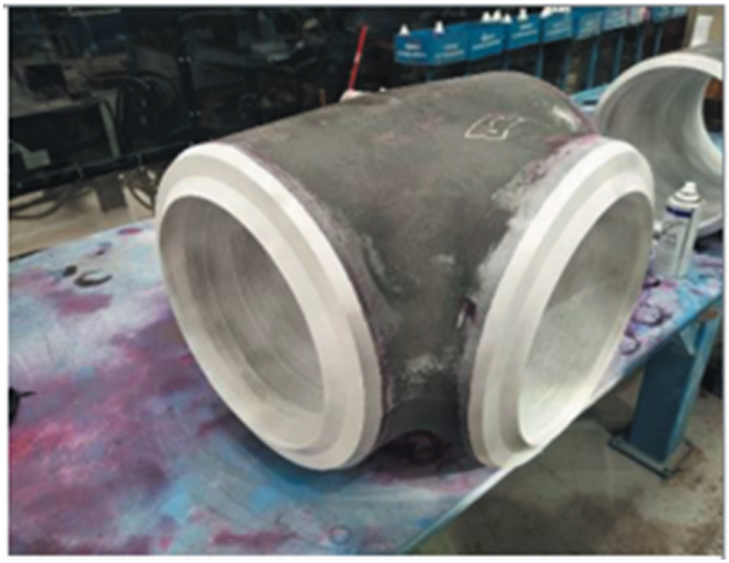


Figure 2. (b). Application of developer after removal of penetrant.

the surface and form through the air over the discontinuity. This is known as leakage of magnetic fields. This leakage can be made visible by applying a ferromagnetic powder. The powder gets accumulated, and a visible indication is observed.

Usually, iron filings are used, but for improved visibility, ferric oxide powder is used. The detection is, however, affected by physical factors, such as thickness (thicker the material stronger the field required), material (different components affect the permeability of the material) and internal stresses (effect on magnetic conductivity) of the component. Some other testing factors, which include distancing between the magnetic poles and speed of the detector, also affect the results of MPI. Test results of MPI are instantaneous and easy to interpret.

Figure 3. Magnetic particle inspection.

3.4. Radiographic Testing

RT is based upon the characteristics of radiation. Radiation is nothing but waves or particles emitted from a source passing through a medium. Radiography is the practice of capturing images by employing the characteristics of radiation. One of the key characteristics of radiation is that it can be absorbed by materials and the amount absorbed varies with the thickness of the material. When radiation is made to pass through a component, it is attenuated by the material. In regions of discontinuities, these rays experience lesser attenuation and travel faster in comparison to rays in uniform regions. The area of the film in the vicinity of the discontinuity experiences more spotlight, and therefore a unique image is obtained, making the discontinuity visible. There are mainly two types of radiations used in radiography, X-radiation and gamma radiation. While gamma rays are produced from a radioactive source and X-rays in a vacuum tube, characteristics of radiation apply to both.

3.5. Ultrasonic Testing

UT is based on the behaviour of sound waves. Sound is nothing but the disturbance of particles in solid, liquid or gas. This disturbance is propagated at different frequencies in different media. In UT, a pulse generated by a piezoelectric material is propagated through a component (as shown in Figure 4(a)). The frequency of this pulse is usually in the order of 106 Hz to facilitate longer transmission distances. As long as the pulse travels through uniform regions, its properties will not be altered. Changes in properties of the pulse (like velocity) are indicative of a present discontinuity. The changes are detected by a transducer which converts sound energy into visual signals that can be displayed on LCD (as shown in Figure 4(b)) or CRT. The transducer can also detect sounds of inaudible frequencies, making UT accurate and reliable. The accuracy of UT depends upon variables like temperature, which can affect the attenuation of waves and even the dimensions of the component. Surface condition is an important variable as

Figure. 4 (a). UT of a pipe segment.

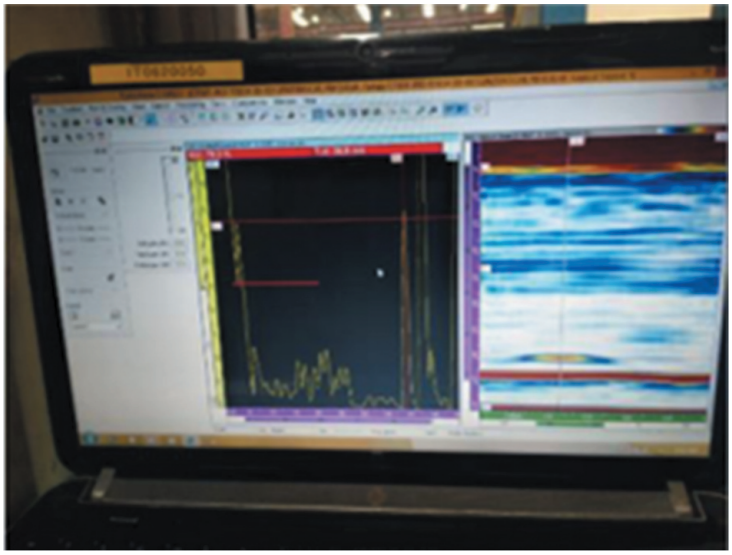


Figure. 4 (b). Real-time feed of UT on the pipe segment.

tighter bound coatings facilitate better pulse transmission. The orientation of the discontinuity is, however, the most important factor in detection by UT.

3.6. Eddy Current Testing

ECT is based on the fluctuation of the magnetic field induced due to eddy currents. When a metal is placed in a fluctuating magnetic field produced by alternating current in a loop, the current is generated in the metal which is swirly in nature. These currents are called eddy currents. A typical ECT set-up consists of two circuits, a primary and a secondary circuit. The primary circuit is connected to an AC supply which sets up a primary fluctuating magnetic field which induces eddy currents in the test specimen. These eddy currents set up a secondary magnetic field that interferes with the primary magnetic field, thereby affecting the current flowing through the coil. Therefore, if eddy currents are altered then the current set-up due to the second magnetic field is also affected, thus

changing the primary current. This change will produce a change in the impedance reading, indicating a discontinuity. There are two types of ECT, pulsed ECT and Remote ECT. Pulsed ECT is used to detect defects in conductive materials, and the latter is used for ECT in remote regions.

3.7. Acoustic Emissions

AE is based on the principle that when a material undergoes deformation, energy is released in the form of elastic waves of high frequency. These waves travel outwards to the surface. Deformations continue to release energy when they are subjected to further loading. These waves are called acoustic emissions. These acoustic emissions are then detected by sensors, and the data is transmitted to a transducer which amplifies and converts the signals into interpretable data. One major difference between AE and other modes of NDT as discussed above is that in AE, the disturbance is originating from the material and not from an external source. Normally, in other modes, a disturbance is generated and transmitted through the specimen, and the response is studied. AE can provide information about the origin of the discontinuity and propagation of the same. Attenuation and wave velocity are two important factors which affect the wave propagation process, ultimately influencing detection.

3.8. Summary

4. Advances in applications of NDT

4.1. Diagnostics and Protection of Cultural Heritage

The requirement for the preservation of historic buildings involves a great knowledge of materials and structures, possible states of damage and their causes. The most arduous problems of unseen situations or detection of the mechanical behaviour of various structural elements like iron tie rods, timber roofs and floors, load-bearing walls, etc., can be solved using the Non-Destructive Testing techniques.

In the instance of masonry, the most often utilised NDTs techniques are thermovision, radar and sonic tests and as Minor Destructive techniques, the flat jack tests. Binda [57] and Suprenat [58] outline the main investigation techniques and their physical principles.

Significant information about the glass manufacturing technique employed can be extracted by using the Non-Destructive Technique of Environmental Scanning Electron Microscopy and Fibre Optics Microscopy with an Energy Dispersive X-Ray analysis on glass fragments obtained from archaeological excavations [59].

A new Non-Destructive Technique which is Adaptable Neural Network Architecture was introduced by Doulamis [60] in 2012 is capable of recognising the artistic styles incorporated in paintings. This technique can be very useful in preserving and restoring cultural paintings.

4.2. On-site Measurement and Monitoring of Concrete by NDT

Concrete is extensively used in various structures. Defects in concrete structures can lead to severe damages and loss of life. Deterioration of Concrete structures often goes

unnoticed, and in many cases, initial damage is not visible until there is huge damage which is beyond repair. Thus, structural failure can be prevented by using Non-Destructive Tests to check for defects during the initial stages only.

To improve the brittle behaviour, steel fibres are introduced into concrete after which chemical coating of the fibres is done in order to attain maximum mechanical properties mainly toughness. For monitoring structures, Acoustic Emission (AE) provides a way to identify between various fracture stages which are pre-peak and post-peak. Acoustic Emission test also reveals the efficiency of chemical coating [61].

NDT can act as a powerful tool providing valuable information about the current condition of a structure. Using NDT during the initial stages of building construction can help in identifying errors beforehand allowing easy repair as the error will be limited to a specific area, as proved by Clausen [62]. This helps in averting grave situations like loss of life and prevent them from happening in future.

From research done by Tatsis, et al. [63] ultrasonic methods of NDT can also be used to estimate the condition of the structural elements and the extent of damages caused by a fire in a building.

Shrinkage cracking is one of the major problems during the early stages in the concrete industry. It occurs primarily due to the usage of low-quality aggregate, which results in volumetric contraction of coarse aggregate. The detection of shrinkage cracking is very important and can be done by using the Acoustic Emission method, which shows some promising results in the identification of shrinkage cracking, according to Ohtsu [64].

4.3. Application of Ultrasonics

Surface waves as described by Rayleigh are waves which move along the solid surface, and their motion is restricted to a region close to the surface whose depth is akin to the wavelength. The surface waves have distinctive characteristics which help in characterisation and detection of anomalies. Depending upon the defect size relative to the wavelength of the surface wave, they are diffracted, scattered and reflected by the surface defects. Surface waves have been used to characterise cracks in aluminium (Al) bars (Edwards et al. [65], Palmer et al. [66], Edwards et al. [67]).

After a specific period of time, fouling deposits start developing on industrial pipes. These deposits are hard to prevent and pose a serious problem to plant operators. There are a lot of varieties of fouling that may occur, entailing different chemical structures and elemental composition. The presence of fouling in pipes results in corrosion, blockage, rupture and deformation, which leads to substantial increase in operation cost and contamination of products. Pipe inspection can be done by involving ultrasonic guided waves, Withers [68], Mc Cements [69], Hay & Rose [70], Lohr & Rose [71], da Silva et al. [72] and da Silva et al. [73]. This technique which is capable of long-range identification with great reliability, has been shown to identify not only fouling but also particular defects in pipes. As experimented by de Carellan [74], due to the potential of conventional ultrasonic cleaning baths to eliminate hard deposits there is a possibility to break the bond between the fouling deposits and the steel pipe by using guided waves of sufficient magnitude.

4.4. Application of Acoustic Emission

Acoustic Emission monitoring is very useful in evaluating the quality of laser cutting, a process of cutting a material using a laser which is defined mainly by the state of the laser cut surface and existence of the re-solidified material occurring as dross at the lower cut edge. Changes in quality of the beam, the surface of materials and gas flow can lead to a reduction in the quality of laser cut. While laser cuts the material, AE signals are emitted due to the interaction between the cutting gas, laser light and plate material. These AE signals can be captured during or immediately after the cutting process, and the dross formation along with the degradation of the laser cut surface can be detected thus helping control the cutting process and ensuring high laser cut quality. Kek [75] proved by experimenting that this method of capturing AE signals to control the laser cutting process is applicable to unalloyed as well as alloyed steel (austenitic stainless steel) with different thicknesses.

Surface rail transportation is one of the most significant and important discoveries of modern technology. With time trains have been improving even if we exclude magnetic levitation technology. Due to the dynamic nature of loading and frequent working of the trains, deformations develop in the wheels and the axles. One of the major factors that lead to the deterioration of the wheel is the emergency braking, which produces a lot of heat due to friction, leading to small or large surfaces of the tyre to be flattened. These flattened surfaces can lead to misalignment, bad rolling quality, fatigue, lower speed, etc. This defect can be detected by using AE monitoring. As the wheels produce transient elastic waves to which AE monitoring is sensitive, thus detecting the defects when the train passes.

Wind turbine blades are one of the most vulnerable and exposed parts of the turbine. They are continuously in contact with gushes of wind and are often exposed to severe weather conditions. This leads to deterioration of the blades. Damaged blades are not acceptable for continued service. Thus, there is a need for monitoring these blades, which can be done by using AE monitoring. Tsopelas [76] summarises the first successful long-term AE monitoring of the full-scale in-service Wind Turbine.

4.5. Other Applications

Ultrasonic methods of NDT can be used in Quality Control of the Ground Improvement for setting up crude oil storage tanks. This method involves the production of vertical impulses by the surface waves on the ground surface. The dispersive characteristic of the incoming wave after being reflected by the ground is calculated, which gives layer stiffness of the layered ground. This procedure permits measuring of shear modulus for profiles of the ground-based on depth, with very small deformations before and after the improvement of the ground. The ratio of shear modulus before and after improvement represents an extent of ground improvement, which is then used as an estimate of successfulness for the conducted groundworks. Using the ultrasonic method of NDT ensures the safety and stability of the oil storage tank [77].

The Radiographic Inspection method of NDT can be used to scan porous construction materials for moisture content. Excess moisture content in construction materials will affect the physical properties of building structures. This leads to dampness and thus

weakening the structure. Radiographic Inspection provides continuous and swift measurement and increased metering on a single sample without the influence of human factor [78].

An application of NDT in quality control can be seen in a technique called Positive Material Identification as shown in Figure 5(a,b) which uses the principle of X-Ray diffraction to determine compositions of alloys in mills post manufacturing.

Aircrafts construction involves usage of carbon fibre laminates which are usually long fibre-enforced composite materials. Quality assurance activities are considered very important because separated fibres are regarded as weak areas. Thus, Structural Health Monitoring is becoming increasingly important nowadays. This structural health monitoring is done using Guided Waves, that is, Ultrasonic method of NDT and also by Acoustic Emission Monitoring technique of NDT [79].

Ground transportation vehicles can also be monitored using Structural Health Monitoring. The wheelset of these vehicles produce sound signals which can be detected by using the Acoustic Emission Monitoring technique of NDT. If there is any damage in the wheel treads such as pitting (formation of cavities or holes), sound signals produced change, and these changes can be detected by integrating AE monitoring system in the wheelset hollow shaft [79].

5. Challenges and future scope

One of the greatest advantages of NDT is that principles utilised in testing have been derived from existing technologies. Development of electronics over the past few decades has largely contributed to advancements in this field. Introduction of LCD in visual testing has facilitated the formation of clearer images, which has made RVT a more reliable technique. Another major advantage of LCD is its portability and durability in the field. Introduction of electronic gauges and sensors has definitely reduced the risk of any misinterpretations and errors. The automated setup used in PT has made this technique not only more suitable for mass manufacturing but has also reduced the risk of human error during the application process. Digital advancements in RT have eliminated the use of films in X-radiography and replaced them with special sensors that are used to process digital images, providing better contrast than films, thus reducing the risk of misinterpretation. Digitalisation in RT has also contributed to better documentation of recorded data. Advancement in electronics has also introduced CT or Computed tomography, which has facilitated the formation of 3D images from 2D figures. Modern UT systems, though manual, are now incorporated with computer-aided systems which have also helped in improving image quality. Another major advancement in the field of NDT is that components can now be simultaneously tested in the field during their operational period. Techniques such as infrared imaging can study stresses induced in a component.

Even though all these advancements have made NDT techniques more favourable, there are certain limitations associated with them. All of the above techniques require highly trained personnel; even then there is no guarantee of zero misinterpretation. Slightest negligence can end in a major catastrophe. Techniques such as VT and RT are especially dependent upon the inspector's knowledge. In RT, sometimes it is possible faults may not be detected due to their orientation with X-Ray beams, it is therefore

Figure. 5 (a). Positive Material Identification (PMI) is an application of X-Ray diffraction. The technique is used to identify materials by checking their content using an X-Ray diffraction machine.

Figure. 5 (b). Material content displayed on the monitor of an X-Ray diffraction machine.

imperative that these techniques require highly trained personnel. It is also important to note that no single technique is sufficient to determine a component's structural integrity.

In some cases, operational costs of equipment used are high, and usually, a minimum of two techniques are required to approve a component safe for operation, making NDT not only time-consuming but expensive as well. VT, PT, RT and UT typically face this issue. Sustainability is another aspect that has to be kept in mind while studying the drawbacks of these techniques. With growing concerns of the rapid depletion of resources, sustainable materials such as composites, glass, fibres, etc., are becoming increasingly popular. Aviation and automobile industries are now switching over to composites as they are lighter and provide better strength than alloys. Composites are

Figure. 6 (a). Often more than two techniques have to be implemented in order to approve a product. Here VT is being done on the internal section of the pipe using a borescope after PT.

Figure. 6 (b). Pinhole defect emerging on the LCD monitor post PT.

also resistant to corrosion and hence provide greater durability. Increasing the use of composites in production would make methods such as ECT and MPI obsolete. Use of liquid penetrants and toxic materials in PT and RT, respectively, also elude NDT technology from achieving sustainability.

Like any other field, manufacturing or non-manufacturing, NDT should also aim for sustainability. This will involve the use of technology that produces minimum life cycle carbon emissions and minimum by-products. Toxic wastes generated in PT and RT pose a risk of contamination. Elimination of the toxic nature of this waste is a scope of interest. Pacana et al. [80] propose methods of choosing fluorescent penetrants in FPI. Although modern radiographic testing technologies have reduced the risk of exposure, waste management still remains a major issue. Another drawback discussed in the previous

paragraph was the employment of highly trained personnel to conduct NDT procedures. However, Artificial Intelligence and Machine Learning show promise in this regard. AI systems can be generated, which can potentially detect flaws that may not be perceptible to an inspector. These systems can particularly hold major advantages in VT, PT, RT and UT, where the risk of misinterpretation is relatively high. Digital Image processing software can be programmed to detect discontinuities in structures through digital images that may be processed in VT, UT, RT or AE. Machine Learning algorithms can also reduce the time of inspection and processing. Robotics also shows promising advancements in NDT. Automated set-ups can be created, which can be integrated with AI systems mentioned previously can greatly improve the effectiveness of NDT techniques.

6. Conclusion

Principle of Non-Destructive testing is simple, to make sure a product conforms to its requirements physically and operationally by conducting a series of tests which do not tamper with the overall viability of the final product. One of the bigger disadvantages of NDT is that a single technique is not sufficient for total satisfaction. However, combining two or more techniques can ensure a greater sense of satisfaction. The underlying working principles of various NDT techniques have proved to be effective in every situation, making NDT a relatively mature field. Scope of development in NDT lies in its application and technology. Development in material technology has broadened the field of application of NDT along with the development of electronics over the last few decades which has provided the field with enhanced and sophisticated equipment (with underlying working principles remaining unchanged) making NDT more reliable and cost-effective. This paper described the various methods involved in NDT and identified as well as compiled the advances in the applications of NDT, showing that the field is not only limited to the identification of an issue but also its classification and rectification followed by preservation.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Mridul Gupta http://orcid.org/0000-0001-9281-3794
Muhsin Ahmad Khan http://orcid.org/0000-0002-0480-3148
Ravi Butola http://orcid.org/0000-0002-6581-4131

References

[1] Cawley P. Non-destructive testing—current capabilities and future directions. J Mater Design Appl. 2001;215:213–223.
[2] Jarmulak J, Eugene JHK, Veen PPV. Case-based reasoning for interpretation of data from non-destructive testing. Eng Appl Artif Intell. 2001;14(4):401–417.
[3] Zhijun L. Non-Destructive Testing of Advanced Composites. Aerosp Mater Technol. 2001.
[4] Park GS, Jang PW, Rho YW. Optimum design of a non-destructive testing system to maximize magnetic flux leakage. J Mag. 2001;6(1):31–35.

[5] Tschelisnig P. Acoustic emission testing (AET) an integral non-destructive testing method. Int J Mater Product Technol. 2001;3(3/4):267–275.
[6] Crostack HA, Reimers W. Evaluation of component integrity by non-destructive testing. Int J Mater Product Technol. 2001;3(2):147–162.
[7] Carlomagnoet GM, Meola C. Comparison between thermographic techniques for frescoes NDT. NDT E Int. 2002;35(8):559–565.
[8] Bollas K, Papasalouros D, Kourousis D, et al. (2011); Acoustic emission monitoring of wheel sets on moving trains, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece, 341–346.
[9] Beard MD, Lowe MJS. Non-destructive testing of rock bolts using guided ultrasonic waves. Int J Rock MechanMining Sci. 2003;40(4):527–536.
[10] Drinkwater BW, Wilcox PD. Ultrasonic arrays for non-destructive evaluation: a review. NDT E Int. 2006;39(7):525–541.
[11] Malcolm K, Lim HC. Combining multiple NDT methods to improve testing effectiveness. Constr Build Mater. 2003;38:950–618.
[12] Luk BL, Alan HSC. Human factors and ergonomics in dye penetrant and magnetic particles nondestructive inspection methods. Engineering Letters. 2007;15(1):163-169.
[13] Taşcă GD, Amza G. Research regarding ultrasonic examination of complex parts. U.P.B. Science Bulletin. 2012;74(2):165-174.
[14] Cau F, Montisci AFA, Testoni P, et al. A signal processing tool for non-destructive testing of inaccessible pipes. Eng Appl Artif Intell. 2006;19(7):753–760.
[15] García-Martín J, Gómez-Gil J, Vázquez-Sánchez E. Non-Destructive techniques based on eddy current testing. Sensors. 2011;11(3):2525–2565.
[16] Pohl R, Erhard A, Montag HJ, et al. NDT techniques for railroad wheel and gauge corner inspection. NDT E Int. 2004;37(2):89–94.
[17] Singh U, Singh M, Singh MK. Analysis study on surface and sub surface imperfections through magnetic particle crack detection for nonlinear dynamic models of some mining components. J Mechan Eng Res. 2012;4(5):185–191.
[18] Bracun D, Polajnar I, Diaci J. Indentation shape parameters as indicators of spot weld quality. Int J Mater Product Technol. 2006;27(3/4):247–257.
[19] Benenson R, Petti S, Fraichard T, et al. Towards urban driverless vehicles. Int J Vehicle Autonom Syst. 2008;6(1/2):4–23.
[20] Dobmann G, Altpeter I, Wolter B. Materials characterization a challenge in NDT for quality management. Int Non-DestructTest Sympos Exhibit 2008. Istanbul, Turkey; 2008. p. 214-223.
[21] Kamal ASM, Boulfiza M. Durability of GFRP rebars in simulated concrete solutions under accelerated aging conditions. J Compos Constr. 2011;15(4):473–481.
[22] Tian GY, Sophian A. Reduction of lift-off effects for pulsed eddy current NDT. NDT E Int. 2005;38(4):319–324.
[23] Shrivastava S, Prakash R. Future research directions with acoustic emission and acousto–ultrasonic technique. Int J Biomed Eng Technol. 2011;6(4):375–386.
[24] Sahama AT, Oloyede A. Near infrared for nondestructive testing of articular cartilage. Nondestruct Test Mater Struct. 2013;6:399–404.
[25] Bo H, Runqiao Y, Zou H. Magnetic nondestructive testing method for thin-plate aluminum alloys. NDT E Int. 2012;47:66–69.
[26] Bieder P, Dobmann G, Kroning M, et al. (2008); Current NDT research & development for NPP inspections; 17th World Conference on Nondestructive Testing, Shanghai, China.
[27] Hasar UC. Non-destructive testing of hardened cement specimens at microwave frequencies using a simple free-space method. NDT E Int. 2009;42(6):550–557.
[28] Wolter B, Dobmann G, Boller C. NDT based process monitoring and control. J Mech Eng. 2011;57(3):218–226.
[29] Banerjee D, Chattopadhyay SK, Chatterjee K, et al. Nondestructive testing of jute–polypropylene composite using frequency-modulated thermal wave imaging. J Thermoplast Compos Mater. 2013;28(4):548-557.

[30] Yunze H, Luo F, Pan M, et al. Defect classification based on rectangular pulsed eddy current sensor in different directions. Sens Actuators A. 2010;157(1):26–31.
[31] Bergmann RB, Bessler FT, Bauer W. Non-Destructive testing in the automotive supply industry. Res Adv Eng. 2007.
[32] Büyüköztürk O, Haupt R, Tuakta C, et al. Remote detection of debonding in FRP-strengthened concrete structures using acoustic-laser technique; Proceedings of NDTMS-2011, Istanbul, Turkey; 2011.
[33] Van Tittelboom K, De Belie N, Lehmann F, et al.; Use of acoustic emission analysis to evaluate the self-healing capability of concrete; Proceedings of NDTMS-2011, Istanbul, Turkey; 2011.
[34] Suzuki T, Naka T, Ohtsu M. Use of acoustic emission method for detection of two-phase flow in service open type water pipeline; Proceedings of NDTMS-2011, Istanbul, Turkey; 2011.
[35] Gür CH. Characterization of steel microstructures by magnetic barkhausen noise technique; Proceedings of NDTMS-2011, Istanbul, Turkey; 2011.
[36] Ribeiro L, Ramos HG, Postolache O. A simple forward direct problem solver for eddy current non-destructive inspection of aluminum plates using uniform field probes. Measurement. 2012;45(2):213–217.
[37] Mazal P, Likuta P, Vlasic F, et al. Use of acoustic emission signal and x-ray diffraction analysis for detection of microstructural changes in cyclically loaded almg alloys; Proceedings of NDTMS-2011, Istanbul, Turkey; 2011.
[38] Zhu Y-K, Tian G-Y, Rong-Sheng L, et al. A review of optical NDT technologies. Sensors. 2011;11(8):7773–7798.
[39] Sato E, Shiwa M, Shinagawa Y, et al. Ultrasonic testing method for detection of planar flaws in graphite material. Mater Trans. 2007;48(6):1227–1235.
[40] Chunguang X, Song W, Pan Q, et al. Nondestructive testing residual stress using ultrasonic critical refracted longitudinal wave. Phys Procedia. 2015;70:594–598.
[41] Mori K, Spagnoli A, Murakami Y, et al. A new non-contacting non-destructive testing method for defect detection in concrete. NDT E Int. 2002;35(6):399–406.
[42] Rainer Pohl A, Erhard H, Montag J, et al. NDT techniques for railroad wheel and gauge corner inspection. NDT E Int. 2004;37(2):89–94.
[43] Avdelidis NP, Almond DP, Hawtin C, et al. Aircraft composites assessment by means of transient thermal NDT. Prog Aerosp Sci. 2004;40(3):143–162.
[44] Ch MR, Röllig AM, Rieck C, et al. Non-Destructive testing application of impulse-thermography for nondestructive assessment of concrete structures. Cem Concr Compos. 2006;28(4):393–401.
[45] Mao Y-M, Que P-W. Application of Hilbert–Huang signal processing to ultrasonic non-destructive testing of oil pipelines. J Zhejiang Univ Sci A. 2006;7(2):130–134.
[46] Gary S, Schajer F, Orhan B. Microwave Non-Destructive testing of wood and similar orthotropic materials. Subsurf Sens Technol Appl. 2005;6(4):293–313.
[47] Davis AG, Lim MK, Petersen CG. Rapid and economical evaluation of concrete tunnel linings with impulse response and impulse radar non-destructive methods. NDT E Int. 2005;38(3):181–186.
[48] Kohl C, Streicher D. Results of reconstructed and fused NDT-data measured in the laboratory and on-site at bridges. Cem Concr Compos. 2006;28(4):402–413.
[49] Shah AA, Ribakov Y. Non-destructive evaluation of concrete in damaged and undamaged states. Mater Des. 2009;30(9):3504–3511.
[50] Orbán Z, Gutermann M. Assessment of masonry arch railway bridges using non-destructive in-situ testing methods. Eng Struct. 2009;31(10):2287–2298.
[51] Rebelo Kornmeier J, Hofmann M, Schmidt S. Nondestructive testing of satellite nozzles made of carbon fibre ceramic matrix composite, C/SiC. Mater Charact. 2007;58 (10):922–927.

[52] Bayraktar E, Antolovich SD, Bathias C. New developments in non-destructive controls of the composite materials and applications in manufacturing engineering. J Mater Process Technol. 2008;206(1–3):30–44.
[53] Chaki S, Bourse G. Guided ultrasonic waves for non-destructive monitoring of the stress levels in prestressed steel strands. Ultrasonics. 2009;49(2):162–171.
[54] Mendikute AA, López-Arraiza A, Lizaranzu M, et al. Comparison and analysis of non-destructive testing techniques suitable for delamination inspection in wind turbine blades. Compos Part B. 2011;42(5):1298–1305.
[55] Van Den Abeele F, Goes P. Nondestructive testing techniques for risk-based inspection. Sust Construct Design. 2011. p. 161-171.
[56] Hoła J, Schabowicz K. State-of-the-art nondestructive methods for diagnostic testing of building structures – anticipated development trends. Arch Civil Mech Eng. 2010;10. DOI:10.1016/S1644-9665(12)60133-2
[57] Binda L, Saisi A, Tiraboschi C. Investigation procedures for the diagnosis of historic masonries. Construct Build Mater Elsevier. 2000;14(4):199–233.
[58] Suprenant B, Schuller M. Non-destructive evaluation & testing of masonry structures. USA: The Aberdeen Group; 1994.
[59] Cheilakou E, Liarokapi N, Koui M. NDT characterization of ancient glass objects from the aegean with an approach of the manufacturing technique, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 63–70.
[60] Doulamis AD, Varvarigou TA. Image analysis for artistic style identification: a powerful tool for preserving cultural heritage, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 71–76.
[61] Aggelis DG, Soulioti DV, Barkoula NM, et al. Influence of the fiber chemical coating on the fracture behavior of steel fiber concrete measured by acoustic emission, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 111–116.
[62] Clausen JS, Zoidis N, Knudsen A. Onsite measurements of concrete structures using impact-echo and impulse response, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 117–122.
[63] Tatsis E, Zoidis N, Vlachopoulos C, et al. Visual inspection and evaluation using NDT testing methods of industrial building after fire, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 131–136.
[64] Ohtsu M, Matsuo T, Nozaki S, et al. AE monitoring of shrinkage process in concrete, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 143–148.
[65] Dixon S, Edwards RS, Jian X. Inspection of rail track head surfaces using electromagnetic acoustic transducers (EMATs). Insight. 2004;46(6):326–330.
[66] Dixon S, Edwards C, Palmer SB. The optimization of lamb and Rayleigh wave generation using wideband-low-frequency emats. Rev Quantitat Nondestruct Evaluat. 2003;22:297–304.
[67] Edwards RS, Dixon S, Jian X. Characterisation of defects in the railhead using ultrasonic. NDT E Int. 2006;39(6):468–475.
[68] Withers PM. Ultrasonics for the detection of fouling. Food Technol Int Eur. 1994;Reprint 768:96–99.
[69] Mc Clements DJ. Advances in the application of ultrasound in food analysis and processing. Trends Food Sci Technol. 1995;6(9):293–299.
[70] Hay TR, Rose JL. Fouling detection in the food industry using ultrasonic guided waves. Food-Control. 2003;14(7):481–488.
[71] Lohr KR, Rose JL. Ultrasonic guided wave and acoustic impact methods for pipe fouling detection. Journal-of-Food-Engineering. 2003;56(4):315–324.
[72] Da Silva J, Wanzeller MG, De Almeida Farias P, et al. Development of circuits for excitation and reception in ultrasonic transducers for generation of guided waves in hollow cylinders

for fouling detection, instrumentation and measurement technology conference, Ottawa, ON, Canada; 2005. p. 467–471.
[73] Da Silva J, Jaidilson J, Lima MNA, et al. Fouling detection using ultrasonic guided waves and parameter estimation, Third IFAC Symposium on System Structure and Control. 2007b.
[74] Ignacio Garcia De C, Catton P, Selcuk C, et al. Methods for detection and cleaning of fouling in pipelines, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 231–236.
[75] Kek T, Grum J. AE monitoring at laser cutting, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 321–326.
[76] Tsopelas N, Kourousis D, Ladis I, et al. Health monitoring of operating wind turbine blades with acoustic emission, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 347–352.
[77] Kovačević MS, Mirčeta A, Marčić D. NDT and safety of crude oil storage tanks, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 477–482.
[78] Skramlik J, Novotny M, Suhajda K. Apparatus for measuring of liquid conductivity, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 489–494.
[79] Meyendorf N, Frankenstein B, Schubert L. Structural health monitoring for aircraft, ground transportation vehicles, wind turbines and pipes—prognosis, Proceedings of the fifth conference on Emerging Technologies in NDT, Ioannina, Greece; 2011. p. 15–22.
[80] Pacana A, Siwiec D, Bednárová L. Method of choice: a fluorescent penetrant taking into account sustainability criteria. Sustainability. 2020;12(14):5854.

International Journal of Pavement Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gpav20

An intelligent approach for predicting the strength of geosynthetic-reinforced subgrade soil

Muhammad Nouman Amjad Raja, Sanjay Kumar Shukla & Muhammad Umer Arif Khan

To cite this article: Muhammad Nouman Amjad Raja, Sanjay Kumar Shukla & Muhammad Umer Arif Khan (2021): An intelligent approach for predicting the strength of geosynthetic-reinforced subgrade soil, International Journal of Pavement Engineering, DOI: 10.1080/10298436.2021.1904237

To link to this article: https://doi.org/10.1080/10298436.2021.1904237

Published online: 01 Apr 2021.

Submit your article to this journal

Article views: 165

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gpav20

INTERNATIONAL JOURNAL OF PAVEMENT ENGINEERING
https://doi.org/10.1080/10298436.2021.1904237

Taylor & Francis
Taylor & Francis Group

Check for updates

An intelligent approach for predicting the strength of geosynthetic-reinforced subgrade soil

Muhammad Nouman Amjad Raja [a,b], Sanjay Kumar Shukla [a,c] and Muhammad Umer Arif Khan [a,d]

[a]Geotechnical and Geoenvironmental Engineering Research Group, School of Engineering, Edith Cowan University, Perth, Australia; [b]Civil Engineering Department, University of Management and Technology, Lahore, Pakistan; [c]Department of Civil Engineering, Delhi Technological University, Delhi, India; [d]Department of Civil Engineering, Mirpur University of Science and Technology, Mirpur, Azad Kashmir, Pakistan

ABSTRACT

In the recent times, the use of geosynthetic-reinforced soil (GRS) technology has become popular for constructing safe and sustainable pavement structures. The strength of the subgrade soil is routinely assessed in terms of its California bearing ratio (CBR). However, in the past, no effort was made to develop a method for evaluating the CBR of the reinforced subgrade soil. The main aim of this paper is to explore and appraise the competency of the several intelligent models such as artificial neural network (ANN), least median of squares regression, Gaussian processes regression, elastic net regularisation regression, lazy *K*-star, M-5 model trees, alternating model trees and random forest in estimating the CBR of reinforced soil. For this, all the models were calibrated and validated using the reliable pertinent historical data. The prognostic veracity of all the tools mentioned supra were assessed using the well-established traditional statistical indices, external model evaluation technique, multi-criteria assessment approach and independent experimental dataset. Due to the overall excellent performance of ANN, the model was converted into a trackable functional relationship to estimate the CBR of reinforced soil. Finally, the sensitivity analysis was performed to find the strength and relationship of the used parameters on the CBR value.

ARTICLE HISTORY
Received 30 December 2020
Accepted 12 March 2021

KEYWORDS
California bearing ratio; geosynthetic reinforcement; subgrade soil; machine learning; intelligent predictive modelling

1. Introduction

The pavement design is highly influenced by the load-carrying capacity of the subgrade soil. Geosynthetic reinforcement provides a sustainable and cost-effective way of soil subgrade improvement. The use of geosynthetic reduces the deformation of the pavement caused by the vehicular load, enhancing its strength and durability (Shukla 2002). Starting in the late 1900s, many studies can be found in the literature that investigate the beneficial effects of geosynthetic reinforcement in road construction projects (e.g. Miura et al. 1990, Perkins 1999, Cuelho et al. 2005, Abu-Farsakh et al. 2016, Chen et al. 2018, Singh et al. 2019). The California bearing ratio (CBR) is considered as one of the key parameters used to ascertain the subgrade capacity to withstand the applied traffic loads (ASTM 2016). To date, many researchers have employed the CBR testing to investigate the effects of geosynthetic reinforcement on subgrade material (Duncan-Williams and Attoh-Okine 2008, Naeini and Ziaie-Moayed 2009, Nair and Latha 2011, Choudhary et al. 2012, Rajesh et al. 2016, Mittal and Shukla 2018, Negi and Singh 2019). The advent of soft computing and data-driven modelling has made many traditional approaches antiquated. In recent times, the use of artificial intelligence (AI)/machine learning (ML) techniques has become very common in solving various complex engineering problems including those related to pavement engineering (Nazemi and Heidaripanah 2016, Daneshvar and Behnood 2020, Ghosh Mondal and Kuna 2020, Han et al. 2020, Olowosulu et al. 2020, Ghorbani et al. 2021). The laborious, costly and time-consuming nature of the CBR, and moreover, due to the complex non-linear relationships between the soil properties, many researchers have utilised ML to predict the CBR of the soil. Taskiran (2010) applied artificial neural network (ANN) and gene expression programming (GEP) techniques to learn the non-linear relationships between CBR and various index properties of the fine-grained soils sourced from Southeast Anatolia, Turkey. The maximum dry unit weight (γ_d), plasticity index (PI), liquid limit (LL), optimum moisture content (OMC) and the content of different soil fractions were identified as the most effective parameters that influence the CBR values of the soils. Yildirim and Gunaydin (2011) used simple multiple regression and ANN to propose correlations for the preliminary estimation of CBR values of different soil, by employing the results of sieve analysis, Atterberg limits, maximum dry density and OMC. Alawi and Rajab (2013) utilised multiple linear regression (MLR) analysis to predict the CBR of the unreinforced subbase soil layer. Recently, similar applications of advanced ML techniques for the prediction of CBR of different unreinforced soils were also conducted by other researchers (Erzin and Turkoz 2016, González Farias et al. 2018, de Souza et al. 2020, Nagaraju et al. 2020, Tenpe and Patel 2020). Similarly, several studies were also carried out to evaluate permanent deformation and resilient modulus of recycled demolition wastes in pavements using ML algorithms (Arulrajah et al. 2013, Ullah et al. 2020, Ghorbani et al. 2020a). Ghorbani et al. (2020b, 2021)

CONTACT Muhammad Nouman Amjad Raja m.raja@ecu.edu.au, noumanamjad@live.com
© 2021 Informa UK Limited, trading as Taylor & Francis Group

successfully developed an ANN model to predict the permanent strain of blends of two recycled waste materials under different stress and temperature levels and shakedown analysis of polyethylene terephthalate (PET) blend with demolition waste material as the subbase layer of pavement, respectively. However, the review of the current literature revealed that the efforts to predict the CBR of geosynthetic-reinforced subgrade soil are extremely limited. In fact, Singh et al. (2020) is the only study that has attempted to predict the CBR of geogrid-reinforced soil using a fuzzy logic-based modelling technqiue. However, to the best of authors' knowledge, there is currently no research in the present literature that provides development, implementation and comprehensive comparison of ML-based solutions to the problem of CBR of geosynthetic-reinforced subgrade soil.

In this paper, an attempt is made to predict the CBR of geosynthetic-reinforced soil (GRS) using the data-driven-based ML models. The main objectives of this research are fourfold: (1) assessment of numerous ML models in predicting the CBR of GRS; (2) comprehensive comparison of all the predictive tools utilised for the same problem; (3) suggestion of ANN-based trackable mathematical formula for estimating the CBR of soil reinforced with geosynthetic layers; and (4) independent validation by conducting new CBR tests on reinforced soil.

2. Material and methods

In this work, eight ML models namely, ANN, least median of squares regression (LMSR), Gaussian processes regression (GPR), elastic net regularisation regression (ENRR), lazy *K*-star (LKS), M-5 model trees, alternating model trees (AMT) and random forest (RF) models, were constructed to predict the CBR of GRS. All the models were simulated using Waikato Environment for Knowledge Analysis (WEKA). It may be noted that in the past, many researchers and scientists have utilised the WEKA framework for function approximation and feature classification (Gao et al. 2019, Moayedi et al. 2020, Olowosulu et al. 2020). Moreover, the new CBR tests were also conducted to ensure the independent validation of the data-driven models and discussed later in the paper.

2.1. Experimental database development and model attributes

In order to calibrate and validate all the data-driven models, the data set of 97 soaked CBR tests is retrieved from the literature. This includes 4 cases reported by Duncan-Williams and Attoh-Okine (2008), 30 cases reported by Vinod and Minu (2010), 16 cases reported by Choudhary et al. (2011), 2 cases reported by Kuity and Roy (2013), 4 cases each reported by Carlos et al. (2016) and Rajesh et al. (2016), 9 cases by Mittal and Shukla (2018, 2019) and 28 cases by Negi and Singh (2019). The reliability of a model depends upon the comprehensiveness of the input data set. In this study, it was ensured by incorporating a wide variety of soils, ranging between sandy soil (SP) and fine-grained soils (ML, MH, CL and CH), as per the Unified Soil Classification System (USCS). These soils cover a range of engineering properties that affect the stiffness of a soil, such as soil index properties and particle size distribution (Youd 1973, Zheng and Hryciw 2016). However, the geotechnical engineering models that are the most effective in predicting the non-linear soil behaviour are based on the important soil parameters that can be obtained via routine tests (Oztoprak and Bolton 2013). The detail of the utilised database is provided in Table 1. For predicting the output, that is, CBR of GRS, the input parameters include LL (X_1), plastic limit (X_2), PI (X_3), dry unit weight (maximum) (X_4), optimum moisture content (X_5), percentage fines (passing sieve No. 200) (X_6), percentage sand (X_7), tensile strength of geosynthetic reinforcement (X_8), number of reinforcement layers (X_9), position of the first reinforcement layer (X_{10}) and position of the subsequent reinforcement layers (X_{11}). All these parameters were selected based on the fact that the current literature shows that they have an effect on the CBR of reinforced soil.

The developed models were also tested for the independent data set obtained by conducting new CBR tests (AS 1289.6.1.1 2014). It is noteworthy that these data are not a part of the actual database utilised to construct the ML models. Non-plastic sandy soil extracted from the pit site located in the northern region of Perth, Australia, was used during the experimental tests. The properties of the soil and geosynthetic (geotextile) are summarised in Table 2. To evaluate the effect of geosynthetic reinforcement on the strength of the subgrade soil, a total of six CBR tests were performed by varying the depth ratio of the first reinforcement layer (u/H), number of reinforcement layers (N) and depth ratio of the subsequent layer (h/H), where H represents the total height of mould. The proposed scheme of the CBR tests is summarised in Table 3.

The reinforcement layer was cut into a circular disk with the diameter slightly smaller than the diameter of the mould. In order to fill the mould, dry weight calculations are done based on the maximum dry unit weight of soil and volume of mould. The soil was mixed thoroughly by adding the water content corresponding to OMC. Thereafter, the mould is filled with the soil by placing the geotextile layer at a predetermined depth as reported in Table 3. The test setup with the schematic diagram of the specimen in the CBR test and the placement of geosynthetic at a predetermined depth in the CBR mould is illustrated in Figure 1. The CBR tests were conducted after soaking the sample in water for 96 h. The surcharge load was applied to the specimen to ensure the effect of the thickness of the overlying layer. Load was applied at 1.25 mm/min (moveable base) and the corresponding penetration was measured through the electronic displacement transducer. The load readings were taken at penetrations ranging from 0.5 to 12.5 mm. The CBR values were estimated by taking the load corresponding to 2.5 or 5.0 mm penetration, whichever is the highest as suggested by the relevant standard. Figure 2 depicts the load–penetration curves obtained for the CBR tests of reinforced soil.

2.2. Methodological background of machine learning models

The first and foremost step before mapping the response of any data-driven model is the database partitioning. For this, the data are needed to be randomly divided into training and

Table 1. Statistical properties of the data set.

Descriptive stats	Liquid limit	Plastic limit	Plastic Index	Dry unit weight	Optimum moisture content	% Fines	% Sand	Tensile strength of geosynthetic	Number of layers	Position of 1st layer	Position of subsequent layers	CBR
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	
Mean	50.746	28.612	22.135	15.782	20.986	85.025	14.268	16.821	1.392	0.382	0.101	6.453
Median	46.000	25.000	23.000	15.100	18.000	94.000	6.000	13.000	1.000	0.330	0.000	4.870
Mode	46.000	23.000	23.000	15.100	18.000	97.000	3.000	30.000	1.000	0.200	0.000	3.460
SD	15.452	8.670	7.787	2.351	7.709	22.826	22.344	11.901	0.701	0.210	0.176	6.183
Kurtosis	1.751	1.565	0.666	−1.143	−0.808	3.783	3.956	−1.207	3.214	−0.478	2.111	10.454
Skewness	−0.837	−0.837	−0.878	0.076	0.805	−2.242	2.285	0.423	1.881	0.780	1.669	3.122
Range	71	39	32	8.25	26.8	83.8	82	39.17	3	0.68	0.66	32.7
Minimum	0	0	0	12.54	8.2	16.2	0	2.81	1	0.15	0	1.3
Maximum	71	39	32	20.79	35	100	82	41.98	4	0.83	0.66	34
Count	97	97	97	97	97	97	97	97	97	97	97	97

SD stands for standard deviation

Table 2. Properties of soil and geosynthetic material

Soil properties	Magnitude	Geosynthetic properties[a]	Magnitude
Maximum dry unit weight (kN/m^3)	17.3	Tensile strength, MD (kN/m)	30
Minimum dry unit weight (kN/m^3)	15.4	Tensile strength, XMD (kN/m)	30
Median grain size (mm)	0.25	Mass (g/m^2)	150
Effective grain size (mm)	0.17	Elongation strain, MD (%)	20
Optimum moisture content (%)	11.8	Elongation strain, XMD (%)	15
Atterberg's limit	NP		
Fines (%)	4.5		
Sand content (%)	95.5		

Note: NP stands for non-plastic; MD stands for machine direction; XMD stands for cross-machine direction.
[a]Provided by the manufacturer.

Table 3. Experimental scheme of the CBR tests

Experiment number	Number of layers (N)	Depth of first layer (u/H)	Depth of subsequent layer (h/H)	CBR (%)
1	1	0.25	0	23.62
2	1	0.50	0	21.98
3	1	0.75	0	17.37
4	2	0.25	0.25	29.3
5	2	0.25	0.5	28
6	2	0.50	0.25	25.88

testing subsets. In this study, 60% of the collected data were randomly chosen for training the ANN, LMSR, GPR, ENRR, LKS, M-5 model trees, AMT and RF models. Thereafter, the predictive veracity of each model was appraised against the remaining 40% of the data set (test data). Moreover, all the data sets have been normalised [−1,1] before feeding it to the model networks, so that each variable would get the same attention during the training process. The brief methodological background of all the data-driven-based modelling techniques utilised to estimate the CBR of reinforced soil is explained in this section. For a more comprehensive understanding, the research scheme employed in this study is also illustrated in Figure 3.

2.2.1. Artificial neural network

ANN is the well-established and widely recognised ML model used for mapping the non-linear response of any system (Moayedi and Hayati 2018). At its core, the ANN model architecture consists of three parts, namely input layer, one or more hidden layers and output layer. Each layer consists of set processing elements called nodes (neurons) which interact with each other through weighted connections. The computation process can be described as follows: (a) data are presented to the model through input layer nodes; (b) in the hidden–output layer, the data are multiplied by the weight matrix and added to the threshold (bias) vector, thereafter the activation function is applied; (c) output of the hidden layer is mapped as the final outcome after passing through the output layer. Mathematically, for n inputs, the output y is computed as follows (Bishop 2006):

$$y = f\left(\sum_{i=1}^{n} X_i w_i + \theta_o\right) \tag{1}$$

where w is the weight connection and θ is the bias/threshold.

In this study, the sigmoid activation function is used in the hidden layer and given as follows (Han and Moraga 1995):

$$\mathrm{sig}(X) = \frac{1}{1 + e^{-x}} \tag{2}$$

For training the neural network, the optimisation procedure suggested by Soleimanbeigi and Hataf (2006) was adopted to select the optimum number of hidden layer nodes. For this, the hidden layer nodes were increased until no further improvement was obtained over the testing data set. Figure 4 illustrates the optimisation process for selecting the number of hidden nodes. It can be observed that the lowest mean absolute error (MAE = 1.233) and root mean square error (RMSE = 1.70) is obtained at nine hidden nodes. However, the model with six hidden nodes is selected as the optimum model as it has less weight connections, but its performance is closer to nine hidden nodes with MAE and RMSE values

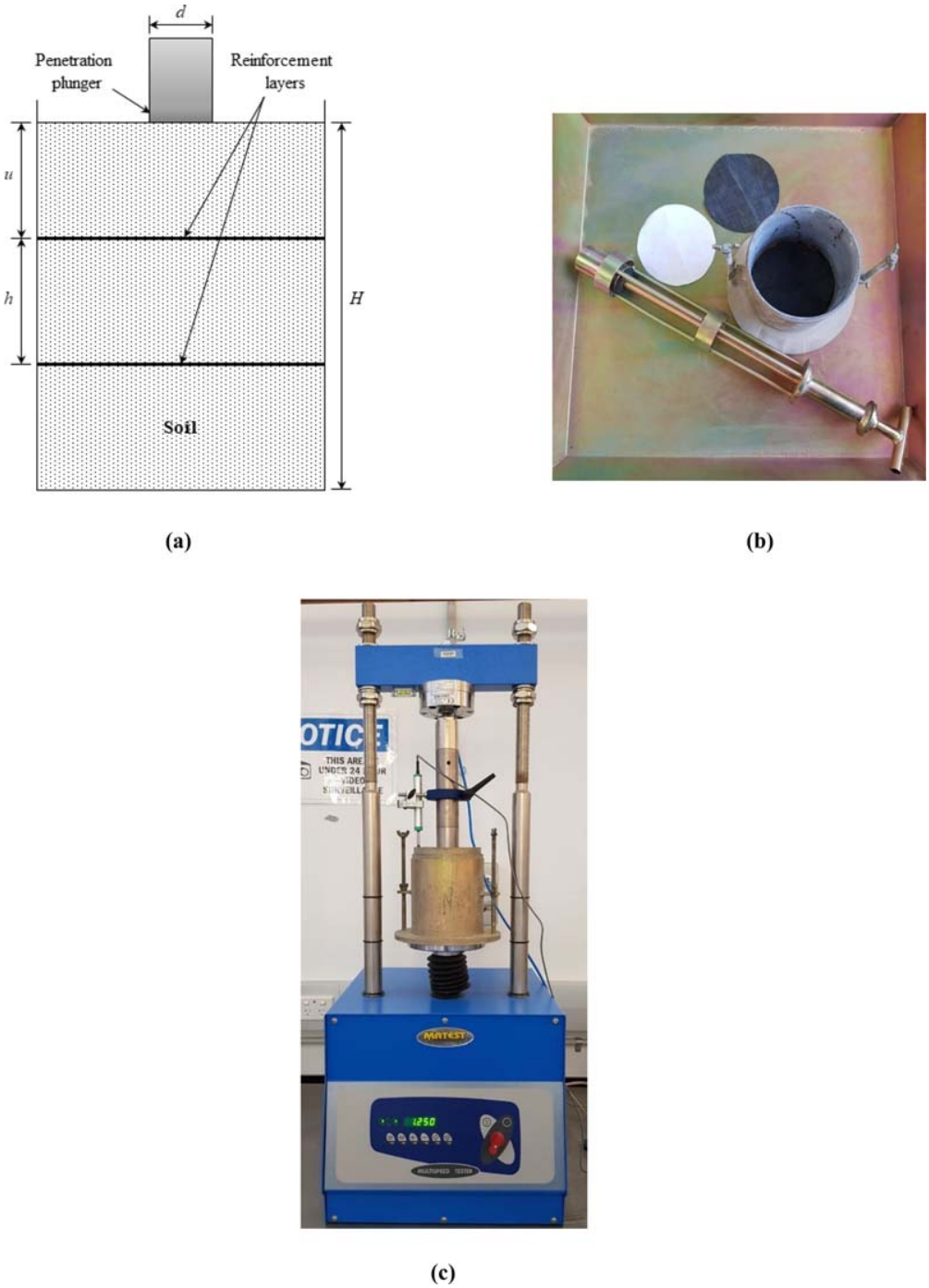


Figure 1. Experimental setup for the CBR tests conducted in this study: (a) schematic diagram of the specimen; (b) layout/placement of reinforcement layer at a predetermined depth in CBR mould; (c) test apparatus.

of 1.31 and 1.74, respectively. The architecture of the optimum ANN model (11–6–1) is given in Figure 5.

2.2.2. Gaussian process regression

In the past, GPR model has been efficiently used in predicting the response of the system (Gao et al. 2019, Zhang et al. 2019, Suthar 2020). GPR is a kernel-based ML model that is founded on Bayesian theory and statistical learning approach. Such model can be completely defined by the mean and covariance function (kernel) as follows (Rasmussen 2006):

$$\mu(x) = \mathrm{E}[g(x)] \tag{3}$$

$$K(x, x') = \mathrm{E}[(g(x) - \mu(x))(g(x') - \mu(x'))] \tag{4}$$

where $\mu(x)$is the mean function, and $k(x, x')$is the covariance function of a real process $g(x)$. The Gaussian process is defined by the set of random parameters, and any finite number of it has the joint Gaussian allocation. Mathematically

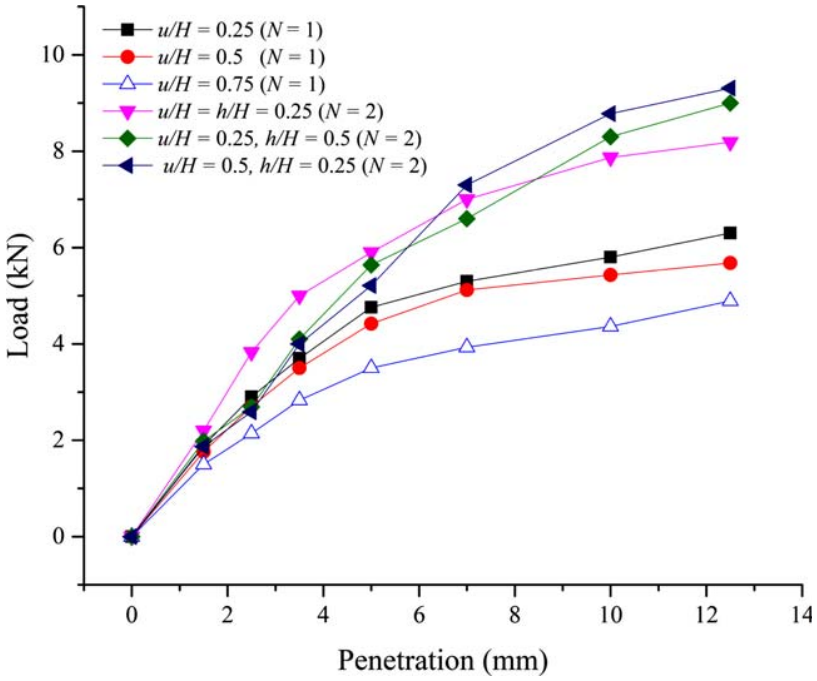


Figure 2: Load–penetration curves of GRS.

(Gao et al. 2019):

$$g(x) \sim \mathrm{GP}(\mu(x), k(x, x')) \tag{5}$$

For detailed derivation, readers are directed to excellent studies available in the literature (Seeger 2004, Rasmussen 2006).

In this study, various kernel functions such as squared exponential, radial bias and Pearson's VII universal kernel functions are used for estimating the CBR of reinforced soil using GPR. The optimum results are obtained by employing the Pearson's VII universal kernel (PUK) function. The mathematical form of PUK is given as follows (Üstün et al. 2006):

$$k(x, x') = 1/[1 + (2\sqrt{\|x - x'\|^2}\sqrt{2^{(1/\omega)} - 1/\sigma)^2}]^{\omega} \tag{6}$$

where $\sigma$and $\omega$are the Person's width, and peak tailing factor, respectively.

2.2.3. Least median of square regression

The LMSRis a semi-parametric quantile regression technique. In contrary to the classical regression model, the sum of least squares is replaced by the median of squared errors (Rousseeuw 1984). The LMSR overcomes the major drawback in the ordinary regression, that is, the sensitivity to the outliers. For a standard univariate linear regression problem, the residuals take the following form (Massart et al. 1986)

$$r_j = y_j - ax_j - b \tag{7}$$

where r is the residual, y is the response (output), x is the input, and a and b are regression coefficients. The principle of least square governs, minimise $\sum_{j=1}^{n} r_j^2$, whereas LMS estimators aims at minimising the median of square errors, that is, minimise $\mathrm{med}\sum_{j=1}^{n} r_j^2$. In this study, the following relationship is obtained for estimating the CBR of reinforced soil by using the LMS regression technique:

$$\begin{aligned}\mathrm{CBR}_{\mathrm{LMSR}} &= 0.0231X_1 - 0.018X_2 + 0.1528X_3 \\ &- 0.015X_5 - 0.6108X_6 + 0.0032X_8 + \\ &0.8856X_9 - 1.3882X_{10} - 1.8478X_{11} + 57.95\end{aligned} \tag{8}$$

2.2.4. Elastic net regularisation regression

The ENRR is a robust regression model which aims at combining the penalties of the least absolute shrinkage and selection operator (LASSO), l_1, and ridge regression technique, l_2(Ogutu et al. 2012). LASSO randomly tends to choose only one attribute and ignore the other, especially when the attributes are highly correlated, whereas the elastic net (EN) overcomes this problem. For a set of data sample with n observations and p predictors, let $\{(x_i, y_i), i = 1, 2 \ldots, n)\}$, where x_i, y_i belongs to R^p. Moreover, if $y = (y_1 \ldots, y_n)^{\mathrm{T}}$ and $X \in R^{n \times p}$ represents the output vector and model matrix, respectively, then the EN can be written as follows (Zou and Hastie 2005):

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} |y - X\beta|^2, \quad \text{subjected to} \quad \xi \\ &= (1 - \lambda)|\beta|_1 + \lambda|\beta|^2 \le s \quad \text{for some } s\end{aligned} \tag{9}$$

where$\beta$is the weight vector, ξ is the EN penalty parameter, that is a combination of l_1 and l_2. For the detail derivation, readers may refer to the research conducted by Zou and Hastie (2005).

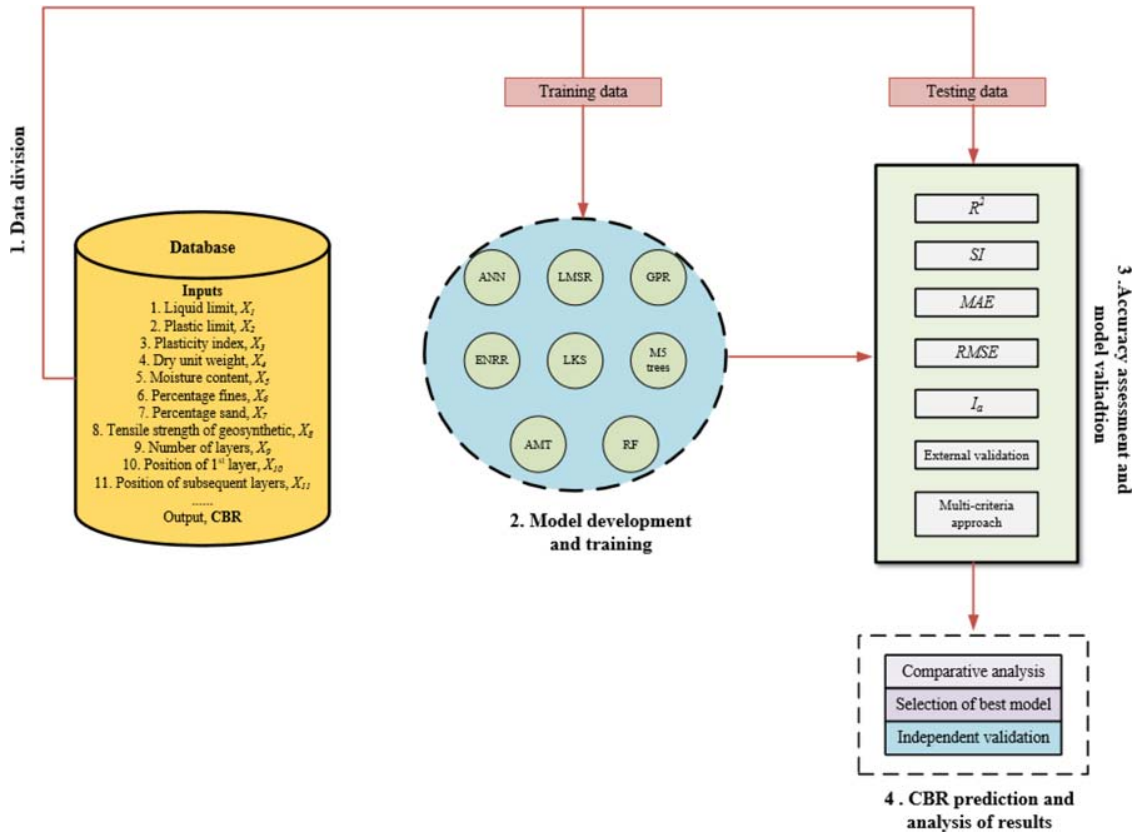


Figure 3. Research scheme employed for estimating the CBR of reinforced soil.

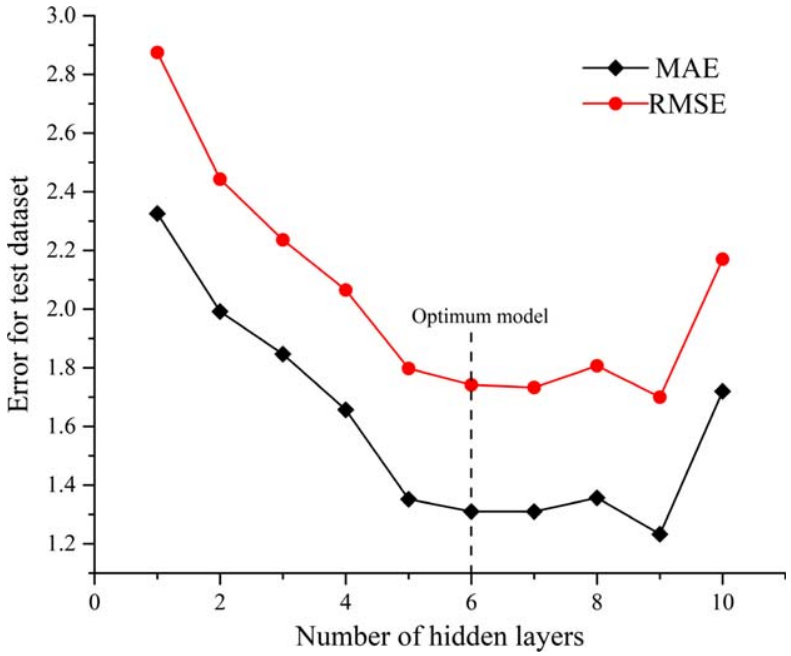


Figure 4. Optimisation process for selecting the number of hidden nodes.

It may be noted that, if $\lambda = 1$, then EN takes the ridge regression shape, and if $\lambda = 0$, then it takes the LASSO form. For $\lambda \in [0, 1)$, the EN encompasses the characteristics of both the ridge and LASSO. In other words, the penalty parameter l_1 enables the automatic selection of variables and l_2 leads to stabilisation by making the problem strictly convex

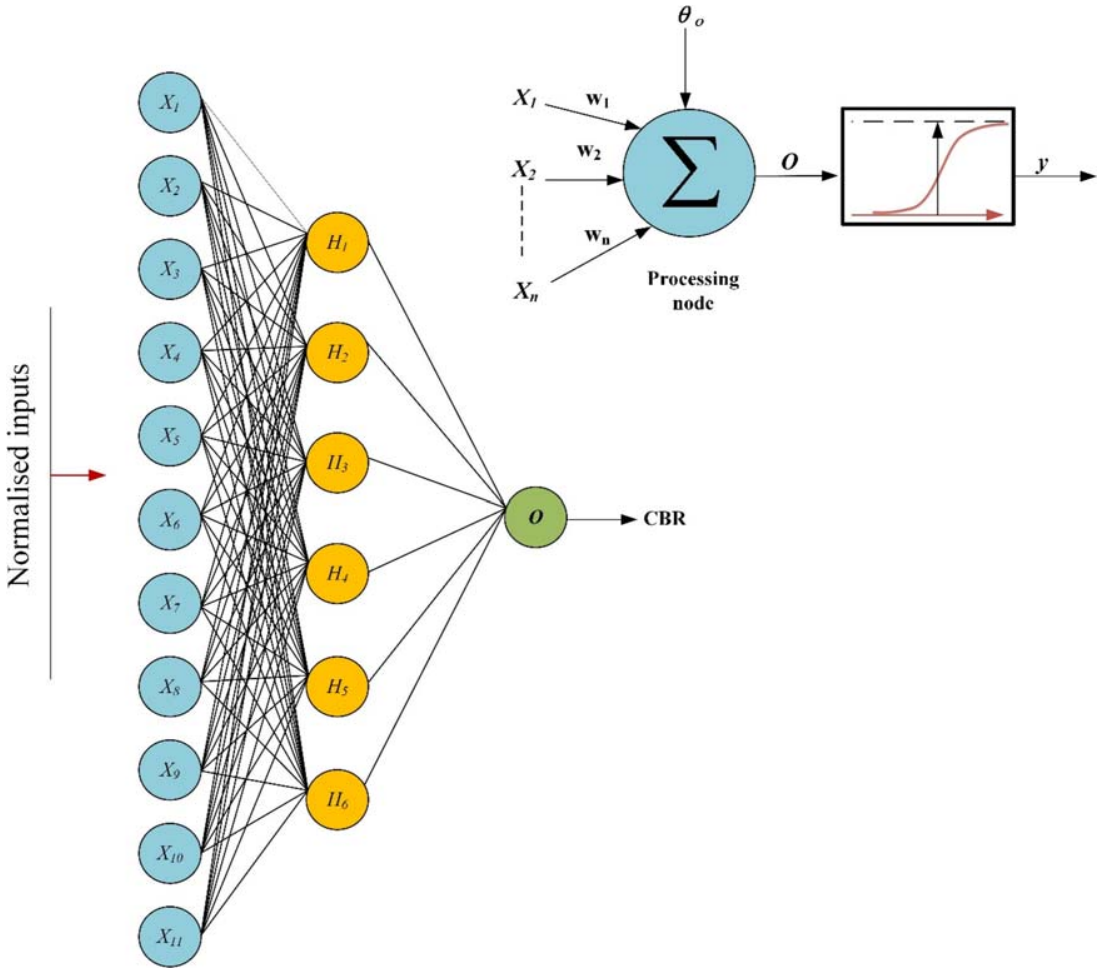


Figure 5. Architecture of the optimum ANN model.

(Gao et al. 2019). For the estimation of CBR, the following relationship is obtained through ENRR regression:

$$\begin{aligned} \mathrm{CBR}_{\mathrm{ENRR}} &= 0.229X_1 + 0..229X_2 - 0.237X_4 - 0.421X_5 \\ &- 0.028X_6 + 0.261X_7 + 0.027X_8 + \\ &0.225X_9 - 0.069X_{10} + 0.039X_{11} - 1.240 \end{aligned} \tag{10}$$

2.2.5. Lazy K-star

Lazy learning is a ML method in which the majority of the computation time is deferred to the consultation time, that is, until a query (call) is made to the system (Webb et al. 2011). LKS is a lazy learning algorithm which performs the generalisation of the training data using instant base learning classification. This means that unlike the ANN or other ML methods, the predictions are not inferred from particular instances in the training data, instead the complete data are stored in the memory and upon call, the response is produced by the nearest neighbour approach. The *K*-star sums all the possible transitions between the two instances and amalgamate them into a single class (Cleary and Trigg 1995). This is achieved by summing the probabilities over all possible transformations between the instances. This entropy-based learning technique has several advantages in comparison to other rule-based schemes, such as better handling of missing values and common attributes. Mathematically, the function *K*-star is defined as follows (Cleary and Trigg 1995, Gao et al. 2019):

$$K^* = -\log P^*(q|p) \tag{11}$$

where P^* represents the probability function, that is, the probability of all the paths from instances p to q.

2.2.6. M-5 model trees

Based on the original research conducted by Quinlan (1992) on the development of decision trees for the regression problem, Wang and Witten (1997) proposed the M-5 model. The M-5 model uses the classical top-down method for growing and pruning decision trees. The data are presented in the form of the mean values and regression function to the leaf nodes; thereafter, the branching/separation is performed at each node (MLR) until the values of response variables reaching a node shows no or negligible change. In the next step, the larger subtrees are replaced by a single larger linear model. For this, the error values inside the inner node of the tree end are compared with those of the tree leaf underlying that node

(Khorrami et al. 2020). Finally, the smoothing function is applied to trim the gaps between the neighbouring (adjacent) leaf nodes. In this way, the final model is obtained by combining all the available linear models along the path from the root node of the tree to each leaf node; thus, effectively producing a linear combination of all the available linear models (Quinlan 1992, Khorrami et al. 2020).

2.2.7. Alternative model trees

AMT is a recently developed algorithm based on the principle of ensemble learning. Originally, proposed by Frank et al. (2015), AMT uses the additive regression technique to grow the trees. Moreover, similar to the decision trees, AMT utilises stage-wise forward additive regression (statistical boosting variant) and cross-validation technique to minimise the square errors and to limit the growth of the trees, respectively (Moayedi et al. 2020). The main difference between the M-5 model trees and AMT model is that the former uses the multivariate linear regression on the leaves, while the latter employs the simple linear regression to obtain the predictive model. For this, AMT utilises two nodes, namely splitter node and predictor node. The splitter node splits the numeric attributes at the median value, and the predictor node predicts the response of the system by using the linear regression technique (Gao et al. 2019, Moayedi et al. 2020). For detailed insight and derivation, the reader may refer to Frank et al. (2015). AMT model proposed for the present study for estimating the CBR of reinforced soil is illustrated in Figure 6.

2.2.8. Random forest

RF is an ensemble learning method originally suggested by Ho (1995) for classification and regression problems. At its core, the RF works by creating a swarm of decision trees and then averaging the output of each tree. The bootstrap aggregating technique helps the RF to obtain a much stable solution, and reduces the chances of overfitting (Breiman 2001). The user control parameters are the number of trees, number of nodes, and number of variables. Generally, the higher number of trees will lead to higher accuracy but require more computation time. As each tree works entirely independently and uses out-of-bag estimates to observe the errors and correlation strength, the abundance of trees will not lead to overfitting of the model. The detailed algorithm is described in Breiman (2001).

2.3. Model assessment criteria

Five statistical matrices were chosen to appraise and compare the predictive strength of all the developed data-driven models. The matrices are as follows: (i) coefficient of determination (R^2); (ii) root means square error (RMSE); (iii) scatter index (SI); (iv) index of agreement (I_a) and (v) mean absolute error (MAE). All these statistical tools have been extensively used in the previous studies to simulate the accuracy of ML-based models (Yaseen et al. 2018, Khorrami et al. 2020, Raja and Shukla 2020, 2021). The mathematical form of these statistical standards is given below

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} - \mathrm{CBR}_{p_i})^2}{\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} - \overline{\mathrm{CBR}_o})^2} \tag{12}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} - \mathrm{CBR}_{p_i})^2} \tag{13}$$

$$\mathrm{SI} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} - \mathrm{CBR}_{p_i})^2}}{\overline{\mathrm{CBR}_o}} \tag{14}$$

$$I_a = 1 - \frac{\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} - \mathrm{CBR}_{o_p})^2}{\sum (|\mathrm{CBR}_{o_i} - \overline{\mathrm{CBR}_o}| + |\mathrm{CBR}_{p_i} - \overline{\mathrm{CBR}_o}|)^2} \tag{15}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{CBR}_{p_i} - \mathrm{CBR}_{o_i}| \tag{16}$$

where n is the number of observations, CBR_{o_i} is the ith observed (measured) value, CBR_{p_i} is the ith predicted value, $\overline{\mathrm{CBR}_o}$ is mean observed value, and $\overline{\mathrm{CBR}_p}$ is the mean predicted value. The range of R^2 is 0–1 and for an ideal model, the value should be close to 1. The I_a ($0 \leq I_a \leq 1$) shows the ratio of mean square error and the prediction error in the system. The value of 1 represents the perfect agreement between the observed and predicted value and 0 represents no agreement (Willmott 1981). RMSE and MAE are widely adopted for assessing the predictive accuracy of the ML models. The MAE represents the average error (equal weightage) over the data set without considering the sign. It means that in MAE, the difference between the observed and predicted values is averaged in a linear manner, while in RMSE, the errors are squared before taking the average (Equation (13)), therefore giving high weightage to the larger errors (Yaseen et al. 2018). For an ideal model, both MAE and RMSE value should be 0. SI represents the ratio of RMSE to the average of observed values. Lower values of SI depict more accuracy and vice-versa (Khorrami et al. 2020)

3. Results and discussion

As mentioned abovethat the main aim of this study is the comprehensive ML-based analysis for modelling the CBR of GRS. Main parameters of all the models are summarised in Table 4. For evaluating the performance of the developed prescient models, namely ANN, LMSR, GPR, ENRR, K-star, M-5, AMT, and RF, the calculated values of all the statistical parameters, namely R^2, RMSE, SI, I_a and MAE are presented in Table 5 for testing (validation) data. The colour intensity coding technique (CICT) has been applied to indicate the strength of each parameter according to its obtained value. It may be noted that this technique has been successfully applied in many previous studies (Nazari et al. 2020, Nguyen et al. 2020). In this way, the parameters which illustrate more

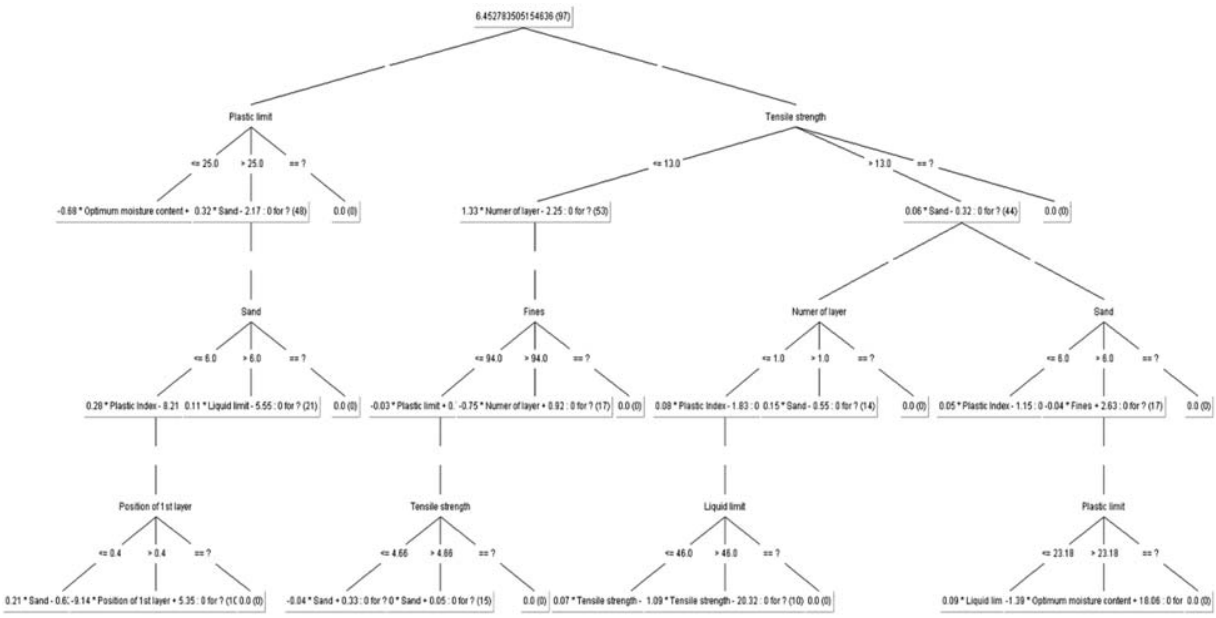


Figure 6. AMT model proposed for estimating the CBR of reinforced soil.

accuracy (higher R^2 and I_a; and lower RMSE, SI and MAE) are given intense colour and vice-versa. For this study, the shades of green colour are used for depicting the predictive accuracy of the developed ML models. The dark green colour cells show more accuracy, and light pale green colour cells depict the lower precision level. Accordingly, each model was scored, and the final ranking for a particular model was obtained by summation of all the partial scores given based on the statistical indices' values.

As Table 5 reports, the LKS and ANN models have shown superior predictive accuracy in estimating the CBR value of reinforced soil. This can be established from the indicated total scores of 40 and 35, respectively, for LKS and ANN models. The calculated values of the statistical indices for ANN, LMSR, GPR, ENRR, LKS, M-5, AMT and RF, such as R^2 (0.944, 0.63, 0.939, 0.807, 0.955, 0.847, 0.927 and 0.932), and MAE (1.27, 5.46, 2.02, 2.19, 1.04, 1.8, 1.30 and 1.29), indicate the best correlation and less absolute error between the actual and simulated values of the CBR for these two models (i.e. LKS and ANN). Also, the SI (0.226, 0.91, 0.51, 0.42, 0.213, 0.44, 0.278 and 0.273) and I_a (0.984, 0.74, 0.907, 0.932, 0.987, 0.948, 0.980 and 0.976) further prove this fact. In concurrence with the same argument, the reliability of AMT and RF trees models can also be established. However, the computed values of MAE (5.46, 2.02, 2.19 and 1.8) and RMSE (10.85, 3.37, and 3.22 and 2.91) for LMSR, GPR, ENRR and M-5 respectively, indicate that the forecasting ability of these models are associated with relatively high bias, in comparison to their counterpart models.

The scholars have argued that the reliability of the ML models should also be assessed using the external validation and/or multi-criteria approach (Gandomi et al. 2013, Naser and Alavi 2020). This gives the realistic evaluation of the model's predictive performance by eradicating/minimising the bias associated with the traditional goodness of fit indices. Therefore, the external model validation criteria by Golbraikh et al. (2003), stabilisation criteria for quantitative structure–activity relationship (QSAR) model by Roy and Roy (2008), and objection function (OBJ) by Gandomi et al. (2013) are also applied to further affirm the accuracy and reliability of the developed data-driven models.

According to Golbraikh et al. (2003), a model must meet the following criteria to be considered reliable. One of the slope regressions lines (k or k') between the observed values (say x_i) and predicted values (y_i) or vice-versa must pass through the origin, and must be close to unity. In terms of the CBR prediction model, it can be written as follows:

$$k = \frac{\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} \times \mathrm{CBR}_{p_i})}{\mathrm{CBR}_{o_i}^2} \tag{17}$$

$$k' = \frac{\sum_{i=1}^{n} (\mathrm{CBR}_{o_i} \times \mathrm{CBR}_{p_i})}{\mathrm{CBR}_{p_i}^2} \tag{18}$$

The value of k or k'must be between 0.85 and 1.15. Additionally, the performance index parameters, that are, m and n

Table 4: Main parameters of the models

Name	Parameters
ANN	Hidden layers = 1; Hidden nodes = 6; Transfer function = Sigmoid; Learning rate = 0.3; Momentum term = 0.2
GPR	Kernel = PuK; $\omega = 1$, $\sigma = 1$
LMSR	$S = 4$, $G = 0$
ENRR	Alpha = 0.001, lambda-seq-threshold = 1.0E-7, number of iteration folds = 10, epsilon = 0.0004
LKS	Global blending parameter (B) = 20
M-5	Minimum no. of instances at leaf nodes (M) = 4.0
AMT	No. of iterations (I) = 100, shrinkage (H) = 1
RF	No. of iterations (I) = 100, max depth (K) = 0, $V = 0.001$

Table 5. Results of colour intensity ranking criteria for all the proposed models in predicting the CBR of reinforced soil

Proposed models	Network results in testing dataset					Ranking the Predicted Models					Total ranking score	Rank
	R^2	RMSE	SI	I_a	MAE	R^2	RMSE	SI	Ia	MAE		
ANN	0.944	1.74	0.226	0.984	1.27	7	7	7	7	7	35	2
LMSR	0.63	10.85	0.91	0.74	5.46	1	1	1	1	1	5	8
GPR	0.939	3.37	0.51	0.907	2.02	6	2	2	2	3	15	6
ENRR	0.807	3.22	0.42	0.932	2.19	2	3	4	3	2	14	7
LKS	0.955	1.52	0.213	0.987	1.04	8	8	8	8	8	40	1
M-5	0.847	2.91	0.44	0.948	1.80	3	4	3	4	4	18	5
AMT	0.927	1.94	0.278	0.98	1.30	4	6	5	6	5	26	4
RF	0.932	1.99	0.273	0.976	1.29	5	5	6	5	6	27	3

should be <0.1 and can be calculated as follows:

$$m = \frac{R^2 - R_o^2}{R^2} \tag{19}$$

$$n = \frac{R^2 - R_o'2}{R^2} \tag{20}$$

whereas R_o^2 and $R_o'2$are the regression coefficients and can be estimated as

$$R_o^2 = 1 - \frac{\sum_{i=1}^{n} \mathrm{CBR}_{o_i}^2 (1-k)^2}{\sum_{i=1}^{n} (\mathrm{CBR}_{p_i} - \overline{\mathrm{CBR}_o})^2} \tag{21}$$

$$R_o'2 = 1 - \frac{\sum_{i=1}^{n} \mathrm{CBR}_{p_i}^2 (1-k')^2}{\sum_{i=1}^{n} (\mathrm{CBR}_{p_i} - \overline{\mathrm{CBR}_p})^2} \tag{22}$$

Ideally, the values of R_o^2or $R_o'2$ should be close to actual R^2, whereas R^2 should be > 0.6. In this way, the model is considered acceptable if it meets all these criteria.

Roy and Roy (2008) established the stabilisation criteria for ensuring the predictability of the developed model. Accordingly, the value of R_m is estimated as stabilisation criterion to measure the reliability of the developed model. Mathematically

$$R_m = R^2 \times \left(1 - \sqrt{|R^2 - R_o^2|}\right) \tag{23}$$

The value of R_m should be > 0.5.

Based on RMSE, R^2 and MAE, the OBJ function contemplates the performance of the model is training and testing data set, simultaneously (Gandomi et al. 2013). The smaller the function value, the better it is and vice-versa. Mathematically

$$\mathrm{OBJ} = \left(\frac{\mathrm{No._{tr}\text{-}No._{ts}}}{\mathrm{No._{tr}} + \mathrm{No._{ts}}}\right) \times \frac{\mathrm{RMSE_{tr}} + \mathrm{MAE_{tr}}}{R_{\mathrm{tr}}^2 + 1} + \left(\frac{2\mathrm{No._{ts}}}{\mathrm{No._{tr}} + \mathrm{No._{ts}}}\right) \times \frac{\mathrm{RMSE_{ts}} + \mathrm{MAE_{ts}}}{R_{\mathrm{tr}}^2 + 1} \tag{24}$$

where subscripts tr and ts represent training and testing data, respectively.

Table 6. Results of external validation and multi-criteria approach used for accessing the accuracy of all the developed ML models

Parameters for accessing the accuracy with respect to external model validation criteria						External validation criteria Golbraikh et al. (2003)			Stabilisation criteria for QSAR model Roy and Roy (2008)		Objective function Gandomi et al. (2013)
ML models	R^2	k	k'	m	n	Condition 1	Condition 2	Condition 3	R_m	R_m >0.5	OBJ
ANN	0.944	0.975	0.997	−0.12	−0.12	✓	✓	✓	0.60	✓	3.40
LMSR	0.630	0.479	1.635	0.24	0.92	✓	x	x	0.27	x	15.43
GPR	0.939	1.262	0.736	0.37	0.17	✓	✓	x	0.38	x	5.38
ENRR	0.807	1.022	0.880	−0.53	−0.45	✓	✓	✓	0.27	x	5.44
LKS	0.955	1.026	0.953	−0.09	−0.09	✓	✓	✓	0.65	✓	2.95
M-5	0.847	1.104	0.839	−0.33	−0.25	✓	✓	✓	0.37	x	4.79
AMT	0.927	1.007	0.957	−0.16	−0.15	✓	✓	✓	0.54	✓	3.52
RF	0.932	1.039	0.925	−0.14	−0.13	✓	✓	✓	0.56	✓	3.56

Note: Conditions 1, 2 and 3 correspond to$R^2 \geq 0.6$; $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$; and $m, n < 0.1$, respectively.

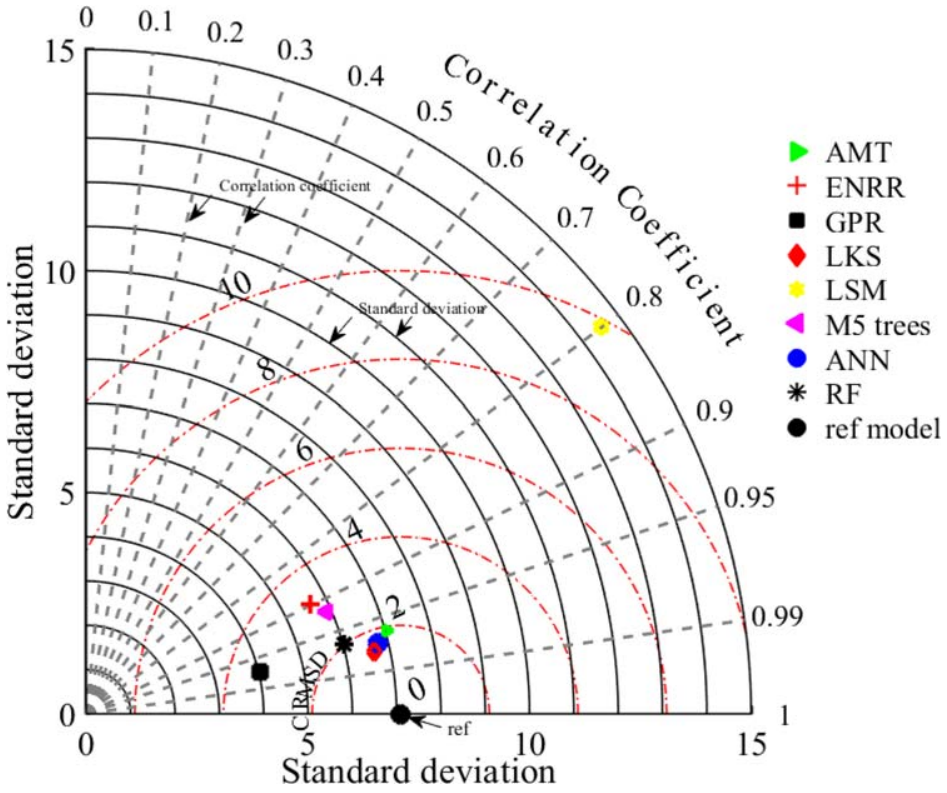


Figure 7. Taylor diagram presenting the visual summary of predictive strength of developed models.

The results of the above-mentioned external validation and multi-criteria approach are summarised in Table 6. From the table, it can be concurred that the ANN, LKS, AMT and RF models have shown excellent prediction ability in calculating the CBR of reinforced soil. This is substantiated by the fact that all these models have met the underlying conditions corresponding to the external validation criteria approach. The ENRR and M-5 models have met all the conditions of Golbraikh et al. (2003) model but failed to meet the model stabilisation condition with R_m values of 0.27 and 0.37, respectively. Based on the values of OBJ function values (3.40, 15.43, 5.38, 5.44, 2.95, 4.79, 3.52 and 3.56), respectively, for ANN, LMS, GPR, ENRR, LKS, M-5 trees, AMT and RF also indicate that the ANN and LKS can be established as the best models for predicting the CBR of reinforced soil. Also, the AMT and RF models performed well by meeting all the criteria, and therefore can be introduced as third- and fourth-best models in the hierarchy. This means that the predictions made by these models are trustworthy and are not a mere coincidence. Additionally, the results of LMS, GPR, ENRR and M-5 models indicate relatively poor performance in comparison to their counterpart models.

Finally, Taylor's diagram originally proposed by Taylor (2001) has been presented in Figure 7. The Taylor diagram presents the visual summary of the predictive power of the data-driven models on a single platform, that is, how closely the actual and simulated responses are related to each other in terms of their correlation and biasness ratio (Taylor 2001, Raja and Shukla 2020). Regarding Figure 7, the solid radial lines (black) represent the standard deviation (SD); thickened dash lines (grey) represent the correlation coefficient (CC); and the dotted radial lines (red) show the centred root mean square deviation (CRMSD) between the simulated (test data) and reference field. The reference model is indicated by the solid black dot with the measured SD of 7.09, CC of unity and zero CRMSD. It can be observed that for LKS and ANN models, the CC, CRMSD and SD are about (0.977, 1.53 and 6.64) and (0.9716, 1.68 and 6.76), respectively. This highlights excellent predictive capability for the developed models followed by AMT model with values of CC, CRMSD and SD of 0.963, 1.92 and 7.02, respectively. For the same parameters, the values for M-5 and RF were (0.921, 2.83 and 5.89) and (0.965, 1.99 and 6.02), respectively. On the contrary, the GPR model has shown little too spatial variability with the SD value of ~4.03, and the LSM model has depicted a large variation in comparison to the observed CBR values with the SD of 14.59. The ENRR model has shown a fair overall performance with the SD value of 5.76; however, the correlation is weak (CC = 0.897), and root mean square error is high (CRMSD = 3.21). Therefore, consistent with the results of the statistical indices and external validation criteria, to this point, it can be established with sufficient trustworthiness, that among all the applied ML models, ANN and LKS models have achieved more accuracy in forecasting the CBR of reinforced soil.

4. Model presentation

For this work, ANN is selected as an appropriate estimator of the CBR values of the GRS, due to its predictive performance

Table 7. Weights and biases of the developed neural network

Weights of input layer – hidden layer, β_{ij}											
1	2	3	4	5	6	7	8	9	10	11	Hidden layer bias θ_j
−0.2536	−0.4607	−0.0521	−0.04	0.2659	−0.2427	−0.4646	1.1754	0.6772	−1.103	0.4061	−0.9167
−0.2914	−0.3918	−0.1618	−0.0108	0.6409	−0.3343	0.0960	0.4698	0.3662	0.3322	0.3849	−0.7143
−0.12738	−0.0038	−0.3035	0.4085	0.6231	−0.3620	0.0739	0.8576	0.4965	0.7954	0.5570	−0.5705
−0.8386	−1.806	0.2339	−0.0798	−1.819	−0.0626	1.2337	1.6369	0.4646	−0.6744	0.4908	−1.1191
−1.4253	−1.7292	−1.0318	1.2547	0.7067	1.8706	−2.8545	−0.7062	−1.0057	0.5092	−0.1882	1.1229
−0.7382	−1.3429	−0.1035	−0.1969	1.1933	−0.9824	0.5970	−0.7195	−0.8305	−0.0584	−0.4887	−0.3609
Weights of hiden-ouput layer, w_{jk}											Output layer bias θ_k
−1.022	−0.322	−0.8921	1.4719	−2.087	−1.234						1.6104

and simplicity. The ANN-based equation is given as follows (Aamir et al. 2020):

$$y = g_n\left(\sum_{j=1}^{h} w_{jk} f_n\left(\theta_j + \sum_{i}^{n} \beta_{ji} X_i\right) + \theta_k\right) \tag{25}$$

where y is the value of output (CBR), g_n is the hidden–output layer transfer function (pureline), f_n is the input-hidden layer transfer function (sigmoid), w_{jk} is the weight connection between the jth node in the hidden layer and single node in the output layer ($k = 1$), β_{ji} is the weight connection between the ith node of the input layer and jth node of the hidden layer, θ_j is the bias of the jth node at the hidden layer and θ_k is the bias at the output layer node. The architecture of the developed single hidden layer neural network is already given in Figure 5, that is, 11 input nodes, 6 hidden nodes and 1 output node. The weights and biases values of the network are reported in Table 7. The values of input parameters should be normalised [−1,1] before feeding to ANN using the following relationship

$$X' = 2\frac{X - X_{\min}}{X_{\max} - X_{\min}} - 1 \tag{26}$$

where $X_{\min}$ is the minimum value of the parameter, and $X_{\max}$ is the maximum value of the parameter, and already given in Table 1.

In order to estimate the CBR of reinforced soil with 11 input parameters, the following relationship is established for the optimum ANN model

$$\mathrm{CBR}'_p = \sum_{j=1}^{11} w_{jk}\mathrm{sig}(\alpha)_j + \theta_k \tag{27}$$

$$\begin{aligned}\alpha_j = {} & \beta_{1j}X'_1 + \beta_{2j}X'_2 + \beta_{3j}X'_3 + \beta_{4j}X'_4 + \beta_{5j}X'_5 + \beta_{6j}X'_6 \\ & + \beta_{7j}X'_7 + \beta_{8j}X'_8 + \beta_{9j}X'_9 + \beta_{10j}X'_{10} + \beta_{11j}X'_{11} \\ & + \theta_j\end{aligned} \tag{28}$$

where CBR'_p is the normalised predicted CBR value [−1,1]; X'_1, X'_2, X'_3, X'_4, X'_5, X'_6, X'_7, X'_8, X'_9, $X'_{10}$and X'_{11} represents the

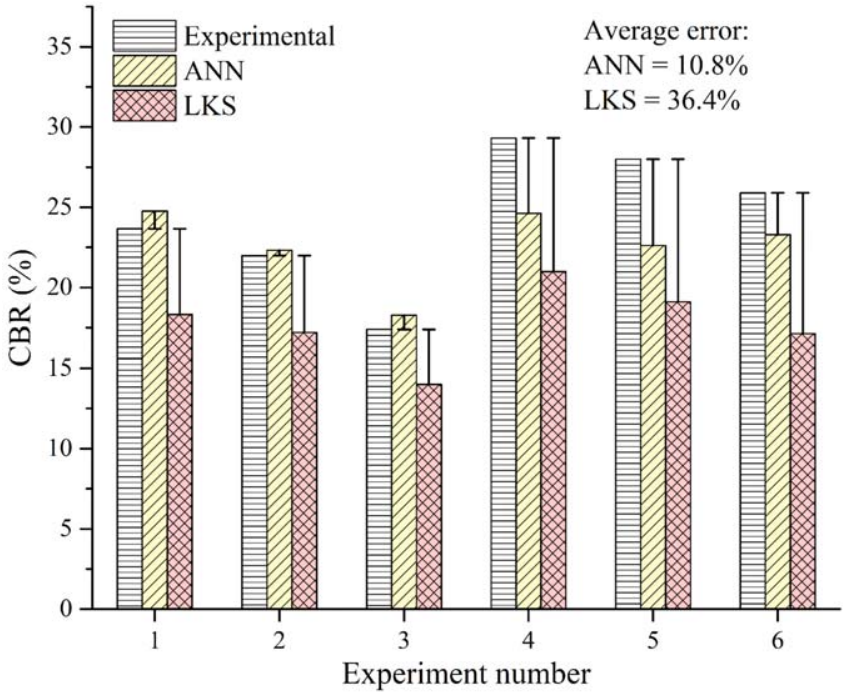


Figure 8. Comparison of the experimental and predicted CBR values for ANN and LKS models.

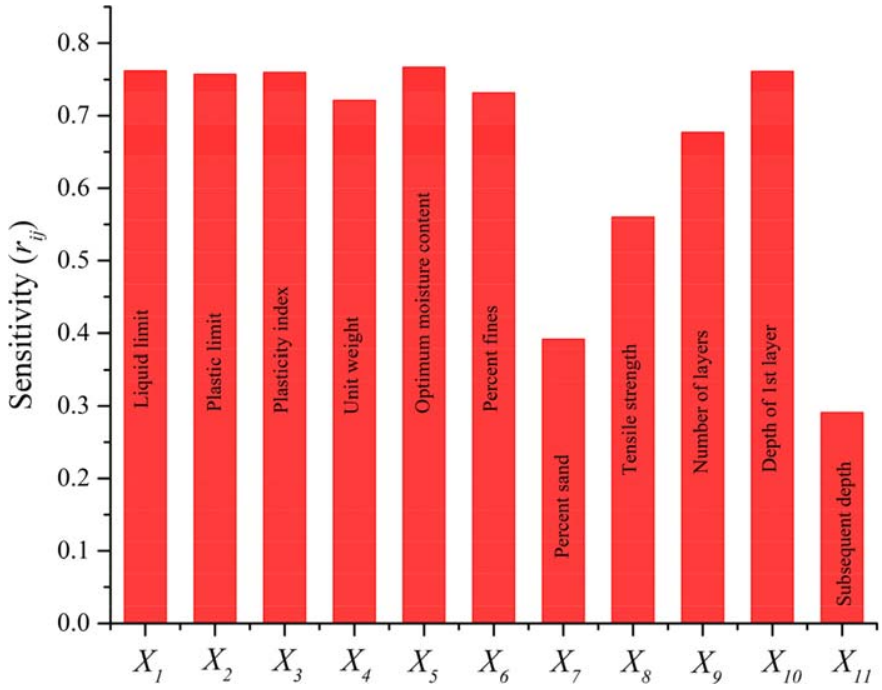


Figure 9. CAM sensitivity analysis for the developed ANN model.

normalised values of all the input parameters. The predicted CBR value should be de-normalised as follows:

$$CBR = (CBR'_p + 1) \times (CBR_{\max} - CBR_{\min})/2 + CBR_{\min} \quad (29)$$

For easy comprehension, the design example is given in the Appendix. It may be noted that the present relation is only calibrated and validated for predicting the soaked CBR of GRSs within the training data range, and therefore, shall not be used for estimating the unsoaked CBR values, and the CBR of unreinforced soil or soil reinforced with other types of reinforcements such as fibres, tire chips, metals, etc.

4.2. Independent validation

LKS and ANN models have shown excellent predictive performance for the data set utilised to establish the ML models. However, in order to establish the supremacy, consistency and reliability of any ML model, its predictive strength should be checked against the entirely new data. Therefore, an experimental study was conducted to establish a new data set. For this, six soaked CBR tests (see Table 3) were conducted and their experimental values were compared with the simulated values obtained from ANN and LKS models.

The vis-à-vis comparison of the experimental and predicted CBR values for ANN and LKS is presented in Figure 8. It can be observed that the ANN predicted the CBR values much closer to the actual values in comparison to LKS with the average absolute error of 10.8% and 36.4%, respectively. Therefore, it is admissible that the ANN model has outperformed its competitive models in predicting the CBR of GRS. Moreover, another main advantage of the ANN network in comparison to other efficient models (say LKS or AMT) is that it can be translated into a trackable functional relationship, and, therefore, can easily be executed without the need for any expensive computer-based program. Moreover, the ANN model can be updated to acquire the better results by presenting more training examples when the new data becomes available.

5. Sensitivity analysis

A sensitivity analysis has also been carried out to determine the relative importance of each parameter affecting the CBR of reinforced soil. It helps in finding the strength of the existing correlation between the input and output dimensions. For this study, the cosine amplitude method (CAM) is used to establish the strength of input parameters with the CBR of reinforced soil.

For CAM, let n data samples in the same region (say X-space), then the data array X can be written as follows (Hasanzadehshooiili et al. 2012):

$$X = \{x_1, x_2, x_3 \ldots, x_m\} \quad (30)$$

Each elementx_i of data array X is the vector (length m) in Equation (30), and is defined as:

$$x_i = \{x_{i1}, x_{i2}, x_{i3} \ldots, x_{im}\} \quad (31)$$

In this way, the correlation strength r_{ij} between the data points $x_i$and$x_j$can be estimated by Equation (32) (Ghorbani et al. 2020c)

$$r_{ij} = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}}, \quad 0 \le r_{ij} \le 1 \quad (32)$$

The result of the CAM sensitivity analysis carried for the ANN model is illustrated in Figure 9. The relative strengths of all the

input parameters (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, $X_{10}$and-X_{11}) with CBR of reinforced soil are 0.761, 0.756, 0.759, 0.721, 0.766, 0.731, 0.40, 0.560, 0.676, 0.761 and 0.30, respectively. This imply that all the input parameters except for the subsequent reinforcement depth (X_{11}) play a significant role in determining the CBR of reinforced soil with r_{ij} approximately ranging from 0.4 to 0.8. Moreover, X_1 (LL), X_5 (OMC) and depth of the first reinforcement layer (X_{10}) has achieved the highest correlation in predicting the CBR of reinforced soil.

6. Conclusions and future outlook

This work presents the comprehensive and detailed comparison of eight data-driven ML-based models, namely ANN, LMSR, GPR, ENRR, LKS, M-5 model trees, AMT and RF for predicting the CBR of subgrade soil reinforced with geosynthetic layers. For this, the pertinent data set was retrieved from previously published scientific studies. Each sample consisted of 11 input variables such as LL, PL, PI, maximum dry unit weight of soil, moisture content, percentage fines, percentage sand, tensile strength of geosynthetic reinforcement, number of reinforcement layers, position of the first reinforcement layer and position of the subsequent reinforcement layers, and one output variable, that is, CBR. The acquired data set was randomly divided into training (i.e. 60% of total data) and 40% testing (i.e. 40% of total data) to calibrate and validate the performance of all the data-driven tools. The performance of all the models was accessed using the five statistical indices, which are, coefficient of determination (R^2), root means square error (RMSE), scatter index (SI), index of agreement (I_a) and mean absolute error (MAE). Based on the results of these indices, a colour intensity coded ranking model was developed. Moreover, the predictive strength of the tools mentioned supra was also appraised using the external validation criteria and multi-criteria approach. The most appropriate models were also tested against the entirely independent validation data obtained by conducting the new CBR tests. Finally, the sensitivity analysis was performed to assess the effect of the input parameters on the CBR value. Based on the acquired results, the following conclusions can be drawn.

- In the CBR approximation, the values of R^2 (0.944, 0.63, 0.939, 0.807, 0.955, 0.847, 0.927 and 0.932), RMSE (1.74, 10.85, 3.37, 3.22, 1.52, 2.91, 1.94 and 2.03), SI (0.226, 0.91, 0.51, 0.42, 0.213, 0.44, 0.278 and 0.273), I_a (0.984, 0.74, 0.907, 0.932, 0.987, 0.948, 0.98 and 0.976), and MAE (1.27, 5.46, 2.02, 2.19, 1.04, 1.80, 1.30 and 1.29) were, respectively, calculated for ANN, LMSR, GPR, ENRR, LKS, M-5 trees, AMT and RF models. Based on the values of these statistical indices, a total ranking score was obtained for all the modelling techniques. The results have shown excellent prediction ability of LKS and ANN with a total score of 40 and 35, respectively. The ranking scores of other models such as LMSR, GPR, ENRR, M-5 trees, AMT and RF were, respectively, 5, 15, 14, 18, 26 and 27.
- Among all the models, the LMSR model has obtained the poorest approximation of the CBR and its insufficiency was depicted by the ranking score (total score = 5) obtained based on the above-mentioned assessment criteria.
- Based on the results of the external validation technique, and multi-criteria assessment approach, the ANN, LKS, AMT and RF models have achieved good prediction ability and model stability in forecasting the CBR of reinforced soil. However, the LKS and ANN have shown superior performance in comparison to their counterpart models with the OBJ function value of 2.95 and 3.40, respectively. Also, among these two models, the latter has predicted the new experimental data (independent data) with more accuracy. Additionally, for this work, the developed ANN model was also converted into trackable mathematical relationship for easy hand or spreadsheet calculations.
- The strength (r_{ij}) of each input variable with respect to the output (CBR) was evaluated by sensitivity analysis. The results revealed that all the parameters have played an important role in determining the CBR of reinforced soil. However, OMC LL and position of the first reinforcement layer with r_{ij} values of 0.77, 0.762 and 0.761, respectively, are the most influential parameters.

In this study, to maximise the modelling efficiency and ease of use, default settings are used in WEKA for most of the individual models. Therefore, future studies should focus on how the parameters in the models can be optimised automatically. This limitation can be explored in future by combining some optimisation scheme with the model network. Most recently, the use of evolutionary algorithm based on metaheuristics (e.g. shuffled frog algorithm, grey wolf optimiser, ant lion optimiser, elephant herd optimisation, etc.) has shown good ability to improve the prediction ability of neural networks by optimising its weights and biases. The future work with the focus on such optimisation techniques can prove to be a useful idea. Moreover, the ensemble learning techniques in which the learning power of the multiple ML models are combined to predict the response of the system might also be applied in the future.

Disclosure statement

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

Funding

The authors would like to thank Edith Cowan University (ECU), Australia and Higher Education Commision, Pakistan, for supporting this research jointly.

ORCID

Muhammad Nouman Amjad Raja http://orcid.org/0000-0001-7463-0601
Sanjay Kumar Shukla http://orcid.org/0000-0002-4685-5560
Muhammad Umer Arif Khan http://orcid.org/0000-0002-3077-2875

References

Aamir, M., et al., 2020. Performance analysis of multi-spindle drilling of Al2024 with TiN and TiCN coated drills using experimental and artificial neural networks technique. *Applied Sciences (Switzerland)*, 10 (23), 1–20.

Abu-Farsakh, M., et al., 2016. Performance of geosynthetic reinforced/stabilized paved roads built over soft soil under cyclic plate loads. *Geotextiles and Geomembranes*, 44 (6), 845–853.

Alawi, M.H., and Rajab, M.I, 2013. Prediction of California bearing ratio of subbase layer using multiple linear regression models. *Road Materials and Pavement Design*, 14 (1), 211–219.

Arulrajah, A., et al., 2013. Resilient moduli response of recycled construction and demolition materials in pavement subbase applications. *Journal of Materials in Civil Engineering*, 25 (12), 1920–1928.

AS 1289.6.1.1, 2014. Methods of testing soils for engineering purposes Soil strength and consolidation tests – Determination of the California Bearing Ratio of a soil – Standard laboratory method for a remoulded specimen.

ASTM, D., 2016. *Standard test method for California bearing ratio (CBR) of laboratory-compacted*. West Conshohocken, PA: ASTM International, 1–9.

Bishop, Christopher, 2006. *Pattern Recognition and machine learning*. New York: Springer-Verlag New York.

Breiman, L, 2001. Random forests. *Machine Learning*, 45 (1), 5–32.

Carlos, D.M., Pinho-Lopes, M., and Lopes, M.L, 2016. Effect of geosynthetic reinforcement inclusion on the strength parameters and bearing ratio of a fine soil. *Procedia Engineering*, 143, 34–41.

Chen, Q., et al., 2018. Performance evaluation of full-scale geosynthetic reinforced flexible pavement. *Geosynthetics International*, 25 (1), 26–36.

Choudhary, A., et al., 2012. Improvement in CBR of the expansive soil subgrades with a single reinforcement layer. *Proceedings of Indian Geotechnical Conference*, 289–292.

Choudhary, A.K., Jha, J.N., and Gill, K.S., 2011. Improvement in CBR values of expansive soil-subgrades using geosynthetics. Punjab: Guru Nanak Dev Engineering College Ludhiana, 155.

Cleary, J.G., and Trigg, L.E, 1995. K*: an instance-based learner using an entropic distance measure. *Machine Learning Proceedings*, 1995, 108–114.

Cuelho, E. V., Perkins, S.W., and Ganeshan, S.K., 2005. Determining geosynthetic material properties pertinent to reinforced pavement design. *Geotechnical Special Publication*, 130–142, 241–252.

Daneshvar, D. and Behnood, A, 2020. Estimation of the dynamic modulus of asphalt concretes using random forests algorithm. *International Journal of Pavement Engineering*.1-11. https://doi.org/10.1080/10298436.2020.1741587

de Souza, W.M., Ribeiro, A.J.A., and da Silva, C.A.U., 2020. Use of ANN and visual-manual classification for prediction of soil properties for paving purposes. *International Journal of Pavement Engineering*.1- 9. https://doi.org/10.1080/10298436.2020.1807546

Duncan-Williams, E., and Attoh-Okine, N.O, 2008. Effect of geogrid in granular base strength – An experimental investigation. *Construction and Building Materials*, 22 (11), 2180–2184.

Erzin, Y., and Turkoz, D, 2016. Use of neural networks for the prediction of the CBR value of some Aegean sands. *Neural Computing and Applications*, 27 (5), 1415–1426.

Frank, E., Mayo, M., and Kramer, S., 2015. Alternating model trees. *In*: *Proceedings of the ACM Symposium on Applied Computing*, 871–878.

Gandomi, A.H., Yun, G.J., and Alavi, A.H, 2013. An evolutionary approach for modeling of shear strength of RC deep beams. *Materials and Structures/Materiaux et Constructions*, 46 (12), 2109–2119.

Gao, W., et al., 2019. Comprehensive preference learning and feature validity for designing energy-efficient residential buildings using machine learning paradigms. *Applied Soft Computing Journal*, 84. https://doi.org/10.1016/j.asoc.2019.105748

Ghorbani, B., et al., 2020a. Experimental investigation and modelling the deformation properties of demolition wastes subjected to freeze–thaw cycles using ANN and SVR. *Construction and Building Materials*, 258. https://doi.org/10.1016/j.conbuildmat.2020.119688

Ghorbani, B., et al., 2020b. Experimental and ANN analysis of temperature effects on the permanent deformation properties of demolition wastes. *Transportation Geotechnics*, 24. https://doi.org/10.1016/j.trgeo.2020.100365

Ghorbani, B., et al., 2020c. Development of genetic-based models for predicting the resilient modulus of cohesive pavement subgrade soils. *Soils Found*, 60, 398–412. doi:10.1016/j.sandf.2020.02.010.

Ghorbani, B., et al., 2021. Shakedown analysis of PET blends with demolition waste as pavement base/subbase materials using experimental and neural network methods. *Transportation Geotechnics*, 27. https://doi.org/10.1016/j.trgeo.2020.100481

Ghosh Mondal, P. and Kuna, K., 2020. An automated technique for characterising foamed bitumen using ultrasonic sensor system. *International Journal of Pavement Engineering*. https://doi.org/10.1080/10298436.2020.1849689

Golbraikh, A., et al., 2003. Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design*, 17 (2–4), 241–253.

González Farias, I., Araujo, W., and Ruiz, G., 2018. Prediction of California bearing ratio from index properties of soils using parametric and non-parametric models. *Geotechnical and Geological Engineering*, 36 (6), 3485–3498.

Han, C., et al., 2020. Intelligent decision model of road maintenance based on improved weight random forest algorithm. *International Journal of Pavement Engineering*. https://doi.org/10.1080/10298436.2020.1784418

Han, J., and Moraga, C, 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 930, 195–201.

Hasanzadehshooiili, H., Lakirouhani, A., and Medzvieckas, J, 2012. Superiority of artificial neural networks over statistical methods in prediction of the optimal length of rock bolts. *Journal of Civil Engineering and Management*, 18 (5), 655–661.

Ho, T.K, 1995. Random decision forests. *In*: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 278–282.

Khorrami, R., Derakhshani, A., and Moayedi, H, 2020. New explicit formulation for shallow foundations' ultimate bearing capacity rested on granular soil using M-5' model tree. *Measurement*, 108032.

Kuity, A., and Roy, T.K, 2013. Utilization of geogrid mesh for improving the soft subgrade layer with waste material mix compositions. *Procedia – Social and Behavioral Sciences*, 104, 255–263.

Massart, D.L., et al., 1986. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, 187 (C), 171–179.

Mittal, A., and Shukla, S, 2018. Influence of geotextile and geogrid reinforcement on strength behaviour of soft silty soil. *Applied Mechanics and Materials*, 877, 264–269.

Mittal, A., and Shukla, S, 2019. Strength improvement of poor subgrade soil reinforced with polyester biaxial geogrid. *Jordan Journal of Civil Engineering*, 13 (2), 214–225.

Miura, N., et al., 1990. Polymer grid reinforced pavement on soft clay grounds. *Geotextiles and Geomembranes*, 9 (1), 99–123.

Moayedi, H., et al., 2020. Feature validity during machine learning paradigms for predicting biodiesel purity. *Fuel*, 262. https://doi.org/10.1016/j.fuel.2019.116498

Moayedi, H., and Hayati, S, 2018. Applicability of a CPT-based neural network solution in predicting load-settlement responses of bored pile. *International Journal of Geomechanics*, 18 (6), 06018009.

Naeini, S.A., and Ziaie-Moayed, R, 2009. Effect of plasticity index and reinforcement on the CBR value of soft clay. *International Journal of Civil Engineering*, 7 (2), 124–130.

Nagaraju, T.V., Prasad, C.D., and Raju, M.J, 2020. Prediction of California bearing ratio using particle swarm optimization. *Advances in Intelligent Systems and Computing*, 1048, 795–803.

Nair, A.M., and Latha, G.M, 2011. Bearing resistance of reinforced soil-aggregate systems. *Proceedings of the Institution of Civil Engineers: Ground Improvement*, 164 (2), 83–95.

Naser, M., and Alavi, A., 2020. Insights into performance fitness and error metrics for machine learning, 25.

Nazari, S., et al., 2020. A proper model to predict energy efficiency, exergy efficiency, and water productivity of a solar still via optimized neural

network. *Journal of Cleaner Production*, 277. https://doi.org/10.1016/j.jclepro.2020.123232
Nazemi, M., and Heidaripanah, A, 2016. Support vector machine to predict the indirect tensile strength of foamed bitumen-stabilised base course materials. *Road Materials and Pavement Design*, 17 (3), 768–778.
Negi, M.S. and Singh, S.K, 2019. Experimental and numerical studies on geotextile reinforced subgrade soil. *International Journal of Geotechnical Engineering*. 1-12. https://doi.org/10.1080/19386362.2019.1684654
Nguyen, H., et al., 2020. Optimizing ANN models with PSO for predicting short building seismic response. *Engineering with Computers*, 36 (3), 823–837.
Ogutu, J.O., Schulz-Streeck, T., and Piepho, H.P, 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6 (Suppl. 2). No. 2, pp. 1-6
Olowosulu, A.T., et al., 2020. Investigating surface condition classification of flexible road pavement using data mining techniques. *International Journal of Pavement Engineering*. 1-12. https://doi.org/10.1080/10298436.2020.1847285
Oztoprak, S., and Bolton, M.D, 2013. Stiffness of sands through a laboratory test database. *Geotechnique*, 63 (1), 54–70.
Perkins, S.W, 1999. Mechanical response of geosynthetic-reinforced flexible pavements. *Geosynthetics International*, 6 (5), 347–382.
Quinlan, J.R., 1992. Learning with continuous classes. *In*: *Australian Joint Conference on Artificial Intelligence*, 343–348.
Raja, M.N.A. and Shukla, S.K, 2020. An extreme learning machine model for geosynthetic-reinforced sandy soil foundations. *Proceedings of the Institution of Civil Engineers – Geotechnical Engineering*, 1–42. https://doi.org/10.1680/jgeen.19.00297
Raja, M.N.A. and Shukla, S.K, 2021. Multivariate adaptive regression splines model for reinforced soil foundations. *Geosynthetics International*, 1–23. https://doi.org/10.1680/jgein.20.00049
Rajesh, U., Sajja, S., and Chakravarthi, V.K, 2016. Studies on engineering performance of geogrid reinforced soft subgrade. *Transportation Research Procedia*, 17, 164–173.
Rasmussen, C.E., 2006. *Gaussian Processes in machine learning*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
Rousseeuw, P.J, 1984. Least median of squares regression. *Journal of the American Statistical Association*, 79 (388), 871–880.
Roy, P.P., and Roy, K, 2008. On some aspects of variable selection for partial least squares regression models. *QSAR and Combinatorial Science*, 27 (3), 302–313.
Seeger, M, 2004. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14 (2), 69–106.
Shukla, S.K, 2002. *Geosynthetics and their Applications*. Geosynthetics and their Applications. Thomas Telford.
Singh, M., Trivedi, A., and Shukla, S.K, 2019. Strength enhancement of the subgrade soil of unpaved road with geosynthetic reinforcement layers. *Transportation Geotechnics*, 19, 54–60.
Singh, M., Trivedi, A., and Shukla, S.K., 2020. Fuzzy-based model for predicting strength of geogrid-reinforced subgrade soil with optimal depth of geogrid reinforcement. *Transportation Infrastructure Geotechnology*.
Soleimanbeigi, A., and Hataf, N, 2006. Prediction of settlement of shallow foundations on reinforced soils using neural networks. *Geosynthetics International*, 13 (5), 218–218.
Suthar, M, 2020. Applying several machine learning approaches for prediction of unconfined compressive strength of stabilized pond ashes. *Neural Computing and Applications*, 32 (13), 9019–9028.
Taskiran, T, 2010. Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. *Advances in Engineering Software*, 41 (6), 886–892.
Taylor, K.E, 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research Atmospheres*, 106 (D7), 7183–7192.
Tenpe, A.R., and Patel, A, 2020. Utilization of support vector models and gene expression programming for soil strength modeling. *Arabian Journal for Science and Engineering*, 45 (5), 4301–4319.
Ullah, S., Tanyu, B.F., and Zainab, B, 2020. Development of an artificial neural network (ANN)-based model to predict permanent deformation of base course containing reclaimed asphalt pavement (RAP). *Road Materials and Pavement Design*, 1–19.
Üstün, B., Melssen, W.J., and Buydens, L.M.C, 2006. Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81 (1), 29–40.
Vinod, P., and Minu, M, 2010. Use of coir geotextiles in unpaved road construction. *Geosynthetics International*, 17 (4), 220–227.
Wang, Y., and Witten, I.H., 1997. Inducing model trees for continuous classes. *In*: *European Conference on Machine Learning (ECML)*, 1–10.
Webb, G.I., et al., 2011. Lazy learning. *Encyclopedia of Machine Learning*, 571–572.
Willmott, C.J, 1981. On the validation of models. *Physical Geography*, 2 (2), 184–194.
Yaseen, Z.M., et al., 2018. Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. *Advances in Engineering Software*, 115, 112–125.
Yildirim, B., and Gunaydin, O, 2011. Estimation of California bearing ratio by using soft computing systems. *Expert Systems with Applications*, 38 (5), 6381–6391.
Youd, T., 1973. Factors controlling maximum and minimum densities of sands. In evaluation of relative density and its role in geotechnical projects involving cohesionless soils. *ASTM International*.
Zhang, Z., et al., 2019. Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression. *Applied Energy*, 247, 270–284.
Zheng, J., and Hryciw, R.D, 2016. Index void ratios of sands from their intrinsic properties. *Journal of Geotechnical and Geoenvironmental Engineering*, 142 (12), 06016019.
Zou, H., and Hastie, T, 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2), 301–320.

Appendix. Design numerical example

Find the CBR of a geosynthetic-reinforced subgrade soil with the following characteristics:

LL (X_1) = 35, plastic limit (X_2) = 25, PI (X_3) = 10, dry unit weight of soil (X_4) = 18.3 kN/m^3, moisture content (X_5) = 11.9, percentage fines (X_6) = 67%, percentage sand (X_7) = 32%, tensile strength of geosynthetic reinforcement (X_8) = 35.1 kN/m, number of reinforcement layers (X_9) = 2, position of the first reinforcement layer (X_{10}) = 0.4 and position of subsequent reinforcement layers (X_{11}) = 0.2.

Solution:

Step 1:

Normalise each input using Equation (26):

$$X'_1, X'_2, X'_3, X'_4, X'_5, X'_6, X'_7, X'_8, X'_9, X'_{10}, X'_{11} = \{-0.01408, 0.2821, -0.375, 0.3964, -0.7238, 0.2124, -0.2195, 0.6487, -0.333, -0.2647, -0.3939\}$$

Step 2:

Estimate normalised CBR utilising Equation (27):

$$\mathrm{CBR}'_p = \sum_{j=1}^{11} w_{jk}\mathrm{sig}(\alpha)_j + \theta_k$$

Compute $\alpha_j$from Equation (28). It may be noted that weights and biases values are given in Table 7.

$$\begin{aligned}\alpha_1 &= (-0.2536 \times -0.01408) + (-0.4607 \times 0.2821)\\ &+ (-0.0521 \times -0.375) + (-0.04 \times 0.3964) + (0.2659 \times -0.7238)\\ &+ (-0.2427 \times 0.2124) + (-0.4646 \times -0.2195) + (1.1754 \times 0.6487)\\ &+ (0.6772 \times -0.333) + (-1.103 \times -0.2647)\\ &+ (0.4061 \times -0.3939) - 0.9167\end{aligned}$$

$$\alpha_1 = -0.5126$$

Similarly

$$\{\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6\} = \{-1.377, -0.8773, 0.1887, 1.868, -1.953\}$$

$$\begin{aligned}\mathrm{CBR}'_p = \sum_{j=1}^{11} (-1.022 \times \frac{1}{1+e^{-(-0.5126)}} + \ldots, -1.234\\ \times \frac{1}{1+e^{-(-1.9537)}}) + 1.6104 = -0.2551\end{aligned}$$

Step 3:

De-normalise using Equation (29):

$$\mathrm{CBR} = (-0.2551 + 1) \times (34 - 1.3)/2 + 1.3 = 13.47\%$$

sustainability

MDPI

Article

An Intelligent Approach to Active and Reactive Power Control in a Grid-Connected Solar Photovoltaic System

Ibrahim Alsaidan [1], Priyanka Chaudhary [2,3,*], Muhannad Alaraj [1] and Mohammad Rizwan [1,2]

[1] Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah 52571, Qassim, Saudi Arabia; Alsaidan@qu.edu.sa (I.A.); Muhannad@qu.edu.sa (M.A.); MR.Khan@qu.edu.sa (M.R.)
[2] Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India
[3] Department of Electrical Engineering, SET, Noida International University, Noida 203201, India
* Correspondence: priyankach.iilm@gmail.com

Abstract: The increasing demand of electrical energy and environmental concerns are invigorating the use of renewable energy resources for power generation. Renewable energy resources can provide an attractive solution for present and future energy requirements. In this scenario, solar photovoltaic systems are becoming prominent and sustainable solutions with numerous advantages. However, the utilization of solar photovoltaic systems in distribution generation makes it mandatory to deploy efficient and organized control measures for integrating solar photovoltaic plants with the grid. In this paper, the control of grid-tied solar photovoltaic systems using a Kalman filter-based generalized neural network is presented with a variable step size perturb and observe-based maximum power point tracking controller to extract the maximum power from a solar photovoltaic plant. The presented system provides power-quality enhancement and supports a three-phase AC grid. The proposed approach extracts the load currents' primary components for efficient harmonics elimination, synchronizes the system with the grid and provides a fast response during rapidly changing conditions. The results of the proposed control technique are also compared with the artificial neural network-based control technique for validation purposes. The proposed algorithm is found more suitable for using a smaller number of unknown weights and training patterns with reduced computational time.

Keywords: renewable energy resources; grid integrated solar PV systems; sustainable power generation; maximum power point tracking; grid reliability and voltage source converter

check for updates

Citation: Alsaidan, I.; Chaudhary, P.; Alaraj, M.; Rizwan, M. An Intelligent Approach to Active and Reactive Power Control in a Grid-Connected Solar Photovoltaic System. *Sustainability* **2021**, *13*, 4219. https://doi.org/10.3390/su13084219

Academic Editors: M. Dolores Esteban, José-Santos López-Gutiérrez and Vicente Negro

Received: 6 March 2021
Accepted: 30 March 2021
Published: 10 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The increasing concern regarding greenhouse gas emissions and the depleting nature of conventional fuels has invigorated the use of renewable energy resources as an alternative and sustainable solutions for the power sector. Solar, wind, biomass and small hydro power are the main renewable energy-based resources which can fulfill the future energy requirements. The solar photovoltaic system (SPV) is more encouraged due to various advantages such as abundance of availability, less environmental pollution, reduced cost and many others. The progress of an SPV system in the existing power system is witnessed for its progress [1]. The integration of an SPV system into the grid causes positive effects, i.e., generating more power, along with some negative effects, i.e., the violation of voltage limitations at common coupling, frequencies disruption and grid stabilization problems, etc. There are set guidelines, codes and standards for grid-connected SPV systems. These standards include IEEE 1547, IEC 61727 and VDE-AR-N4105 [2]. The use of these standards is encouraged and imposed to maintain the stability and power quality related to the grid. The large-scale exploitation of distributed generation make it mandatory to deploy efficiency and organized control measures for integrating and measuring problems. The control algorithms are helpful for the accommodation and facilitation of the integration of an SPV in the distribution grid. Single-stage and two-stage systems have been used for

the integration of SPV systems into the grid, as shown in previous works. It has also been shown by various researchers that for a three-phase system, a single-stage configuration is more advantageous [3]. However, conventional single-stage SPV systems suffer from the drawback of converters being idle during the unavailability of power from the SPV system. To overcome the abovementioned problems, the proposed system is designed in such a manner so that it operates as a distribution static compensator (DSTATCOM) while SPV power is not available and restores the operation after recovering the SPV power generation.

The electrical power distribution system has been designed to distribute active power pertaining a smaller number of harmonics to the consumers. DSTATCOM integrated as a shunt compensator with a distribution system effectively controls harmonics, fulfils the demand of reactive power of consumers and improves voltage regulation with balanced and unbalanced loads [4]. The researchers have reported various configurations of DSTATCOM in the literature [5–8]. For efficiently integrating distributed a generation system that compensates for the reactive power and harmonics for utility grids at common coupling points, several converter topologies [9–11] and control strategies [12–18] were reported in the previous work. Additionally, Jain et al. [19] presented a grid-integrated solar energy conversion setup with a movable direct current link voltage to vary the voltages at common pairing points with two-stage circuitry methodology and an associated double-frequency second-order generalized integrator (DFSOGI)-dependent control strategy for controlling the multi-functional voltage source converter (VSC) in abrupt load modification at common connecting points.

Varma et al. [19] have proposed a solar photovoltaic static compensator for improving the power transfer capabilities of the system to transmit real power to utilities with an existing converter system. Humid et al. [20] proposed a strategy that depends on a power conditioning unit located parallel to the plant that works in feed-forward mode to compensate the distorted current photovoltaic output to decrease the harmonic of current from a PV system. Kannan et al. [21] presented icos Φ control technique for a distribution static compensator that gives uninterrupted harmonic degradation, compensates re-active power and compensates the loads along with comparing performances of fuzzy-logic controllers to ordinary proportional integral (PI) controllers. Mishra et al. [22] implemented an optimistic control technique which optimizes the proportional integral's coefficients and photovoltaic fed distribution static compensator's filtration parameters.

To enhance the performance of grid the integrated SPV plant, various control techniques were adopted by the researchers. There are some performance indicators such as less computational time, high accuracy and less complexity, etc. A huge number of techniques depending upon artificial neural networks (ANN) have been widely adopted by a large number of journalists for distribution static compensators implementing dynamic loading [23–27], whereas photovoltaic fed distribution static compensator systems have been adopted by very few. Singh et al. [28] presented a grid interfaced SPV generation plant implementing a neural network-based control technique that utilizes the least mean square (LMS) algorithm, termed adaline (adaptive linear element), for estimating the reference source current. Artificial neural network-based approaches are complex and need large computational times. Further, generalized neural network (GNN)-based control strategies can be used to overcome the abovementioned issues. Applications of GNN are widely available in load forecasting, solar irradiance/energy forecasting, load frequency control in power systems, power system stabilizers, electrical machines and control system-related problems [29–36]. There are few works are reported in the literature related to the application of GNN as a control strategy for grid-tied SPV systems. GNN-based models reduce the training time as well as improve the performance of the system. Considering this, a GNN-based control strategy has been developed for a PV DSTATCOM system in the present work.

The converting capability of a solar photovoltaic plant is comparatively small; thus, to increase the capability of solar photovoltaic plants, it is essential to keep a track of the maximum

power point. Maximum Power Point Tracking (MPPT) is complicated because of the non-linear character of SPV plants due to changing meteorological factors, i.e., solar irradiance, ambient temperature, wind velocity and relative humidity, etc. A lot work has been reported in the literature related to MPPT, including techniques such as perturb and observe, incremental conductance, constant voltage, open-circuit voltage, short-circuit current, extremum seeking control and hybrid, etc. Further, there are a number of intelligent techniques also available for MPPT such as artificial neural networks (ANN), fuzzy logic, genetic algorithms, etc. [37–39]. A changeable step size perturb and observe (P&O) MPPT technique is utilized in this research to track the MPP of a solar photovoltaic plant.

Further, to improve the performance of the system, an extended Kalman filter (EKF) [40–42]-based method is adopted to estimate and update the weights of the GNN model. The main contributions of this work are:

- Active power feeding to the connected loads and grid with mitigation of power quality issues. A generalized neural network (GNN)-based approach plays the role of primary control strategy and decides the switching pattern of the voltage source converter (VSC).
- Further, the performance of the proposed algorithm has been improved with the help of EKF for GNN weight estimations.
- The performance of the proposed setup is validated using simulation results implemented in the MATLAB/ Simulink platform.
- The developed system obtains acceptable limits of harmonics in utility currents and voltage fluctuations according to the IEEE-519 and IEEE-1547 standards.

The developed system functions very well with an EKF GNN-based approach and gives a very fast response. Moreover, single-stage topology is able to reduce the losses in semiconductor devices and increase the overall efficiency. The proposed control approach performs with more flexibility in training the network under dynamic conditions. The proposed technique possesses various advantages such as its speed, using a single layer, less mathematical calculations and easy implementation on hardware.

The organization of paper is as follows: Section 2 provides details about the system configuration, followed by Section 3 which explains the developed control technique. The simulation and hardware results are provided in Section 4. The conclusion is given in Section 5, followed by acknowledgements and references.

2. System Description

The proposed system deploys a single-stage circuit topology and consists of a solar PV array, voltage source converter (VSC), ripple filter, loads and a three-phase utility grid, as shown in Figure 1. A solar PV array is a combination of several PV modules in a series and parallel according to the requirements. A PV cell serves as the basic unit for a solar PV module. A single diode solar PV cell is considered for this system, and mathematical modeling is given [43]. The developed system has been implemented with a minimum number of sensors with a variable step size perturb and observe (P&O) MPPT algorithm to achieve maximum power point operation of the solar PV system for different meteorological parameters. An extended Kalman filter (EKF)-based GNN methodology is used for controlling the VSC switching to synchronize the system with the grid and the main task of separating the weights of the primary real and reactive parts of three-phase load currents to obtain reference grid currents. The developed system design specifications are given below:

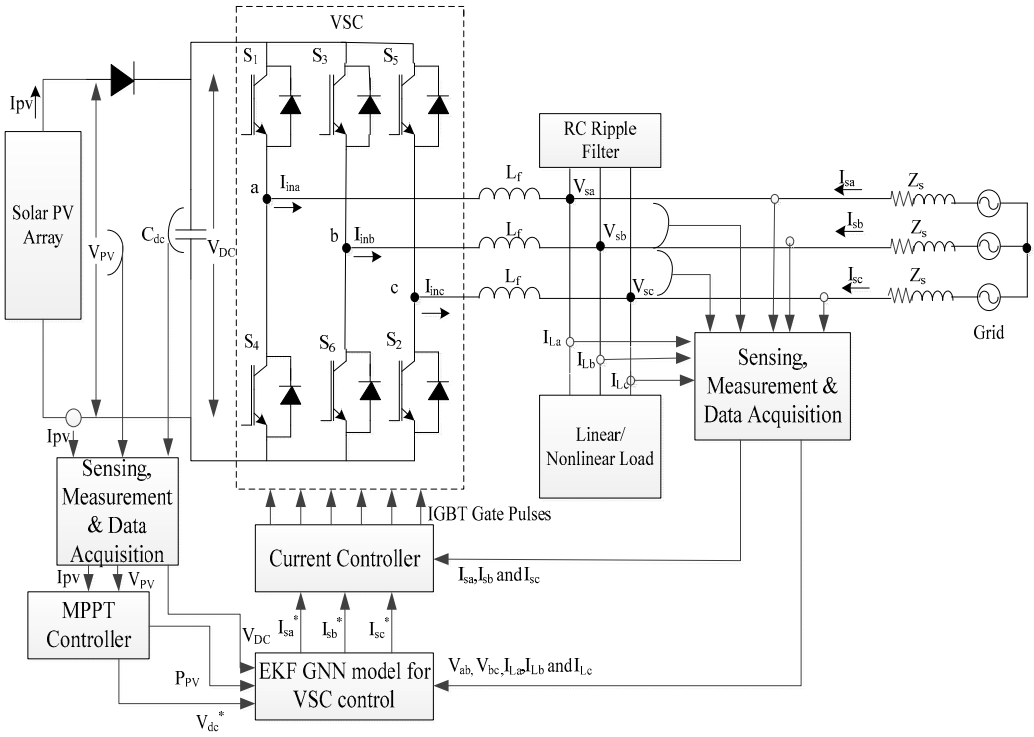


Figure 1. Schematic representation of the system used in this research.

Grid description = three-phase, 415 Volts, 50 Hz, solar PV array ratings = (V_{MPP}) = 700 Volts, current (IMPP) = 13.5 Amp, solar photovoltaic power at maximum power point (PMPP) = 10 kW, rating of interfacing inductors (L_{fa} = L_{fb} = L_{fc}) = 2.6 mH, rating of DC link capacitor (C_{DC}) = 10 mF, DC link voltage (VDC) = 700 V.

3. Extended Kalman Filter-Based GNN Control Algorithm

3.1. Maximum Power Point Tracking Control

A variable step size perturb and observe (P&O)-based MPPT algorithm is employed in this research for the extraction of maximum power from the SPV system. The main operation of the given algorithm is dividing the SPV system's dP_{pv}/dV_{pv} curve into three independent regions. Region 0 denotes the SPV system power's closeness to the highest power point. Region 0 uses the normal value of tracking step size, whereas region 1 and region 2 needed a larger value of step size as compared to region 0 for achieving the high tracking speed. Figure 2 explains the performed functions of the proposed technique. ΔVref0, ΔVref1 and ΔVref2 represent the step size of tracking for region 0, region 1 and region 2, respectively.

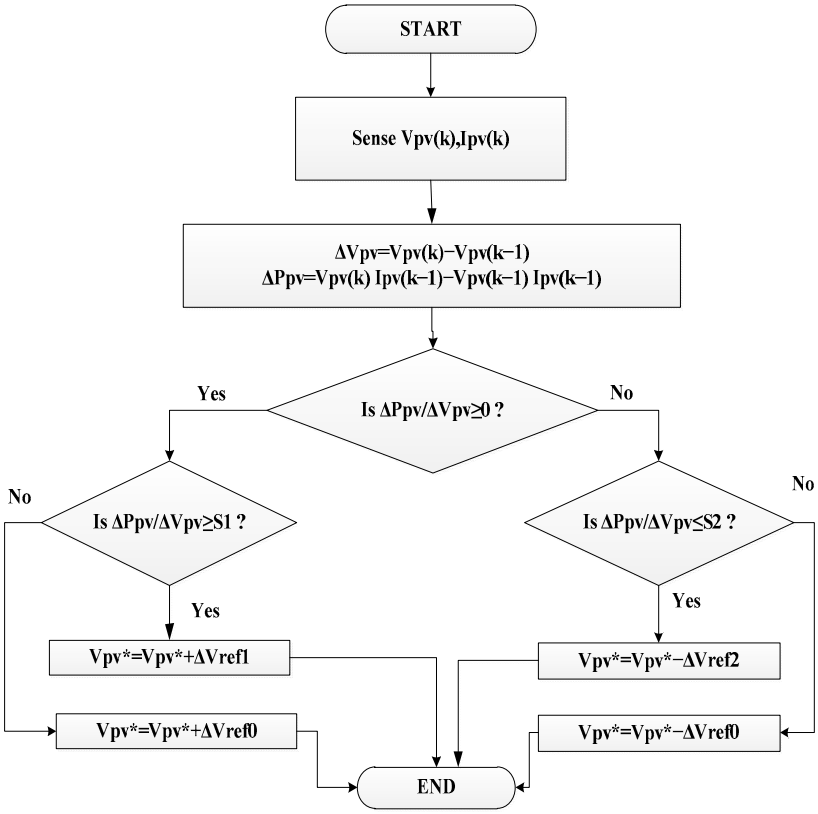
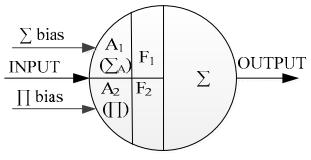


Figure 2. Flowchart of variable step size perturb and observe (P&O) algorithm.

3.2. Extended Kalman Filter-Based GNN Control Algorithm

A generalized neural network works like a multi-layer feed-forward network, where every node implements a precise function on all signals that come to the node, and the parameters are set referring to the node. Figure 3 shows an aggregate-type format of a GNN model, with Σ and Π being the aggregated functions. Σ is an aggregated function, which is adopted with the sigmoidal characteristic function f_1, whereas the Π aggregation function is adopted with the Gaussian function f_2. The derivation of active and re-active load current components is performed by assuming load current (i_{La}, i_{Lb}, i_{Lc}) as the input to the summation (Σ_A) and product (Π) neurons having undefined weights of ($W_{\Sigma i}$) and ($W_{\Pi i}$), respectively.



(**a**)

Figure 3. *Cont.*

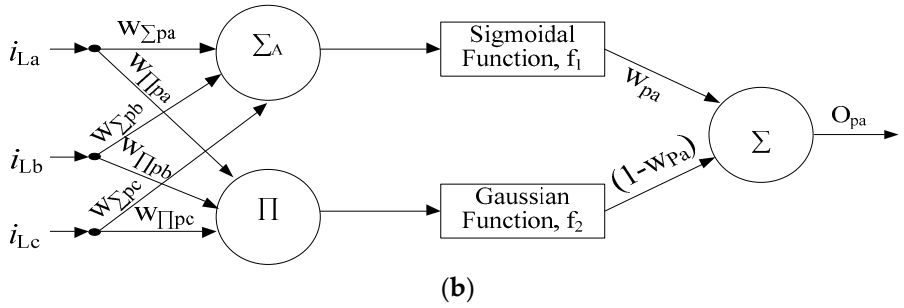


Figure 3. (**a**) Summation-type generalized neural network (GNN) model, (**b**) a summation-type GNN structure to determine the fundamental active element of a load current.

3.2.1. Estimation of Amplitude of Terminal Voltage and Unit Templates

With the help of sensed line voltages (v_{ab}, v_{bc}, v_{ca}) at the point of common coupling (PCC), the amplitude of phase voltages (v_a, v_b and v_c) are calculated as follows:

$$v_a = \frac{2v_{ab} + v_{bc}}{3},\ v_b = \frac{-v_{ab} + v_{bc}}{3}\ ,\ v_c = \frac{-v_{ab} - 2v_{bc}}{3}\ \Sigma \tag{1}$$

The terminal voltage amplitude at PCC can be estimated as

$$V_t = \sqrt{\left[\frac{2({v_a}^2 + {v_b}^2 + {v_c}^2)}{3}\right]} \tag{2}$$

$$u_{pa} = \frac{v_a}{V_t}\ , u_{pb} = \frac{v_b}{V_t}\ ,\ u_{pc} = \frac{v_c}{V_t} \tag{3}$$

Further, quadrature unit templates can be calculated using in-phase unit templates as

$$u_{qa} = \frac{u_{pc}}{\sqrt{3}} - \frac{u_{pb}}{\sqrt{3}}\ , u_{qb} = \frac{\sqrt{3}u_{pa}}{2} + \frac{u_{pb} - u_{pc}}{2\sqrt{3}}, u_{qc} = \frac{u_{pb} - u_{pc}}{2\sqrt{3}} - \frac{\sqrt{3}u_{pa}}{2} \tag{4}$$

3.2.2. Terminal Voltage Amplitude and Unit Templates

The DC link voltage (V_{dc}) is sensed and compared with the reference DC link voltage (V_{dc}^*) in order to determine the active loss component.

$$V_{dce}(k) = V_{dc}^*(k) - V_{dc}(k) \tag{5}$$

k is the number of iterations. Further, the error (V_{dce}) of reference V_{dc}^* and sensed V_{dc} are processed through a PI controller, the output of which is an active current component $\left(w_{pdc}\right)$ used to regulate the DC link. The controller output at k^{th} iteration is estimated as

$$w_{pd}(k) = w_{pd}(k-1) + k_{pd}\{V_{dce}(k) - V_{dce}(k-1)\} + k_{id}V_{dce}(k). \tag{6}$$

During no sunshine conditions ($P_{PV} = 0$), the set point DC link voltage is set to reference the DC link voltage of DSTATCOM so that the system operates in the power quality improvement mode (as a DSTATCOM).

The error V_{te} has been considered between actual and set terminal voltage, V_t and V_t^*, at PCC to calculate the reactive loss component. Further, this error is passed through a PI controller for minimization. The calculated reactive loss component is considered to maintain AC terminal voltage constant and close to its reference value.

$$V_{te}(k) = V_t^*(k) - V_t(k) \tag{7}$$

The controller output at k^{th} instant is

$$w_{qt}(k) = w_{qt}(k-1) + k_{pt}\{V_{te}(k) - V_{te}(k-1)\} + k_{it}V_{te}(k) \quad (8)$$

where w_{qt} is a part of the reactive loss current component, and k_{pt} and k_{it} are proportional and integral gains, respectively.

A feed-forward weight is calculated and incorporated into the controller to achieve a fast dynamic response, which can be written as:

$$w_{PV}(k) = \frac{2P_{PV}(k)}{3V_t} \quad (9)$$

where P_{PV} is solar power.

3.2.3. Fundamental Active and Reactive Component of Load Current

A generalized neural network (GNN) of a summation-type network is used in the present work for the estimation of the fundamental active and reactive components of the load current. The output calculations have been divided into two sections: forward calculations and reverse calculations.

With the help of Figure 3, the below mentioned equation is used to obtain the output of the summation part of the GNN:

$$O_{\Sigma} = f_1(\Sigma W_{\Sigma i}X_i + X_{o\Sigma}) \quad (10)$$

Similarly, the output of the product part of the GNN can be calculated as

$$O_{\Pi} = f_2(\Pi W_{\Pi i}X_i + X_{o\Pi}) \quad (11)$$

The output of the final GNN will be the sum of the summation part and product part which can be expressed as

$$\boldsymbol{O_i} = \boldsymbol{O_{\Sigma}} * \boldsymbol{W_{\Sigma}} + \boldsymbol{O_{\Pi}}(\mathbf{1} - \boldsymbol{W_{\Sigma}}) \quad (12)$$

In the above equation, O_{Σ} denotes the output of the summation part of the neuron, O_{Π} shows the output of the product part of the neuron, and W depicts the weights. The proposed control strategy using GNN utilizes a GNN model to determine the fundamental active current components of the sensed load currents (i_{La}, i_{Lb}, i_{Lc}). The sensed load currents are the input to the proposed model which are further multiplied with weights ($w_{\Sigma pa}, w_{\Sigma pb}, w_{\Sigma pc}$) and ($w_{\Pi pa}, w_{\Pi pb}, w_{\Pi pc}$) at summation ($\Sigma_A$) and product ($\Pi$) neurons, respectively. At the initial level, these weights are referred as unknown weights and calculated by considering the in-phase unit templates (u_{pa}, u_{pb}, u_{pc}) as the reference weights. A summation-type GNN structure for the calculation of fundamental active current component of a phase is given in Figure 3b.

The output calculations can be divided into two parts: forward calculations and reverse calculations. The aggregation function of Σ_A and Π of the forward mode calculations for phase "*a*" is shown in Equations (8) and (9), respectively. Similarly, the aggregation function of Σ_A and Π of the forward mode calculations of another two phases, "*b*" and "*c*", can be calculated.

$$\Sigma_{A_{Pa}} = i_{La}w_{\Sigma pa} + i_{Lb}w_{\Sigma pb} + i_{Lc}w_{\Sigma pc} + \Sigma_{bias} \quad (13)$$

$$\Pi_{pa} = i_{La}w_{\Pi pa} * i_{Lb}w_{\Pi pb} * i_{Lc}w_{\Pi pc} * \Pi_{bias} \quad (14)$$

Σ_{bias} and Π_{bias} represent the initial bias of the Σ_A and Π part of the structure, respectively. The proposed developed neuron has both Σ and Π aggregation functions. The Σ_A aggregation function has been used with the sigmoidal characteristic function f_1, while the Π aggregation function has been used with the Gaussian function f_2 as a characteristic

function. The output of the Σ_A part with a sigmoidal characteristic transfer function after threshold can be calculated as

$$O_{\Sigma_{Apa}} = f_1(\Sigma_{Apa}) = \frac{1}{1 + e^{-\lambda_{\Sigma p} * \Sigma_{A_{Pa}}}} \tag{15}$$

The output of the Π part is threshold by using Gaussian transfer function and can be written as

$$O_{\Pi pa} = f_2(\Pi_{pa}) = e^{-\lambda_{\Pi p} * {\Pi_{pa}}^2} \tag{16}$$

where $\lambda_{\Sigma p}$ and $\lambda_{\Pi p}$ are the gain scaling parameters of the Σ_A and Π part of the network, respectively, and are considered as unity here to avoid complexity. Similarly, output of the Σ_A and Π part of the forward mode calculations of other phases, "*b*" and "*c*", are calculated as

$$O_{\Sigma_{Apb}} = f_1\left(\Sigma_{Apb}\right) = \frac{1}{1 + e^{-\lambda_{\Sigma p} * \Sigma_{A_{Pb}}}} \tag{17}$$

$$O_{\Pi pb} = f_2\left(\Pi_{pb}\right) = e^{-\lambda_{\Pi p} * {\Pi_{pb}}^2} \tag{18}$$

$$O_{\Sigma_{Apc}} = f_1(\Sigma_{Apc}) = \frac{1}{1 + e^{-\lambda_{\Sigma p} * \Sigma_{A_{Pc}}}} \tag{19}$$

$$O_{\Pi pc} = f_2(\Pi_{pc}) = e^{-\lambda_{\Pi p} * {\Pi_{pc}}^2} \tag{20}$$

The final output of the GNN will be the function of two outputs and related to the weights W and $(1 - W)$, respectively, through the linear transfer function which can be written as

$$O_{pa} = O_{\Pi pa}(1 - W_a) + O_{\Sigma Apa} W_a \tag{21}$$

$$O_{pb} = O_{\Pi pb}(1 - W_b) + O_{\Sigma Apb} W_b \tag{22}$$

$$O_{pc} = O_{\Pi pc}(1 - W_c) + O_{\Sigma Apc} W_c \tag{23}$$

where $W_a,\ W_b,\ W_c$ are weights associated with the phase "*a*", "*b*" and "*c*", respectively. The mean active component of load currents (w_{Lp}) is obtained by averaging the final output of the GNN for each phase and is given as

$$w_{Lp} = \frac{\left(O_{pa} + O_{pb} + O_{pc}\right)}{3} \tag{24}$$

However, the estimation of the fundamental reactive current component under the load current has been calculated by implementing the proposed GNN model. The sensed load currents (i_{La}, i_{Lb}, i_{Lc}) with their unknown weights, $\left(w_{Aqa}, w_{Aqb}, w_{Aqc}\right)$ and $\left(w_{\Pi qa}, w_{\Pi qb}, w_{\Pi qc}\right)$, are processed as inputs to Σ_A and Π neurons. The aggregation function of Σ_A and Π of the forward mode calculations for phase "*a*" is given in Equations (23) and (24), respectively. Similarly, the aggregation function of Σ_A and Π of the forward mode calculations of the other remaining phases "*b*" and "*c*" are calculated.

$$\Sigma_{A_{qa}} = i_{La} w_{\Sigma qa} + i_{Lb} w_{\Sigma qb} + i_{Lc} w_{\Sigma qc} + \Sigma_{bias} \tag{25}$$

$$\Pi_{qa} = i_{La} w_{\Pi qa} * i_{Lb} w_{\Pi qb} * i_{Lc} w_{\Pi qc} * \Pi_{bias} \tag{26}$$

where Σ_{bias} and Π_{bias} denote the initial bias of the Σ_A and Π part of the network, respectively. Similarly, as explained above while calculating the fundamental active load current component, here also the Σ_A aggregation function has been adopted along with the sigmoidal characteristic function f_1, whereas the Π aggregation function has been adopted

along with the Gaussian function f_2, taken as a characteristic function. The output of the Σ_A part with a sigmoidal characteristic transfer function after threshold can be obtained as

$$O_{\Sigma_{Aqa}} = f_1(\Sigma_{Aqa}) = \frac{1}{1 + e^{-\lambda_{\Sigma q} * \Sigma_{Aqa}}} \tag{27}$$

The output of the Π part becomes threshold by using the Gaussian transfer function and can be written as given:

$$O_{\Pi qa} = f_2(\Pi_{qa}) = e^{-\lambda_{\Pi q} * {\Pi_{qa}}^2} \tag{28}$$

where $\lambda_{\Sigma q}$ and $\lambda_{\Pi q}$ are the gain scaling parameters of the Σ_A and Π part of the network, respectively, and are considered as unity here to avoid complexity. Similarly, the output of the Σ_A and Π part of the forward mode calculations of the other two phases "*b*" and "*c*" are calculated as

$$O_{\Sigma_{Aqb}} = f_1\left(\Sigma_{Aqb}\right) = \frac{1}{1 + e^{-\lambda_{\Sigma q} * \Sigma_{A_{qb}}}} \tag{29}$$

$$O_{\Pi qb} = f_2\left(\Pi_{qb}\right) = e^{-\lambda_{\Pi q} * {\Pi_{qb}}^2} \tag{30}$$

$$O_{\Sigma_{Aqc}} = f_1(\Sigma_{Aqc}) = \frac{1}{1 + e^{-\lambda_{\Sigma q} * \Sigma_{Aqc}}} \tag{31}$$

$$O_{\Pi qc} = f_2(\Pi_{qc}) = e^{-\lambda_{\Pi q} * {\Pi_{qc}}^2} \tag{32}$$

The output of the GNN model as a function of weights W and $(1 - W)$ can be written as

$$O_{qa} = O_{\Pi qa}(1 - W_{a1}) + O_{\Sigma Apa} W_{a1} \tag{33}$$

$$O_{qb} = O_{\Pi qb}(1 - W_{b1}) + O_{\Sigma Aqb} W_{b1} \tag{34}$$

$$O_{qc} = O_{\Pi qc}(1 - W_{c1}) + O_{\Sigma Aqc} W_{c1} \tag{35}$$

where W_{a1}, W_{b1} and W_{c1} are the weights associated with the Σ_A and Π part of the developed GNN model for the estimation of the fundamental reactive load current components. The fundamental reactive current component of the load current (w_{Lq}) is determined by taking the average of output of the GNN model for the reactive current component of the load currents and given as:

$$w_{Lq} = \frac{\left(O_{qa} + O_{qb} + O_{qc}\right)}{3} \tag{36}$$

3.2.4. GNN Weight Prediction and Updating Using Extended Kalman Filter (EKF)

The Kalman Filter (KF) is a recursive algorithm used for estimating the state of a dynamic system in the case of less availability of data because of the presence of noise, etc. The KF algorithm utilizes the prior knowledge to predict the past, present and future state of the given system. The main advantage of the KF-based approach is less memory space requirements because the data are updated in each and every iteration. The basic KF theory is based on the probability of the hypothesis of the predicted state of the system under consideration by hypothesis of prior state and then using the available data from measurement sensors to correct the hypothesis to obtain the best estimation for each iteration. Two basic assumptions are made to derive the basic equations for KF to be optimal in the sense of mean square error, which should be described by a model of linear state space; the noises are white and Gaussian with zero mean, uncorrelated with each other.

$$x_{k+1} = F_{k+1,\,k}\, x_k \; + \; q_k \tag{37}$$

$$y_{k+1} = H_{k+1}\, x_{k+1} \; + \; r_{k+1} \tag{38}$$

Equation (37) is known as a process equation. x_k is a system state vector, a minimal set of data that uniquely defines the behavior of a system, and k depicts discrete time. $F_{k+1,\,k}$ is the transition matrix to take the state x_k from time k to $k+1$, and q_k is the additive process noise, white and Gaussian, with a zero mean and possessing a covariance matrix Q_k. The measurement step is shown in Equation (38), where H_{k+1} is the measurement matrix, y_{k+1} is observable at time $k+1$, and r_{k+1} is the additive process noise, white and Gaussian, with a zero mean and possessing a covariance matrix R_k. Both noises are uncorrelated with each other. The KF algorithm works in two repeated functional steps:

The prediction step (time update): This step is to compute the estimation of state and error covariance.

$$\hat{x}^{-}{}_{k+1} = F_{k+1,\,k}\,\hat{x}_k \tag{39}$$

$$P_{k+1} = F_{k+1,\,k}\,P_k\;F^{T}{}_{k+1,\,k} \;+\; Q_k \tag{40}$$

where P_k is there error covariance matrix.

Correction step (measurement update): This step is to correct the estimated state according to the previous step with the help of y_{k+1}

$$K_{k+1} = P^{-}{}_{k+1}\,H^{T}{}_{k+1}\;\left[H_{k+1}P^{-}{}_{k+1}\,H^{T}{}_{k+1}\;+R_{\,k+1}\right]^{-1} \tag{41}$$

$$\hat{x}_{k+1} = \hat{x}_k{}^{-}\,K_{k+1}\,\left(\,y_{k+1} - H_{k+1}\,\hat{x}^{-}{}_k\right) \tag{42}$$

$$P_{k+1} = (I - \;K_{k+1}\;\;H_{k+1}\,)P_{k+1}{}^{-} \tag{43}$$

where $K_{\,k+1}$ is the Kalman gain matrix.

The GNN model is a nonlinear system, so the basic KF approach should be extended by using the linearization process and is known as the extended Kalman Filter (EKF). The basic difference between KF and EKF is the linearization of the nonlinear system function performed by using the Jacobian matrix in EKF, and then rest of the KF steps can be applied. The nonlinear dynamic system can be defined by using the following equations:

$$x_{k+1} = x_k \;+\; q_k \tag{44}$$

$$y_{k+1} = h(x_{k+1},\,u_{k+1}) +\; r_{k+1} \tag{45}$$

Equation (44) depicts the state of a stationary process corrupted with process noise q_k, and state x_k consists of network weights. x_{k+1} is the weight vector matrix and input vector depicted by u_{k+1}. The noise covariance matrix can be written as $R_{k+1} = E\left[r_{k+1}r^{T}{}_{k+1}\,\right]$ and $Q_{k+1} = E\left[q_{k+1}q^{T}{}_{k+1}\,\right]$

In EKF, the linearization of measurement equation is carried out at each time step around the newest state estimation by using the first-order Taylor approximation. The GNN training problem is taken as a problem to find the state estimation of x_{k+1} for minimizing the least square errors by implementing the previous calculations. K_{k+1} Kalman gain can be written as

$$K_{k+1} = \;P_{k+1}H^{T}{}_{k+1}\;\left[H_{k+1}\,P_{k+1}H^{T}{}_{k+1}\;+R_{\,k+1}\right]^{-1} \tag{46}$$

$$\hat{x}_{k+1} = \hat{x}_{k+1} +\;K_{k+1}[\,y_{k+1}\;-h(\,\hat{x}_{k+1} + u_{k+1})] \tag{47}$$

$$P_{k+1} = P_k - K_k H_k\,P_k \;+\; Q_k \tag{48}$$

3.2.5. Reference Current Calculation

The fundamental real power current component of the load current can be obtained as

$$I_{Lp} = w_{pd} + w_{Lp} -\; w_{PV} \tag{49}$$

The reference active components of the grid currents can be written as

$$i_{psa} = I_{Lp} * u_{pa},\; i_{psb} = I_{Lp} * u_{pb}, i_{psc} = I_{Lp} * u_{pc} \tag{50}$$

Similarly, the fundamental reactive current component of a load current can also be obtained as

$$I_{Lq} = w_{qt} + w_{Lq} \tag{51}$$

The grid currents' active reference components can be written as

$$i_{qsa} = I_{Lq} * u_{qa},\ i_{qsb} = I_{Lq} * u_{qb}, i_{qsc} = I_{Lq} * u_{qc} \tag{52}$$

The net fundamental reference grid currents ($i^*_{sa},\ i^*_{sb},\ i^*_{sc}$) can be obtained as

$$i^*_{sa} = i_{psa} + i_{qsa},\ i^*_{sb} = i_{psb} + i_{qsb},\ i^*_{sc} = i_{psc} + i_{qsc} \tag{53}$$

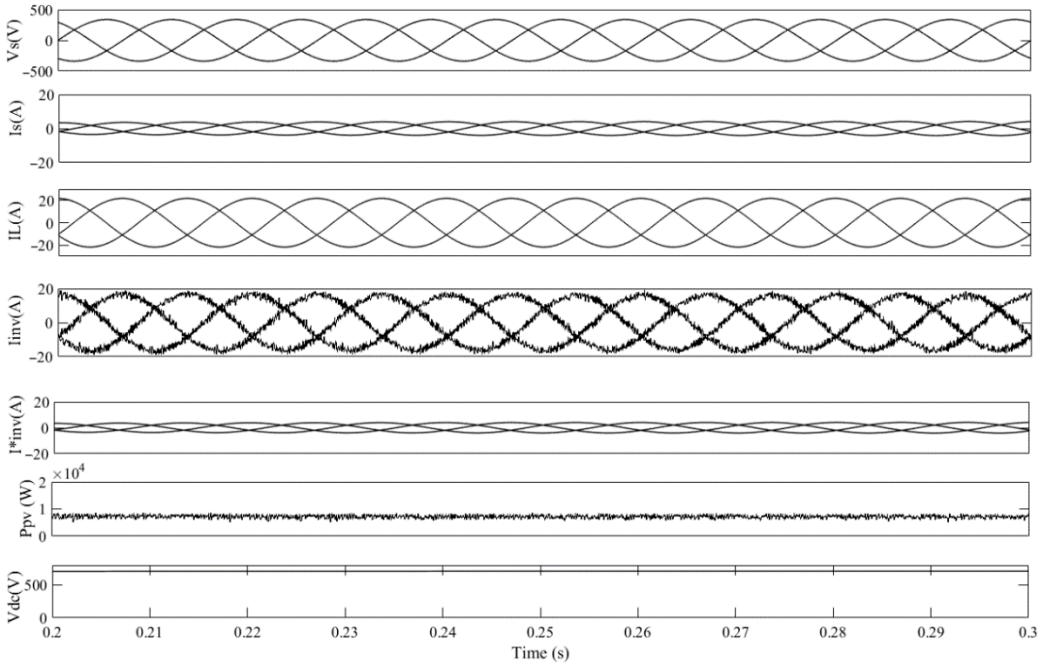
VSC gating signals have been generated by comparing reference ($i^*_{sa},\ i^*_{sb},\ i^*_{sc}$) and actual ($i_{sa}, i_{sb}, i_{sc}$) grid currents for an individual phase, and the error is processed through the PI current regulator. A hysteresis current regulator is used for indirect current control.

4. Results

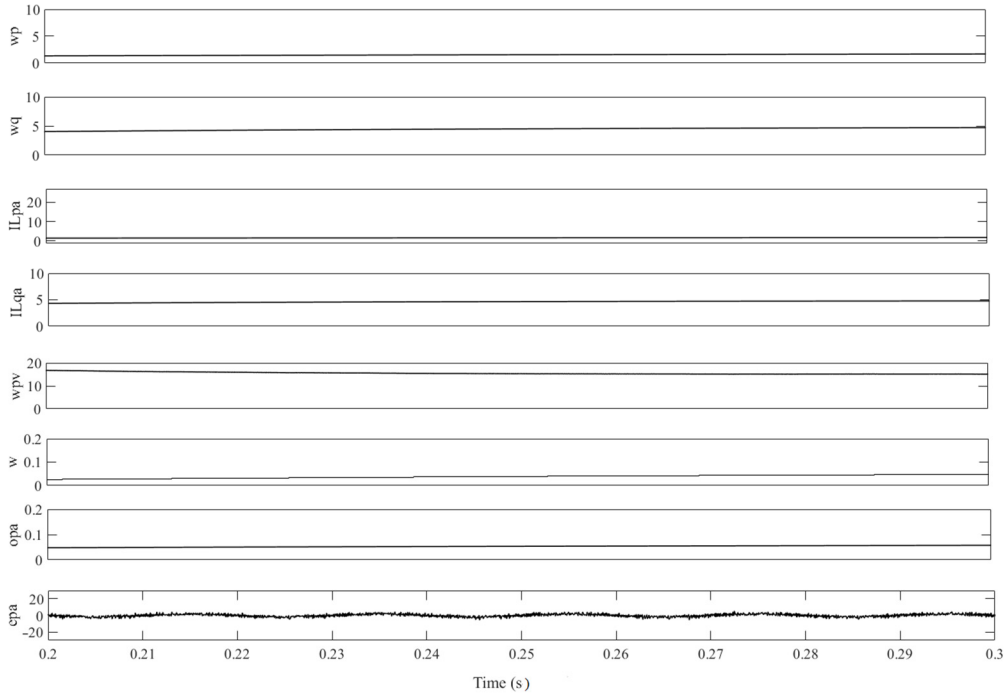
The SPV array is designed for a maximum power rating of 10 kW. The given system is designed, developed and simulated using the MATLAB/Simulink platform. The performance evaluation of the proposed system is carried out under a linear and nonlinear load with dynamic changes and a changing solar irradiance scenario.

4.1. Performance Analysis of Proposed Controller for Linear Load for PFC

The performance of the proposed EKF-based GNN controller is evaluated with a combined structure of the summation and product neurons collectively in single layer. In the proposed GNN network, each processed neuron consists of different weights with the inputs as measured load currents (i_{La}, i_{Lb}, i_{Lc}). With the help of this, the summation weights ($w_{\Sigma pa}, w_{\Sigma pb}, w_{\Sigma pc}$) and product weights ($w_{\Pi pa}, w_{\Pi pb}, w_{\Pi pc}$) are estimated. The proposed GNN is trained to calculate the unspecified weights and EKF filter which is used to predict and update the weights of the GNN network. The output of the proposed system is evaluated under the influence of several performance parameters, i.e., grid voltage (v_s), grid current (i_s), load current (i_L), voltage source converter current (i_{vsc}), grid real power (P_g), grid re-active power (Q_g), PCC terminal voltage (V_t), SPV power (P_{pv}), SPV array current (I_{pv}) and voltage at direct current link (V_{dc}), provided in Figure 4a,b in steady-state linear load (at t = 0.2 to 0.3 s) conditions. Intermediate signals are given in Figure 4b, which depicts the performance parameters and corresponding weights with a variation in error between the active fundamental components of the load currents. The SPV system remains in operation for feeding the maximum value of power onto the load and grid at unity power factor (UPF).



(**a**)



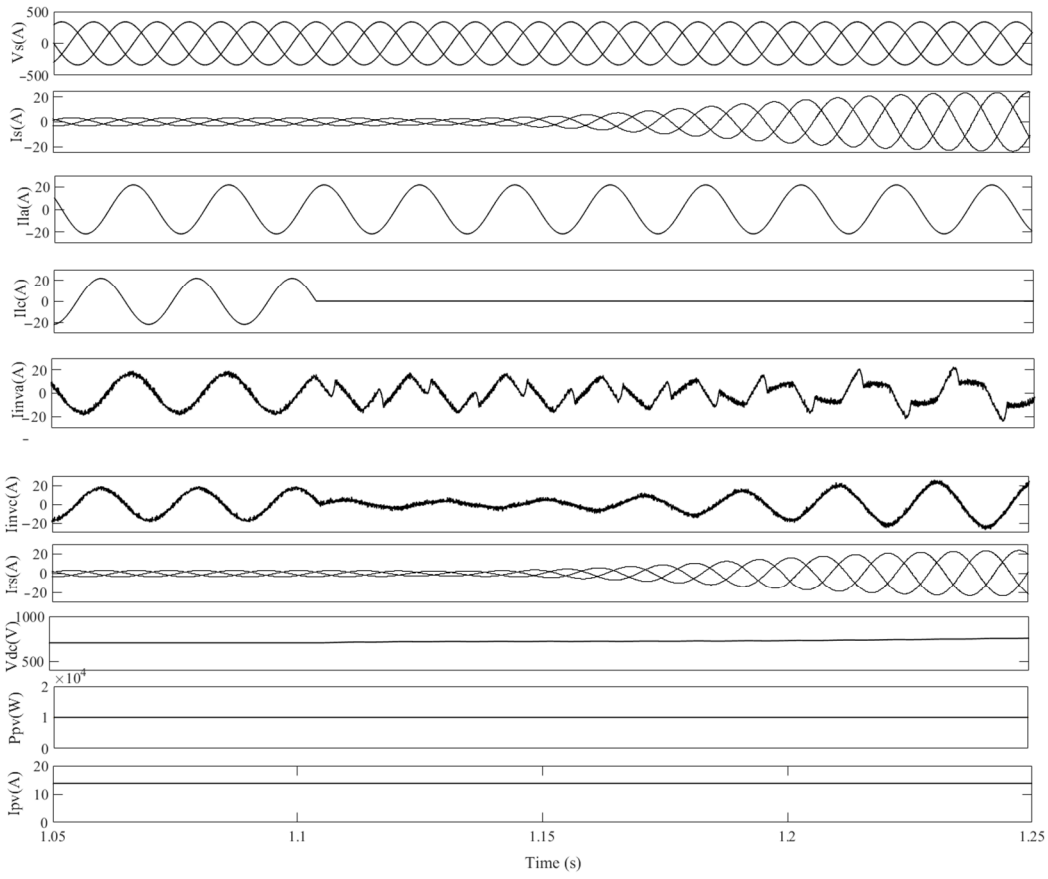
(**b**)

Figure 4. (**a**) Performance parameters, (**b**) estimated weights under steady-state linear load conditions in PFC mode.

4.2. Performance Analysis of Proposed Controller Considering Dynamic Linear Load for Zero Voltage Regulation (ZVR)

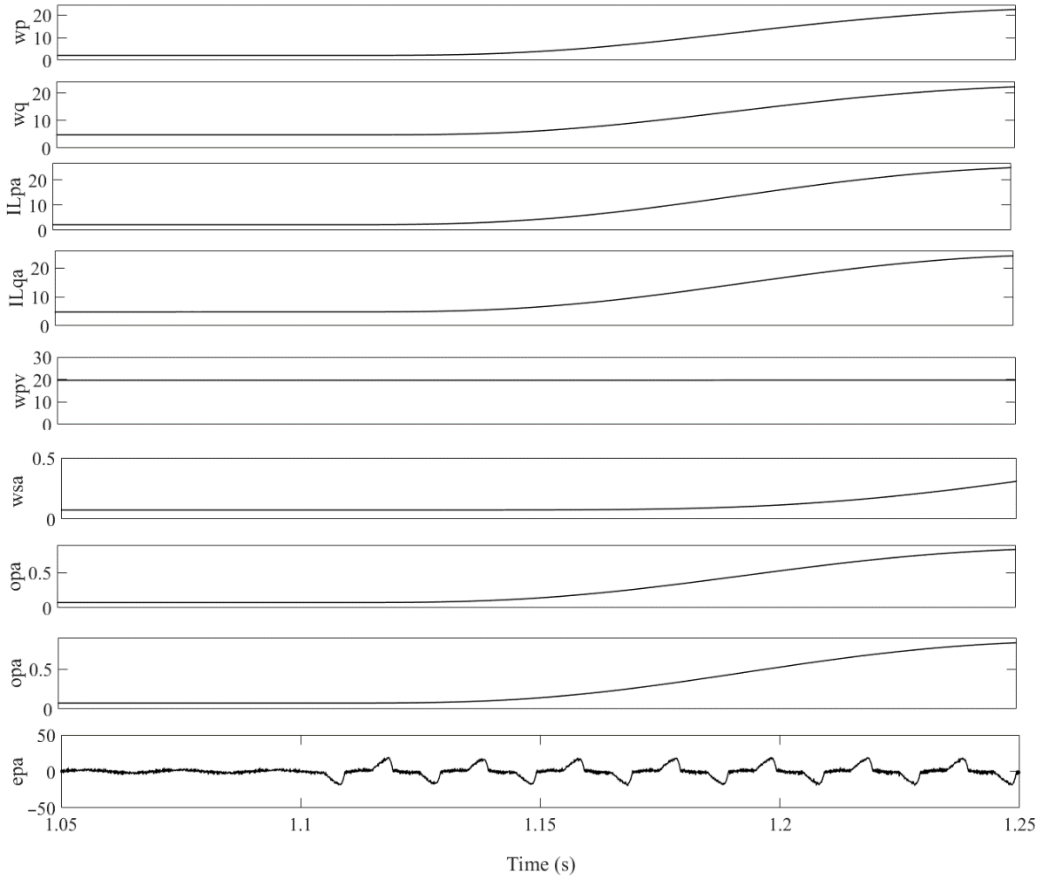
If the load is drawing reactive power with unbalanced behavior, then a change in the terminal voltage (V_t) occurs at the point of common coupling. Figure 5a,b shows the behavior of the given system for dynamic linear loading conditions under a zero voltage regulation (ZVR) mode. It can be observed from the results after phase "*c*" of the load is taken out from the supply at 1.1 s, the grid currents (i_s) remain sinusoidal with the help of a voltage source converter. The PCC voltage (V_t) and voltage at the DC junction are

maintained at set values, which are 415 V and 700 V, respectively, without any fluctuations. Intermediate signals are provided in Figure 5b, which shows the performance parameters and corresponding weights with a variation in error between the active fundamental components of the load currents. The reactive power (Qg) taken from the utility grid is approximately equivalent to 0, for compensating the linear loads using reactive power.



(**a**)

Figure 5. *Cont.*

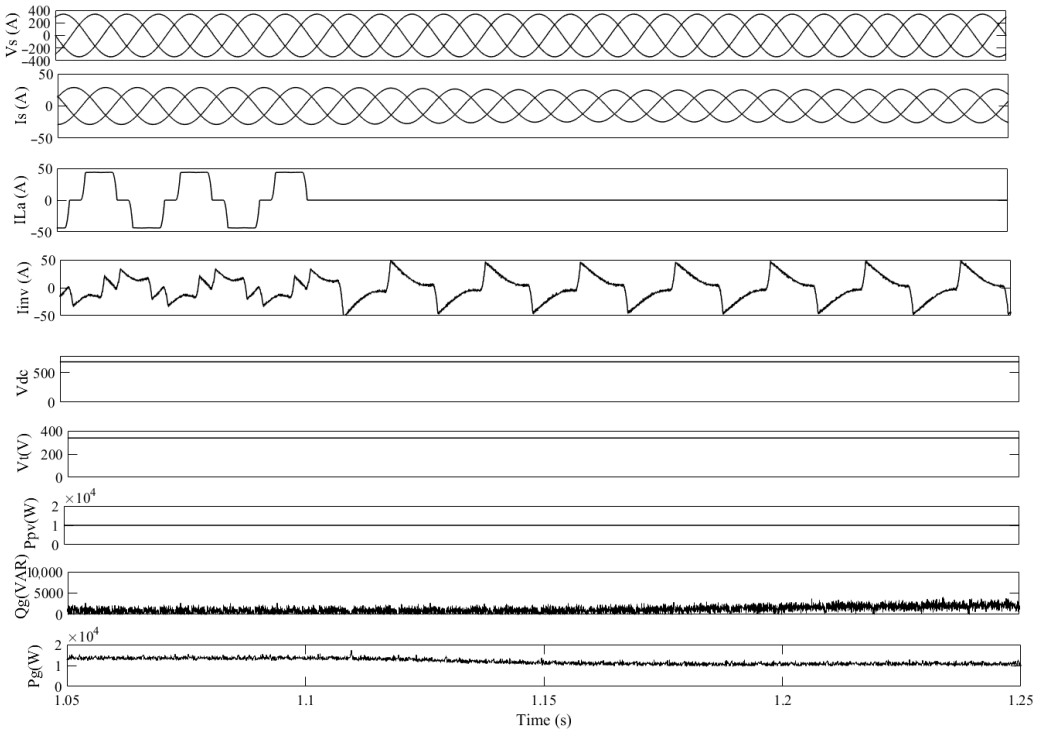


(**b**)

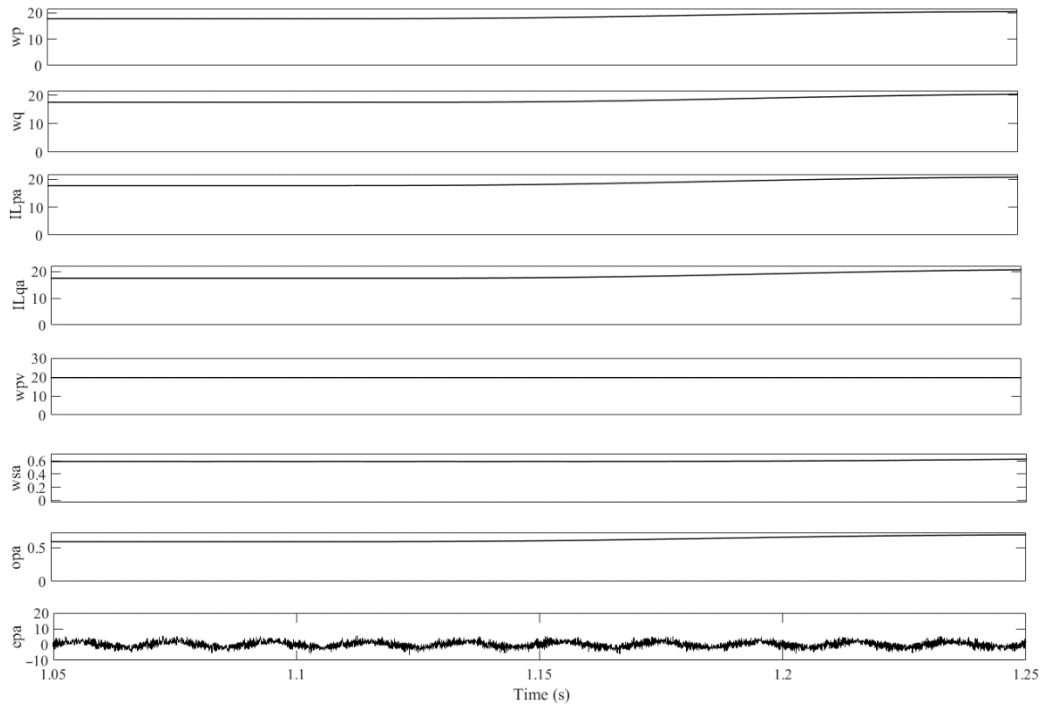
Figure 5. (**a**) Performance parameters, (**b**) estimated weights under dynamic linear load conditions for zero voltage regulation (ZVR).

4.3. Performance under a Nonlinear Load

Figure 6a,b show the behavior of a developed system for dynamic nonlinear loads under the zero voltage regulation (ZVR) mode. By using the given control algorithm, the grid current (i_s) is sinusoidal in nature, when the load of phase "*c*" is extracted at 1.1 s. Intermediate signals are given in Figure 6b for corresponding weight signals with a variation in error between the active fundamental components of the load currents. W_{pa} of the linear transfer function Σ and the estimated GNN output of the active current component of the load current are represented by O_{pa}, and its actual value is achieved by scaling and depicted by I_{Lpa}. This indicates that the calculated weights change according to the load condition. Figure 7a,c show the total harmonic distortion (THD) at the point of common coupling for phase "*a*" of a load current, grid current and grid voltage, respectively, and obtained as 37.48%, 1.04% and 1.07%, respectively, which is within acceptable limits of harmonics in utility currents and voltage fluctuations according to the IEEE-519 and IEEE-1547 standards. THD analysis of the given system is provided in Table 1 for a nonlinear load.

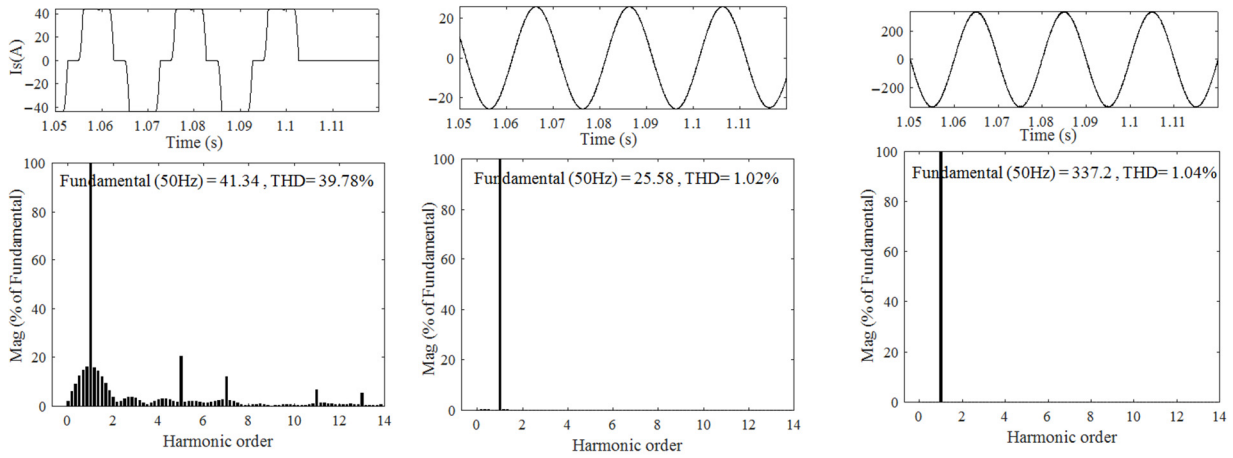


(**a**)



(**b**)

Figure 6. (**a**) Performance indices, (**b**) weights with training parameters for phase "*a*" with steady-state loading in PFC mode.



(**a**) (**b**) (**c**)

Figure 7. Total harmonic distortion (THD) for (**a**) load current of phase "*a*", (**b**) grid current and (**c**) grid voltage in ZVR mode.

Table 1. Parameters of the developed system for the proposed controller.

Operating Mode	Parameters	GNN Based Control Algorithm	EKF GNN Based Control Algorithm
ZVR	Grid voltage (V), %THD at PCC	333.02 V, 1.89%	337.2 V, 1.04%
	Grid current (A), %THD at PCC	23.17 A, 2.54%	25.58 A, 1.02%
	Load current (A), %THD at PCC	40.58 A, 40.62%	41.34 A, 39.78%

4.4. Performance Analysis at Varying Solar Irradiance

Dynamic behavior of the developed system is also observed under varying solar irradiance conditions. The solar irradiance (S) is raised to 1000 W/m^2 from 600 W/m^2 at 0.5 s. The SPV feeds the connected load and excess amount of power output provided to the utility grid. As shown in the results, the grid power (P_g) decreases after 0.5 s because of an increase in power (P_{pv}) output from SPV, given in Figure 8. The proposed system under variable solar irradiance remains operating at the maximum power point of the SPV array and works at unity power factor. The grid side current is maintained to be sinusoidal, and the voltage at DC link is also maintained at a set point.

4.5. Comparative Study of Developed Algorithm with Other Conventional Approaches

The proposed algorithm is analyzed for different scenarios and found accurate as per the desired performance. The parameters of proposed EKF GNN-based controller is presented in Table 1 and compared with the GNN approach. The developed EKF GNN based control approach is found better than ANN conventional approaches such as ADALINE and multilayer perceptron neural network (MLPN). A comparative analysis is made and presented in Table 2 under dynamic nonlinear load conditions on the basis of various performance parameters. The smaller number of unknown weights required and a single layer enhanced the performance of the controller. Additionally, the weight update time, settling period and maximum change in DC voltage is less for the proposed technique.

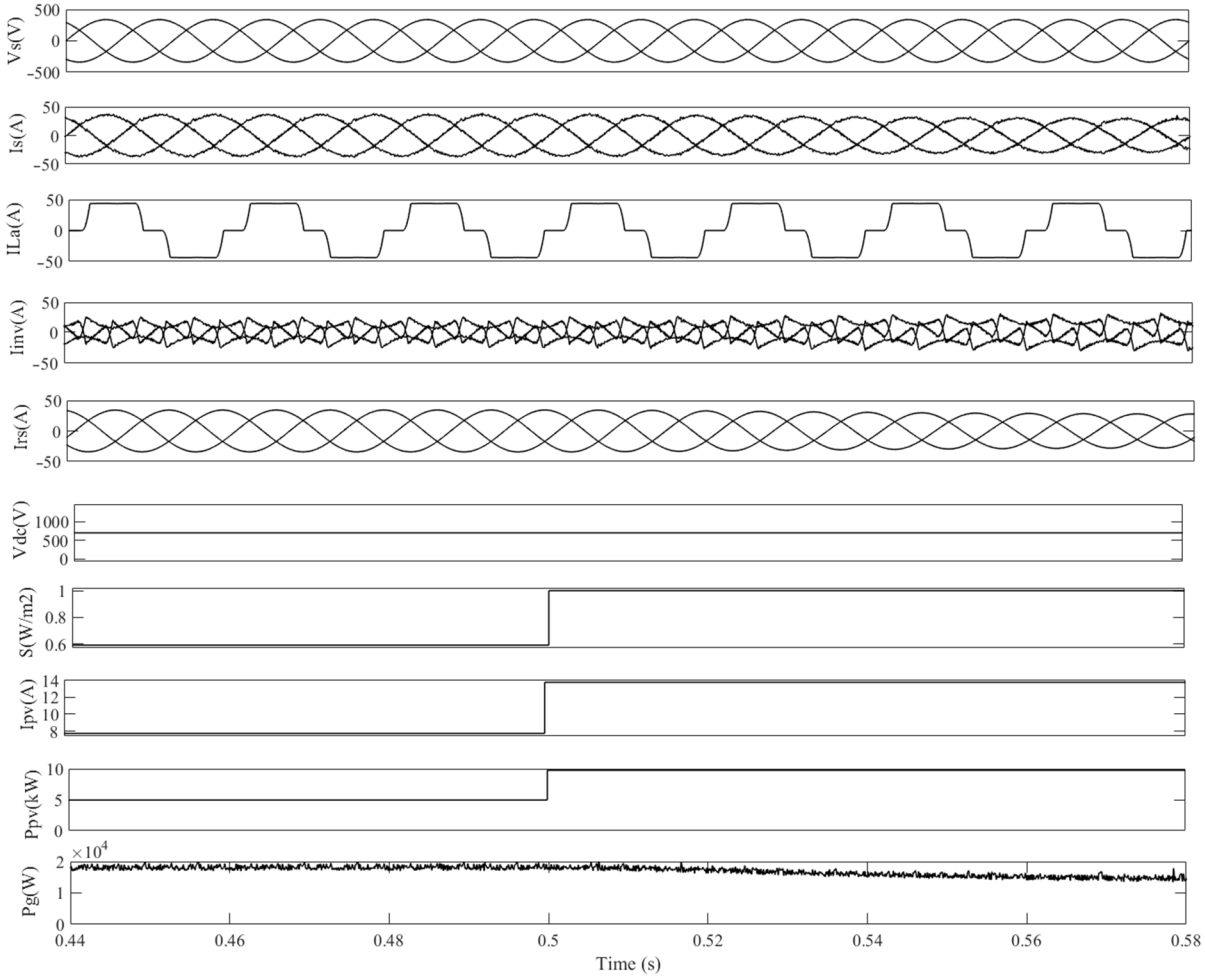


Figure 8. Analysis of performance for a non-linear load with variable solar irradiance.

Table 2. Comparative performance analysis of the proposed technique.

Performance Parameters	ADALINE	MLPN	GNN
Training	Least Mean Square	Back Propagation	EKF
Learning	Gradient decent (GD)	GD/GDM	GDM
Layers	Two	Three	One
Training pattern	Online Training	Offline stochastic training	Online training
Estimation nature	Linear	Nonlinear	Both types
Transfer function	Linear	Sigmoidal	All
Weight update Time	15.7 µs	82 µs	6 µs
Settling period of DC link	1 cycle	2 $\frac{1}{2}$ cycle	1 cycle
Max change in DC link voltage	4.5 V	10 V	3.7 V

Figure 9 shows the performances by experimenting the given system in power factor correction mode by implementing linear and nonlinear loads. It depicts the voltage source current (I_{vsc}), load current (I_{La}), grid current (I_g) and direct current link voltage (V_{dc}). Phase "a" gid voltage and grid current are shown in Figure 10; it also shows the performances of the controller under PFC mode.

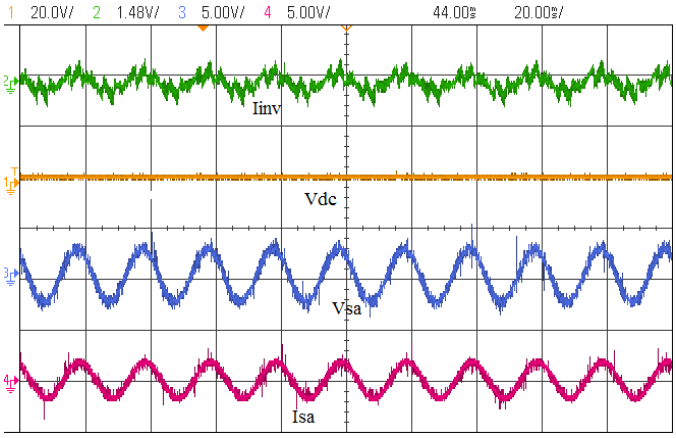


Figure 9. Performance parameters under linear load for the extended Kalman filter (EKF) GNN.

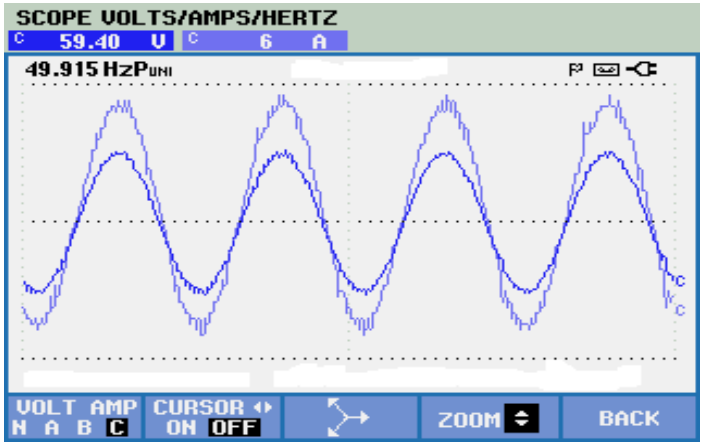


Figure 10. Source side voltage and current of phase "*a*" under the power factor correction mode.

Figure 11 provides the performance of the SPV plant for nonlinear loads, whereas Figure 12a depicts non-linear current and Figure 12b depicts the source current and voltages for phase "*a*". The THD of the load current is given in Figure 13.

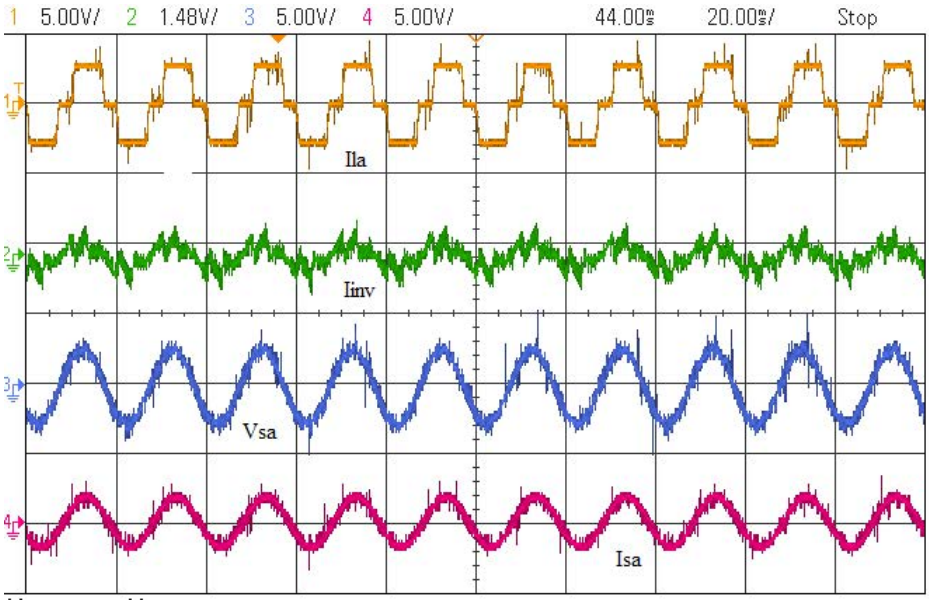


Figure 11. Performance analysis considering non-linear load conditions for EKF GNN.

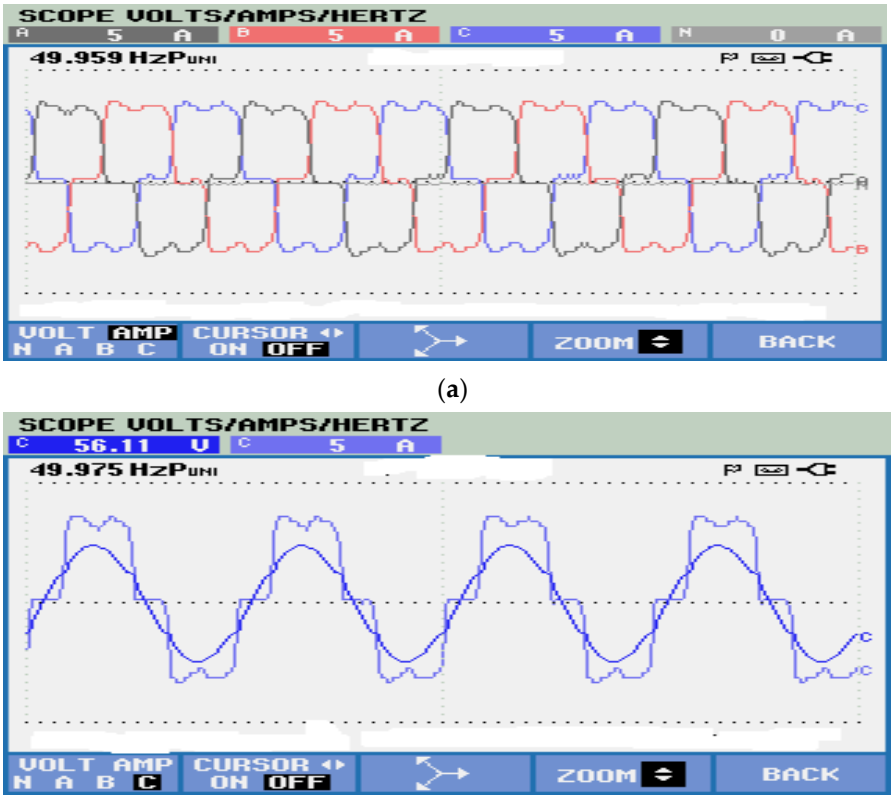


Figure 12. (**a**) Non-linear current, (**b**) current and voltage of phase "*a*".

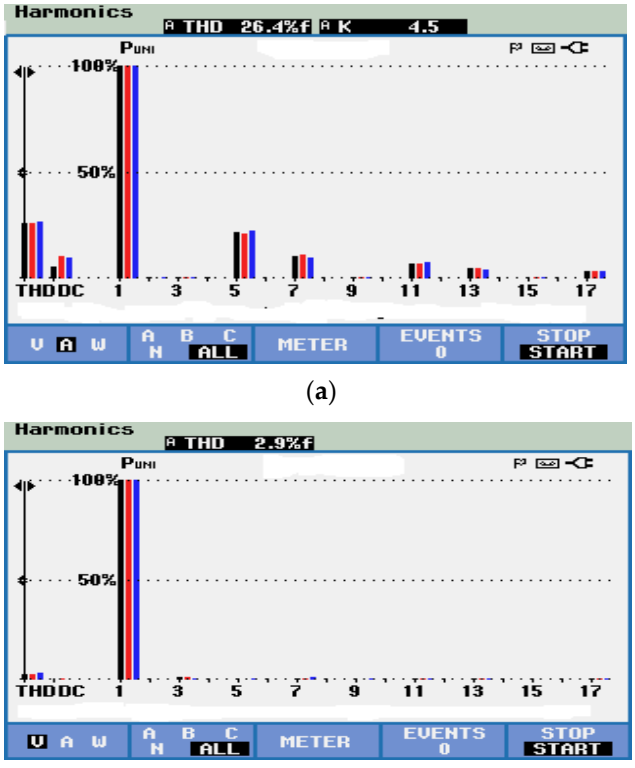


Figure 13. (**a**) THD spectrum of nonlinear load, (**b**) improved THD for grid current under nonlinear load.

The performance of the proposed controller under nonlinear dynamic load is shown in Figure 14; when load of phase "*b*" is removed, the grid current is reduced. Further, the performance of intermediate signals for different weights is also shown in Figure 15. Figures 16 and 17 show the controller performance under variable solar irradiance.

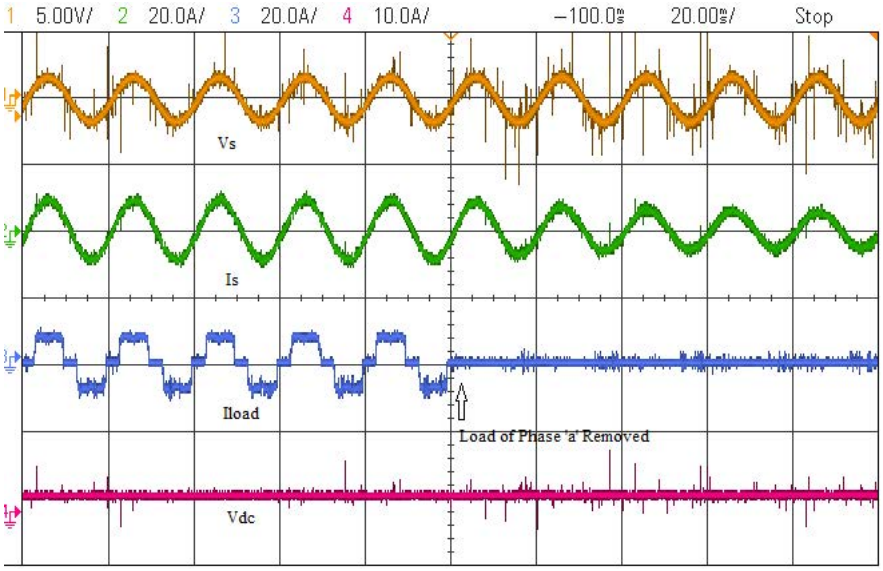


Figure 14. Performance parameters under dynamic nonlinear load for EKF GNN.

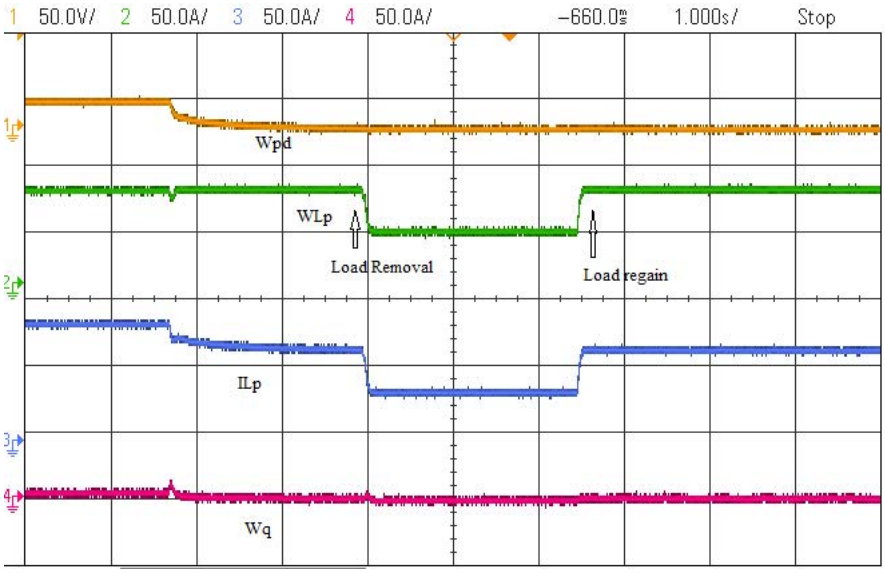


Figure 15. Performance of intermediate weight signals under dynamic nonlinear load for EKF GNN.

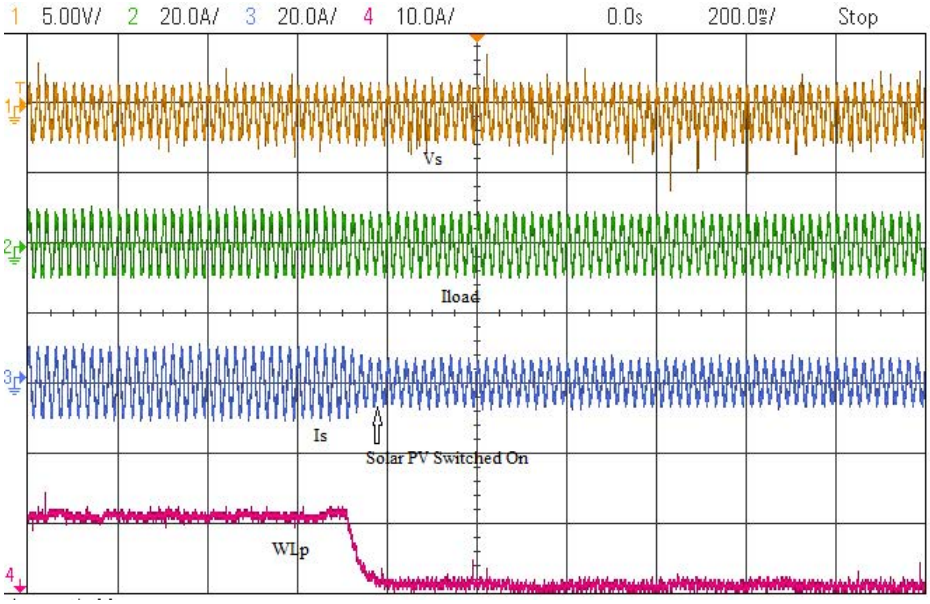


Figure 16. Performance parameters under nonlinear load with solar PV ON for EKF GNN.

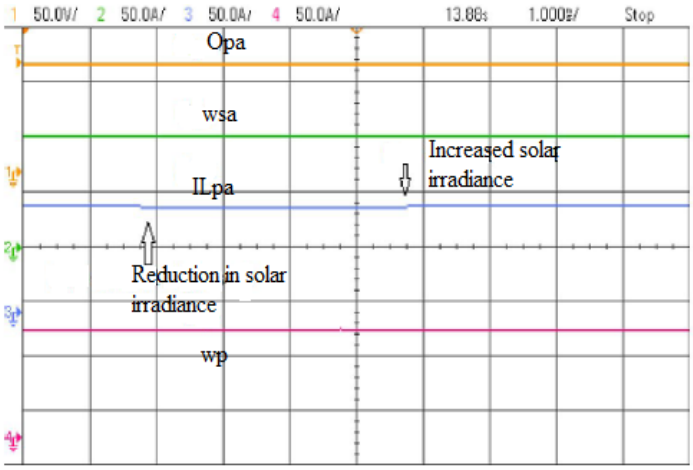


Figure 17. Performance parameters under variable solar irradiance for EKF GNN.

5. Conclusions

The proposed system was developed and tested using MATLAB simulation environment, and the performance of the EKF GNN-based VSC control approach was validated under linear and nonlinear loads for static and dynamic scenarios. The proposed EKF GNN approach has successfully improved the function of the developed SPV system. The performance of the developed system was found satisfactory under load unbalancing and varying solar irradiance. Additionally, it continues to operate in a UPF mode of operation, providing reactive power compensation, load balancing, MPP extraction and harmonics compensation. A changeable step size perturb and observe (P&O) MPPT approach was utilized in this research and enabled fast tracking of MPP and convergence. Moreover, single-stage topology was able to reduce the losses in semiconductor devices and increase the overall efficiency.

The proposed control approach performs with more flexibly in training the network under dynamic conditions and provides improved online learning. This control technique needs a smaller number of training patterns and unknown weights and hence reduces the complexity and the computational time. The developed system obtains acceptable limits of harmonics in utility currents and voltage fluctuations according to the IEEE-519 and IEEE-1547 standards.

Author Contributions: Conceptualization, P.C., M.R. and I.A.; methodology, P.C., M.A. and M.R.; validation, P.C. and I.A.; formal analysis, P.C., I.A., M.A. and M.R.; resources, M.R. and I.A.; writing—original draft, P.C. and M.R.; writing—review and editing, P.C., I.A., M.R. and M.A.; visualization, P.C.; supervision, M.R.; funding acquisition, I.A., M.A. and M.R. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data are used here.

Acknowledgments: Researchers would like to thank the Deanship of Scientific Research, Qassim University for funding publication of this project.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

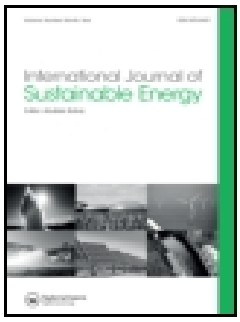
C	DC link capacitor (μF)
f_1	Sigmoidal characteristic function
f_2	Gaussian characteristic function
I_{inv}	Inverter output current
I^*_{inv}	Reference Inverter output current
i_{La}, i_{Lb}, i_{Lc}	Load currents of phase '*a*', '*b*' and '*c*' respectively (Ampere)
I_{Lp}	Fundamental active current component
I_{Lpa}	Fundamental active current component for phase '*a*'
I_{Lq}	Fundamental reactive current component
I_{Lqa}	Fundamental reactive current component for phase '*a*'
I_{MPP}	PV current at maximum power point (Ampere)
i_{psa}	Active reference component of grid current
I_{PV}	PV array output current (Ampere)
i_{qsa}	Reactive reference component of grid current
I_{rs}	Reference component of grid current
i_{sa}, i_{sb}, i_{sc}	Grid currents of phase '*a*', '*b*' and '*c*' respectively (Ampere)
i^*_{sa}, i^*_{sb}, i^*_{sc}	Reference currents of phase '*a*', '*b*' and '*c*' respectively (Ampere)
K_i	Integral gain of PI controller
K_p	Proportional gain of PI controller
L_f	Interfacing inductor (mH)
O_i	Final output of generalized neuron
O_{pa}	output of the GNN model for active component
O_{qa}	output of the GNN model for reactive component
O_Σ	Output of Σ_A part network
O_Π	Output of Π part network
P_{pv}	PV power (W)
P_g	Grid Active power (W)
Q_g	Grid Reactive power (vAR)
S	Solar irradiance (W/m^2)
u_{pa}, u_{pb}, u_{pc}	In phase unit templates of phase voltages
u_{qa}, u_{qb}, u_{qc}	Quadrature unit templates of phase voltages
V_a, V_b, V_c	Phase voltage of utility grid (Volts)
v_{ab}, v_{bc}, v_{ca}	Line voltage of utility grid (Volts)
V_{dc}	DC link voltage (Volts)
V^*_{dc}	Reference DC link voltage (Volts)
V_{MPP}	PV voltage at maximum power point (Ampere)
V_{pv}	PV output voltage (Volts)
V_t	Voltage at point of common coupling (Volts)

V^*_t	Reference voltage at point of common coupling (Volts)
V_{te}	Error between sensed and reference voltage at point of common coupling (Volts)
w	Weights of GNN
w_{ap}, w_{bp}, w_{cp}	Updated weights for hidden layer of active components
w_{aq}, w_{bq}, w_{cq}	Updated weights for hidden layer of reactive components
w_{Lp}	Mean active component of load current
w_{Lq}	Mean reactive component of load current
w_p	Active current component
w_{pdc}	Active current component phase c
w_{pv}	Feed forward weight function of solar power
w_q	Reactive current component
w_{qt}	Function of reactive loss current component
w_{sa}	weight of summation neuron
W_Σ	Weights of summation part of GNN
ΔW	Change in weights of GNN
Greek Symbols	
α	Learning rate
μ	Micro
Ω	Ohm
ϕ	Phase
Σ_A	Aggregation function used with sigmoidal characteristic function
Π	Aggregation function used with Gaussian characteristic function
η	Learning rate

References

1. Hudson, R.; Heilscher, G. PV Grid Integration—System Management Issues and Utility Concerns. *Energy Procedia* **2012**, *25*, 82–92. [CrossRef]
2. IEEE. *IEEE Recommended Practices and Requirement for Harmonic Control on Electric Power System*; IEEE Std.: New York, NY, USA, 1992; p. 519.
3. Wu, T.-F.; Chang, C.-H.; Lin, L.-C.; Kuo, C.-L. Power loss comparison of single and two-stage grid-connected photovoltaic systems. *IEEE Trans. Energy Convers.* **2011**, *26*, 707–715. [CrossRef]
4. Arya, S.R.; Niwas, R.; Bhalla, K.K.; Singh, B.; Chandra, A.; Al-Haddad, K. Power quality improvement in isolated distributed power generating system using DSTATCOM. *IEEE Trans. Ind. Appl.* **2015**, *51*, 4766–4774. [CrossRef]
5. Singh, B.; Jayaprakash, P.; Kothari, D.P.; Chandra, A.; Al-Haddad, K. Comprehensive study of DSTATCOM configurations. *IEEE Trans. Ind. Inform.* **2014**, *10*, 854–870. [CrossRef]
6. Singh, B.; Solanki, J. A comparison of control algorithms for DSTATCOM. *IEEE Trans. Ind. Electron.* **2009**, *56*, 2738–2745. [CrossRef]
7. Kumar, C.; Mishra, M.K. A Multifunctional DSTATCOM Operating Under Stiff Source. *IEEE Trans. Ind. Electron.* **2013**, *61*, 3131–3136. [CrossRef]
8. Ahmad, M.T.; Kumar, N.; Singh, B. Generalised neural network-based control algorithm for DSTATCOM in distribution systems. *IET Power Electron.* **2017**, *10*, 1529–1538. [CrossRef]
9. Bag, A.; Subudhi, B.; Ray, P.K. A combined reinforcement learning and sliding mode control scheme for grid integration of a PV system. *CSEE J. Power Energy Syst.* **2019**, *5*, 498–506.
10. He, J.; Li, Y.W. Hybrid Voltage and Current Control Approach for DG-Grid Interfacing Converters with *LCL* filters. *IEEE Trans. Ind. Electron.* **2012**, *60*, 1797–1809. [CrossRef]
11. Tomar, A.; Mishra, S.; Bhende, C.N. Modified MISO DC-DC converter based PV water pumping system. In Proceedings of the 2016 IEEE 7th Power India International Conference (PIICON), Bikaner, India, 25–27 November 2016; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2016; pp. 1–6.
12. Alfaris, F.E.; Bhattacharya, S. Control and Real-Time Validation for Convertible Static Transmission Controller Enabled Dual Active Power Filters and PV Integration. *IEEE Trans. Ind. Appl.* **2019**, *55*, 4309–4320. [CrossRef]
13. Rengasamy, M.; Gangatharan, S.; Elavarasan, R.M.; Mihet-Popa, L. The Motivation for Incorporation of Microgrid Technology in Rooftop Solar Photovoltaic Deployment to Enhance Energy Economics. *Sustainability* **2020**, *12*, 10365. [CrossRef]
14. Jain, C.; Singh, B. A Three-Phase Grid Tied SPV System with Adaptive DC Link Voltage for CPI Voltage Variations. *IEEE Trans. Sustain. Energy* **2016**, *7*, 337–344. [CrossRef]
15. Agarwal, R.K.; Hussain, I.; Singh, B. Three-phase single-stage grid tied solar PV ECS using PLL-less fast CTF control technique. *IET Power Electron.* **2017**, *10*, 178–188. [CrossRef]
16. Singh, B.; Jain, C.; Goel, S.; Chandra, A.; Al-Haddad, K. A multifunctional grid-tied solar energy conversion system with ANF-based control approach. *IEEE Trans. Ind. Appl.* **2016**, *52*, 3663–3672. [CrossRef]
17. Vidal, H.; Rivera, M.; Wheeler, P.; Vicencio, N. The Analysis Performance of a Grid-Connected 8.2 kWp Photovoltaic System in the Patagonia Region. *Sustainability* **2020**, *12*, 9227. [CrossRef]

18. Campanhol, L.B.G.; da Silva, S.A.O.; de Oliveira, A.A.; Bacon, V.D. Single-stage three-phase grid-tied PV system with universal filtering capability applied to DG systems and AC microgrids. *IEEE Trans. Power Electron.* **2017**, *32*, 9131–9142. [CrossRef]
19. Tomar, A.; Mishra, S. CMPVI-Based MIDO Scheme under SSE for Optimum Energy Balance and Reduced ROI. *IEEE Trans. Sustain. Energy* **2017**, *9*, 1318–1327. [CrossRef]
20. Huang, P.; Zhang, X.; Copertaro, B.; Saini, P.; Yan, D.; Wu, Y.; Chen, X. A Technical Review of Modeling Techniques for Urban Solar Mobility: Solar to Buildings, Vehicles, and Storage (S2BVS). *Sustainability* **2020**, *12*, 7035. [CrossRef]
21. Varma, R.K.; Rahman, S.A.; Vanderheide, T. New control of PV solar farm as STATCOM (PV-STATCOM) for increasing grid power transmission limits during night and day. *IEEE Trans. Power Deliv.* **2014**, *30*, 755–763. [CrossRef]
22. Hamid, M.I.; Jusoh, A.; Anwari, M. Photovoltaic plant with reduced output current harmonics using generation-side active power conditioner. *IET Renew. Power Gener.* **2014**, *8*, 817–826. [CrossRef]
23. Tomar, A.; Mishra, S.; Bhende, C.N. AOMH–MISO based PV–VCI irrigation system using ASCIM pump. *IEEE Trans. Ind. Appl.* **2018**, *54*, 4813–4824. [CrossRef]
24. Kannan, V.K.; Rengarajan, N. Investigating the performance of photovoltaic based DSTATCOM using I cos Φ algorithm. *Int. J. Electr. Power Energy Syst.* **2014**, *54*, 376–386. [CrossRef]
25. Panigrahi, R.; Mishra, S.K.; Srivastava, S.C.; Srivastava, A.K.; Schulz, N.N. Grid Integration of Small-Scale Photovoltaic Systems in Secondary Distribution Network—A Review. *IEEE Trans. Ind. Appl.* **2020**, *56*, 3178–3195. [CrossRef]
26. Mazumdar, J.; Harley, R.G. Recurrent neural networks trained with backpropagation through time algorithm to estimate nonlinear load harmonic currents. *IEEE Trans. Ind. Electron.* **2008**, *55*, 3484–3491. [CrossRef]
27. Mukundan, N.; Singh, Y.; Naqvi, S.B.Q.; Singh, B.; Pychadathil, J. Multi-Objective Solar Power Conversion System with MGI Control for Grid Integration at Adverse Operating Conditions. *IEEE Trans. Sustain. Energy* **2020**, *11*, 2901–2910.
28. Singh, B.; Arya, S.R. Back-Propagation Control Algorithm for Power Quality Improvement Using DSTATCOM. *IEEE Trans. Ind. Electron.* **2013**, *61*, 1204–1212. [CrossRef]
29. Jayachandran, J.; Sachithanandam, R.M. Neural Network-Based Control Algorithm for DSTATCOM Under Nonideal Source Voltage and Varying Load Conditions. *Can. J. Electr. Comput. Eng.* **2015**, *38*, 307–317. [CrossRef]
30. Janpong, S.; Areerak, K.L.; Areerak, K.N. A literature survey of neural network applications for shunt active power filters. *World Acad. Sci. Eng. Technol.* **2011**, *5*, 273–279.
31. Chaudhary, P.; Rizwan, M. QNBP NN-based I cos ϕ algorithm for PV systems integrated with LV/MV grid. *Soft Comput.* **2021**, *25*, 2599–2614. [CrossRef]
32. Chaudhary, P.; Rizwan, M. Intelligent approach-based hybrid control algorithm for integration of solar photovoltaic system in smart grid environment. *IET Smart Grid* **2019**, *2*, 445–454. [CrossRef]
33. Chaturvedi, D.K.; Malik, O.P.; Kalra, P.K. Generalised neuron-based adaptive power system stabiliser. *IEEE Proc. Gener. Transm. Distrib.* **2004**, *151*, 213–218. [CrossRef]
34. Chaturvedi, D.K. *Soft Computing Techniques and Its Applications in Electrical Engineering*; Springer: Berlin/Heidelberg, Germany, 2008.
35. Rizwan, M.; Jamil, M.; Kothari, D.P. Generalized Neural Network Approach for Global Solar Energy Estimation in India. *IEEE Trans. Sustain. Energy* **2012**, *3*, 576–584. [CrossRef]
36. Kulkarni, R.V.; Venayagamoorthy, G.K. Generalized neuron: Feedforward and recurrent architectures. *Neural. Netw.* **2009**, *22*, 1011–1017. [CrossRef]
37. Reisi, A.R.; Moradi, M.H.; Jamasb, S. Classification and comparison of maximum power point tracking techniques for photovoltaic system: A review. *Renew. Sustain. Energy Rev.* **2013**, *19*, 433–443. [CrossRef]
38. Almutairi, A.; Abo-Khalil, A.; Sayed, K.; Albagami, N. MPPT for a PV Grid-Connected System to Improve Efficiency under Partial Shading Conditions. *Sustainability* **2020**, *12*, 10310. [CrossRef]
39. Salah, C.B.; Ouali, M. Comparison of fuzzy logic and neural network in maximum power point tracker for PV systems. *Electr. Power Syst. Res.* **2011**, *81*, 43–50. [CrossRef]
40. Williams, R.J. Training recurrent networks using the extended Kalman filter. In Proceedings of the 1992 IJCNN International Joint Conference on Neural Networks, Baltimore, MD, USA, 7–11 June 1992; IEEE: New York, NY, USA, 2003; pp. 241–246.
41. de Oliveira, M.A. An application of neural networks trained with kalman filter variants (ekf and ukf) to heteroscedastic time series forecasting. *Appl. Math. Sci.* **2012**, *6*, 3675–3686.
42. Bishop, G.; Welch, G. An introduction to the kalman filter. *Proc. SIGGRAPH Course* **2001**, *8*, 23175–27599.
43. Chaudhary, P.; Rizwan, M. Hybrid control approach for PV/FC fed voltage source converter tied to grid. *Int. J. Hydrogen Energy* **2018**, *43*, 6851–6866. [CrossRef]



International Journal of Sustainable Energy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gsol20

Analysing the impact of various incentives on solar tariff in India

Abhishek Sharma, Priya Mahajan & Rachana Garg

To cite this article: Abhishek Sharma, Priya Mahajan & Rachana Garg (2021): Analysing the impact of various incentives on solar tariff in India, International Journal of Sustainable Energy, DOI: 10.1080/14786451.2021.1900184

To link to this article: https://doi.org/10.1080/14786451.2021.1900184

Published online: 29 Mar 2021.

Submit your article to this journal

Article views: 39

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gsol20

INTERNATIONAL JOURNAL OF SUSTAINABLE ENERGY
https://doi.org/10.1080/14786451.2021.1900184

Taylor & Francis
Taylor & Francis Group

Check for updates

Analysing the impact of various incentives on solar tariff in India

Abhishek Sharma, Priya Mahajan and Rachana Garg

Department of Electrical Engineering, Delhi Technological University, Delhi, India

ABSTRACT

In India, solar tariff has been witnessing a continuous fall since the announcement of revised targets for the National Solar Mission (NSM) in the year 2015. This paper discusses components that decide the solar tariff and demonstrates the role of various incentives being provided in bridging the tariff gap between solar and conventional power. The paper concludes with a comparison of various incentives (viability gap funding (VGF), accelerated depreciation (AD), generation-based incentive (GBI), etc.) based on the cost-effectiveness index. Based on the analysis, it is recommended that as levelised cost of energy (LCOE) is coming down, the government may think of pulling back some of the incentive schemes such as VGF, AD, and GBI while some hidden incentive schemes such as waiver of inter and/or intrastate transmission charges and losses, etc may continue.

ARTICLE HISTORY
Received 6 December 2020
Accepted 27 February 2021

KEYWORDS
Levelised cost of energy; solar tariff; viability gap funding; generation-based incentive; accelerated depreciation; cost-effectiveness index

Nomenclature

AD	accelerated depreciation
AT_{PV}	applicable tax
AWC_{PV}	annual working capital
APPC	average power purchase cost
C_{PV}	life cycle cost
CAPEX	capital expenditure
CC_{PV}	capital cost
CG_{PV}	contingency cost
CPSUs	central public sector undertakings
CUF	capacity utilisation factor
CW_{PV}	civil and general work cost
DF_{pvi}	discount factor
DoP	Department of power
DR_{PV}	discount rate
E_{PV}	lifetime energy production
FOR	forum of regulators
GBI	generation-based incentive
GNCTD	Government of National Capital Territory of Delhi
GoI	Government of India
IDC	Interest during construction
IOL_{PV}	Interest on loan
IWC_{PV}	Interest on working capital
JNNSM	Jawaharlal Nehru National Solar Mission
LCOE	levelised cost of energy
LC_{PV}	land cost
MC_{PV}	module cost
MOD	merit order dispatch

CONTACT Abhishek Sharma abhiabhi30@gmail.com

© 2021 Informa UK Limited, trading as Taylor & Francis Group

MOM_{PV}	monthly operation and maintenance expenses
MSS_{PV}	mounting and supporting structure cost
MS_{PV}	maintenance spares
NPV	net present value
NSM	National Solar Mission
O&M	operation and maintenance
PCU_{PV}	power conditioning unit cost
PoC	point of connection
PPC	power purchase cost
PV	photovoltaics
RE	renewable energy
REC_{PV}	receivable energy charge
RES-E	renewable energy sources
ROE_{PV}	return on equity
$RROE_{PV}$	rate of return on equity
R&M	repair and maintenance
VGF	viability gap funding
WACC	weighted average cost of capital

1. Introduction

In recent years, the major factor that has contributed to the decline of the solar tariff is the reduction in the cost of solar cells and other associated components, adoption of more efficient production techniques, government impetus through various incentives, and enhanced competition and scale of projects. Globally, in 2019, the solar PV market increased by an estimated 12% to around 115 GW, increasing the total capacity of solar PV installation to approximately 627 GW (REN21 2020). Figure 1 illustrates the progress of solar PV capacity globally from 2009 to 2019 (REN21 2020).

The growth of the solar PV market also brings economic and technical challenges to the grid due to the diurnal and stochastic nature of solar energy. This requires a revolutionary change in modern power systems development, which needs to incorporate better energy storage, intelligent communication networks, power electronic converters, demand response, and advanced controllers (Lai et al. 2017). Not many countries in the world are as susceptible to the effects of climate change as India, owing to its large agriculture-dependent population, its vast coastal lines, and the presence of the Himalayan region and islands, and this effect has become more pronounced due to rapid depletion of the ozone layer. Even today, coal is a predominant source of power generation and, therefore, a major source of carbon dioxide emissions in India. Solar energy holds great potential

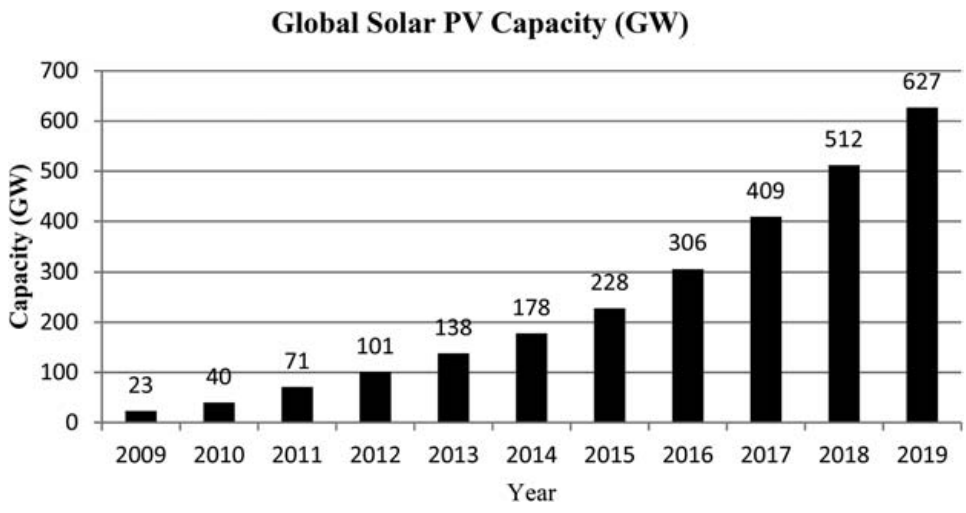


Figure 1. Progress of solar PV capacity globally from 2009 to 2019.

for providing a solution to the long-standing energy issues continuously faced by developing countries (Singh, Sood, and Deepak 2019). The everlasting search for achieving energy security through clean and sustainable energy source for a country like India can be best answered through the sun's energy that is abundantly available and also offers many benefits over the conventional energy sources, including no carbon dioxide production, a low impact on the environment and climate, no global warming emission, improved public health, in-exhaustive, and vast energy supply, creating more job opportunity, stability in energy prices, more resilient and reliable energy sources, etc. (Dawn et al. 2016).

As per the Paris accord on climate change, India pledged that by 2030, 40% of installed power generation capacity shall be based on clean sources. Accordingly, the Government of India (GoI) has set an ambitious target of achieving 175 GW of renewable energy generation by 2022. This includes 100 GW from solar, 60 GW from wind, 10 GW from bio-power, and 5 GW from small hydropower (RUMS 2018). With sunshine hours ranging between 2300 and 3200 per year, depending upon locations, India is well placed to harness its enormous solar potential and is, therefore, all set to comfortably achieve 100 GW of solar energy capacity by 2022 (Sharma 2011; Sharma and Chandel 2013; Ansari et al. 2013; MNRE 2018).

There is a high potential for the generation of renewable energy from various sources – wind, solar, biomass, small hydro, and cogeneration bagasse in India. The total potential for renewable power generation in the country as of 31 March 2018 is estimated at 1096 GW. This includes solar power potential of 749 GW (68.33%), wind power potential of 302 GW (27.58%) at 100 m hub height, small-hydro power potential of 19.75 GW (1.80%), Biomass power of 17.54 GW (1.60%), 5 GW (0.46%) from bagasse-based cogeneration in sugar mills, and 2.55 GW (0.23%) from waste to energy as shown in Figure 2. Furthermore, the distribution of the estimated potential of renewable power reveals that Rajasthan has the highest share of about 15% (162 GW), followed by Gujarat with 11% share (122 GW) and Maharashtra with 10% share (114 GW), mainly on account of solar power potential (MoSPI 2019).

As of 31 October 2020, India's total renewable energy installed capacity (excluding hydropower above 25 MW) had reached over 89.63 GW. During the last 6 years, India has witnessed the fastest rate of growth in renewable energy capacity addition among all large economies, with renewable energy capacity growing by 2.5 times and solar energy expanding by over 13 times. Renewable energy now constitutes over 24% of the country's installed power capacity and around 11.62% of the electrical energy generation. If large hydro is included, the share of renewable energy in electric installed capacity would be over 36% and over 26% of the electric energy generation. Around 49.59 GW renewable energy capacity is under installation, and an additional

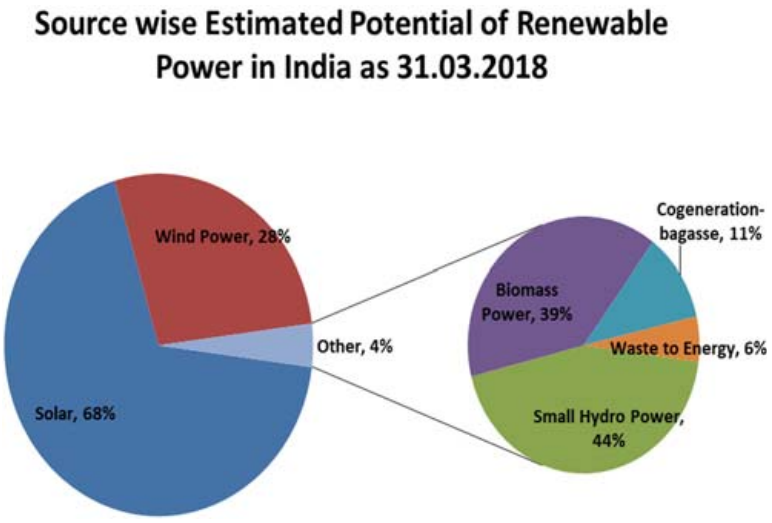


Figure 2. Source wise estimated potential of renewable power in India as on 31 March 2018 (MoSPI 2019).

27.41 GW capacity has been tendered. This makes the total capacity that is already commissioned and in the pipeline about 166.63 GW (PIB 2020). The paper (Raghuwanshi and Arya 2019) also describes the potential of RES and region-wise installed capacity in India. In order to regulate renewable energy tariffs, the levelised tariff of various technologies of generation including solar is determined by Central Electricity Regulatory Commission (CERC) with and without accelerated depreciation (AD). AD allows greater deductions in the earlier years of the life of an asset than the traditional straight-line method. In the year 2010–2011, when the National Solar Mission (NSM) was first announced, the levelised tariff for solar as issued by CERC stood at Rs. 17.91/kWh without AD and Rs. 14.95/kWh with AD for 25 years and it decreased by 68% to Rs. 5.68/kWh without AD and by 66% to Rs. 5.08/kWh with AD in 2016–2017. Moreover, the capital cost of solar PV projects has also reduced by 68% during this period (CERC 2010, 2012, 2015, 2016a) and competitive solar bids from solar projects fell from Rs. 12.16/kWh in the year 2010 to the levelised tariff of Rs. 3.30/kWh for 25 years in 2017 marking a whopping 73% drop (RUMS 2018).

In this paper, an attempt has been made to discuss various direct and indirect incentives available for the promotion of solar PV installation in India. The effect of various incentives on solar LCOE is also compared on the basis of the cost-effectiveness index. Although the study is done keeping in mind the solar tariff scenario in India however, the policymakers of other countries can take benefit of the outcome of this study during the formulation of policies and incentive schemes for their respective countries.

The paper proceeds as follows, analysis of the literature survey, and motivation for the study is carried in section 2. Key assumptions used in the study are discussed in section 3. After a discussion of various components of solar tariff in section 4, the determination of LCOE is carried in section 5. Sections 6 and 7 provide a comparison of solar and coal tariffs and reviews the role of various direct and indirect benefits in bridging the tariff gap respectively. Section 8 discusses the rating of incentives based on the cost-effectiveness index. Section 9 provides a discussion on the international financing support scenario. Section 10 provides the conclusion of the study along with recommendations for future work.

2. Literature survey and motivation for the study

For making solar energy a mainstream source of power generation, the Indian government has taken several initiatives in the last decade. Presently, the deployment of solar power projects in India is governed by both central and state governments under various schemes (Rohankar et al. 2016). The importance of government support can be judged by fact that China has got ahead of other countries in the solar industry due to the technological production of solar panels and governmental supports (Liu 2018).

Shrimali and Rohra reviewed NSM and provided an analysis of the Jawaharlal Nehru National Solar Mission (JNNSM) in the Indian institutional context – in particular, in the context of the power sector reforms. They identified the policy implementation challenges likely to be encountered in the case of grid-connected, roof-top, and off-grid applications as well as in the areas of research, development, and technology transfer (Shrimali and Rohra 2012).

Branker and Pathak have studied methodology for evaluation of the levelised cost of energy to determine the economic feasibility of PV projects. The paper reviews the methodology for precisely calculating the LCOE for solar PV, thereby correcting various misconceptions. Then, a template is provided for better reporting of LCOE results for PV that are needed to influence policy mandates or make investment decisions (Branker, Pathak, and Pearce 2011).

Several studies have pointed out many barriers acting as a hindrance to achieving the target set by NSM (Khanna et al. 2010; Shrimali and Nekkalapudi 2014; Hairat and Ghosh 2017; Palchak et al. 2017; Rathore et al. 2018) however, the central/state governments and regulatory bodies have taken many path-breaking initiatives like AD, VGF, GBI, etc. to address the

impediments to the NSM. These policy instruments led to a rapid growth in the Indian renewable energy sector, at over 20% during the period 2010–2017. The share of renewables in terms of total installed capacity is 18% and in terms of grid penetration, it is about 7% (Thapar, Sharma, and Verma 2018).

Papadelis, Stavrakas, and Flamos (2016) presented an approach to facilitate on-going assessments of the efficiency of support measures for renewable energy sources (RES-E) deployment and also demonstrated the proposed approach using the feed-in tariff (FIT) support policy in Greece as a case study.

In the paper Anagnostopoulos, Spyridaki, and Flamos (2017) studied the profitability reduction in various PV investments due to modified legislative and regulatory framework in Greece over the years from 2009 to 2014. To quantify the impacts, a fully parametric analysis was conducted. The analysis shows that the impact on the profitability of PV stations is largely mediated by parameters such as the size of the station and the PV activation date (affecting installation costs and respective profit margins) as well as the option of subsidy taking.

Tariff determination is a complex phenomenon and has not been elaborated in a way to understand its nuances in any of the studies conducted so far. Moreover, studies done so far are focused on the need for incentives for the growth of the Indian solar energy sector. Furthermore, none of these studies has pointed out the effect of various incentive schemes on solar LCOE. Another shortcoming in the literature surveyed so far is the absence of a comparison of various incentive schemes. Such type of comparisons is required for decision making by the policymakers, solar power developers, and other stakeholders involved.

Taking into account the above gaps in the existing literature, this paper highlights the various components of solar tariff. Furthermore, various incentives being provided by central and state governments for the promotion of solar PV installations in India are discussed and their effect on solar tariff has been analysed.

3. Key assumptions

The key assumptions taken during analysis are

(1) The GBI has been assumed to be disbursed in equal amounts at Rs. 2 per kWh for a generation of more than 1100 units in accordance with disbursal criteria notified in the solar policy of Government of National Capital Territory of Delhi (GNCTD) through its Department of Power (DoP) (DoP, GNCTD 2016).
(2) Disbursal of AD benefit at 40% for 1st year and at 80% for 2nd year onwards has been considered in accordance with the Ministry of New and Renewable Energy (MNRE) guidelines.
(3) To have maximum impact on the tariff VGF ceiling of Rs. 94 lakh/MW has been considered and it is assumed that the complete amount is disbursed in the initial year.
(4) To obtain power purchase cost (PPC) of thermal energy on year to year basis, PPC without renewable energy share has been taken from Forum of Regulators (FOR) report (FOR 2018). The genesis of assumption that PPC cost without RE significantly denotes thermal generation is that majority of the share of energy generation in India, even as on date (>62%) is coal generation (CEA 2020).
(5) The rating of various direct incentives and schemes in the cost-effectiveness index is indicative and depends on the values of parameters taken in analysis and are thus case-specific, for example, if the rate of depreciation, GBI, or the amount of initial VGF are changed, the results will be different.
(6) The value of capacity utilisation factor (CUF) is taken as 19% and the life of the plant is taken as 25 years as per CERC guidelines (CERC 2016a).

4. Components of solar tariff

In India, the tariff of electricity from solar PV plant is calculated in accordance with CERC regulation 'Terms and Conditions for Tariff determination from renewable energy sources regulations' (CERC 2010, 2012, 2015, 2016a). The tariff for solar photovoltaic (PV) plant depends on the following basic cost components

(1) Interest on loan (IOL_{PV}) capital
(2) Return on equity
(3) Depreciation
(4) Interest on working capital
(5) Operation and maintenance expenses

The values of each component as per the CERC tariff order 2016–2017 is given in Table 1 (CERC 2016a). The tariff period is for 25 years and the ratio of debt to equity is 70:30. Operation and maintenance (O&M) expenses comprise of repair and maintenance (R&M) expenses, establishment expenses (including employee expenses), and administrative and general expenses. These expenses are taken as Rs. 7.00 lacs with an escalation of 5.72% per annum. The depreciation of asset shall be allowed up to a maximum permissible value of 90% of the capital cost and salvage value shall be considered as 10%. Depreciation shall be chargeable from the first year of commercial operation and for part of the year it shall be charged on a pro-rata basis. The rates have been considered as 5.83% for the first 12 years and 1.54% from the twelfthth year onwards. The rate of interest on the loan is determined from the average base rate prevalent during the first six months of the

Table 1. Values of various solar PV project parameters.

Parameter	Details	Values
Financial assumptions	1. Tariff period	25 years
	2. Capital cost/MW	Rs. 530.02 lakh
	3. Debt: Equity	Rs. 371.02 lakh (@70%)
	a. Debt	Rs. 159.01 lakh (@ 30%)
	b. Equity	
O&M expenses	1. O&M expenses (2016–2017)	Rs. 7.00 lakh
	2. O&M expenses escalation	@5.72% per annum
Depreciation	1. Depreciation rate for first 12 years	@5.83%
	2. Depreciation rate 13th year onwards	@ 1.54%
Interest on term loan (IOL_{PV})	1. Loan repayment period is 12 years	@ 12.76%
Interest on working capital (IWC_{PV})	For fixed charges	
	1. O&M charges	1 month
	2. Maintenance spares (as % of O&M expenses)	15%
	3. Receivables for debtors	2 months
	4. For variable charges	13.26%
Return on equity (ROE_{PV})	1. (ROE_{PV}) for first 10 years	@20.00% *p*.a.
	2. (ROE_{PV}) 11th Year onwards	@24.00% *p*.a.

previous year plus 300 basis points, for example, during the first six months of FY 2015–2016, the base rate was 9.76% that has resulted in interest rate at 12.76% for FY 2016–2017. The RE tariff regulations provide a method for determination of annual working capital requirements of the PV projects as under (a) one-month O&M expenses;(b) two months receivables energy charges; (c) maintenance spares at 15% of O&M expenses.

The return on equity is taken as 20% per annum for the first 10 years, and 24% per annum from the eleventh year onwards.

Based on the parameters given in Table 1 (CERC 2016a), the solar tariff for the year 2016–2017 as issued by CERC can be calculated as given in Appendix B (Table B.1.). In recent years, the solar tariffs have fallen to new levels and thus, have re-invigorated the power sector of India. Table 2 is showing the trends of cost components and effective levelised tariff as issued by CERC from FY 2010 to FY 2017 (CERC 2010, 2012, 2015, 2016a). The levelised tariff is calculated by carrying out levelization for the 'useful life' of PV technology considering the discount factor for the time value of money. Discount factor denotes the factor by which a future cash flow must be multiplied to obtain present values.

It can be seen that capital cost has reduced from Rs. 1690 lakh/MW in the year 2010–2011 to Rs.530.02 lakh/MW in the year 2016–2017, a reduction of around 69%. It is mainly because of a decrease in the cost of PV modules in international markets over the years which form a major portion of the capital cost.

In the year 2016–2017, the PV module component share was around 62% of capital cost as can be seen in Figure 3 (CERC 2016b). The CERC annually reviews benchmark cost norms on the basis of comments of all stakeholders involved.

5. Determination of the levelised cost of energy

There are several methods to evaluate the economic viability of distributed generation projects. The capital cost of assets, the operation and maintenance costs, and the fuel costs must be considered in a systematic way so that a comparison can be made. One of the most commonly used metrics is the levelised cost of electricity (LCOE). LCOE aims to provide a comparison of various technologies with respect to different project sizes, lifetime, different capital cost, return, risk, and capacities. The LCOE can also be regarded as the minimum cost at which electricity must be sold to achieve break-even over the lifetime of the project (Lai and McCulloch 2017). LCOE can be calculated using Equation (1) (Branker, Pathak, and Pearce 2011; Lai and McCulloch 2017; Lai et al. 2017; Lai et al. 2019).

$$\mathrm{LCOE} = \frac{\text{life cycle cost}(C_{\mathrm{PV}})}{\text{lifetime energy production}(E_{\mathrm{PV}})}. \tag{1}$$

Table 2. Summary of CERC generic tariff orders for solar PV technologies from FY 2010 to FY 2017.

Period	FY 2010–2011	FY 2012 -2013	FY 2015–2016	FY 2016–2017
Capital cost (Rs. lakh/MW)	1690	1000	605.85	530.02
Discount rate	15.97%	10.62%	10.81%	10.64%
Depreciation rate for first 12 years	7.00%	5.83%	5.83%	5.83%
Depreciation rate 13th year onwards	1.33%	1.54%	1.54%	1.54%
Interest on working capital	12.89%	12.80%	13.50%	13.26%
O&M expenses (Rs. lakh/ Annum)	9.51	11	13	7
Levelised tariff (Rs./unit)	17.91	10.39	7.04	5.68
AD levelised benefit (Rs./unit)	2.96	1.04	0.69	0.60
Effective tariff (Rs./unit)	14.95	9.35	6.35	5.08

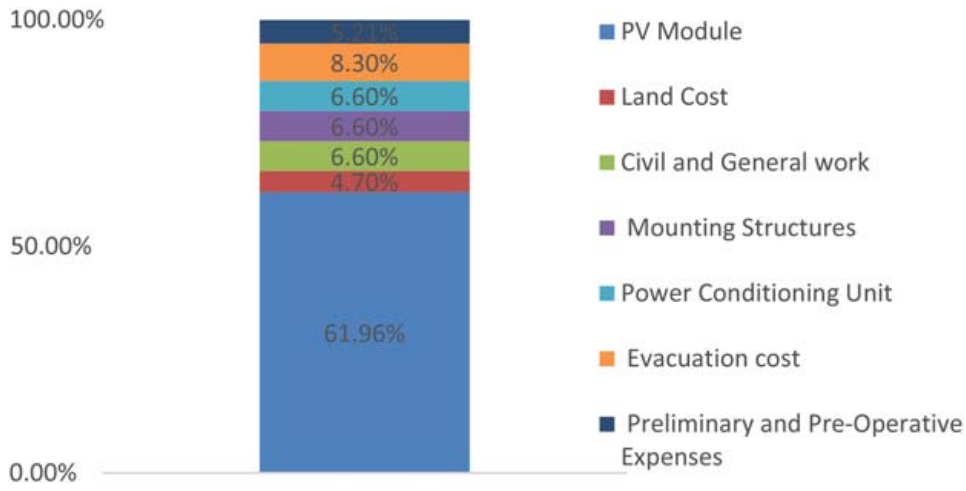


Figure 3. Bifurcation of components of the capital cost of solar PV based power projects for FY 2016–2017.

The levelised tariff is calculated by carrying out levellisation for the 'useful life' of each technology (25 years in the case of PV plants). The LCOE is dependent on the following factors and has been estimated using a spreadsheet (Microsoft Excel) based programme using Equations (2) to (6):

(a) Capital cost: The various components of capital cost and their weightage are shown in Figure 3 (CERC 2016b). It can be seen that capital cost (CC_{PV}) can be broadly divided into module cost and non-module cost components. PV module cost (MC_{PV}) forms a significant component of the overall capital cost of the PV based power plant and varies from 60% to 65% of overall capital cost and is highly dependent on PV module prices in the international market. The non-module cost component that typically varies from 30% to 40% of the overall capital cost comprises of power conditioning unit cost (PCU_{PV}), land cost (LC_{PV}), mounting and supporting structure cost (MS_{PV}), civil and general work cost (CW_{PV}), transportation/evacuation cost (TC_{PV}), and preliminary and pre-operative expenses including interest during construction (IDC) and contingency cost (CG_{PV}). The expression presenting capital cost is given by Equation (2).

$$CC_{PV} = MC_{PV} + LC_{PV} + CW_{PV} + MS_{PV} + PCU_{PV} + TC_{PV} + CG_{PV}. \quad (2)$$

(b) Discount rate (DR_{PV}): The present value is determined by using an assumed interest rate, usually referred to as a discount rate is a rate at which an investment's revenues and costs are discounted to calculate its present value and it is used to calculate discount factor. In India, in compliance with renewable energy (RE) tariff regulation, the discount factor is considered equal to the post-tax weighted average cost of the capital on the basis of normative debt: equity ratio (70:30) and is given by Equation (3).

i For the loan component, interest rate considered is 12.76%.

ii For the equity component, rate of return on equity ($RROE_{PV}$) is considered at post-tax ROE of 16%.

$$DR_{PV} = \{(LP_{PV} \times RIOL_{PV} \times (1\text{-}AT_{PV})\} + \{EP_{PV}x\ RROE_{PV}\}, \quad (3)$$

where LP_{PV} is the percentage at which loan is obtained for PV project, AT_{PV} is the applicable tax in % (current value is 34.61%), EP_{PV} is the equity percentage and $RIOL_{PV}$ is the rate of interest on loan capital. A realistic discount rate can also be determined using the Weighted Average Cost of Capital (WACC) that represents the weighted mix of debt and equity costs for the investor (Papadelis, Stavrakas, and Flamos 2016). The WACC is

considered as the overall combined effect of the cost of debt, cost of equity, share of CAPEX, and the corporate tax rate (Lai et al. 2019) and has a very high influence on LCOE.

The discount rate considered in India as per CERC comes out to be 10.64%.

(c) Annual working capital (AWC_{PV}): As per RE tariff regulations, the annual working capital requirements of the solar PV projects is given by Equation (4).

$$AWC_{PV} = MOM_{PV} + 2(REC_{PV}) + MS_{PV}, \tag{4}$$

where MOM_{PV} are monthly O&M expenses, REC_{PV} is receivable energy charge and MS_{PV} are maintenance spares (15% of O&M expenses).

d) Interest on working capital (IWC_{PV}): The interest on working capital is an interest rate equivalent to the average base rate prevalent during the first six months of the previous year plus 350 basis points for the period. In India, the average base rate of the State Bank of India is considered. During the first six months of FY 2015–2016, the base rate was 9.76% that has resulted in interest on working capital at 13.26% for FY 2016–2017. Therefore, the expression for calculating interest on working capital is given by Equation (5).

$$IWC_{PV} = 0.1326 \times AWC_{PV}. \tag{5}$$

5.1. Levelised cost of energy

The levelised tariff is calculated by carrying out levellisation for 'useful life' considering the discount factor for the time value of money. For the levelised tariff computation, the discount factor equivalent to post-tax WACC is considered. Therefore, LCOE can be calculated from Equation (1).

$$\text{LCOE} = \frac{C_{PV}}{E_{PV}}, \tag{6}$$

Since we are taking the example of a regulated plant whose total cost of the project and return on equity is regulated and reimbursed annually through fixed cost. Component of fixed cost is return on equity, interest on loan, depreciation, interest on working capital and operation, and maintenance expenses so the life cycle cost of the complete project shall be the net present value of these fixed cost/cash flows.

$$C_{PV} = \text{NPV}(FC_{PV1} + FC_{PV2} + \ldots\ldots\ldots + FC_{PV25}). \tag{7}$$

$$C_{PV} = \sum_{i=1}^{25} \{FC_{PVi}/\ (1+d)^i\ \}, \tag{8}$$

where d is the discount rate.

Similarly, lifetime energy production shall net present value of energy generated by plant annually (kWh).

$$E_{PV} = \text{NPV}\ (kWh_{PV1} + kWh_{PV2} + \ldots\ldots\ldots\ldots + kWh_{PV25}) \tag{9}$$

$$E_{PV} = \sum_{i=1}^{25} \{\ kWh_{PVi}/(1+d)^i\} \tag{10}$$

Therefore,

$$\mathrm{LCOE} = \frac{\sum_{i=1}^{25} \{FC_{PVi}/(1+d)^i\}}{\sum_{i=1}^{25} \{\ kWh_{PVi}/\ (1+d)^i\}}. \tag{11}$$

whereas $1/(1+d)^i$ is Discount factor, that is, $\mathrm{DF_{PVi}}$.

$$\mathrm{LCOE} = \frac{\sum_{i=1}^{25} \{(\mathrm{FC_{PVi}} \times \mathrm{DF_{PVi}})/\mathrm{kWh_{PVi}}\}}{\sum_{i=1}^{25} \{\mathrm{DF_{PVi}}\}}. \tag{12}$$

$$\mathrm{LCOE} = \sum_{i=1}^{25} \{\mathrm{CUE_{PVi}} \times \mathrm{DF_{PVi}})\}/\sum_{i=1}^{25} \{\mathrm{DF_{PVi}}\}, \tag{13}$$

where $\mathrm{CUE_{PV}}$ is the cost per unit of energy and $\mathrm{DF_{PV}}$ is the discount factor in a financial year.

The project details are placed in Appendix A and the sheet for determination of solar tariff for the year 2016–2017 is placed in Appendix B (Table B.1.). It can be observed from Appendix B that the total cost of generation is varying from Rs. 7.09 per unit to Rs. 3.7 per unit. The calculated discount rate is Rs. 10.64% that resulted in the levelised tariff of Rs 5.68 per unit for 25 years.

6. Coal tariff vs solar tariff

It is interesting to see the trend of power purchase cost from coal generation vis-à-vis solar PV generation. Tariffs from thermal energy are plant-dependent and mainly depends on coal quality and its mix, thus to get an idea of the PPC of thermal energy on year to year basis, PPC without renewable energy share has been taken for Delhi from forum of regulators (FOR) report as an illustration (FOR 2018). Table 3 depicts the variation of power purchase cost without renewable energy share (FOR 2018) and Figure 4 shows the variation of PPC without RE vis-à-vis generic CERC levelised cost of a solar PV plant with and without AD (FOR 2018; CERC 2010, 2012, 2015, 2016a).

From Table 3 (FOR 2018), it can be observed that power purchase cost is being steadily increasing from FY 12 to FY 17.

7. Bridging the tariff gap

Various direct and indirect factors have contributed to bridging the tariff gap, that is, to bring solar PV LCOE at par with PPC of thermal power (Chandrasekar and Kandpal 2005; Sarzynski, Larrieu, and Shrimal 2012; Lee, Hong, and Koo 2012; Yang et al. 2015). It is evident from Figure 4 (FOR 2018; CERC 2010, 2012, 2015, 2016a) that initially in the year 2012, the gap between generic solar tariff and PPC without RE was around Rs 7.77 that has narrowed to Rs. 2.09 in the year 2017. To bridge the tariff gap, incentivize solar power generation, and promote the setting up of a large number of solar power projects, policymakers have considered various incentives. The various incentives available have been categorized into direct/apparent benefits and indirect/hidden benefits. The following sections will explain the various alternatives and their effect on a generic tariff. The main aim of this paper is to study the effect of these benefits on solar LCOE. The flowchart of the methodology adopted for the study is shown below in Figure 5

Table 3. Variation of power purchase cost without renewable energy.

Year	FY 12	FY 13	FY 14	FY 15	FY 16	FY 17
PPC without RE (Rs./kWh)	2.62	2.79	2.97	3.16	3.37	3.59

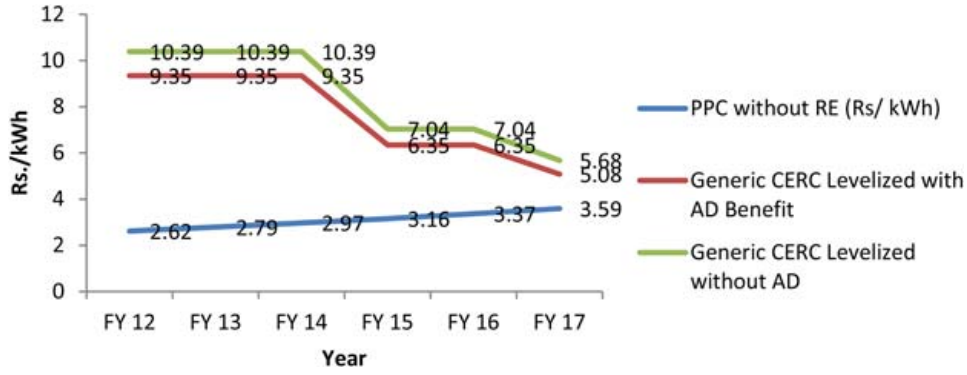


Figure 4. Variation of PPC without RE vis-à-vis generic CERC levelised cost with and without AD.

7.1. Effect of direct factors in bridging the tariff gap

The various direct factors that contributed in bridging the tariff gap are given below:

7.1.1. Viability gap funding (VGF) scheme

VGF scheme is one of the most popular schemes in the early stage of the promotion of PV plant. It is an economic instrument used by any government with the motive of supporting projects that are economically justified but are not financially viable. In the case of renewable energy projects, 'VGF' is the fraction of capital cost required to achieve the target LCOE. VGF is granted by the government through the reverse bidding process. The lowest bidder of the VGF amount is required to deliver output at a predefined cost. The scheme does not focus on long-term performance but rather is a direct financial incentive that requires upfront payments so; it helps in mitigating risk chances of lenders, reduction in tariff, and overall promotion of the sector. VGF payments are generally spread over few initial period of project (say 6 years), with an aim to elicit project performance. VGF amounts are decided through a competitive process, and the payments are not specifically linked

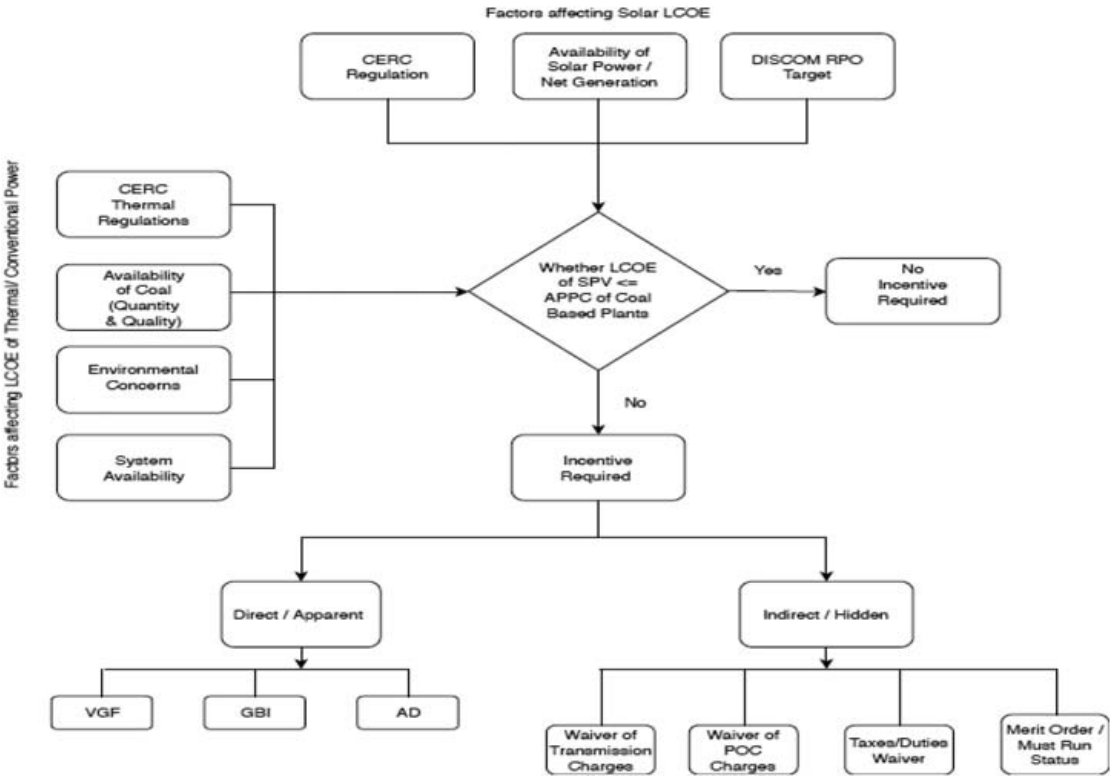


Figure 5. Flow chart of methodology adopted (APPC = average power purchase cost).

to the amount of power generated. However, to monitor project development, VGF amount is generally released in three tranches as follows:

- 25% amount at the time of delivery of 50% (at least) of the major equipment at the site.
- 50% amount on successful commissioning (full capacity) of the plant.
- 25% amount after one year of operation meeting requirements of generation.

The effect of the VGF scheme on LCOE has been shown in Figure 6. VGF ceiling of Rs. 94 lakh/MW has been considered and it is assumed that the complete amount is disbursed in the initial year itself so that it has maximum impact on tariff. VGF of Rs. 94 lakh has reduced LCOE from Rs. 5.68/unit to Rs. 5.43/unit. However, it can be argued that the same amount could be disbursed periodically based on performance or milestones achieved but then, the overall impact on tariff would be lesser.

7.1.2. Generation-based incentive (GBI) scheme

Generation-based incentives are provided to support small grid solar power projects connected to the distribution network under the solar GBI scheme (Akoijam and Krishna 2017). GBI scheme aims to reduce payback time and increase solar PV adoption. GBI is an output/performance-linked incentive scheme and solely depends on the number of units generated over and above a minimum value. Under the scheme for GBI in solar energy, a fixed amount per unit of energy generated is provided to support small grid solar power projects connected to the distribution grid below 33 kV. This financial incentive is not upfront and does not burden the government like the VGF fund. Since this incentive is linked with the generation output of the plant, it encourages proper operation of the plant and promotes the use of quality solar generation equipment.

GBI is basically a tail end incentive scheme that provides tariff comfort and therefore offers meagre a financial advantage. However, it can be designed in such a way that if the GBI amount is paid to the concerned utility then it will motivate utility to buy more and more solar power and also help them to make timely payments to generators.

The effect of the GBI scheme on LCOE has been shown in Figure 7. GBI benefit at Rs. 2 per kWh for a generation of more than 1100 units has been considered. The maximum limit of generation for GBI disbursement has been capped to 1500 units. The GBI benefit is applicable for the first 8 years of operation. Taking the base case of CERC generic tariff for the year 2016–2017, the GBI benefit has resulted in decreasing LCOE to Rs. 4.59 from base LCOE of Rs. 5.68.

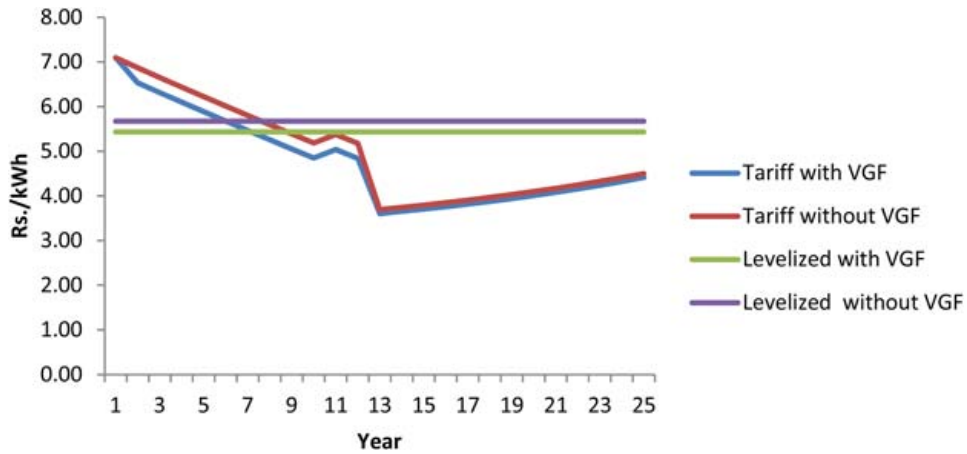


Figure 6. Effect of VGF on solar tariff.

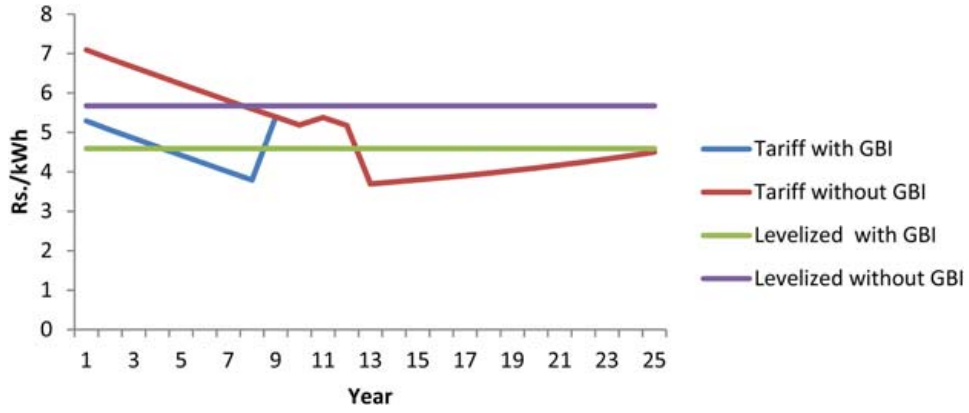


Figure 7. Effect of GBI on solar tariff.

7.1.3. Accelerated depreciation (AD)

In AD, the depreciation is deliberately allowed more in the initial years of its useful life (say 40% in the first year), as the vendor has to pay lesser taxes if the asset gets depreciated earlier. Thus, accelerated depreciation accounts for major relief in the upfront cost of PV by providing a tax break in the first year of operation. This method does not have any substantial or direct impact on the government although the government loses some money in terms of taxes. Although this method has the ability to attract large investors, the limitation with this method is that it is not linked with the operational performance of the plant. However, it can be designed to link with the plant's performance (e.g. a bank guarantee may be deposited by the vendor who wants to avail AD benefit, which can be released periodically on a minimum performance basis annually).

The effect of AD on LCOE has been shown in Figure 8. AD benefit at 40% for first year and @ 80% for second year onwards has been considered. AD benefit has resulted in decreasing LCOE to Rs. 5.14 from base LCOE of Rs. 5.68.

7.1.4. Combined effect of AD, VGF, and GBI benefit on solar tariff

A typical case is considered where the effect of already discussed benefits viz. AD, GBI, and VGF are considered on the solar tariff. The effect of AD, GBI, and VGF on solar tariff has been shown in Figure 9 and it can be observed that levelised tariff after considering the combined effect of VGF, GBI, and AD is Rs. 3.81/- against baseline levelised tariff of Rs. 5.68/-.

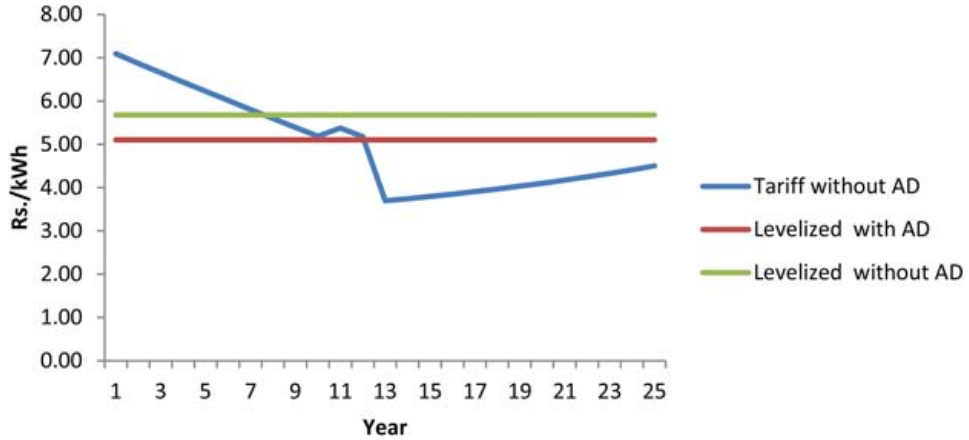


Figure 8. Effect of AD on solar tariff.

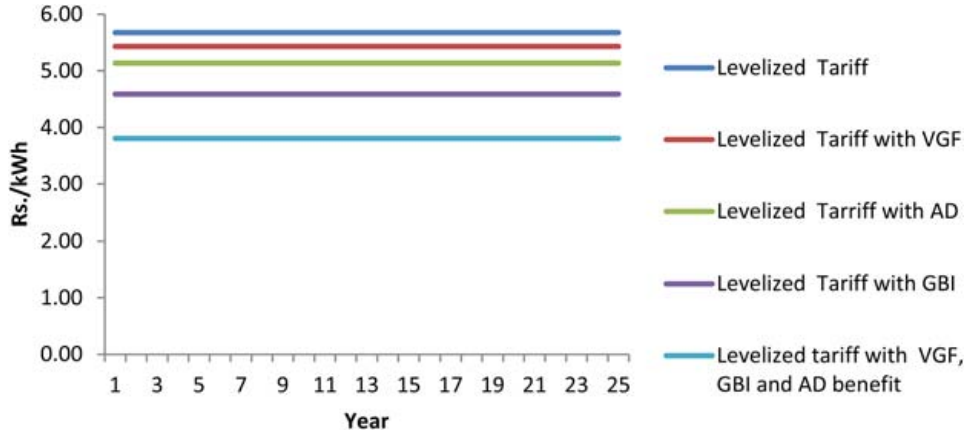


Figure 9. Effect of AD, VGF and GBI on solar tariff.

7.2. Contribution of indirect factors in bridging the tariff gap

Besides various direct factors explained above, there are some indirect factors that have given impetus to the solar sector and helped in bringing solar PV LCOE at par with PPC of thermal power. The various hidden alternatives available are waiver of transmission losses and point of connection charges, import duty waiver on imported PV modules, exclusion of solar generation from merit order dispatch/must-run Status, and various government polices especially the announcement of NSM. The following sections will explain some of these alternatives that have resulted in a further decrease in the solar tariff.

7.2.1. Waiver of transmission losses and point of connection (PoC) charges

Ministry of Power, GoI has been promoting the solar PV sector by incentivizing and thereby making the cost of electricity from these plants at par with that of conventional thermal generation plants, which are still the base load plants. Table 4 shows some of the latest discovered rates of electricity from PV parks (RUMS 2018). If we take a closer look at tariff rates discovered through competitive bidding for the solar park, it can be seen that levelised tariff has come down as low as Rs. 2.97/– for the Rewa plant for the first year and a levelised tariff of Rs. 3.30/– for 25 years.

In the case of the Rewa solar park project, solar generation at Rewa, Madhya Pradesh is injected into Delhi (withdrawal point) for utilisation to Delhi Metro Rail Corporation (DMRC) network. Recently, Power System Operation Corporation Limited (POSOCO), GoI enterprise has issued an amendment to sharing regulations for Inter-state Transmission System (ISTS) charges and losses and waived ISTS charges and losses otherwise DMRC may have to bear an additional nearly additional 91 paise per unit cost duty to ISTS charges and losses (POSOCO 2020). That would have resulted in an increase of tariff from Rs. 3.30 per unit to nearly Rs. 4.21 per unit. Thus, it can be said that waiver of ISTS charges and losses played a key role in reducing the tariff rates.

7.2.2. Exclusion from merit order dispatch/must-run status

Solar power projects enjoy must-run status, that is, DISCOMs must procure electricity from solar power plants even if that results in partial scheduling or complete shutdown of conventional thermal power plants. Since CUF of a typical solar PV project in India is only 15–19%, if scheduling is

Table 4. Recent rates discovered through competitive bidding in India.

Solar park/plant	Capacity (MWp)	Levelised tariff for 25 years (Rs./ kWh)
Rewa PV	750	3.30 (Rs. 2.97 for first year)
Bhadla PV	420	4.34
Pavagada PV	500	4.78

capped for these plants on basis of merit, the CUF shall go even down and shall increase LCOE as a total annual generation will reduce. Thus, by providing an exclusion from merit order dispatch, it is ensured that rates of solar generation are kept under check.

7.2.3. Government policies

Government policies have played a crucial role in bringing down solar tariffs. GoI has announced its commitment to achieve 40% of generation by 2030 through RE and accordingly initial target of 20 GW solar energy under NSM in 2010, was revised to 100 GW by 2022 in 2015. At the time of the launch of NSM in 2010, the country had only 160 MW solar generation and that has increased to 24 GW till September 2018. Several schemes namely (i) Defence scheme (ii) Central Public Sector Undertakings (CPSUs) scheme (iii) Bundling scheme (iv) Canal bank/Canal top scheme (v) VGF Scheme (vi) Solar park scheme (vii) Solar rooftops have been initiated/launched by the Ministry under NSM which are under implementation. The effect of government push can be gauged from the fact that tariff in one recent bid came down to Rs. 3.30/- per unit, whereas that approved by the CERC for solar power was about Rs. 15.39 per unit in 2010–2011 when the mission was launched.

8. Cost-effectiveness index

One of the possible ways to rate the direct incentive scheme as per some criteria is through the cost-effectiveness index that is basically the ratio of cost reduction to consumers and the total cost to government (both direct and indirect) as can be seen in the last column of Table 5. The LCOE (column c) is derived from Appendix B (Table B.2.) for respective schemes. Reduction in LCOE (column d) is considered over the baseline LCOE of Rs. 5.68. The cost reduction for the consumer over a period of 25 years (column e) is considered. The direct cost to the government is the net present value of the amount disbursed in the GBI and VGF scheme while the tax loss due to the provision of AD benefit will comprise indirect cost to the government (column f & g). The total cost to the government will then be the sum of direct and indirect cost (column h). Cost-effectiveness index is the ratio of cost reduction for consumers to the total cost to the government, hence, the higher the ratio, the better the financial effectiveness of that policy option.

From the study, it can be concluded that GBI is scoring high among VGF, GBI, and AD. Initially, we have taken the case of disbursal of GBI for the initial eight years but even if the analysis is carried for a more practical duration of the initial two years, the effect on the cost-effectiveness index is negligible. The analysis given above mainly depends on the values of parameter taken in analysis

Table 5. Cost-effectiveness index considering direct factors.

Calculation of cost-effectiveness Index for AD, GBI and VGF								
1000 kW solar plant (a)	Capital expenditure (CAPEX) (b)	LCOE (c)	Reduction in LCOE from baseline (d)	Cost reduction for consumer over lifetime (25 yrs.) (e)	Direct cost to Govt. (f)	Indirect cost to Govt. (income tax loss due to AD benefit) (g)	Total cost to Govt. (direct & indirect) (h)	Cost-effectiveness index (i)
Units	Rs. Lakh	Rs./ kWh	Rs./kWh	Rs. lakh (Net Present Value)	Rs. lakh (Net Present Value)	Rs. lakh (Net Present Value)	Rs. lakh (Net Present Value)	Rs. lakh (Net Present Value)
Baseline	530.02	5.68	–	–	–	–	–	–
AD	530.02	5.14	0.54	224.48	0.00	154.82	154.82	1.45
GBI- 2 Yrs.	530.02	5.32	0.36	147.82	57.00	0	57.00	2.59
GBI- 8 Yrs.	530.02	4.59	1.09	452.08	173.03	0	173.03	2.61
VGF	530.02	5.43	0.24	101.11	94.00	0	94.00	1.08

and are thus case-specific, for example, if the rate of depreciation or the amount of initial VGF is changed, the results will be different.

9. Economics of solar energy: international scenario

With the solar market becoming more and more competitive and subsidy-independent, financiers are becoming more comfortable with solar market risks. India and China have a strong and sophisticated domestic financing ecosystem, capable of meeting their investment requirements while countries such as Kenya or Myanmar do not have such a strong domestic financing ecosystem, therefore, they rely more on international financing for scaling up their solar growth. Germany, the USA, and China relied predominantly on domestic financing, India is moving from international to domestic financing. The solar market in countries such as Kenya, Nigeria, Ethiopia, and Myanmar is based on a higher share of international finance and in Mexico domestic commercial banks provide most of the debt to both utility-scale and rooftop solar projects and a large part of the equity financing is provided from international developers (TFE Consulting 2018).

Among all renewable power generation technologies, PV has experienced the most rapid cost declines, the LCOE of utility-scale solar PV projects (>1 MW) commissioned in 2018 fell to USD 85/MWh, 13% lower than in 2017 mainly as module prices have fallen by more than 90% since 2010. The variation of LCOE among various regions in 2018 is shown in Figure 10 (REN21 2019).

In India, the country-average LCOE is around USD 0.06/kWh and it excludes the impact of government support policies through various incentives/schemes, if considered, that will result in further lowering of LCOE.

10. Conclusion

This study discusses the role of various direct potential incentives like VGF, AD, and GBI and indirect benefits like waiver of transmission losses and point of connection charges, exclusion of solar generation from merit order dispatch/must-run status and various government policies etc. in bridging the tariff gap between solar and conventional power. The summary of results achieved after consideration of various direct incentives is given in Table B.2 (Appendix B). The paper also proposed a criterion to rate the direct incentive scheme through the cost-effectiveness index. Various direct incentive schemes have been compared based on assumptions made, and it has been observed from the cost-effectiveness index matrix that GBI is scoring high among VGF, GBI, and AD (refer to Table 6). Despite the fact that results are case-specific as the decrease in value of LCOE from the base value of Rs. 5.68/kWh mainly depends on the initial value of VGF taken, rate, and duration of

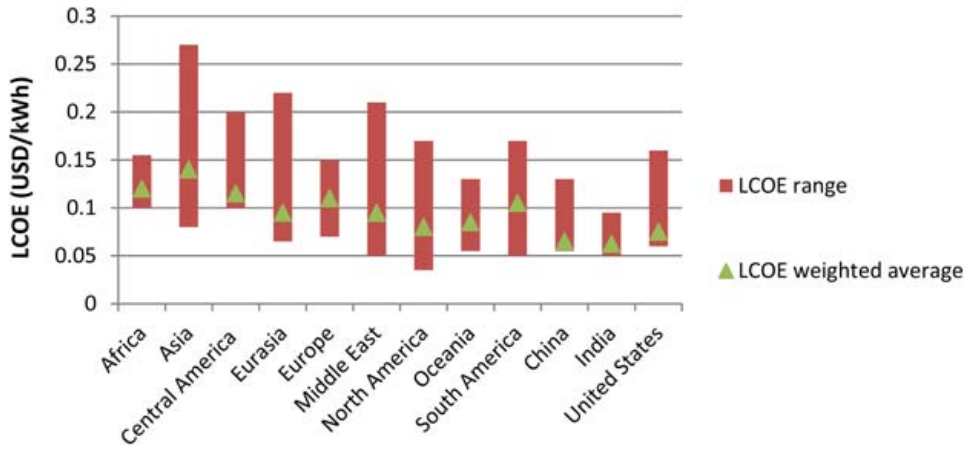


Figure 10. Variation of LCOE in different regions in 2018.

disbursal of GBI, and on the rate of depreciation assumed during analysis, the results strongly indicate the role of various incentives and schemes in narrowing the tariff gap. It has been felt that on one hand, the solar industry is facing the challenge of how to bring LCOE down to keep consumers and discom's interested, and on the other hand, it has to take steps to safeguard the interest of investors. LCOE alone is not the major factor for the adoption of solar PV power, financial indicators such as IRR, NPV, and ROE are the basic indicators of profitability of the project. Distribution companies in India, which are already under financial stress foresees the high cost of electricity from solar PV as the prime deterrent for adoption. In the past few years, the Indian government has rightly targeted the right tool of directly subsidising the technology through VGF, AD, and GBI, etc. to bring it at par with conventional generation tariffs. This has led to a decrease in the LCOE of the solar PV projects for the consumer and thus paved the way for its faster adoption and at the same time from the investor perspective, CERC provides a high rate of return on equity of 20% per annum for the first 10 years, and 24% per annum from the 11th year onwards for solar PV which is one of the highest in the sector. A high rate of ROE not only means more profitability for the investor and but it also brings more investment in the sector. Solar power capacity has increased by more than 14 times in the last five years from 2630 MW to 37505 MW in December 2019 (MNRE 2020). Moreover, with enhanced installation, more module manufacturers have entered the fray resulting in a substantial reduction of module cost and with the reduction of module cost, LCOE of the projects has come down drastically and, direct and indirect government subsidies are resulting in over subsidising the projects.

Based on this study, it is recommended that as LCOE is coming down, the government may think of pulling back some of the direct incentive schemes and let the PV projects compete in the market without any kind of economic stimuli. Furthermore, some hidden incentive schemes such as waiver of inter and/or intrastate transmission charges and losses, waiver of import duty on modules, the bestowment of must-run status or exemption from merit order dispatch (MOD) are still needed for the sustenance of solar PV plants. The results of the study have further demonstrated that AD has no linkage to operational performance, as the incentive is due to investments being made. Similar to AD, VGF is also a capacity linked subsidy and does not focus on long-term performance. While, GBI is an output/performance-linked incentive and hence has very limited possibilities of misuse. With a 175 GW renewable energy generation target by the year 2022 insight, The GoI has an enhanced target for producing renewable energy to 450 GW by 2030 (IEA 2020). Therefore, the government may further consider introducing more incentives tools like GBI apart from non-financial support options to further scale higher solar targets. GBI can further redesigned to be very effective to incentivize discoms to buy renewable energy and bring down their cost of RE procurement. Given solar energy's sustainability and cost-efficiency, its inclusiveness and scalability are going to be the fulcrum on which its wider acceptance as a source of power is going to be based.

Furthermore, MNRE is providing various incentives and schemes for grid-connected rooftop PV systems, solar water pumping systems, and solar parks. A similar analysis can be conducted on the economics of these systems as a part of future work. Given the quest for a sustainable 24×7 supply solution for a country like India, the effect of energy storage on the levelised cost of energy for photovoltaic systems is also a topic of considerable importance for future research.

Disclosure statement

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

References

Akoijam, A. S., and V. V. Krishna. 2017. "Exploring the Jawaharlal Nehru National Solar Mission (JNNSM): Impact on Innovation Ecosystem in India." *Institution of Electronics and Telecommunication Engineers Technical Review* 9 (5): 573–585.

Anagnostopoulos, P., N. A. Spyridaki, and A. Flamos. 2017. "A "New-Deal" for the Development of Photovoltaic Investments in Greece?" *A Parametric Techno-Economic Assessment. Energies* 10: 1173.
Ansari, M. F., R. K. Kharb, S. Luthra, S. L. Shimmi, and S. Chatterji. 2013. "Analysis of Barriers to Implement Solar Power Installations in India Using Interpretive Structural Modeling Technique." *Renewable and Sustainable Energy Reviews* 27: 163–174.
Branker, K., M. J. M. Pathak, and J. M. Pearce. 2011. "A Review of Solar Photovoltaic Levelized Cost of Electricity." *Renewable and Sustainable Energy Reviews* 15: 4470–4482.
Central Electricity Authority. 2020. "All India Installed Capacity (in MW) of Power Stations." Accessed July 1, 2020. https://cea.nic.in/reports/monthly/installedcapacity/2020/installed_capacity-03.pdf.
Central Electricity Regulatory Commission. 2010. "Determination of Levelized Generic Tariff for FY 2010-11." Accessed July 1, 2020. http://cercind.gov.in/2010/ORDER/February2010/53-2010_Suo-Motu_RE_Tariff_Order_FY2010-11.pdf.
Central Electricity Regulatory Commission. 2012. "Determination of Levelized Generic Tariff for FY 2012–13." Accessed July 1, 2020. http://cercind.gov.in/2012/orders/RE_35_2012.pdf.
Central Electricity Regulatory Commission. 2015. "Determination of Levelized Generic Tariff for FY 2015–16." Accessed July 1, 2020. https://cercind.gov.in/2015/orders/SO4.pdf.
Central Electricity Regulatory Commission. 2016a. "Determination of Levelized Generic Tariff for FY 2016–17." Accessed July 1, 2020. https://www.cercind.gov.in/2016/orders/SORE.pdf.
Central Electricity Regulatory Commission. 2016b. "Determination of Benchmark Capital Cost Norm for Solar PV Power Projects and Solar Thermal Power Projects Applicable during FY 2016-17." Accessed July 1, 2020. http://www.cercind.gov.in/2016/orders/so17.pdf.
Chandrasekar, B., and T. C. Kandpal. 2005. "Effect of Financial and Fiscal Incentives on the Effective Capital Cost of Solar Energy Technologies to the User." *Solar Energy* 78 (2): 147–156.
Dawn, S., P. K. Tiwari, A. P. Goswami, and M. K. Mishra. 2016. "Recent Developments of Solar Energy in India: Perspectives, Strategies and Future Goals." *Renewable and Sustainable Energy Reviews* 62: 215–235.
Department of Power, Government of National Capital Territory of Delhi. 2016. "Delhi Solar Policy." Accessed July 1, 2020. http://web.delhi.gov.in/wps/wcm/connect/38d9f3804dab96b0a615ee998ec321bd/Solar.pdf?MOD=AJPERES&lmod=1149868863&CACHEID=38d9f3804dab96b0a615ee998ec321bd.
Forum of Regulators. 2018. "Assessment of Achievable Potential of New and Renewable Energy Resources in Different States during 12th Plan Period and Determination of RPO Trajectory and Its Impact on Tariff." Accessed July 1, 2020. https://www.forumofregulators.gov.in/Data/Reports/Final_Report_FOR_RPO_Study.pdf.
Hairat, M. K., and S. Ghosh. 2017. "100 GW Solar Power in India by 2022 – A Critical Review." *Renewable and Sustainable Energy Reviews* 73: 1041–1050.
IEA (International Energy Agency). 2020. "India 2020 Energy Policy Review." Accessed February 1, 2021. https://niti.gov.in/sites/default/files/2020-01/IEA-India%202020-In-depth-EnergyPolicy_0.pdf.
Khanna, A., G. Sargsyan, N. Kulichenko, C. Subramaniam, and A. Garg. 2010. "Report on Barriers for Solar Power Development in India." *South Asia Energy Unit, Sustainable Development Department, The World Bank*, Washington, DC. Accessed July 1, 2020. https://www.esmap.org/sites/esmap.org/files/The%20World%20Bank_Barriers%20for%20Solar%20Power%20Development%20in%20India%20Report_FINAL.pdf.
Lai, C. S., Y. Jia, L. L. Lai, Z. Xu, X. Li, J. Cao, and M. D. McCulloch. 2017. "Levelized Cost of Electricity for Photovoltaic/Biogas Power Plant Hybrid System with Electrical Energy Storage Degradation Costs." *Energy Conversion and Management* 153: 34–47.
Lai, C. S., Y. Jia, L. L. Lai, Z. Xu, M. D. McCulloch, and K. P. Wong. 2017. "A Comprehensive Review on Large-scale Photovoltaic System with Applications of Electrical Energy Storage." *Renewable and Sustainable Energy Reviews* 78: 439–451.
Lai, C. S., G. Locatelli, A. Pimm, Y. Tao, X. Li, and L. L. Lai. 2019. "A Financial Model for Lithium-Ion Storage in a Photovoltaic and Biogas Energy System." *Applied Energy* 251: 113179.
Lai, C. S., and M. D. McCulloch. 2017. "Levelized Cost of Electricity for Solar Photovoltaic and Electrical Energy Storage." *Applied Energy* 190: 191–203.
Lee, M., T. Hong, and C. Koo. 2012. "An Economic Impact Analysis of State Solar Incentives for Improving Financial Performance of Residential Solar Photovoltaic Systems in the United States." *Renewable and Sustainable Energy Reviews* 16: 6317–6332.
Liu, Z. 2018. "What is the Future of Solar Energy? Economic and Policy Barriers." *Energy Sources, Part B: Economics, Planning, and Policy* 13 (3): 1–4.
Ministry of New and Renewable Energy. 2018. "Renewable Energy News." Accessed July 1, 2020. https://mnre.gov.in/img/documents/uploads/670406a017f54c9386fcde911ee5abe6.pdf.
Ministry of New and Renewable Energy. 2020. "Annual Report 2019–2020." Accessed July 1, 2020. https://mnre.gov.in/img/documents/uploads/file_f-1585710569965.pdf.
MoSPI (Ministry of Statistics and Programme Implementation). 2019. "Energy Statistics 2019." Accessed February 1, 2021. http://mospi.nic.in/sites/default/files/publication_reports/Energy%20Statistics%202019-finall.pdf?download=1.

Palchak, D., J. Cochran, R. Deshmukh, A. Ehlen, R. Soonee, S. Narasimhan, M. Joshi, et al. 2017. "Greening the Grid: Pathways to Integrate 175 gigawatts of Renewable Energy into India's Electric Grid, Vol. I–National Study. Accessed July 1, 2020. https://www.nrel.gov/docs/fy17osti/68530.pdf.

Papadelis, S., V. Stavrakas, and A. Flamos. 2016. "What do Capacity Deployment Rates Tell Us about the Efficiency of Electricity Generation from Renewable Energy Sources Support Measures in Greece?" *Energies* 9: 38.

Power System Operation Corporation Limited. 2020. "Amendment to Sharing Regulations 2020 in line with Ministry of Power Order dated 05.08.2020." Accessed November 30, 2020.

Press Information Bureau. 2020. "MNRE: Year End Review-2020." Accessed February 1, 2021. https://pib.gov.in/Pressreleaseshare.aspx?PRID=1685046.

Raghuwanshi, S. S., and R. Arya. 2019. "Renewable Energy Potential in India and Future Agenda of Research." *International Journal of Sustainable Engineering*, 6:1–12.

Rathore, P. K. S., S. Rathore, R. P. Singh, and S. Agnihotri. 2018. "Solar Power Utility Sector in India: Challenges and Opportunities." *Renewable and Sustainable Energy Reviews* 81: 2703–2713.

REN21. 2019. "Renewables in Cities 2019 Global Status Report." Accessed July 1, 2020. https://www.ren21.net/wp-content/uploads/2019/05/gsr_2020_full_report_en.pdf.

REN21. 2020. "Renewables 2020 Global Status Report." Accessed July 1, 2020. https://www.ren21.net/wp-content/uploads/2019/05/gsr_2020_full_report_en.pdf.

Rohankar, N., A. K. Jain, O. P. Nangia, and P. Dwivedi. 2016. "A Study of Existing Solar Power Policy Framework in India for Viability of the Solar Projects Perspective." *Renewable and Sustainable Energy Reviews* 56: 510–518.

RUMS (Rewa Ultra Mega Solar Ltd.). 2018. "Rewa: A Case Study." Accessed July 1, 2020. https://niti.gov.in/writereaddata/files/Manu%20Srivastava.pdf.

Sarzynski, A., J. Larrieu, and G. Shrimal. 2012. "The Impact of State Financial Incentives on Market Deployment of Solar Technology." *Energy Policy* 46: 550–557.

Sharma, A. 2011. "A Comprehensive Study of Solar Power in India and World." *Renewable and Sustainable Energy Reviews* 15: 1767–1776.

Sharma, V., and S. S. Chandel. 2013. "Performance Analysis of a 190 kWp Grid Interactive Solar Photovoltaic Power Plant in India." *Energy 2013* 55: 476–485.

Shrimali, G., and V. Nekkalapudi. 2014. "How Effective has India's Solar Mission Been in Reaching its Deployment Targets?" *Economic Political Weekly* 49 (42): 54–63.

Shrimali, G., and S. Rohra. 2012. "Review Paper, India's Solar Mission: A Review." *Renewable and Sustainable Energy Reviews* 16 (8): 6317–6332.

Singh, A. D., B. Y. R. Sood, and C. Deepak. 2019. "Recent Techno-economic Potential and Development of Solar Energy Sector in India." *Institution of Electronics and Telecommunication Engineers Technical Review* 36 (3): 223–224.

TFE Consulting. 2018. "India's Solar Leap – Financing a Mature Market." Accessed July 1, 2020. https://www.tfeconsulting.com/_website/wp-content/uploads/2017/12/TFE_Report-SolarFinancing_web.pdf.

Thapar, S., S. Sharma, and A. Verma. 2018. "Analysing Solar Auctions in India: Identifying Key Determinants." *Energy for Sustainable Development* 45: 66–78.

Yang, D., H. Latchman, D. Tingling, and A. A. Amarsingh. 2015. "Design and Return on Investment Analysis of Residential Solar Photovoltaic Systems." *Institute of Electrical and Electronics Engineers Potentials* 34: 11–15.

Appendices
Appendix A

Table A. Project summary sheet.

Project capacity (MW)	Life of project (years)	Capacity utilisation factor of the project (%)	Annual energy production (lakh kWh)	Total project cost (Rs. lakh)	Equity to be invested (Rs. lakh)	Loan component (Rs. lakh)	Loan repayment period (years)	Return on equity for 1–10 years (A.)	Return on equity for 11–25 years (A.)	Depreciation till loan repayment (%)	Depreciation after loan repayment (%)	Total O&M expenses for 1st year (Rs. lakh)	Escalation in O&M expenses (%)	Discount rate (%)
1	25	19	16.644	530.02	159	371	12	20%	24%	5.83%	1.54%	7	5.72%	10.64%

Appendix B

Table B.1. Tariff Calculation Summary sheet.

	Years																								
Tarrif calculations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Energy Sold	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6	16.6
O&M Expenses	7.0	7.4	7.8	8.3	8.7	9.2	9.8	10.3	10.9	11.5	12.2	12.9	13.6	14.4	15.3	16.1	17.0	18.0	19.1	20.1	21.3	22.5	23.8	25.2	26.6
Depriciation	30.9	30.9	30.9	30.9	30.9	30.9	30.9	30.9	30.9	30.9	30.9	30.9	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2	8.2
Interest on Term Loan	45.53	41.6	37.6	33.7	29.8	25.8	21.9	17.9	14.0	10.0	6.1	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Interest on Working Capital	2.82	2.62	2.55	2.47	2.40	2.33	2.26	2.19	2.12	2.05	2.13	2.07	1.53	1.56	1.59	1.62	1.65	1.69	1.73	1.76	1.80	1.85	1.89	1.94	1.99
Return on Equity	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2
Total Fixed Cost	118.1	114.3	110.7	107.1	103.6	100.1	96.6	93.1	89.7	86.3	89.5	86.2	61.5	62.3	63.2	64.1	65.0	66.0	67.1	68.2	69.4	70.7	72.0	73.4	74.9
Cost per Unit of Electricity	7.09	6.87	6.65	6.44	6.22	6.01	5.80	5.60	5.39	5.19	5.38	5.18	3.70	3.74	3.80	3.85	3.91	3.97	4.03	4.10	4.17	4.25	4.33	4.41	4.50
Tariff with VGF	7.1	6.5	6.3	6.1	5.9	5.7	5.5	5.3	5.1	4.9	5.0	4.8	3.6	3.7	3.7	3.8	3.8	3.9	3.9	4.0	4.1	4.2	4.2	4.3	4.4
Tariff with GBI	5.29	5.07	4.85	4.63	4.42	4.21	4.00	3.79	5.39	5.19	5.38	5.18	3.70	3.74	3.80	3.85	3.91	3.97	4.03	4.10	4.17	4.25	4.33	4.41	4.50
Discount Rate for Net present Value	0.1																								
Discount Factor	1.00	0.90	0.82	0.74	0.67	0.60	0.55	0.49	0.45	0.40	0.36	0.33	0.30	0.27	0.24	0.22	0.20	0.18	0.16	0.15	0.13	0.12	0.11	0.10	0.09

Table B.2. Summary of results after consideration of various incentive schemes.

Levelised Tariff (Rs./kWh)	Levelised Tariff with VGF (Rs./kWh)	Levelised Tariff with GBI - 2 year (Rs./kWh)	Levelised Tariff with GBI -8 year (Rs./kWh)	Levelised Tariff with AD (Rs./kWh)
5.68	5.43	5.32	4.59	5.14

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

Analysis and Design of Series LC Partial-Resonance-Pulse Based ZCS Current-Fed Push-Pull Converter

Swati Tandon, *Student Member, IEEE* and Akshay Kumar Rathore, *Fellow, IEEE*
Electrical and Computer Engineering, Concordia University, Montreal, Canada

***Abstract*— This paper analyzes a partial-resonance-pulse, created through a series LC resonant circuit, on the current-fed push-pull converter topology to achieve the zero-current-switching (ZCS) of the semiconductor switches. During switching overlap of the semiconductor switches, referred as energy storage mode, the resonance occurs and offers the natural commutation of the semiconductor devices. During this period, the current is smoothly transferred from one switch to the other and outgoing switch commutates naturally with ZCS. It solves the historical problem of the voltage spike across the devices at their turn-off and thus, eliminates the requirements of the snubber across switches or the clamping circuit. This reduces the hardware complexity and makes the converter snubberless, compact, and cost-effective. Variable frequency with constant switching overlap time is adopted for the load voltage regulation and to maintain the ZCS of the devices for a given range of source voltage and load current. A comprehensive study of the steady-state operation and analysis of the proposed topology is reported along with the design and selection of the resonant tank elements. A hardware proof-of-concept prototype rated at 500W is developed and tested to validate the analysis and claims and demonstrate the converter performance. An efficiency of 96.2% is recorded in laboratory on the prototype at full-load.**

Index terms – Zero current switching, current-fed dc/dc converter, partial resonance pulse, series resonance

I. INTRODUCTION

Clean and sustainable energy opened opportunities for advanced power electronics and fast deployment of solar/fuel cells integrated hybrid microgrid or nanogrid system delivering high performance while still reducing CO_2 emissions [1]-[2]. However, the effective integration of these alternative energy sources still remains a challenge due to their unregulated, discontinuous and low voltage output issuing a requirement of a dedicated power conversion unit. To revolutionize the way these sources are interfaced with a DC microgrid or to the conventional grid as shown in Fig. 1 [3]-[4], dc/dc converters are expected to be as efficient and effective as possible in terms of price and performance.

Today, most PV/Fuel cell enabled systems are 24V or 48V to deliver higher power level, reduced cost and scalability. Moreover, they can be designed to replace 48V standby batteries in several applications offering higher power density and reduction of weight and size. Over the years, a number of non-isolated and isolated dc-dc converter topologies have been published for distributed energy applications. High device

switching frequency has been preferred in the dc/dc converters to realize high-density and low-cost solutions [5]. Soft-switching if the semiconductor devices has been implemented in order to achieve better device utilization, limit the device switching losses, and reduce the size and cost of the heat-sink associated with them for the thermal stability.

However, voltage-fed converter topologies fully depend upon the high-frequency (HF) transformer to obtain the required voltage gain due to buck-derived nature. Current-fed converter topologies are boost derived and offer natural voltage gain reducing the burden on the HF transformer [6]. However, they are less explored and picked up interest recently. The main reasons are the commutation of the semiconductor devices and the snubber requirement owing to the turn-off voltage overshoot across the devices. It introduces complexity in circuit design and compromises on the power density, efficiency, cost, and gain capability of the converter. In addition, soft-switching is not that straightforward as voltage-fed topologies.

Traditionally, passive (dissipative) snubber-based topologies have been proposed, studied and analyzed. These topologies are less efficient, and compromise on the density and modularity of the converter. Active-clamping has been proposed to effectively clamp the device voltage simultaneously achieving the zero-voltage-switching (ZVS) of the devices [7]-[11]. This solution has been the most efficient

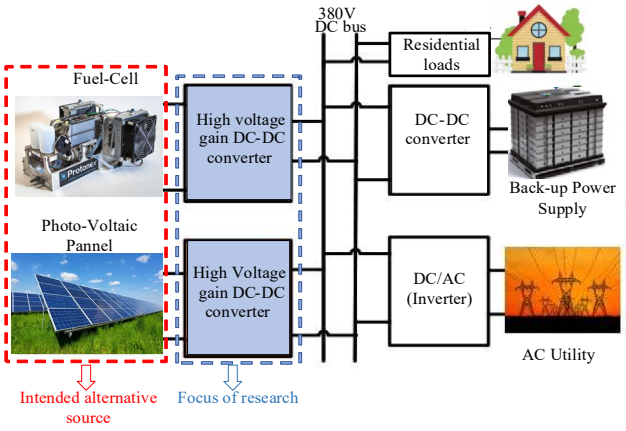


Fig. 1. A typical architecture of hybrid microgrid system.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

TABLE I: COMPARISON OF THE PROPOSED TOPOLOGY WITH OTHER STATE-OF-THE ART CONVERTERS

Description	[11]	[12]	[17]	[26]	[24]	Proposed
Semiconductor Count	3 MOSFETs + 4 Diode	8 MOSFETs + 0 Diode	6 MOSFETs + 0 Diode	3 MOSFETs + 2 Diode	2 MOSFETs + 2 Diode	2 MOSFETs + 2 Diode
Inductor/capacitor count	1 Input inductor +1 resonant inductor +1 clamp capacitor+ 2 output capacitor	1 Input inductor	1 Input inductor + 2 output capacitor	1 Input inductor +2 inductor+ 2 output capacitor+ 1 resonant capacitor	1 Input inductor + 2 output capacitor+ 1 resonant cap	1 Input inductor + 2 output capacitor+ 1 resonant cap
Switch Voltage Stress	$\frac{V_o}{2n} + V_{Lr}$	$\frac{V_{in}}{1-D}$	$\frac{V_o}{2n}$	$\frac{V_{in}}{1-D}$	$\frac{2V_o}{n}$	$\frac{V_o}{n} + \frac{2V_{cr_{pk}}}{n}$
Control Technique	Variable duty	Variable duty	Variable duty	Phase shift modulation	Variable frequency	Variable Frequency
Soft Switching Method	Active clamping	Secondary modulation	Load side switches	Quasi resonant	Parallel LC impulse resonance	Series LC quasi resonance
Duty Cycle Loss	High	-	Low	Moderate	High	Low
Resonant Energy	Fixed	-	-	Fixed	Fixed	Adaptive
Circuit complexity	Complex	Simple	Complex	Complex	Simple	Simple
Circulating currents	High	Low	Low	High	High	Low
Input/output voltage	36V/400V	12V/280V	(20~30)V/400	30~50V/350V	(42~48)V/380V	(42~48)V/380V

among all other snubber based solutions. However, it requires additional floating switches, their control, and thus, requires additional PCB footprints. It also reduces the gain of the converter. Innovation of snubberless solutions eliminated the historical requirement of snubbers and solved the problem of device voltage turn-off spike with natural ZCS turn-off. These converters utilize secondary side modulation to attain soft-switching of the semiconductor devices [12]-[15] with higher efficiency, high-density, while preserving the original circuit features and inherent boosting ability. However, owing to the presence of all active devices and their control, it is more suited for bidirectional power transfer application. For unidirectional applications such as solar and fuel cells, it creates additional burden of control and switching the second set of devices (load or secondary side) whereas diodes could do the conversion task simply [16]-[18].

To further reduce the semiconductor device count and maintain extended soft-switching, resonant tanks have been adopted exploiting transformer leakage inductance and other circuit parasitic [19]-[20]. Conventional resonant converters encounter poor voltage regulation, higher circulating currents and loss of soft-switching at light-loads resulting in lower light-load efficiency [21]-[22]. On the contrary, partial-resonant converters are potentially meritorious as they only utilize resonant feature for a fraction of switching period to facilitate natural commutation of semiconductor devices resulting in reduced peak and RMS current allowing relatively lower conduction losses [23]. Impulse-commutation assisted ZCS is analyzed for various single-phase and three-phase current-fed topologies with parallel LC tank in [24], [25]. These converters were critically analyzed and designed for ZCS and voltage clamping of the devices through a short-resonance-impulse. Nevertheless, high peak in switch and resonant current were observed at light-load due to fixed resonant energy causing high circulating current. Furthermore, these converters experience drastic variation in switching frequency even for a narrow change in source voltage [26]. To deal with these challenges, a series LC based adaptive resonant energy is analyzed, and implemented in current-fed half-bridge converter topology to attain ZCS and device voltage clamping [27]. This concept is extended for current-fed push-pull topology in [28] because it requires a single inductor at the input with similar number of semiconductor devices and simple control circuitry. This paper focuses on experimental validation of this topology as shown in Fig. 2 The major contributions of this article are elaborated as follows:

1) Series resonant tank energy being load adaptive helps in keeping minimum circulating currents under all operating conditions. This further limits the rms current resulting in higher overall efficiency.
2) ZCS and suppressing over voltage spikes in semiconductor devices.
3) Semiconductor devices share ground with the source which helps in reducing EMI issues/leakage current problem as well as reduces gate drive requirement to cater cost and space limitations.

Table I corroborates the merits of the proposed soft-switching converter compared to the existing state-of-the-art converters reported in introduction. The layout of the paper

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

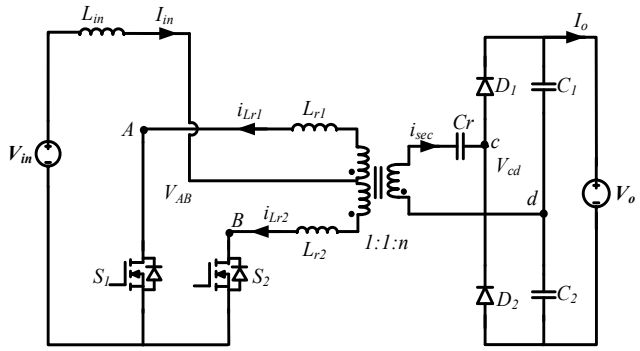


Fig. 2. Schematic of the proposed current-fed push-pull converter topology.

is as follow: Steady-state operation and analysis of the proposed converter with partial-resonance-pulse is studied in Section II. Detailed design to determine the resonant tank to achieve desired objectives is explained in Section III. Experimental results from testing a proof-of-concept hardware prototype are reported in section IV to demonstrate the performance and verify the analysis, design and the claims. Loss distribution and efficiency analysis are given and discussed in detail in section V.

II. Proposed Topology and Staedy State Analysis

The proposed partial-resonance-pulse operated current-fed push-pull converter with simple series-resonant tank is shown in Fig. 2. It is operated with variable frequency modulation to regulate output voltage and maintain ZCS. Resonant tank utilizes a combination of transformer leakage inductance (L_{lk}) and a series connected HF capacitor at the secondary side. The load dependent resonant energy due to series placed capacitor assists in ZCS with reduced circulating energy at all conditions. Operating principle and detailed mathematical analysis are studied for each mode of operation during half switching period. The operation remains same for the other half-cycle with other symmetrical devices conducting. Fig. 3 shows the steady state operating waveforms of the converter for $D > 0.5$ with 180º phase-shifted gating signals. The equivalent circuit and power flow during each mode of operation is laid out in Fig. 4. Analytical expressions for each mode facilitate estimation of working parameters and also validate theoretical converter performance. Following assumptions are considered for simplified analysis:

1. All the components are assumed ideal and PCB parasitic are neglected.
2. The input inductor is large enough to maintain stiff dc current through it.
3. Output voltage is maintained constant with sufficiently large output capacitors.
4. Transformer magnetizing inductance is infinitely large and so neglected.

a) *Mode 1 (t_0-t_1): (Power transfer mode refer Fig. 4(a))*

Assuming switch S_1 to be conducting with a constant current I_{in} through leakage L_{r1}. The energy from the supply is

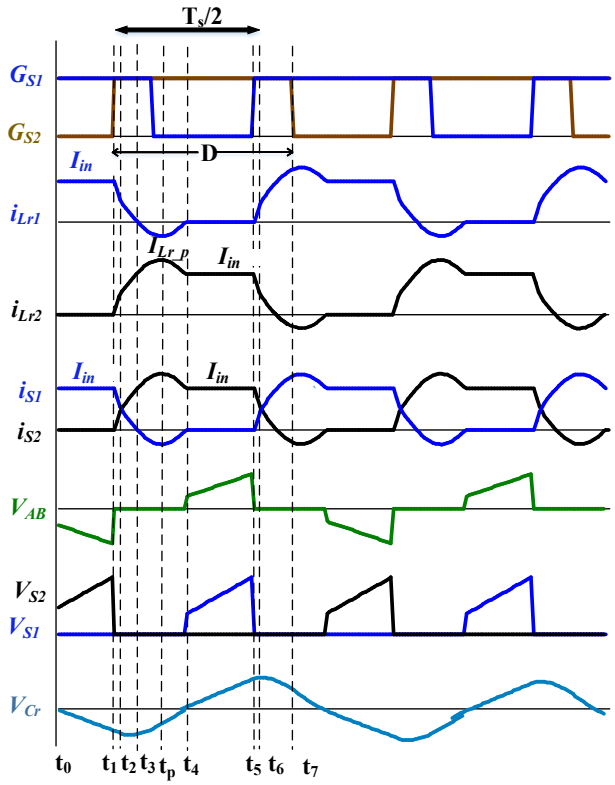


Fig. 3. Steady-state operating waveforms of the proposed converter topology.

transferred to the load through diode D_2 and C_1 as shown in Fig. 4(a). During this time resonant capacitor C_r is discharged by the negative current resulting in linearly decreasing bridge output voltage V_{AB}.

$$i_{S1}(t) = I_{in}, \qquad i_{Lr1}(t) = I_{in} \tag{1}$$

$$i_{S2}(t) = i_{Lr2}(t) = 0 \tag{2}$$

$$V_{cr}(t) = V_{cr}(t_o) + \frac{1}{C_r}\int_{t_o}^{t_1} -nI_{in}\,dt = V_{cr}(t_o) - \frac{nI_{in}}{C_r}(t - t_0) \tag{3}$$

$$V_{AB}(t) = -2\left(\frac{V_o}{2} + V_{cr}(t_1)\right)\frac{1}{n} \tag{4}$$

b) Mode 2 (t_1-t_2): (Transition mode refer Fig. 4(b))

With the turn-on of switch S_2 at instant t_2, voltage V_{AB} is clamped to zero while the device capacitance C_2 discharges in a short period of time. Positive voltage appears across L_{r2} therefore, current through switch S_2 starts to increase linearly while current through S_1 begins to fall with the same slope. Converter equations during this mode are as follows:

$$i_{S1}(t) = i_{Lr1}(t) = I_{in} - \frac{1}{nL_r}\left(\frac{V_o}{2} + V_{cr}(t_1)\right)(t - t_1) \tag{5}$$

$$i_{S2}(t) = i_{Lr2}(t) = \frac{1}{nL_r}\left(\frac{V_o}{2} + V_{cr}(t_1)\right)(t - t_1) \tag{6}$$

At the end of this mode:

$$i_{Lr1}(t_2) = i_{S1}(t_2) = \frac{I_{in}}{2} \tag{7}$$

$$i_{Lr2}(t_2) = i_{S2}(t_2) = \frac{I_{in}}{2} \tag{8}$$

$$V_{AB} = 0 \tag{9}$$

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

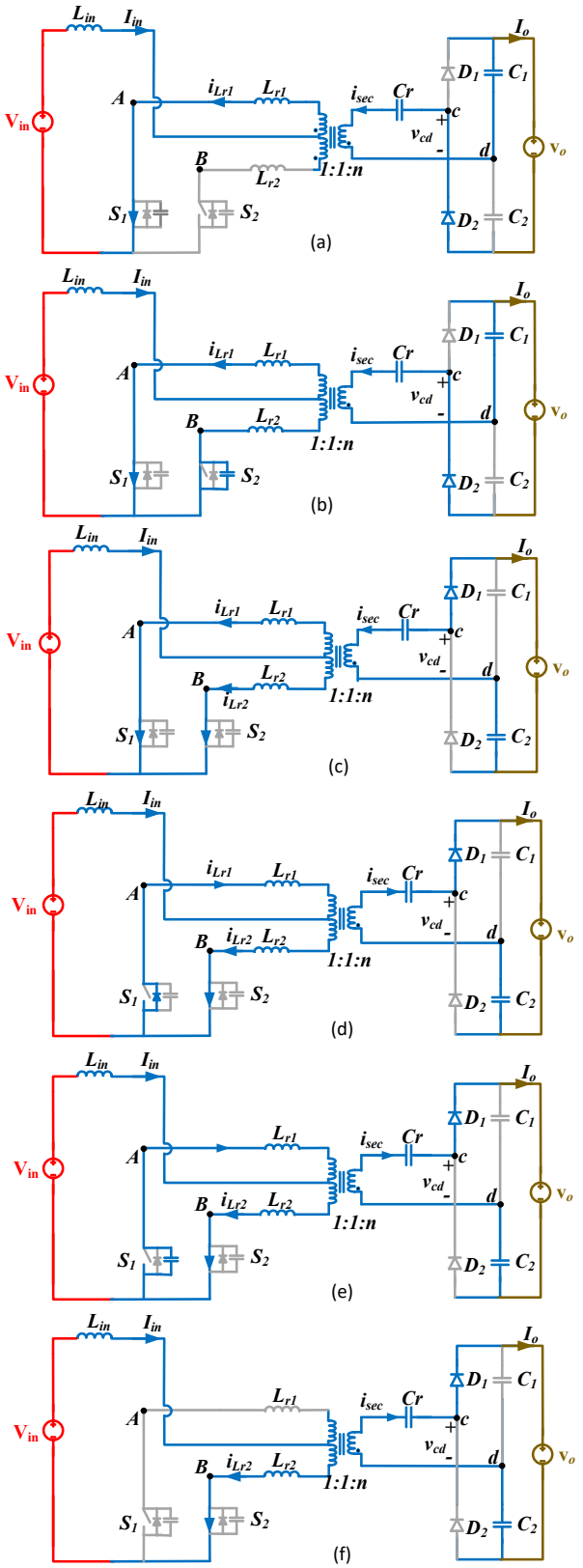


Fig. 4. Steady-state operating waveforms of the proposed converter topology.

$$V_{cr}(t_2) = V_{cr}(t_1) = -V_{cr_pk} \quad (10)$$

From (5), (7) and (8), duration of this mode can be written as:

$$T_{21} = \frac{nI_{in}L_r}{2 * \left(\frac{V_o}{2} + V_{cr}(t_1)\right)} \quad (11)$$

c) Mode 3 (t_2-t_3): (Resonance mode Fig. 4(c))

Switch S_1 and S_2 are conducting during this mode and resonance commences between $(L_{r1} + L_{r2})$ and C_r resulting in sinusoidal current through transformer primary and switches. Therefore, the switch and series inductor currents are given as.

$$i_{S1}(t) = i_{Lr1}(t) = \frac{I_{in}}{2} - \frac{1}{nZ_o}\left(\frac{V_o}{2} + V_{cr_pk}\right) sin\omega_r(t - t_2) \quad (12)$$

$$i_{S2}(t) = i_{Lr2}(t) = \frac{I_{in}}{2} + \frac{1}{nZ_o}\left(\frac{V_o}{2} + V_{cr_pk}\right) sin\omega_r(t - t_2) \quad (13)$$

At the end of this mode, current i_{Lr2} increases to I_{in} while the current through switch S_1 and L_{r1} reaches zero.

$$i_{S2}(t_3) = i_{Lr2}(t_3) = I_{in}\,, \qquad i_{S1}(t_3) = i_{Lr2}(t_3) = 0 \quad (14)$$

From (12) and (14), duration of this mode can be given as:

$$T_{32} = \frac{1}{\omega_r}\sin^{-1}\left(\frac{nI_{in}Z_o}{2 * \left(\frac{V_o}{2} + V_{cr_pk}\right)}\right) \quad (15)$$

d) Mode 4 (t_3-t_4): (ZCS mode Fig. 4(d) and 4(e))

Resonance continues during this mode, with resonant current (i_{Lr2}) surpassing I_{in}. By virtue of this additional current, internal diode conducts before the gating is removed resulting in soft turn-off of the targeted switch. Resonance pulse terminates at instant t_4 when current i_{Lr2} again reaches I_{in}. During this mode, i_{Lr2} reaches its peak value (I_{Lr2_p}) at instant t_p which can be estimated as:

$$I_{S2_p} = I_{Lr2_p} = i_{Lr2}(t_x) = \frac{I_{in}}{2} + \frac{1}{nZ_o}\left(\frac{V_o}{2} + V_{cr_peak}\right) \quad (16)$$

The critical condition for ZCS of the primary switches is given as,

$$I_{Lr2_p} > I_{in}$$

$$\frac{I_{in}}{2} + \frac{1}{nZ_o}\left(\frac{V_o}{2} + V_{cr_peak}\right) > I_{in} \quad (17)$$

With the above knowledge, time duration of this mode can be computed as:

$$T_{42} = \frac{1}{\omega_r}\left(\pi - \sin^{-1}\left(\frac{nI_{in}Z_o}{2 * \left(\frac{V_o}{2} + V_{cr_pk}\right)}\right)\right) \quad (18)$$

$$T_{43} = T_{42} - T_{32} \quad (19)$$

e) Mode 5 (t_4-t_5): (refer Fig. 4(f))

At t_4, current i_{Lr2} reaches I_{in} with switch S_2 conducting and power is transferred from source to the load through diode D_1 and C_2 as shown in Fig. 4(f). During this mode, constant current flows through switch S_2 resulting in zero voltage across series inductor while the series capacitor charges linearly.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

$$I_{Lr1,2_{rms}} = I_{Sw1,2_{rms}} = \left[I_{in}^2 \left(\frac{1}{2} - \frac{f_s}{f_r} \left(\frac{1}{4} - \frac{1}{8}\left(\frac{1}{k}\right)^2 + \frac{k}{6\pi} - \left(\frac{\sqrt{1-k^2}}{8\pi k}\right) \right) + \frac{1}{\pi}\left(\frac{1}{8}\left(\frac{1}{k}\right)^2 - \frac{1}{4} \right) sin^{-1}k \right) \right]^{\frac{1}{2}} \quad (26)$$

$$i_{S1}(t) = i_{Lr1}(t) = 0 \quad (20)$$

$$i_{S2}(t) = i_{Lr2}(t) = I_{in} \quad (21)$$

Duration of this mode is denoted as:

$$T_{10} = T_{54} = \frac{T_s}{2} - T_{41} \quad (22)$$

III. Design Procedure of the Proposed Converter

This section presents the detailed design of the proposed converter suited for low input voltage applications (e.g. PV, fuel-cell), ranging from 42–48V feeding a high voltage dc bus at 380V. The key objectives are 1) keeping stiff dc voltage at the output and 2) soft-switching of the semiconductor devices against fluctuations in source voltage and load. ZCS is achieved by exploiting sinusoidally falling switch current due to series resonance. Design of the working parameters significantly affect the losses in active and passive components which therefore influence the efficiency of the converter. Therefore, it is imperative to design converter parameters keeping minimum stresses on the components. Design parameters are depicted in Table II.

A. Voltage gain of the proposed converter

The voltage gain dynamics of the converter is governed by the number of variables. Therefore, DC voltage gain expression can be derived in terms of variables n, r_n and f_n.

$$M = f(n, f_n, r_n) = \frac{n\left(1 - \frac{f_n r_n X}{\pi}\left(1 - \sqrt{1-k^2}\right)\right)}{\left(1 - f_n\left(\frac{1}{2} - \frac{1}{2\pi}sin^{-1}k + \frac{k}{2\pi}\right)\right)} \quad (23)$$

Where, n: Turn ratio of HF transformer, $k = \frac{M}{nr_nX}$;

f_n: Normalized frequency, $= \frac{f_{sw}}{f_r}$

r_n: Normalized resistance, $= \frac{R_{FL}}{Z_o}$

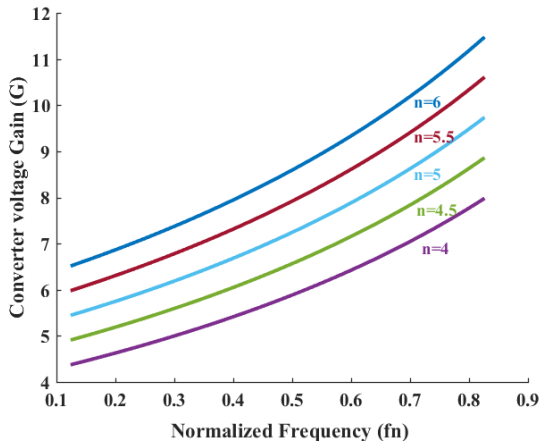


Fig. 5. Proposed series-LC resonant-pulse current-fed push-pull converter.

Fig. 5 exemplifies the gain characteristics plotted for a range of f_n with different turns ratio n. Gain variation occurs for different values of source voltage and load which is fulfilled by modulating switching frequency in order to maintain stiff output voltage for the given converter specifications.

B. Maximum device voltage and current stresses

Components rating can be determined by the maximum voltage/current appearing across them. Voltage across MOSFETs depends on primary referred peak resonant capacitor charge and output voltage.

$$V_{DS,max} = 2\left(\frac{V_{cr_{pk}}}{n} + \frac{V_o}{2n}\right) \quad (24)$$

Diode voltage rating is determined by

$$V_{D1_max} = V_{D2_max} = V_o \quad (25)$$

As device voltage inversely depends on the turns ratio as evident from (24), selecting a lower value of turns ratio results in high voltage rating of the switches with higher $R_{Ds,on}$ for the same current rating. As a result, higher conduction losses occur on primary side devices. RMS current helps in determining the current stresses on converter components. Therefore, (26) represents RMS resonant current and switch current whereas diode current rating depends on the secondary current given as,

$$I_{D1-2,avg} = I_o \quad (27)$$

C. Selection of HF Transformer Turns Ratio

Converter voltage gain (M) has a direct correlation with turns ratio (n) of HF transformer and normalized frequency f_n as shown in Fig. 5. These parameters have a prominent effect on the voltage and the current stress which then aggravate the conduction losses on the high current side. It is therefore vital to minimize conduction losses in order to accomplish desired voltage gain, by choosing a right combination of turns ratio and switching frequency. It can be construed that higher value of turns ratio could lower the voltage stress across the semiconductor devices while increasing the currents on the primary side resulting in higher conduction losses. Whereas, smaller turns ratio would require higher voltage rating switches while minimizing current stresses on the low voltage side. Considering previously mentioned issues, the turns ratio has been chosen as 1:1:4.5. Thereafter, maximum switching frequency can be computed for the minimum input voltage. For this design example, the switching frequency band is 80 kHz up to 200 kHz for V_{in} = 48 V down to 42 V.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

D. ZCS boundary Condition

Ideally, design of the resonance pulse must ensure sufficient energy to maintain ZCS for extended range of load and source voltage while limiting the circulating current and conductions losses in the converter. Therefore, the following conditions are satisfied to confirm ZCS for all conditions.

a) $I_{Lr1_p}, I_{Lr2_p} > I_{in}$

This condition helps in determining suitable value of characteristics impedance, Z_o of the resonant tank to have a balance between high peak and ZCS range. The limit for Z_o can be estimated using the following expression:

$$Z_o < \frac{2\left(\frac{V_o}{2} + V_{cr_{peak}}\right) * R_{FL} * V_{in}}{nV_o{}^2} \quad (28)$$

b) $T_{32} < T_{overlap} < T_{42}$

In order to preserve ZCS for an extensive load range, retaining a constant overlap period to account for the variation in switching frequency becomes essential. This can be done by slight variation in duty cycle or corresponding change in frequency.

E. Resonant tank design

Resonant tank design follows a procedure:

1. Select switching frequency f_{sw} for a maximum gain condition with V_{in} = 42V at full load.
2. Estimate f_r using (23) for a maximum gain condition with transformer turns ratio and switching frequency f_{sw}.
3. Maximum characteristic impedance $Z_{o,maz}$ is computed for minimum source voltage condition.
4. A value lower than $Z_{o,max}$ that satisfies (28) is selected for smooth commutation with limited circulating current.

$$f_r = \frac{1}{2\pi\sqrt{(L_{r1} + L_{r2})\, C_{r_ref}}}, \qquad Z_o = \sqrt{\frac{(L_{r1} + L_{r2})}{C_{r_ref}}} \quad (29)$$

where C_{r_ref} is the primary referred equivalent for series capacitance (C_r). Tank parameters L_{r1}, L_{r2} and C_r are computed as *2.5μH* and *8.5nF* respectively using (29).

F. Boost inductor calculation

To deliver maximum power rated at 500W with minimum source voltage of 42V, a maximum input current is computed to be 11.9A. For a 10% ripple in the input current inductor value is computed. In this scenario, ripple content depends on the resonance pulse width. Therefore, input boost inductor for the proposed design is calculated using:

$$L_{in} = \frac{V_{in}\,(t_4 - t_1)}{\Delta I_{in}} \quad (30)$$

IV. Experimental Validation and Results

A scaled-down 500W laboratory prototype has been developed for verification and testing of the proposed design as shown in Fig 6. The hardware implementation has been done for specifications given in Table III and Table IV. TMS320F28335 DSP is adopted to generate gating signals for the semiconductor devices to certify the steady-state performance in open-loop. The experimental test results are recorded on digital storage oscilloscope. The converter performance is observed for variation in the fuel-cell stack voltage under different loading condition.

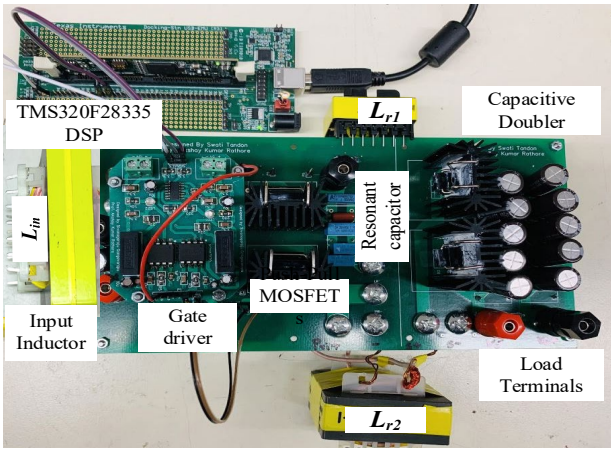


Fig. 6. A proof-of-concept 500W prototype of the proposed converter.

Fig. 7(a) and (b) demonstrate the steady-state behavior under full-load (500W) condition for V_{in} = *42V* and for V_{in} = *48V* respectively. Constant current region is witnessed when

TABLE II: DESIGN PARAMETERS

Parameter	Value
Resonant Inductors, L_{r1}, L_{r2}	2.5μH
Resonant Capacitor, C_r	8.5nF
Input inductance, L_{in}	300μH
Output capacitance, C_1, C_2	80μF
Turns ratio, 1:1: n	1:1:4.5
Maximum Duty cycle, $\boldsymbol{D_{max}}$	0.6

TABLE III: CONVERTER SPECIFICATIONS

Parameter	Value
Source voltage, $\boldsymbol{V_{in}}$	42-48 V
Maximum duty, $\boldsymbol{D}$	0.6
Output power, $\boldsymbol{P_o}$	500W
Output voltage, $\boldsymbol{V_o}$	380 V
Switching frequency, $\boldsymbol{f_{sw}}$	80-200 kHz

TABLE IV. EXPERIMENTAL SETUP COMPONENT SPECIFICATIONS

Component	Specifications
Power switches, $\boldsymbol{S_1, S_2}$	SCT3060ALGC11-ND, 650V, 39A, 60mΩ
Input Boost inductors, L_{in}	55 x 28 x 21, EE Ferrite cores, 300μH
Resonant Inductors, L_{r1} and L_{r2}	EE ferrite core, $Lr1$=1μH and L_{r2}= 0.7μH
Resonant capacitors, $\boldsymbol{C_r}$	HF Film capacitor, 8.5nF, 1000V
HF Centre Taped Transformer	EE ferrite core, Primary turns, =10, 10 secondary turns = 45, L_{lk1}= *1.55μH*, L_{lk2}= *1.8μH*
Diodes, $\boldsymbol{D_{1-2}}$	STPSC20065D, 650V, 20A,
Output capacitor, $\boldsymbol{C_1, C_2}$	80μF 450V electrolytic capacitor, 10nF film capacitor
Gate driver	HCNW3120

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

only one of the switches is conducting while sinusoidal waveforms are observed during gating overlap due to the resonance feature. Soft-commutation of the MOSFETs (S_1 and S_2) occur due to naturally falling switch current to zero before the gating signal is removed. Switch voltage remains zero while the switch current goes negative during switching overlap, which confirm anti-parallel diode conduction. Switching frequency is decreased from its maximum value ($f_{sw,max}$ = 200kHz) to regulate the output voltage for V_{in}= 48V.

Fig. 8(a) and (b) exhibit the experimental results confirming *ZCS* at turn-off and over voltage clamping of the semiconductor devices for half-load (250W*)* condition with two extreme source voltage points *i.e* V_{in} = 42V and V_{in} = 48V respectively.

Similarly, Fig. 9(a) and (b) validate the experimental steady-state behavior under light-load (100W) condition for V_{in} = 42V and for V_{in} = 48V respectively. It is witnessed that the ratio of peak current to the input RMS current (I_{Lrp}/ I_{in}) is quite low for loads below rated power, resulting in considerably lower circulating energy even at light load. Therefore, soft-switching can be achieved without power loss due to leakage current. It is also observed that switch voltage stress decreases with decrease in output power further reducing the switching losses. The peak voltage stress across primary side MOSFETs are limited to ≈220V which is relatively higher compared to the state-of-the-art converters. Fig. 7(c), 8(c) and 9(c) verify the terminal voltage V_{AB}, the resonant inductor current I_{Lr}, and the resonant capacitor voltage V_{cr} in accordance with the theoretical waveform. It should also be noted that V_{AB} is bipolar and balanced without any dc-offset. However, a slight ringing in the switch voltage waveform can be seen due to device intrinsic snubber capacitor and PCB parasitic.

Fig. 10 presents the bar chart with varying resonant capacitor charge for different output power at nominal input voltage. This chart clearly elucidates the load dependent

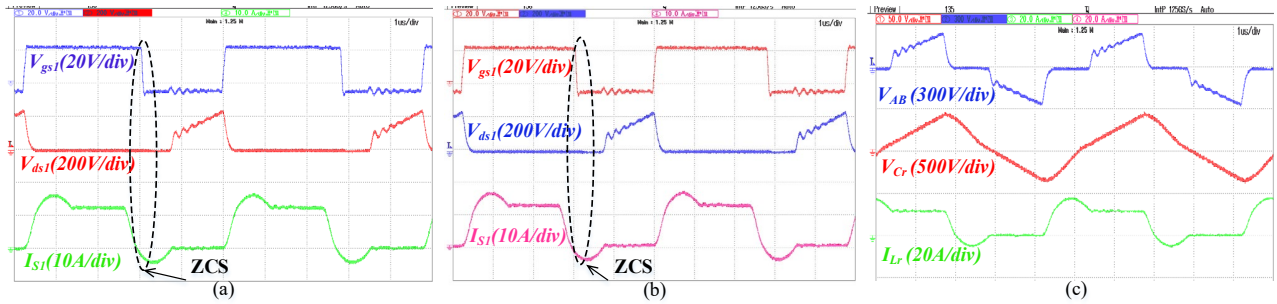


Fig. 7. Experimental steady state waveforms for (a) gate-source voltage V_{gs}, drain-source voltage V_{ds} , switch current I_{S1} with V_{in} = 42V, f_{sw}=200kHz *and* D=0.63 (b) gate-source voltage V_{gs}, drain-source voltage V_{ds}, switch current I_{S1} with V_{in}= 48V, f_{sw}=185kHz *and* D=0.63 (c) voltage $V_{AB,}$ resonant capacitor voltage V_{cr}, and series resonant tank current I_{Lr} at full load (500W).

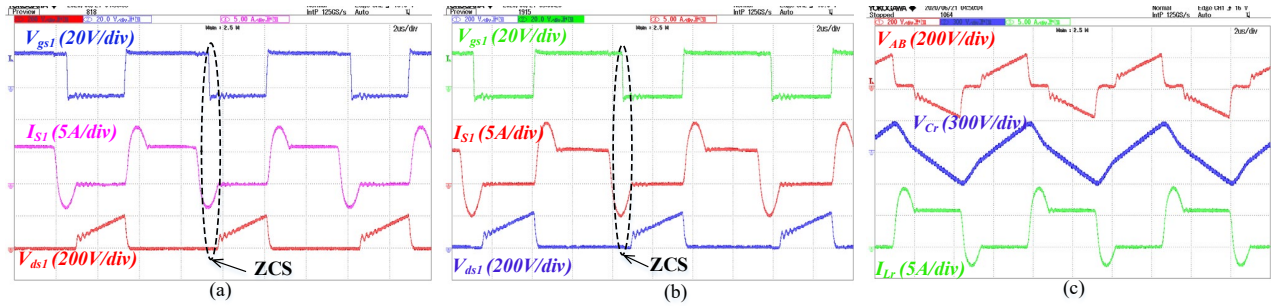


Fig. 8. Experimental steady state waveforms for (a) gate-source voltage V_{gs}, drain-source voltage V_{ds} , switch current I_{S1} with V_{in} = 42V, f_{sw}=130kHz *and* D=0.63 (b) gate-source voltage V_{gs}, drain-source voltage V_{ds}, switch current I_{S1} with V_{in}= 48V, f_{sw}=115kHz *and* D=0.63 (c) voltage $V_{AB,}$ resonant capacitor voltage V_{cr}, and series resonant tank current I_{Lr} at full load (250W).

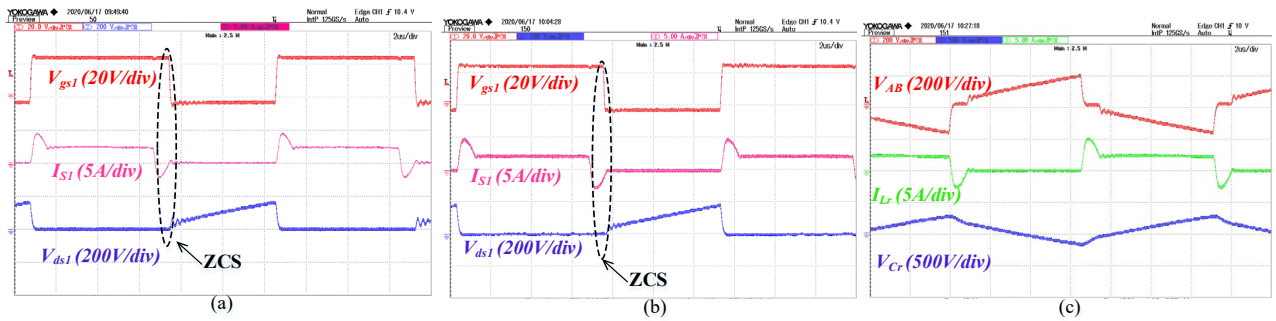


Fig. 9. Experimental steady state waveform for (a) gate-source voltage V_{gs}, drain-source voltage V_{ds} , switch current I_{S1} with V_{in}= *42V*, f_{sw}=90kHz *and* D=0.56 (b) gate-source voltage V_{gs}, drain-source voltage V_{ds}, switch current I_{S1} with V_{in}= *48V* , f_{sw}=80kHz *and* D=0.56 (c) voltage $V_{AB,}$ resonant capacitor voltage V_{cr}, and series resonant tank current I_{Lr} at light-load (100W).

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

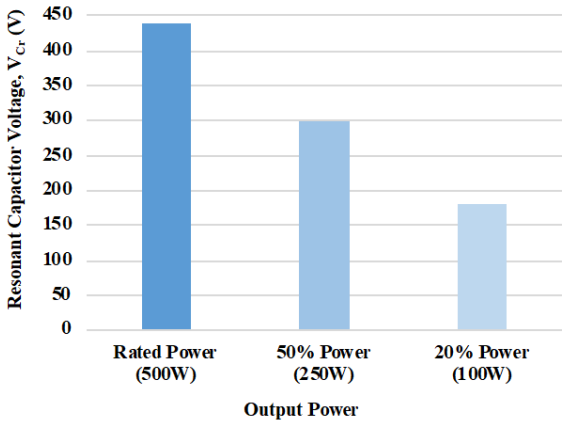
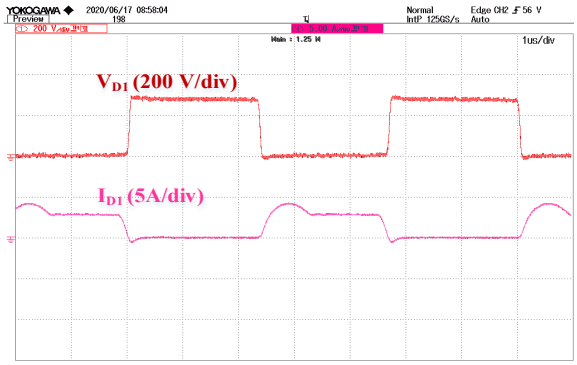
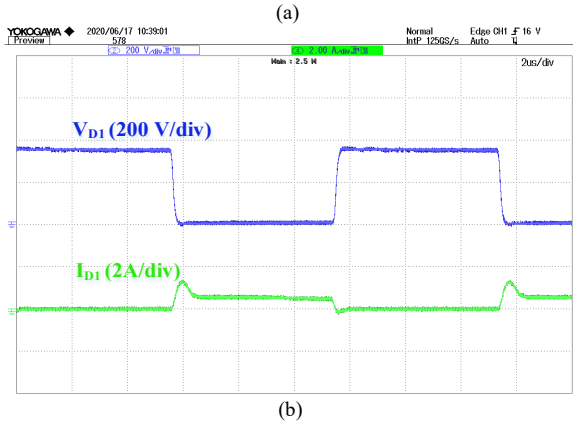


Fig. 10. Resonant capacitor charge for different output power at nominal input voltage.

resonant energy, therefore allowing lower resonant energy to bring soft switching without compromising on the conversion efficiency. In addition, the experimental results confirm stiff voltage regulation under varying load and source voltage owing to the frequency modulated gating control. Fig. 11(a) and (b) depict experimental waveforms for secondary-side rectifier diode in one leg for nominal source voltage (V_{in}=42V) at full-load and light load respectively. ZCS turn-off of the diodes should be observed with blocking voltage equal to output voltage V_o. Fig. 12 demonstrates the theoretical and measured switching frequency trend for the



(a)



(b)

Fig. 11. Experimental results for rectifier diode current I_{D1} and voltage V_{D1} for V_{in} = 42V (a) at full-load and (b) at light-load.

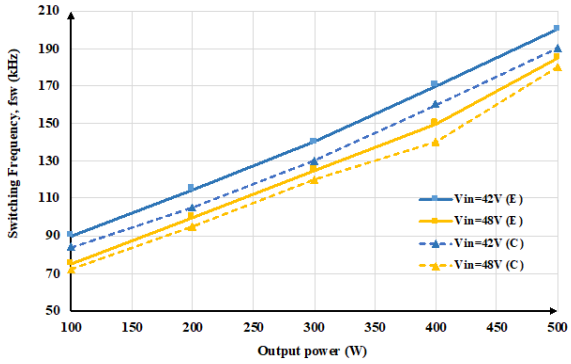


Fig. 12. Switching frequency trend with varying output power for two extreme input voltages (C: Calculated, E: Experimental value).

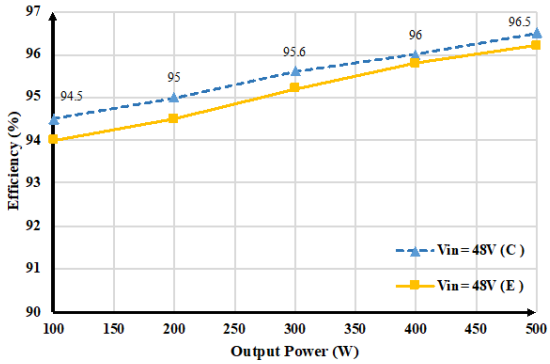


Fig. 13. Experimental efficiency curve with output power variation for nominal input voltage (C: Calculated, E: Experimental efficiency).

two extremities in source voltage with different loading condition. The experimental value matches closely with the theoretically calculated value with less than 10% error. The experimental switching frequency limit is recorded to be 185–200 kHz for rated-load. It is witnessed from Fig. 12 that frequency deviation is subtle for variation in input voltage as compared to the load change making the converter more suitable for constant load applications. From the experimental results and discussion, it is clear that the proposed converter can operate safely from full-load down to 20% load with switching frequency band of 80-200 kHz. However, no-load condition has not been considered in this experiment, which may require additional auxiliary control. Experimentally recorded efficiency curves for different loading condition at two extreme source voltages are depicted in Fig. 13. Maximum full-load efficiency of 96.2% for V_{in}= 48V and minimum efficiency of 93.2% for V_{in}= 42V are observed on the hardware prototype. Therefore, it can be established from the experimental results that the proposed design is able to achieve ZCS and switch overvoltage clamping allowing safe and efficient integration of alternative energy sources having inherent variability in DC voltage.

V. Loss Evaluation and Influence on Efficiency

In this section, the loss evaluation of the proposed converter is carried out. To correctly estimate the theoretical efficiency of the proposed design, various losses for active and passive components are computed using analytical

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

TABLE V: CONVERTER ITEMIZED LOSSES AND THEORETICAL EFFICIENCY

Output power	500	100	500	100	W
Input voltage, V_{in}	48	48	42	42	V
Input current (rms)	10.41	2.08	11.90	2.38	A
Duty cycle, D	0.6	0.56	0.6	0.56	
Loss Type					
MOSFET conduction loss	9.0	1.1	10.80	1.0	W
MOSFET switching loss	1.40	0.85	1.38	0.80	W
Rectifier diode loss	2.10	1.60	2.50	1.50	W
Total semiconductor losses	12.50	3.55	14.58	3.30	W
Input inductor loss	2.50	0.80	3.50	0.95	W
HF Transformer loss	1.50	0.30	1.90	0.52	W
Gate drive loss	0.116	0.048	0.136	0.04	W
Auxiliary loss	0.70	0.90	0.704	1.33	W
Total power losses	17.50	5.6	20.8	6.2	W
Efficiency	96.5	94.4	95.8	93.8	%

expressions and the datasheet specifications of the hardware components for rated power and 20% rated output power. Table V depicts the individual component losses for extreme input voltages at rated load as well as light-load. It can also be inferred that conduction loss are dominant with losses predominantly occurring in semiconductor switches, diodes, HF transformer and input inductor. For this experiment, MOSFET with low on-state resistance, $R_{ds,on}$ and extremely low output capacitance is chosen in order to limit switch turn-on and conduction losses. SiC Schottky diodes are carefully chosen for the secondary side rectifier to assist in smooth turn-off with negligible reverse recovery/ turn-off loss. In addition, EE ferrite core is utilized to design center-taped HF transformer to limit core loss. Besides, converter efficiency is further improved by developing in-lab optimized gate drive circuit with separate turn-on and turn-off control.

VI. CONCLUSION

Partial-resonance-pulse operation is studied and analyzed in current-fed push-pull dc/dc converter to attain soft-commutation and voltage clamping of the devices. Push-pull configuration and voltage doubler facilitate the least number of semiconductor devices with simple gate control requirement. ZCS turn off and smooth current transfer from an outgoing switch to the incoming switch occurs through a resonance pulse created during switching overlap by a series LC circuit. By incorporating short duration series LC resonance-pulse, load adaptive ZCS is possible with lower circulating and peak currents through the components. This in turn allows lower power loss due to circulating energy at light load compared to traditional resonant converters and therefore offers better efficiency. As a result, the proposed topology benefits from reduced conduction losses and lower rating of the primary-side devices with low $R_{ds,on}$ minimizing heat sink requirements and overall footprints. The proposed converter proves to be an effective and simple solution to resolve the fundamental problem of device voltage spike in current-fed converter without compromising on the converter performance, cost and size. Proof-of-concept hardware prototype is built in laboratory and tested to demonstrate the accuracy of the analysis and design along with converter performance. Hence, considering aforementioned merits, the proposed push-pull design is well suited for medium to high power voltage boost applications (150W to 1000W). Many such diverse applications include PV enabled back-up power supply, micro inverter etc. with inherent source voltage disturbances.

REFERENCES

[1] A. T. Ghareeb, A. A. Mohamed, and O. A. Mohammed, "DC microgrids and distribution systems: An overview," *in Proc. IEEE Power Energy Soc. Gen. Meeting*, Vancouver, BC, USA, 2013, pp.1–5.

[2] J. D. van Wyk and F. C. Lee, "On a future for power electronics," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 1, no. 2, pp. 59–72, Jun. 2013.

[3] R. Sharma and Hongwei Gao, "Low cost high efficiency DC-DC converter for fuel cell powered auxiliary power unit of a heavy vehicle," in *IEEE Transactions on Power Electronics*, vol. 21, no. 3, pp. 587-591, May 2006.

[4] F. L. Tofoli, D. C. Pereira, W. J. Paula, and D. S. O. Junior, "Survey on non-isolated high-voltage step-up dc–dc topologies based on the boost converter," *IET Power Electron.,* vol. 8, no. 10, pp. 2044–2057, Oct. 2015.

[5] A. K. Rathore, A. K. S. Bhat, R. Oruganti, "A comparison of soft-switched DC–DC converters for fuel-cell to utility-interface application", *Proc. IEEE Power Convers. Conf.*, pp. 588-594, 2007.

[6] A. K. Rathore and U. R. Prasanna, "Comparison of soft-switching voltage-fed and current-fed bi-directional isolated dc/dc converters for fuel cell vehicles," *in Proc. IEEE Int. Symp. Ind. Electron*., 2012, pp. 3212–3219.

[7] R. Watson and F. C. Lee, "A soft-switched, full-bridge boost converter employing an active-clamp circuit," *in Proc. IEEE Power Electron. Spec. Conf.*, 1996, pp. 1948–1954.

[8] V. Yakushev, V. Meleshin, and S. Fraidlin, "Full-bridge isolated current fed converter with active clamp," *in Proc. 14th IEEE Appl. Power Electron. Conf. Expo.*, 1999, pp. 560–566.

[9] E. S. Park, S. J. Choi, J. M. Lee, and B. H Cho, "A soft-switching active clamp scheme for isolated full-bridge boost converter," *in Proc. IEEE Appl. Power Electron. Conf. Expo*., 2004, vol. 2, pp. 1067–1070.

[10] J.-T. Kim, B.-K. Lee, T.-W. Lee, S.-J. Jang, S.-S. Kim, and C.-Y. Won, "An active clamping current-fed half-bridge converter for fuel-cell generation systems," *in Proc. IEEE Power Electron. Spec. Conf*., 2004, pp. 4709– 4714.

[11] Q. Wu, M. Wang, W. Zhou, C. Huang, G. Liu and X. Wang, "High frequency active-clamped zero-current switching current-fed push-pull converter for micro-converter applications," *2020 IEEE Energy Conversion Congress and Exposition (ECCE)*, Detroit, USA, 2020, pp. 1273-1278.

[12] A. K. Rathore and U. R. Prasanna, "Analysis, design, and experimental results of novel snubberless bidirectional naturally clamped ZCS/ZVS current-fed half-bridge dc/dc converter for fuel cell vehicles," *IEEE Trans. Ind Electron.*, vol. 60, no. 10, pp. 4482–4491, Oct. 2013.

[13] P. Xuewei, A. K. Rathore, "Novel Bidirectional snubberless naturally commutated soft-switching current-fed full bridge isolated DC/DC Converter for fuel cell vehicles", *IEEE Trans. Ind. Electro.*, vol. 61, no. 5, pp. 2307-2315, May 2014.

[14] P. Xuewei and A. K. Rathore, "Naturally clamped zero-current commutated soft-switching current-fed push-pull dc/dc converter: Analysis, design, and experimental results," *IEEE Trans. Power. Electron.*, vol. 30, no. 3, pp. 1318–1327, Mar. 2015.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3074109, IEEE Transactions on Industry Applications

[15] D. Chakraborty, A. K. Rathore, E. Breaz and F. Gao, "Parasitics assisted soft-switching and naturally commutated current-fed bidirectional push-pull voltage doubler," *IEEE Industry Applications Society Annual Meeting,* Addison, TX, USA, 2015, pp. 1-8.

[16] D. R. Nayanasiri, G. H. B. Foo, D. M. Vilathgamuwa and D. L. Maskell, "A Switching Control Strategy for Single- and Dual-Inductor Current-Fed Push–Pull Converters," *in IEEE Transactions on Power Electronics,* vol. 30, no. 7, pp. 3761-3771, July 2015

[17] R. Kosenko, A. Chub and A. Blinov, "Full-soft-switching high step-up bidirectional isolated current-fed push-pull DC-DC converter for battery energy storage applications," *IECON 2016 -42nd Annual Conference of the IEEE Industrial Electronics Society,* Florence, Italy, 2016, pp. 6548-6553.

[18] S. D. Nugraha, O. A. Qudsi, D. S. Yanaratri, E. Sunarno and I. Sudiharto, "MPPT-current fed push pull converter for DC bus source on solar home application," *International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE),* Yogyakarta, Indonesia, 2017, pp. 378-383

[19] M. J. Ryan, W. E. Brumsickle, D. M. Divan, and R. D. Lorenz, "A new ZVS LCL-resonant push-pull DC-DC converter topology," *IEEE Trans. Ind. Appl.,* vol. 34, no. 5, pp. 1164–1174, Sep./Oct. 1998.

[20] D. Sha, F. You and X. Wang "A high-efficiency current-fed semidual-active bridge DC-DC converter for low input voltage applications," *IEEE Trans. Ind. Electron.,* vol. 63, no. 4, pp. 2155–2164, April. 2016

[21] Y. Zhang, L. Zhou, M. Sumner and P. Wang, "Single-Switch, Wide Voltage-Gain Range, Boost DC–DC Converter for Fuel Cell Vehicles," in *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 134-145, Jan. 2018

[22] W. Chen, Z. Lu, X. Zhang, and S. Ye, "A novel ZVS step-up push-pull type isolated LLC series resonant dc-dc converter for UPS systems and its topology variations," *in Proc. IEEE APEC, 2008*, pp. 1073–1078.

[23] R.-Y. Chen, T.-J. Liang, J.-F. Chen, R.-L. Lin, and K.-C. Tseng, "Study and implementation of a current-fed full-bridge boost DC-DC converter with zero-current switching for high voltage applications," *IEEE Trans. Ind. Appl*., *vol. 44, no. 4, pp. 1218–1226, Jul./Aug. 2008.*

[24] K. R. Sree and A. K. Rathore, "Impulse commutated zero current switching current-fed push-pull converter: Analysis, design and experimental results," *IEEE Trans. Ind. Electron*., vol. 632, no. 1, pp. 363–370, Jan. 2015.

[25] K. R. Sree and A. K. Rathore, "Analysis and design of impulse-commutated zero-current-switching single-inductor current-fed three-phase push–pull converter*," IEEE Trans. On Ind. Appl*., *vol. 53, no. 2, Mar./Apr. 2017.*

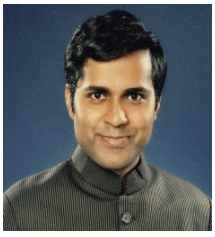
[26] Q. Wu, Q. Wang, J. Xu, H. Li and L. Xiao, "A High-Efficiency Step-Up Current-Fed Push–Pull Quasi-Resonant Converter With Fewer Components for Fuel Cell Application," *in IEEE Transactions on Industrial Electronics*, vol. 64, no. 8, pp. 6639-6648, Aug. 2017

[27] S. Tandon and A. K. Rathore, "Analysis and Design of Series LC Resonance-Pulse Based Zero-Current-Switching Current-Fed Half-Bridge DC-DC Converter," *in IEEE Transactions on Industrial Electronics,* doi: 10.1109/TIE.2020.3005104.

[28] S. Tandon and A. K. Rathore, "Partial Series LC Resonance-Pulse Assisted Zero Current Switching Current-Fed Push-Pull Converter," *2020 IEEE Industry Applications Society Annual Meeting*, Detroit, MI, USA, 2020, pp. 1-6, doi: 10.1109/IAS44978.2020.9334783.

Swati Tandon (Student Member, IEEE) received her master's degree in power system from the Delhi Technological University, Delhi, India in 2013. She is currently working towards her Ph.D. degree at the department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada. Her primary research interests include soft-switching techniques, development of high frequency resonant based current-fed DC-DC converter topologies for microgrid or distributed energy systems.

From 2013 to 2017, she has worked as an Assistant Manager at Design and development center, Electrical & Automation, Larsen & Toubro limited, Mumbai, India.



Akshay Kumar Rathore (IEEE Fellow) received his M.Tech. degree in electrical machines and drives from the Indian Institute of Technology (BHU) Varanasi, India, in 2003, and the Ph.D degree in power electronics from the University of Victoria, BC, Canada in 2008.

He had two subsequent postdoctoral research appointments with the University of Wuppertal, Wuppertal, Germany, and the University of Illinois at Chicago, Chicago, IL, USA. From November 2010 to February 2016, he was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada. He has published about 270 research articles in international journals and conferences, including 90 IEEE TRANSACTIONS. His research is mainly focused on current-fed converters and multilevel inverters.

He is serving as the IEEE Industry Applications Society (IAS) Awards Department Chair, an IAS Prominent Lecturer and Industrial Electronics Society (IES) Member-at-Large. Dr. Rathore received the Gold Medal during his M.Tech. degree for securing the highest academic standing among all electrical engineering specializations. He is a recipient of the 2013 IEEE IAS Andrew W. Smith Outstanding Young Member Achievement Award, the 2014 Isao Takahashi Power Electronics Award, the 2017 IEEE IES J David Irwin Early Career Award, 2020 IEEE IAS Outstanding Area Chair Award and the 2020 IEEE Bimal Bose Award for Industrial Electronics Applications in Energy Systems. He is also the Paper Review Chair of the IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS FOR RENEWABLE AND SUSTAINABLE ENERGY CONVERSION SYSTEMS. He is an Associate Editor of IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS and an editor of IEEE Transactions on Sustainable Energy. He was a Distinguished Lecturer from 2017-18 and an IAS Executive Board Member-at-Large from 2017-19 of the IAS Society. He chaired the IEEE IAS Industrial Automation and Control Committee from 2018-19 and the IEEE IES Technical Committee on Transportation Electrification from 2016-17. He served as the Editor-in-Chief of the IEEE Industrial Electronics Technology News (ITeN) from 2016-18 and received recognition award in 2019.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:59:23 UTC from IEEE Xplore. Restrictions apply.

Silicon
https://doi.org/10.1007/s12633-021-01103-6

ORIGINAL PAPER

Analysis of a Novel Nanoscale Vacuum Channel TF-FinFET

Mridul Prakash Kashyap [1] · Sanmveg Saini [1] · Rishu Chaujar [1]

Received: 28 November 2020 / Accepted: 31 March 2021
© Springer Nature B.V. 2021

Abstract
Concerned work is solely dedicated to the optimized characteristics of Nanoscale vacuum channel TF (Truncated fin)-FinFET at gate length of 7 nm. NVCTF-FinFET has its own benefit, due to high immune to noise and temperature as compared to TF-FinFET. The working of device stood perfectly well in case of NVCTF-FinFET. When tested, we end up with 3 times less current variation w.r.t temperature, 3 times more efficiency, 10^4 times more amplifying power analysed at corresponding peak values. These results evident the performance enhancement of NVCTF-FinFET, if seen from analog point of view. When studied from the perspective of linearity for RFIC designs, we end up with enhancement in some figure of merits such as 3.5 times in 1-dB compression point, 4 times more IP3, 10^2 times in HD3 and around 10^5 times in IMD3 at corresponding peak values. As NVCTF-FinFET provided descent switching ratio and V_{th} at input voltage (<2 V), which could lead these transistors beyond Moore's law. These simulation difference ultimately made NVCTF-FinFET, a significant candidate to high speed and noise reduction System on chip (SOCs) operations in outer space.

Keywords TF-FinFET · NVCTF-FinFET · Vacuum channel · Electric field · Linearity · Analog

1 Introduction

THESE days FinFET is used as major fundamental block in high performance chips because of their small sizes and attractive performance. However, as the speed of processing is being increased because of present demand aspects, it is again required to improve upon the parameters of FinFET. For example, reducing the leakage current and Energy delay product of transistor will reduce the overall power consumption of the chip and increasing the Gain-bandwidth product of the them will surely increase the amplifying ability over large bandwidth of input signal. Nanoscale vacuum channel transistors have been acting as promising devices, since they possess a very small OFF-state current and they show very high rigidity towards temperature. In fact, they don't suffer from the problem of impact ionization in the channel at harsh temperatures or radiations. So, accordingly, there has been a great demand of such devices at the scale of few nanometres, which ultimately, induces a more intense research in this particular field. Nevertheless, conventional vacuum tubes had been acting as boon from the perspective of high noise immunity and faster operable rates, vanished because of inclination towards high power consumption, sparse lifetime and arduous implementation methodology [1–6].

As the study was carried out, it was found out that not only the OFF-state current has been hugely decreased by orders of magnitude but also, most of the other parameters such as gain-bandwidth product, energy-delay product and some RF linearity parameter has been affected positively [7–11]. The usage of constant electric field scaling theory has been adopted for the enhancement of semiconductor devices, features such as gate length, height and operable voltages scaled down in accordance with this scaling theory. Scaling promotes higher circuit density and performance of overall IC onto the semiconductor chip. These parameters are intensively discussed in details here. In this paper, Silvaco TCAD has been used as semiconductor device modelling and simulation to devise and

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Mridul Prakash Kashyap
mridul.prakash2012@gmail.com

Sanmveg Saini
sanmveg@gmail.com

[1] Department of Applied Physics, Delhi Technological University, New Delhi, India

Published online: 22 April 2021

study the influence of various structural parameters on the electrical performance of an active device [12–15].

It has already been investigated that a Novel FinFET structure (i.e., TF-FinFET) stands far better than the conventional FinFETs in terms of lower power dissipation, lower operable voltage, more signal processing strength and more immune to temperature variation at 3 nm node [2]. Now, innovating something better than better sound good, if seen from the perspective of growing advancement in the field of nanoelectronics. Keeping these parameters in mind, the NVCTF-FinFET has been introduced at 7 nm node, as it has never been studied beyond 10 nm scale due to structural limitations. NVCTF-FinFET is designed to amalgamate all gaining properties of TF-FinFET and nanoscale vacuum channel FET together at 7 nm gate length with inherit improvements of TF-FinFET such as 74.6% decline in substrate current at ON-state, 35.8% decrease in Subthreshold Swing (S.S), 4 times increment in intrinsic voltage gain and 10.6 times enhancement in early voltage, in comparison with conventional FinFET.

2 Device Structure and Simulation Methodology

Both the devices (i.e., NVCTF-FinFET and TF-FinFET) were kept same from structural point of view in Fig. 1(a). Silicon channel has been replaced by vacuum channel with same dimensionality shown in Fig. 1(b & c) and gate electrode material was replaced from molybdenum (Mo) in TF-FinFET to Tantalum Nitride (TaN) in NVCTF-FinFET. The fin part is divided into upper and lower parts in such a way that the lower one is 1/3rd of its upper one, in terms of width as well as in height [16–18]. Rest all the physical parameters remained same and could also be seen from Table. 1. Moreover, the analysed tunnelling phenomena is different in both the devices as showed in Fig. 1(d & e).

The following physical models are used to incorporate quantum correction. *BQP model* is taken in account to introduce a position dependent quantum potential (Q), which indicates to potential energy of a given carrier type. As, Bohr Quantum Potential (BQP) model can be used to address a similar problem as density gradient model does, but with better convergence and better calibration to Schrodinger-Poisson model. *Doping dependent Shockely Read-Hall (SRH) recombination model* is also considered to examine the phonon transition occurs in the presence of traps within the forbidden gap of semiconductor. *Concentration and field dependent Mobility models* are introduced to consider the transverse field and doping dependent parts of the mobility and expressed by Mathiessen's rule. *Crowell and Sze impact ionization model* also considered for TF-FinFET only, to report the relationship between electric field and the ionization rate. As the electric field in the device is high enough to accelerate free carriers up to a point, where they can generate more free carriers on colliding with atoms of crystal [16, 17]. *Klassen Band to Band Tunnelling (BBT) and FN-tunnelling models* are also introduced for TF-FinFET and NVCTF-FinFET respectively, to examine the electron to tunnel, by internal field emission from valence band to corresponding conduction band. As the device has sufficiently high electric field, the band bending may result into the electron tunnelling.

As, the device integration is now been emerging as the most crucial parameter in designing a FET with more compact size, more lifetime and less power consumption. So, accordingly, to overcome the thirst, the idea of TF-FinFET has been collaborated with the properties of vacuum channel FETs in

Table 1 Parameters for Device Simulation

Parameters	Dimensions of Simulated Devices	
	TF-FinFET (nm)	NVCTF-FinFET (nm)
Gate length (L_g)	7	7
Fin height (H_f)	8	8
Lower fin width (W_{fl})	2	2
Upper fin width (W_{fu})	6	6
Effective oxide thickness (T_{ox})	0.4	0.4
Gate metal workfunction	4.55 eV (Mo)	4.30 eV (TaN)
Substrate width (W_s)	2	2
Upper fin height (H_{fu})	6	6
Lower fin height (H_{fl})	2	2
Device Length (L_D)	25	25
Source Length (L_s)	9	9
Drain Length (L_d)	9	9
Device Height (H_D)	11	11

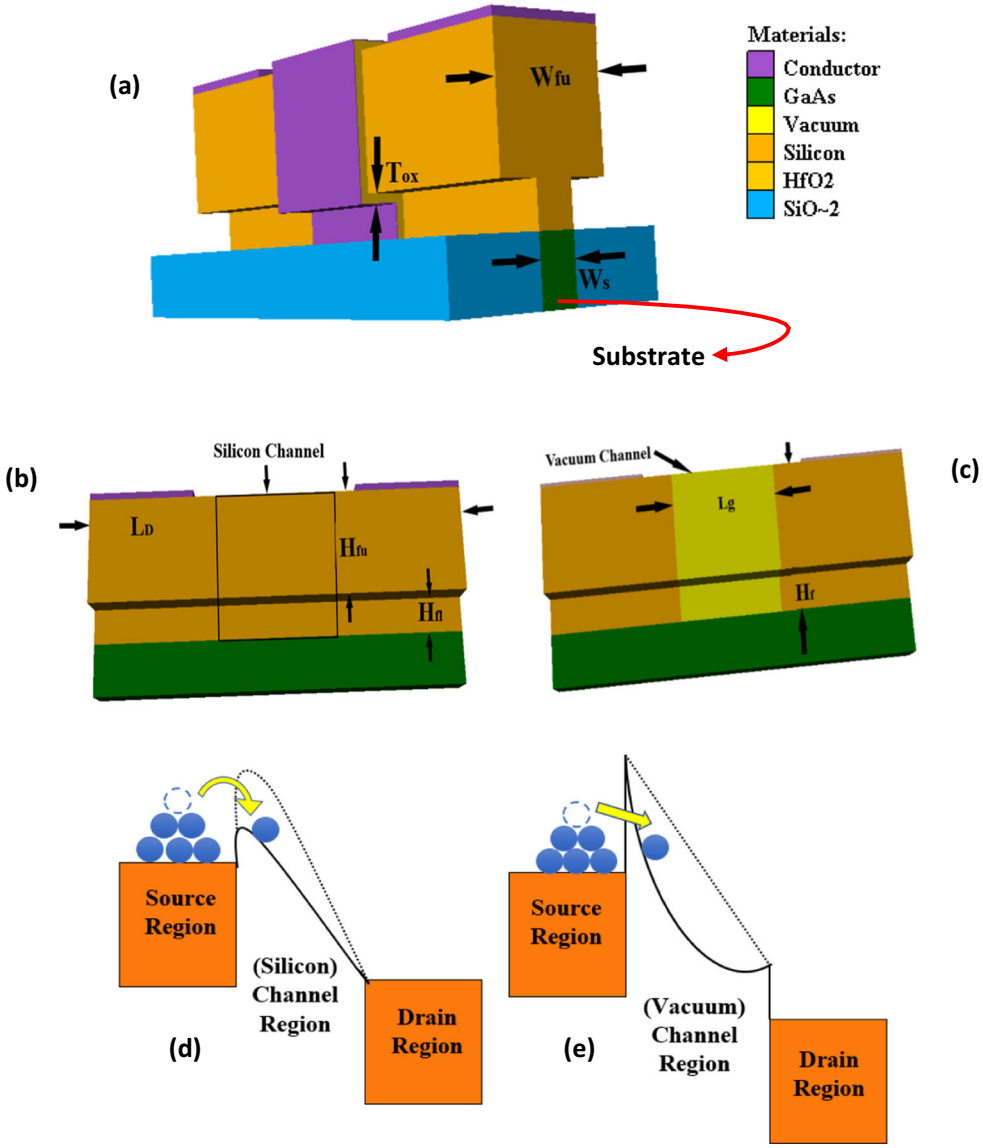


Fig. 1 (**a**) 3-D view of TF and NVCTF-FinFET. (**b**) & (**c**) shows 3-D view of TF-FinFET and NVCTF-FinFET respectively, without gate and STI. (**d**) & (**e**) shows tunneling mechanism in TF-FinFET and NVC-FinFET respectively

order to intrude into an optimized device with enhance performance in the concerned field of designing.

During the simulation process, an appropriate physical model was required to show the conduction mechanism in channel of NVCTF-FinFET. As, its already known that the solid-state devices have exceptional inclination towards the drift-diffusion mechanism, whereas, it is not suitable in case of vacuum channels due to absence of intrinsic charge carriers in the channel. Thus, the Fowler-Nordheim tunnelling model was introduced in the simulation methodology to study and analyse the conduction in case of NVCTF-FinFET. As far as the transfer characteristics of devices are concerned, the tunnelling parameter comes into the picture. FN-tunnelling is entirely reliant on the high electric field in the source-channel region that has been proved theoretically and experimentally as well [19–22]. The tunnelling is responsible to show the movement of electrons from fermi level of silicon (source) to vacuum and the concerned density produced with this phenomenon can be analysed by Eq. 1.

$$J = \frac{e^3 F^2}{8\pi h \phi t} \exp\left[-\frac{8\pi \times \sqrt{2m} \times \phi^{1.5}}{3heF} v(y)\right] \tag{1}$$

Where, (ϕ) is the emitter (source) workfunction and (F) is electric field due to emitter. The above equation can be reduced to Eq. 2.

$$J = \frac{A\,\beta^2 F^2}{\phi} \exp\left[-\frac{B\,\phi^{1.5}}{\beta\,F}\right] \tag{2}$$

Where, (β) is called as enhancement factor, the values of constants (A) and (B) are 1.54×10^6 and 6.83×10^7

respectively. Now, incorporating the physical parameters as $J = I/\alpha$ and $F = Vd$. We get,

$$I = aV^2 \exp\left[-\frac{b}{V}\right] \tag{3}$$

Where, (α) is area of charge emission from source region and (d) is the vacuum channel width formed during the conduction between source and drain. Moreover, the constants like (a) and (b) are decided as follow:

$$a = \frac{\alpha\beta^2\left(1.56x10^{-6}\right)}{\phi\, d^2} \quad b = \frac{\phi^{1.5}\left(6.83x10^{9}\right)}{\beta\, d}$$

Hence, we get the final Eq. 4, which determines the occurrence of FN-tunnelling in vacuum channel.

$$\ln\frac{1}{v^2} = \ln a - \frac{b}{v}. \tag{4}$$

To have better accuracy while dealing with the vacuum channel, the FN-tunnelling model is incorporated at both the interfaces (Source-Channel and Drain-Channel) very precisely. Just to ensure the accuracy, the meshing precision has been taken as 0.2 nm from source to drain which divides these interfaces into calculable segments.

FN-tunnelling is evidently occurring in NVCTF-FinFET at low voltages, the reason behind it can also describe itself. As the reduction in gate length, reduces the channel length significantly and in turn, the electric field generated at low voltage stood comparable with the electric field induced in the vacuum channel with much greater length operated at high voltages (i.e., >2 V). The inverse relationship of electric field with the dimensional parameter of NVCTF-FinFET is responsible for the mentioned phenomena. Moreover, the study conducted in [23], also evident that the electric field increases in the channel while going towards shorter channel lengths. The electric field in NVCTF-FinFET depicted from Fig. 5, is sufficient to let the electron transfer from source to drain through vacuum channel. The same has been tried to show with Eq. 1.

3 Calibration and Validation of Models

Figure. 2 shows the calibration of the simulated work (when reproduced on the original Silvaco TCAD EDA tools), to the experimental (and modelled too) work performed in the (A.B. Sachid et. al). After applying the quantum confinement and other models discussed in the concerned paper [24], extracted very similar type of input characteristics, V_{th}, and S.S. Like I_{ON} as 450 uA/um at $V_{GST} = 0.65$ V and I_{OFF} as 189.5 pA/um at $V_{GST} = -0.3$ V, which reflected around 4.65% and 1% deviation as the trend analysed by respective paper [24]. Moreover, extracted V_{th} as 0.318 V and S.S as 69.98 mV/dec, which reflected around 0.99% and 6% deviation from original values of the experimental work [24].

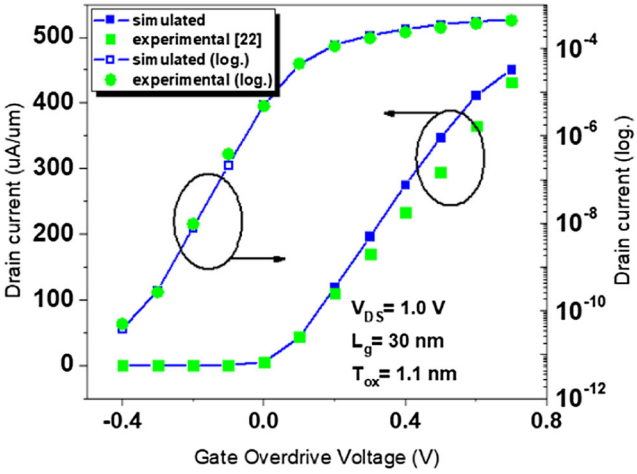


Fig. 2 Calibration of our simulated work with the experimental published work [24] at gate length of 30 nm

4 Resuts and Discussion

4.1 Correlation of Tunneling Current with Workfunction of Source (Emitter)

From Eq. 3, one can depict the inverse relationship Between current from the source (emitter) and the local workfunction (ϕ) of electron emitting surface. For the analysis, the combined equation has been formed as Eq. 5 for simple triangular barrier tunnelling.

$$I = \frac{\alpha\beta^2\left(1.56x10^{-6}\right)}{\phi\, d^2} V^2 \exp\left[-\frac{\phi^{1.5}\left(6.83x10^{9}\right)}{\beta\, d\, v}\right] \tag{5}$$

For the sufficiently large electric field, due to application od appropriate voltage in the dimensions of few nanometres, the silicon emission surface can be assumed as the degenerate in spite of n-type bulk doping [21]. So, at this instance, silicon could be qualitatively considered as metal and the same has been tried to show from Eq. 1.

4.2 Reason of Selecting Tantalum Nitride (TaN) as Gate Metal Electrode in NVCTF-FinFET

As studied from [25, 26], the performance of device is deeply dependent on workfunction difference of channel material and gate electrode. If the difference of workfunction between channel material and gate electrode is higher, higher would be the threshold voltage (V_{TH}), and improvement in S.S and switching ratio, possessed by the simulated devices. The standard workfunction of silicon is generally taken as 4.1 eV and the gate metal electrode with workfunction 4.55 eV (Mo) has

been used to attain the appropriate V_{th} at 7 nm node. Whereas, the vacuum channel is taken into account in case of NVCTF-FinFET, the gate metal electrode TaN (with workfunction 4.30 eV) has been used, just to make large difference between their work functions. Due to which improvement in $V_{TH,}$ S.S, switching ratio and other performance parameters is observed. The large barrier must be employed in case of NVCTF-FinFET, due to condition of FN-tunneling (by overcoming triangular barrier). Higher the barrier between source and drain, low is the probability of having conduction between them at OFF-state.

The metal gate workfunction is crucial parameter for the device integration, as it directly affects the V_{TH} and hence overall performance of the active device. By considering Fig. 3(a) in comparison with Fig. 4(a), it can be seen that the NVCTF-FinFET with Molybdenum (Mo) gate electrode possesses I_{OFF} (OFF-state current) in range of 10^{-15} A/um and in the range of 10^{-18} A/um when Tantalum Nitride (TaN) gate electrode has been implemented at the gate of NVCTF-FinFET at 300 K, which ultimately indicates the 10^3 times improvement with TaN gate electrode. Whereas, if seen at 450 K, Mo gate electrode in case of NVCTF-FinFET possess 530 times more I_{OFF} from TaN gate electrode NVCTF-FinFET. This optimization owes the switching ratio enhancement of NVCTF-FinFET with TaN gate electrode of about 500 times and 10 times at 300 K and 450 K respectively in Fig. 3(b).

Another aspect raised by considering the effects of different channel material while having the same gate electrode material, if the plots of Fig. 3 would be analysed separately. In Fig. 3(a), NVCTF-FinFET marked its I_{ON} as 4.07×10^{-8} A/um and 1.543×10^{-7} A/um at 300 K and 450 K respectively, whereas TF-FinFET marked its I_{ON} as 1.047×10^{-5} A/um and 6.130×10^{-5} A/um at 300 K and 450 K respectively. Hence, TF-FinFET with molybdenum gate metal and silicon channel material showed faded results in terms of temperature tolerance (3.78 times I_{ON} increment), from device's reliability point of view, as compared to NVCTF-FinFET with vacuum channel and Mo as gate metal electrode (5.85 times I_{ON} increment). From this analysis, one can grasp the robustness of

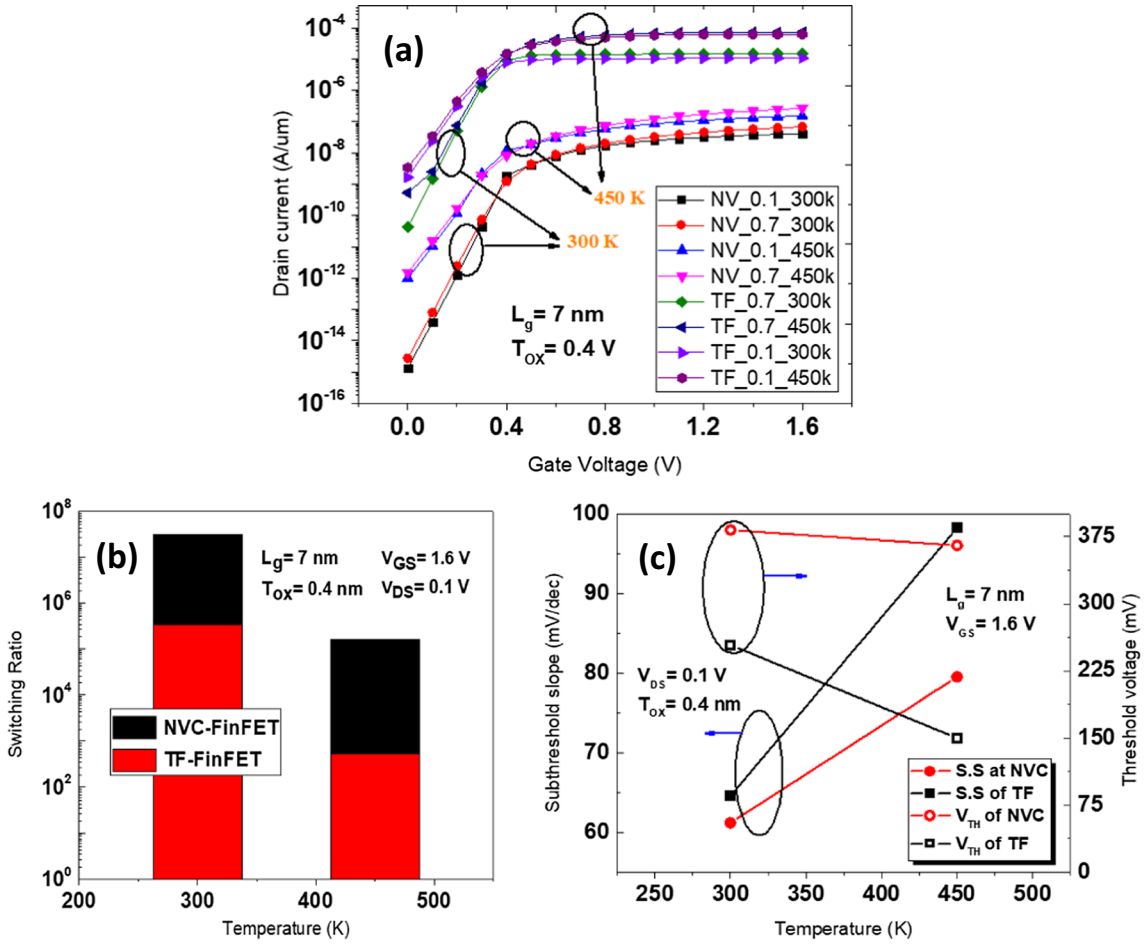


Fig. 3 (**a**) Dependency of NVCTF-FinFET Drain current on gate input voltage at 300 K and 450 K, (**b**) Switching ratios of NVCTF-FinFET and (**c**) Variation of Subthreshold slope and threshold voltage w.r.t temperature, when (Mo) gate electrode is implemented on both the devices

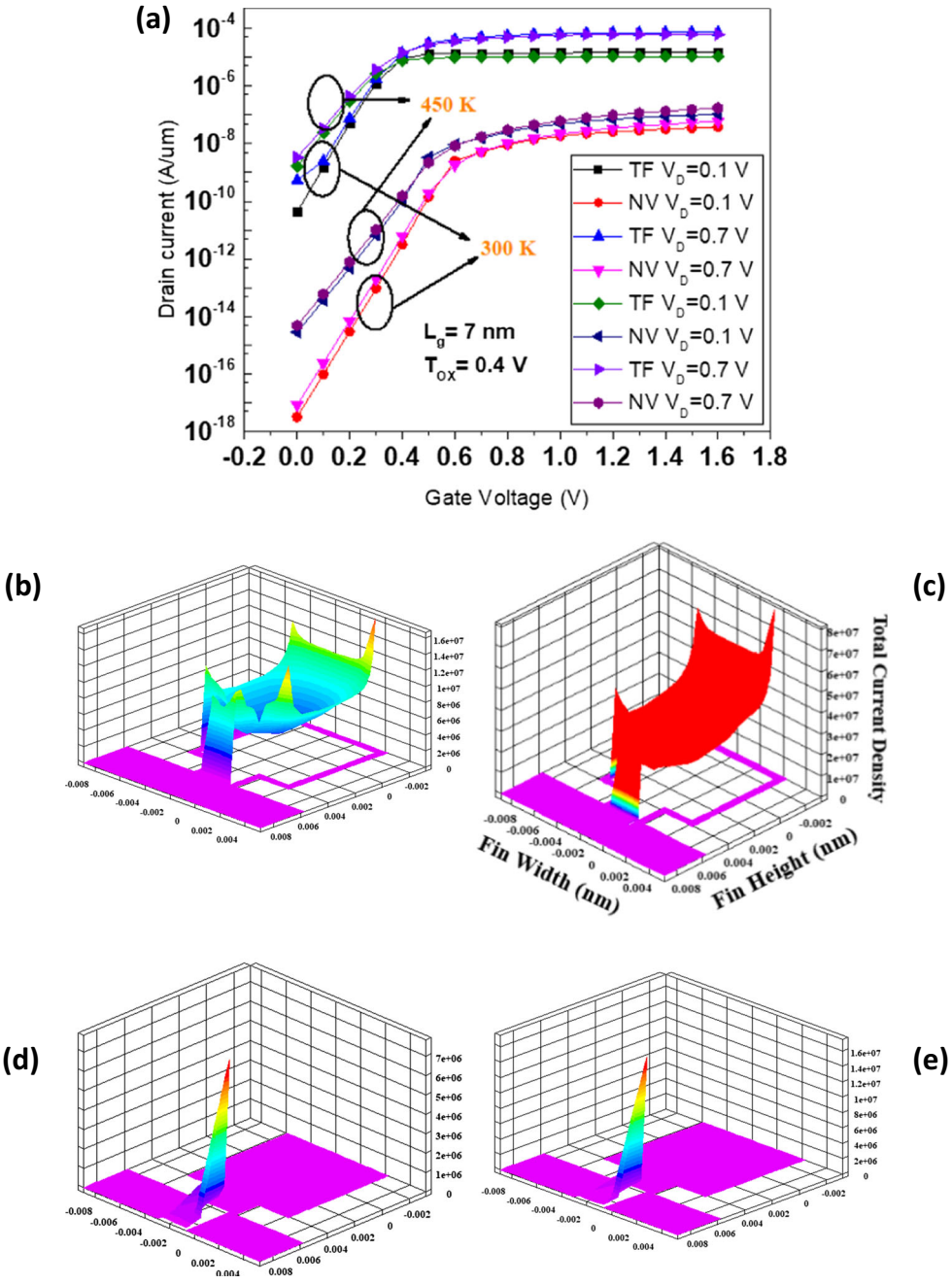


Fig. 4 (**a**) Variation of drain current w.r.t V_{GS} on different temperatures, (**b** & **c**) shows the current density in 3-D lateral cross-sectional view of TF-FinFET at 300 K and 450 K respectively and at V_{DS} = 0.7 V, (**d** & **e**) shows the current density in 3-D lateral cross-sectional view of NVCTF-FinFET at 300 K and 450 K respectively and at V_{DS} = 0.7 V when different gate materials were applied on both the devices

NVCTF-FinFET over TF-FinFET from 2 replacements (silicon channel with vacuum channel and Mo gate metal with TaN gate metal).

Just to indicate the necessity of channel material, switching ratios of both the devices has also been investigated. In Fig. 3(b), NVCTF-FinFET claimed its optimization by recording 90.5 times increment (3.83×10^5 for TF-FinFET and 3.0605×10^7 for NVCTF-FinFET) in switching ratio from TF-FinFET at 300 K and around 300 times increment (527.94 for TF-FinFET and 1.574×10^5 for NVCTF-FinFET) if seen from the perspective of 450 K.

Apart of affecting OFF-state current and hence the switching ratio, the replacement of gate metal electrode from Mo to TaN has also impacted the subthreshold slope (S.S) and Threshold voltage of NVCTF-FinFET as shown in Fig. 3(c). The S.S values for NVCTF-FinFET with Mo electrode stood at 61.3 mV/dec and 79.55 mV/dec at 300 K and 450 K respectively, which indicates the deterioration of about 1.6% and 50.1% at 300 K and 450 K respectively if compared from NVCTF-FinFET with TaN gate electrode. Additionally, if V_{TH} is taken into account, the amount of deterioration has observed about 33.8% and 2.5% at 300 K and 450 K

respectively if compared from NVCTF-FinFET with TaN gate electrode over the fin. The properties discussed above are some basic fundamental characteristics which insisted us to use TaN as gate metal electrode for NVCTF-FinFET instead of using Molybdenum over there.

Again, by exploring the effects of changing channel material only, the improvements occurred in the values of S.S and V_{th} has been observed at 2 different temperatures as shown in Fig. 3(c). The values of subthreshold slope for NVCTF-FinFET were investigated as 61.24 mV/dec and 79.55 mV/dec at 300 K and 450 K respectively, whereas they stood as 64.6 mV/dec and 98.3 mV/dec at 300 K and 450 K respectively for TF-FinFET. Hence, one can say that NVCTF-FinFET with Mo gate metal has more decremental change of drain current while tapping down the corresponding gate voltage, by the factor of 10. On the other side, the threshold voltage values of the devices, which describes the reliability if seen from electron travelling ease across the channel in subthreshold region, found out to be 382.42 mV and 365.3 mV at 300 K and 450 K respectively for NVCTF-FinFET and 254 mV and 149.9 mV at 300 K and 450 K respectively for TF-FinFET. The analysed parameters are well sufficient in itself to reflect the robustness of NVCTF-FinFET with both type of gate metals.

4.3 Usage of TaN Gate Metal Electrode on NVCTF-FinFET for Optimization

The input characteristics of both the transistors are well efficient to evident their behaviour in harsh conditions as well as at normal. Since, at higher temperature, atoms of Si get ionized and devise an electron-hole pair in channel region, which ultimately can collapse the overall functionality of the device. By considering Fig. 4(a), it could be easily analysed that NVCTF-FinFET is far better than its counterpart in terms of durability at high temperature. Here, in Fig. 4(b & c) the peak current values could be depicted to draw difference. The peak current densities in TF-FinFET at 300 K and 450 K are 1.6×10^7 A/cm^2 and 7.8×10^7 A/cm^2 respectively, whereas, it is 3.3×10^6 A/cm^2 and 7.0×10^6 A/cm^2 for NVCTF-FinFET in Fig. 4(d & e). So, accordingly, the current densities varying about 5 times in TF-FinFET on increasing temperature, far worse than NVCTF- FinFET [18, 27–29].

As far as efficiency and gain of the devices are concerned, here also NVCTF-FinFET stood in firm enhancement to TF-FinFET. As depicted from Fig. 5(a), the efficiency, showed with TGF (Transconductance generator factor), found better for NVCTF-FinFET from $V_{GS} = 0.1$ V onwards. However, it is only slightly better at higher voltages [30, 31]. The reason is optimization in the ratio of g_m to I_d, in case of NVCTF-FinFET. Moreover, intrinsic gain, being a quantity to specify the gain in device performance, also inclined towards the NVCTF-FinFET if analysed from $V_{GS} = 0.45$ V to upper constraint. The same analysis of TGF has also been conducted at $V_{DS} = 0.7$ V, that could be depicted from Fig. 5(b). At higher applied drain to source voltage as well, the NVCTF-FinFET stood extremely well with good difference (of about 16.5 times enhancement at $V_{GS} = 0.4$ V) if compared with TF-FinFET. The context itself infer the robustness of NVCTF-FinFET over TF-FinFET at higher drain voltages.

Since, the dielectric oxide has significantly low thickness in both the simulated device, the trade-Off has been observed between the gate controllability on carrier transfer and the amount of leakage current through the gate terminal. The trade-Off found quite heavy in case of NVCTF-FinFET, if compared from TF-FinFET as shown in Fig. 5(b). The gate current stood more in NVCTF-FinFET, by the factor of 2, than TF-FinFET at $V_{GS} = 1.6$ V. So, as a matter of fact, the appropriate parameters consideration must be taken into account, from the perspective of good logic circuit designing, to achieve a balanced performance of NVCTF-FinFET onto the wafer.

The current transfer characteristics if NVCTF-FinFET also changes while going from 300 K to 450 K as shown in Fig. 4. The reason behind is the thermionic emission occurred from the source region when the thermal energy of charges overcome the workfunction of emitter (source) material. Moreover, from Eq. 1, we can see that current density changes with F and ϕ hence they are the design parameters. It has been shown that work function of N-type Silicon semiconduction vary relatively less, with respect to change in temperature [32, 34]. In case of NVCTF-FinFET, we are changing the temperature from 300 K to 450 K and in this range, the work function of channel remains practically constant. It is clearly evident from Fig. 6 that increase in emitter electric field at 450 K (from 300 K) is appreciable due to which gate leakage current increases in NVCTF-FinFET.

It is being evident by Fig. 7 that TF-FinFET has much relatively larger cornered electric field, which is denoting reduction of reliability due to higher amount of electric field in gate oxide near drain region. This induces high impact-ionization rate in TF-FinFET. Nevertheless, this phenomenon is significantly absent in case of NVCTF-FinFET, hence increasing its reliability [21]. For reliability perspective, the switching ratio could be considered as a crucial parameter, signifying the overall gap between I_{ON} and I_{OFF}. In spite of low I_{ON} in case of NVCTF-FinFET in Fig. 7(d), it has max. Ratio because of its optimization from I_{OFF} point of view. So, TF-FinFET is relatively leaky at room as well as at harsh temperature.

Figure 8 (a) showed the output characteristics of the devices. Here, Early voltage is claiming the betterment of NVCTF-FinFET over TF-FinFET. Since, it symbolizes the amplification capacity of the amplifying device. Hence, amplification power of NVCTF-FinFET is more due to minimal

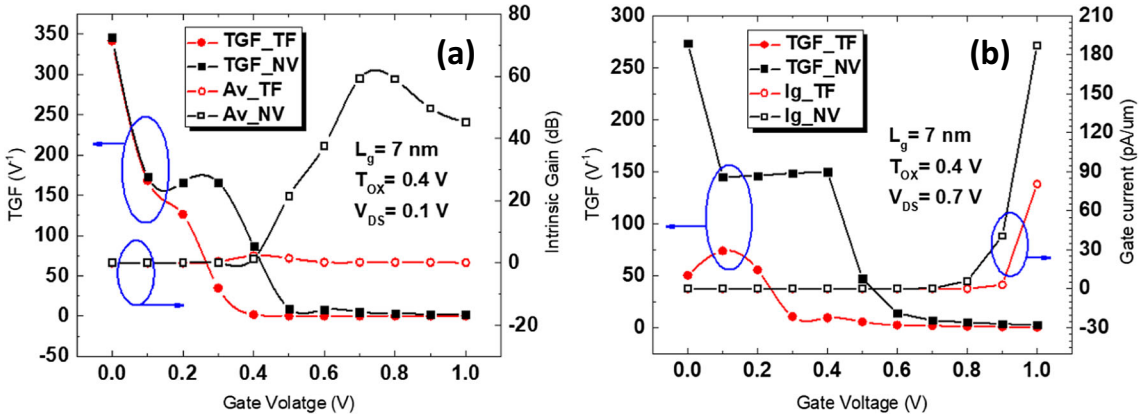


Fig. 5 (**a**) Dependency of TGF and Intrinsic gain of simulated devices on V_{GS}, (**b**) Dependency of TGF and Gate leakage current of both the simulated devices on V_{GS} when different gate materials were used (TaN for NVCTF-FinFET and Molybdenum for TF-FinFET)

drain-source conductance in this case. If seen from I_D w.r.t drain voltage point of view, TF-FinFET is relatively less stable between $V_{DS} = 0.1$ V to $V_{DS} = 0.3$ V. This ultimately denotes the superiority of NVCTF-FinFET. In Fig. 8(b), the GBP (Gain Bandwidth product) is specified to determine the max. Gain that can be extracted from the device in bandwidth of 1 GHz to a few THz and vice-versa as well. Being a device with least C_{gd}, GBP is max. For NVCTF-FinFET at $V_{GS} =$ 0.35 V and so forth, which ultimately denotes the feasibility of it, at this high frequency as well. Nonetheless, the product of average energy and gate delay (EDP) is inclined towards TF-FinFET because of its higher value of I_{ON} [19, 33].

Figure 9 is dedicated to show the working of devices with the help of scattering parameters. Here, both the devices are operated at non-zero values of V_{GS} and V_{DS} however, NVCTF-FinFET, showed in Fig. 9(b), acted as a symmetric

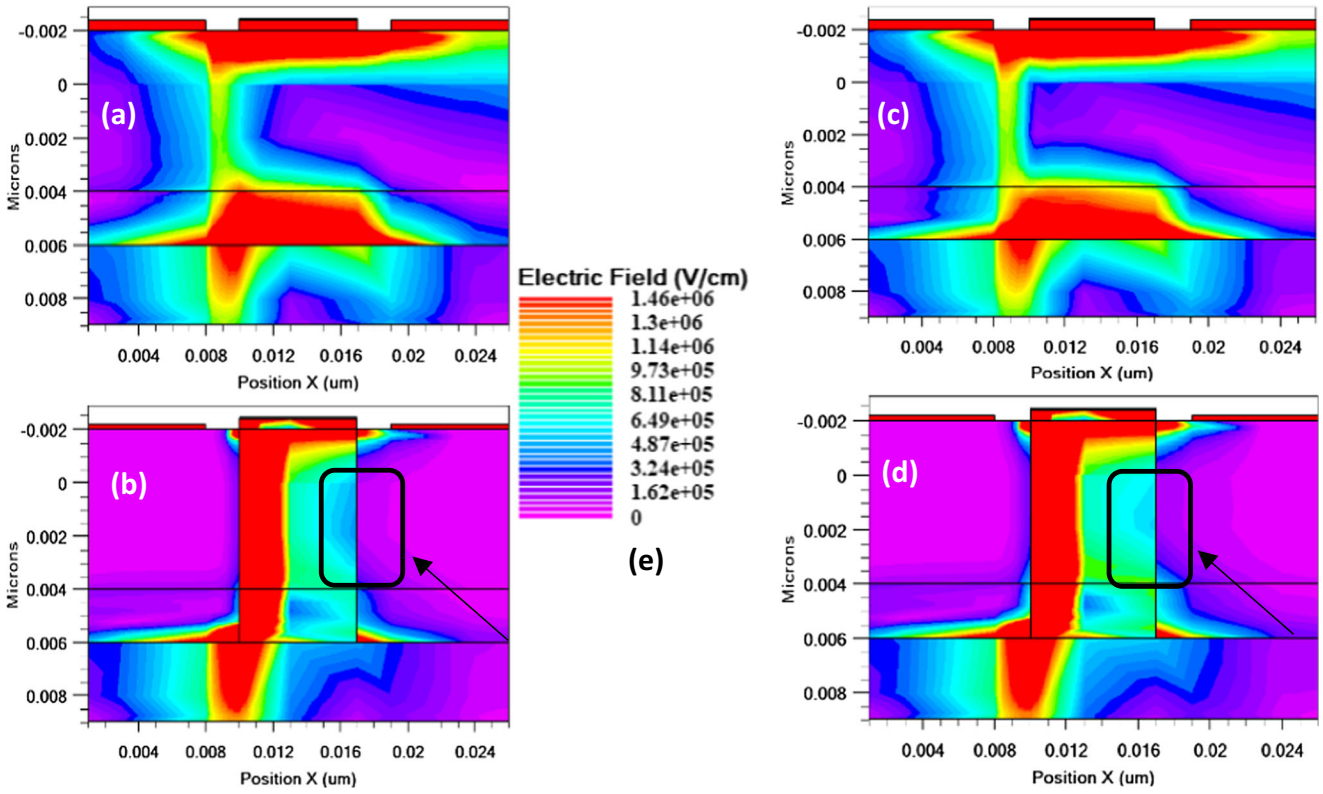


Fig. 6 (**a** & **b**) Variation of electric field in TF-FinFET and NVCTF-FinFET respectively at $V_{GS} = 1.6$ V and $V_{DS} = 0.7$ V at 300 K, if seen from longitudinal cross-sectional area, (**c** & **d**) Variation of electric field in TF-FinFET and NVCTF-FinFET respectively at 450 K, (**e**) Values of electric field used in this analysis when different gate materials were used on the devices (TaN for NVCTF-FinFET and Mo for TF-FinFET)

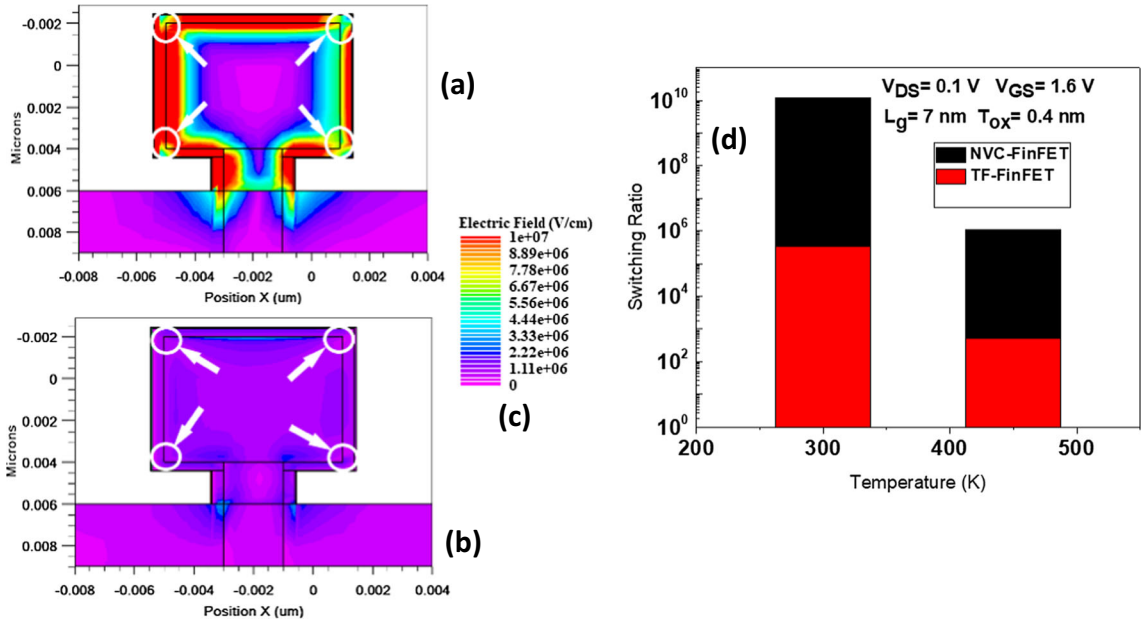


Fig. 7 (**a** & **b**) Variation of electric field in TF-FinFET and NVCTF-FinFET respectively at $V_{GS} = 1.6$ V and $V_{DS} = 0.7$ V, if seen from lateral cross-sectional area, (**c**) Values of electric field used in this analysis, (**d**) Change in Switching Ratio w.r.t temperature, when different gate materials have been used for both the devices

device (i.e., $S_{12} = S_{21}$) due to absence of appropriate current flow. As depicted from Fig. 9(a), TF-FinFET stood at ON-state, indicating transitive gain (i.e., forward gain is higher than the inverse gain). This due to higher through mechanism (S_{21}) as compared to feedback mechanism (S_{12}) and hence device is considered asymmetric.

Before going to the advance figure of merits, the major capacitances (C_{gd} and C_{gg}) w.r.t gate voltage have been obtained, at input signal frequency of 1 GHz, with the help of Silvaco TCAD capacitance module. The plot is analysed in order to predict the transitional behaviour of the simulated devices. Higher the parasitic capacitance in the device, higher is the time taken by the same device to process a signal (high or low). In Fig. 10, the C_{gd} and C_{gg} have been extracted to analyse the electric energy storing ability between the concerned terminals. More is the electron density between gate and drain at ON-state, more is the gate -drain capacitance between the 2 terminals. From the figure, one could depict the significantly low electron density between gate-drain in case of NVCTF-FinFET, as FN-tunnelling is the major source of conduction. As far as, C_{gg} is concerned, it stood lower in NVCTF-FinFET in spite of same gate dielectric material in both the devices. But, due to the pinch-off channel near gate-drain interface reduces the area of channel length under gate and hence, C_{gg} of TF-FinFET found greater.

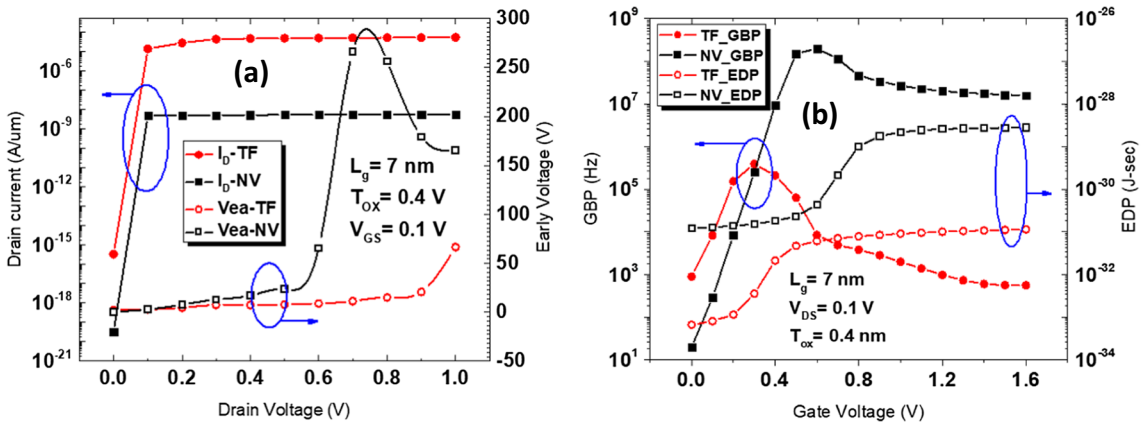


Fig. 8 (**a**) Dependency of drain current and Early voltage of simulated devices on V_{DS}. (**b**) Variation of GBP and EDP of simulated devices w.r.t V_{GS} when different gate materials were used (Mo for TF-FinFET and TaN for NVCTF-FinFET)

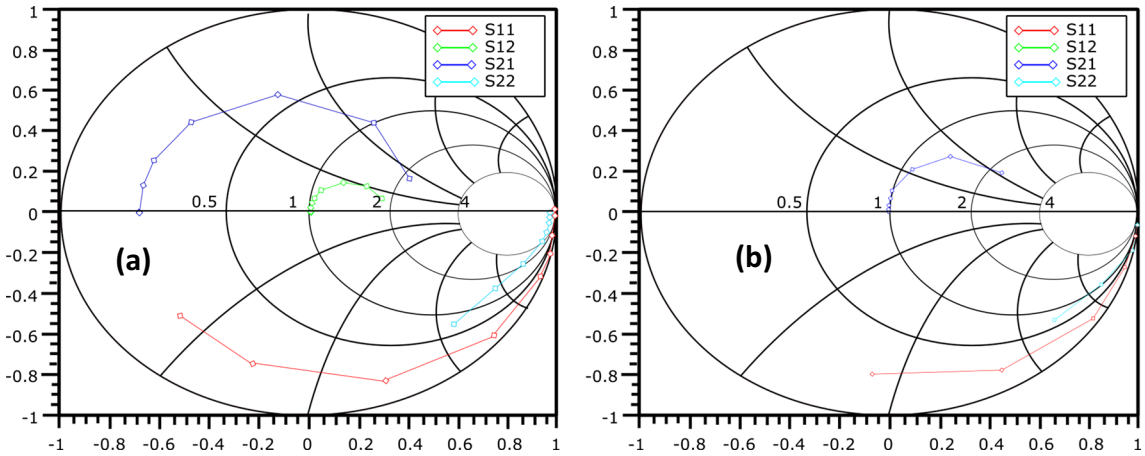


Fig. 9 (**a**) Variation of scattering parameters of TF-FinFET at $V_{DS} = 0.4$ V and $V_{GS} = 0.3$ V. (**b**) Variation of scattering parameters of NVCTF-FinFET at $V_{DS} = 0.4$ V and $V_{GS} = 0.3$ V while having different gate materials for both FinFETs

As per the simulation, the enhanced values of S.S (Subthreshold Swing) and V_{th} (Threshold voltage) were obtained at 300 K and 450 K. Values of S.S stood at 64.6 mV/dec and 98.3 mV/dec at 300 K and 450 K respectively for TF-FinFET, whereas it stood 60.3 mV/dec and 53 mV/dec at 300 K and 450 K respectively for NVCTF-FinFET. As far as V_{th} is concerned, it stood at 254 mV and 149.9 mV at 300 K and 450 K respectively for TF-FinFET, whereas stood 575 mV and 373 mV at 300 K and 450 K respectively for NVCTF-FinFET. Here are some equations which were used during analysis [27].

$$EDP = \left(C_{gg}V_D^2\right)\left(C_{gg}V_D/I_{on}\right) \tag{6}$$

$$GBP = (g_m)/\left(2\pi \times 10 \times C_{gd}\right) \tag{7}$$

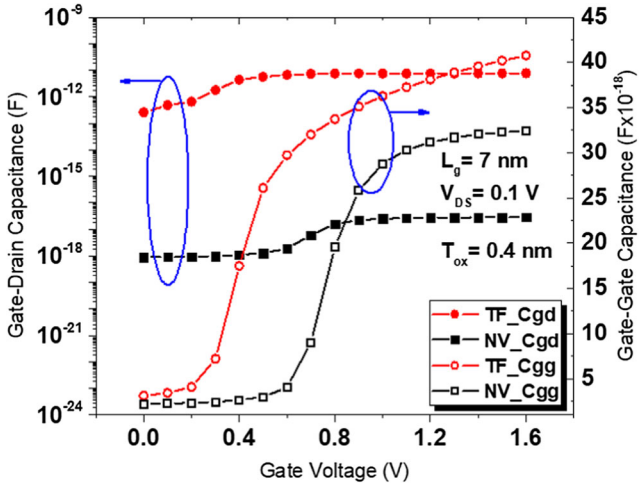


Fig. 10 Variation of Gate-Drain Capacitance (Cgd) and gate-Gate Capacitance (Cgg) w.r.t gate voltage for different gate materials applied on both the devices

$$1\text{-}dB\ Compression\ point == 0.22 \times \sqrt{{}^{g_{m1}}/_{g_{m3}}} \tag{8}$$

$$IP3 = 10log\left({}^{g_{m1}}/_{g_{m3}}\right) + 12.2 \tag{9}$$

$$VIP3 = \sqrt{24 \times {}^{g_{m1}}/_{g_{m3}}} \tag{10}$$

$$IMD3 = \left({}^{9}/_{2} \times (VIP3)^3 \times g_{m3}\right)^2 \times R_s \tag{11}$$

$$HD3 = 0.25V_a{}^2 \frac{\left({}^{d^2 g_{m1}}/_{d\ V_{GT}{}^2}\right)}{6g_{m1}} \tag{12}$$

As depicted from Fig. 11(a), the compression points of NVCTF-FinFET at 1-dB, stood larger than TF-FinFET for V_{GS} greater than 0.5 V. Being a quantity, denoting point where distortion starts and compression occurs. This enhancement is due to less increment of g_{m3} in NVCTF-FinFET which can also be analysed from Eq. (8).

From analog and mixed-signal circuit perspective, linearity in CMOS is highly desirable. Same has been shown with the help of IP3 and VIP3 from Eqs. (9 & 10). One can easily analyse from Fig. 11(b) that NVCTF-FinFET has higher value of IP3, all the way from $V_{GS} = 0.6$ V. Moreover, from VIP3 variation, it's quite insisted to observe that TF-FinFET has relatively less extrapolated input voltage at which 1st and 3rd harmonics are equal. The reason here is the higher ratio of 1st order harmonic (g_m) to 3rd order harmonic (g_{m3}) [20].

As far as Inter-modulated harmonics power (IMD3) is concerned, it stood relatively maximum for TF-FinFET from $V_{GS} = 0.5$ V onwards in Fig. 12(a), showed in Eq. (11). It is due to the fact that NVCTF-FinFET has lower extrapolated power at which 1st and 3rd harmonic powers are equal. Moreover, harmonic distortion plays an important role in determining the distortion parameters in an

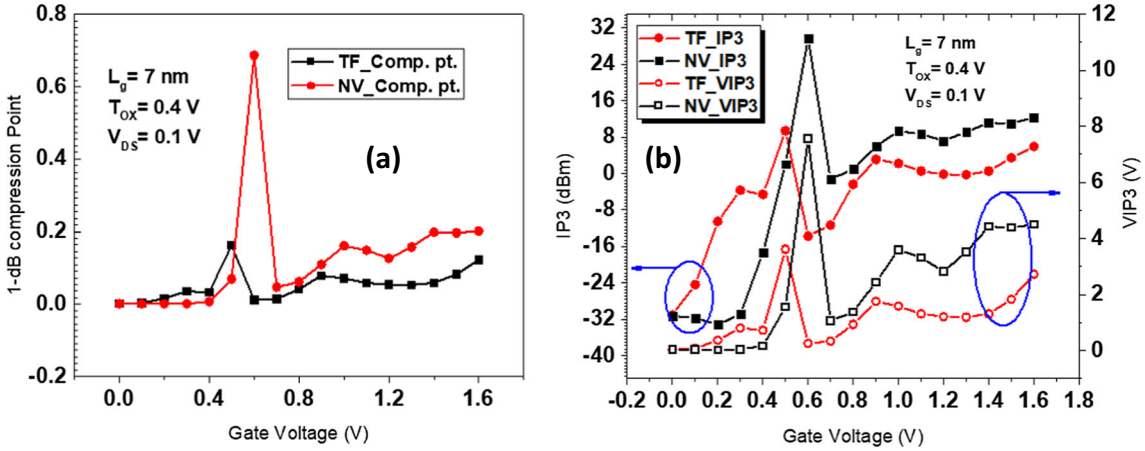


Fig. 11 (**a**) Variation of 1-dB compression point w.r.t gate voltage, (**b**) Variation of IP3 and VIP3 w.r.t gate voltage of simulated devices with different gate materials

amplifying device, same has been shown w.r.t to its 3rd order (HD3) in Eq. (12) as well. Here, even NVCTF-FinFET has relatively lower values due to minimal ratio of g_m to g_{m3} in itself. Additionally, just to present the effects of channel medium replacement and Gate metal replacement, both the devices have also been investigated with same gate metal electrode (i.e., Mo) in Fig. 12(b). Here also, NVCTF-FinFET stood far away from TF-FinFET if analysed from the perspective of same distortional parameters (i.e., HD3 and IMD3). Hence NVCTF-FinFET has relatively more capacity to preserve the processing signal from getting distorted.

5 Conclusion

A comprehensive study on the overall performance of the TF-FinFET and NVCTF-FinFET has been conducted in order to determine the optimized characteristics of the best between them and to show the viability of nanoscale vacuum channel transistors at such a miniature gate length of 7 nm. All the simulation work was accomplished on Silvaco Atlas TCAD EDA tools. NVCTF-FinFET provided superior variation of IP3, VIP3, IMD3, HD3 over its counterpart. Moreover, from analog perspective, NVCTF-FinFET stood firm at higher slot too, as TGF, Early voltage and Intrinsic gain were proved more inclined towards it. With the combination of Vacuum channel and TaN as gate electrode material in TF-FinFET, NVCTF-FinFET stood out to be more durable and reliable at harsh environments (i.e., at extremely high and low temperature and more immune to hazardous radiations too) as well. As NVCTF-FinFET provided significantly high switching ratio at input voltage (<2 V), which could be considered as a strong alternative to the semiconductor transistors beyond Moore's Law.

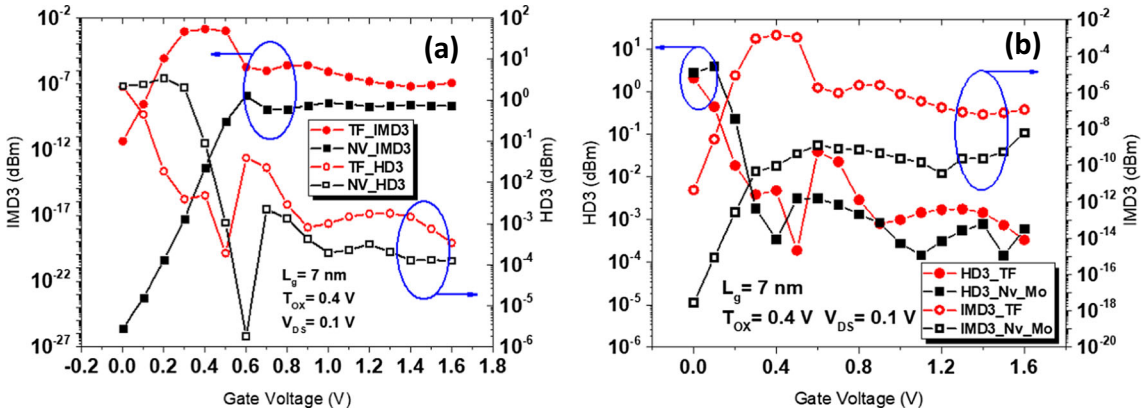


Fig. 12 (**a**) Dependency of IMD3 and HD3 on gate voltage for both the simulated devices with different gate materials (TaN for NVCTF-FinFET and Molybdenum for TF-FinFET), (**b**) Dependency of IMD3 and HD3 on gate voltage for both the simulated devices with same gate material (Molybdenum)

Acknowledgments All the authors would like to show the extreme regards to Microelectronics Research Laboratory, Department of Applied Physics, Delhi Technological University, for supporting this research work.

Code Availability The relevant code with the manuscript is also available and would be available, if will be asked to do so later.

Authors' Contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mridul Prakash Kashyap, Sanmveg Saini and Rishu Chaujar. The first draft of the manuscript was written by Mridul Prakash Kashyap and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors are grateful to Microelectronics Research Laboratory, Department of Applied Physics, Delhi Technological University for supporting this research work. However, there has been '***no funding***' grants from the organization.

Data Availability The above-mentioned authors have all the relevant data associated with this research work and will be dedicated to share that, if will be asked to do so in future.

Declarations

Ethics Approval ***Not Applicable*** (as *the results of studies does not involve any human or animal*).

Consent to Participate ***Not Applicable*** (as *the results of studies does not involve any human or animal*).

Consent for Publication ***Not Applicable*** (as *the results of studies does not involve any human or animal*).

Conflict of Interests/ Competing Interests 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. Boukortt N, Patanè S, Crupi G (2019) 3D investigation of 8-nm tapered n-FinFET model. Silicon 12:19–20. https://doi.org/10.1007/s12633-019-00253-y
2. Kashyap MP, Chaujar R (2020) "temperature investigation of a novel 3nm TF-bulk FinFET for improved performance," 2020 IEEE 20th international conference on nanotechnology (IEEE-NANO). Montreal, QC, pp 382–387. https://doi.org/10.1109/NANO47656.2020.9183594
3. Kumar A, Tripathi MM, Chaujar R (2017) Investigation of parasitic capacitances of In2O5Sn gate electrode recessed channel MOSFET for ULSI switching applications. Microsystem Technol 23(12):5867–5874. https://doi.org/10.1007/s00542-017-3348-2
4. Kumar A, Gupta N, Chaujar R (2016) Analysis of novel transparent gate recessed channel (TGRC) MOSFET for improved analog behaviour. Microsyst Technol 22(11):2665–2671. https://doi.org/10.1007/s00542-015-2554-z
5. Yu E, Heo K, Cho S (2018) Characterization and optimization of inverted-T FinFET under Nanoscale dimensions. IEEE Trans Electron Devices 65(8):3521–3527. https://doi.org/10.1109/TED.2018.2846478
6. Sachid AB, Manoj CR, Sharma DK, Ramgopal Rao V (2008) Gate fringe-induced barrier lowering in Underlap FinFET structures and its optimization. IEEE Electron Device Letters 29(1):128–130. https://doi.org/10.1109/LED.2007.911974
7. Han MH, Chang CY, Bin Chen H, Wu JJ, Cheng YC, Wu YC (2013) Performance comparison between bulk and SOI junctionless transistors. IEEE Electron Device Lett 34(2):169–171. https://doi.org/10.1109/LED.2012.2231395
8. Huang WT, Li Y (2015) Electrical characteristic fluctuation of 16-nm-gate trapezoidal bulk FinFET devices with fixed top-fin width induced by random discrete dopants. Nanoscale Res Lett 10(1):1–8. https://doi.org/10.1186/s11671-015-0739-0
9. Zheng P, Liao Y, Damrongplasit N, Chiang M-H, Liu K, Tsu-Jae (2014) Variation-aware comparative study of 10-nm GAA versus FinFET 6-T SRAM performance and yield. IEEE Trans Electron Devices 61(12):3949–3954. https://doi.org/10.1109/TED.2014.2360351
10. Kashyap MP, Chaujar R (2020) Gate oxide variability analysis of a novel 3 nm truncated fin–FinFET for high circuitry performance. Silicon. https://doi.org/10.1007/s12633-020-00734-5
11. Kranti A, Armstrong GA (2007) Design and optimization of FinFETs for ultra-low-voltage analog applications. IEEE Trans Electron Devices 54(12):3308–3316. https://doi.org/10.1109/TED.2007.908596
12. Liu C, Zheng F, Sun Y, Li X, Shi Y (2017) Novel tri-independent-gate FinFET for multi-current modes control. Superlattice Microst 109:374–381. https://doi.org/10.1016/j.spmi.2017.05.025
13. Sachid AB, Chen MC, Hu C (2016) FinFET with high-κ spacers for improved Drive current. IEEE Electron Device Lett 37(7):835–838. https://doi.org/10.1109/LED.2016.2572664
14. Musalgaonkar G, Sahay S, Saxena RS, Kumar MJ (Feb. 2019) An impact ionization MOSFET with reduced breakdown voltage based on Back-gate misalignment. IEEE Trans Electron Devices 66(2):868–875. https://doi.org/10.1109/TED.2018.2887168
15. ITRS 2.0 Publication, "International Technology Roadmap for Semiconductors, 2015," Available online at http://www.itrs2.net/
16. PH Drive, S. Clara, Atlas User's Manual, Device Structure Editing and Simulation Software; 2017. Pp. 1–1715. Available: www.silvaco.com
17. Han J-W, Moon D-I, Meyyappan M (2017) Nanoscale Vacuum Channel Transistors. Nano Letters *17*(4):2146–2151. https://doi.org/10.1021/acs.nanolett.6b04363
18. Ghosh P, Haldar S, Gupta RS, Gupta M (2012) An investigation of linearity performance and intermodulation distortion of GME CGT MOSFET for RFIC design. Electron Devices IEEE Trans 59:3263–3268. https://doi.org/10.1109/TED.2012.2219537
19. Ellis RK (1982) Fowler-Nordheim emission from non-planar surfaces. IEEE Electron Device Letters 3(11):330–332. https://doi.org/10.1109/EDL.1982.25590
20. Forbes RG (1999) Refining the application of Fowler–Nordheim theory. Ultramicroscopy 79(1–4):11–23. https://doi.org/10.1016/S0304-3991(99)00097-2
21. Lenzlinger M, Snow EH (1968) Fowler-Nordheim tunneling into thermally grown SiO2. IEEE Trans Electron Devices 15(9):686–686. https://doi.org/10.1109/T-ED.1968.16430
22. Forbes R, Fischer A Mousa M (2012) Improved approach to Fowler-Nordheim plot analysis J Vacuum Sci Technol B vol 31. DOI: https://doi.org/10.1116/1.4765080
23. Xu J, Qin Y, Shi Y, Shi Y, Yang Y, Zhang X (2020) Design and circuit simulation of Nanoscale Vacuum Channel transistor. Nanoscale Advances 2:3582–3587. https://doi.org/10.1039/D0NA00442A

24. Sachid AB, Chen MC, Hu C (2017) Bulk FinFET with low-κ spacers for continued scaling. IEEE Trans Electron Devices 64(4):1861–1864. https://doi.org/10.1109/TED.2017.2664798
25. Ana F, Hakim N-U-D (2011) Gate Workfunction engineering for deep sub-Micron MOSFET's: motivation, Features and Challenges. Electron Commun Japan (Part III Fundamental Electronic Science) 2:29–35
26. Erben E, Hempel K, Triyoso D (2018) Work function setting in high-k metal gate devices, complementary metal oxide semiconductor. Kim Ho Yeap and Humaira Nisar, IntechOpen. https://doi.org/10.5772/intechopen.783
27. Kang T, Yang Y, Chien F (2014) Poly Si nanowire thin film transistors with vacuum gap design. IEEE Trans Electron Devices 61(6):2113–2118. https://doi.org/10.1109/TED.2014.2318706
28. Kang S, Choi B, Kim B (2003) Linearity analysis of CMOS for RF application. Microwave Theory and Techniques. IEEE Trans 51: 972–977. https://doi.org/10.1109/TMTT.2003.808709
29. Han J-W, Oh J, Meyyappan M (2012) Vacuum nanoelectronics: Back to the future? Gate insulated nanoscale vacuum channel transistor. Appl Phys Letters. 100. https://doi.org/10.1063/1.4717751
30. Garg N, Pratap Y, Gupta M, Kabra S (2020) Reliability assessment of GaAs/Al_2O_3 Junctionless FinFET in the presence of interfacial layer defects and radiations. IEEE Trans Device Mater Reliabil 20(2):452–458. https://doi.org/10.1109/TDMR.2020.2991662
31. Ballabio A, Bietti S, Scaccabarozzi A, Esposito L, Gurioli M, Isella G, Sanguinetti S (2019) GaAs epilayers grown on patterned (001) silicon substrates via suspended Ge layers. Nature Scientific Reports. DOI: 9. https://doi.org/10.1038/s41598-019-53949-x
32. Smith AH (1949) Temperature Dependence of the Work Function of Semiconductors. Phys Rev 75(6):953–959. https://doi.org/10.1103/PhysRev.75.953
33. Kumar A, Tiwari B, Singh S, Mohan Tripathi M, Chaujar R (2018) Radiation analysis of N-channel TGRC-MOSFET: an X-ray dosimeter. IEEE Trans Electron Devices 65(11):5014–5020. https://doi.org/10.1109/TED.2018.2869536
34. Lima L, Moreira M, Cioldin F, Diniz JA, Doi I, Pavanello M, Claeys C, Martino JA (2010) Tantalum nitride as promising gate electrode for MOS technology. ECS Trans 31:319–325. https://doi.org/10.1149/1.3474175

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

WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Analysis of Carbon Fiber

Siddharth Garg [a],Tanishq Gupta[a], Mukesh Dagde[b]

[a]UG Student, Department of Mechanical Engineering, Delhi Technological University Bawana Road, Delhi110042, India.
[b]Professor, Department of Mechanical Engineering, Delhi Technological University Bawana Road , Delhi110042, India.

A B S T R A C T

For many years, scientists worldwide have been looking for a wonder material that can fulfill all the needs of "raw" material needed in many industries, and all fingers after decades of research points towards Carbon Fiber. Carbon atoms bonded along the chain are the base material of Carbon fiber. They show various properties like being lightweight, robust, and a good conductor of electricity, and many more hence are used in many processes as basic building blocks. They come in various forms as the primary building material such as weaves, braids, yarns, and uni-directional, to name a few, which are used to create composite parts of various materials. With the weight that being of plastic and strength that being of steel, Carbon Fiber has proved itself to be the future raw material of various substances. Its high strength to weight ratio also makes it the most sought-after material in recent times.

Keywords: Carbon Fiber, Textile, Material, Yarn

1. Introduction

Carbon Fibers are made up of carbon atoms and are about 5-10mm in thickness. Although its efficient mass production is still a topic of ongoing research, making it a little bit more expensive than other types of fibers like glass and plastic, it still finds itself being used in multidimensional fields like aviation, automobile, telecommunication, and civil industries due to it being light, having high stiffness and high tensile strength along with high chemical resistance and temperature tolerance.

Figure 1: Threads of Carbon Fiber wrapped together

(Image Source: www.blendernation.com)

* *Corresponding author.*
E-mail address: *agentsid786@gmail.com*

2. Classification and Types

Carbon fibers can be classified on the basis of modulus, strength, and final heat treatment temperature:

2.1 On the basis of Carbon Fiber properties-

- Super high-tensile
- Ultra high modulus
- High modulus
- Intermediate modulus
- Low modulus and High tensile

2.2 On the basis of precursor fiber materials-

- PAN (Polyacrylonitrile) based carbon fibers.
- Pitch-based carbon fibers.
- Mesophase pitch based carbon fibers.
- Isotropic pitch based carbon fibers.
- Rayon based carbon fibers.
- Gas phase grown carbon fibers.

2.3 On the basis of final heat treatment temperature-

- High Heat Treatment carbon fibers (HTT), where final heat treatment temperature should be more than 2000°C and can be related with high-modulus type fiber.
- Intermediate Heat Treatment carbon fibers (IHT), where final heat treatment temperature should be around or more than 1500°C and can be related with high-strength type fiber.
- Low Heat Treatment carbon fibers (LHT), where final heat treatment temperatures is not more than 1000°C. These are low modulus and low strength materials.

3. Manufacturing Process of making Carbon Fibers from Polyacrylonitrile (PAN):

3.1 Raw Materials used-

Carbon Fiber is made up of raw material called precursor. The main component of Carbon Fiber of Polyacrylonitrile (around 90%) and the rest is made up of petroleum pitch or rayon. Carbon Fibers are organic molecules and are characterised by a long chain of molecules bound by carbon atoms. Various liquids and gases are used during the manufacturing process depending upon the properties desired in the final product.

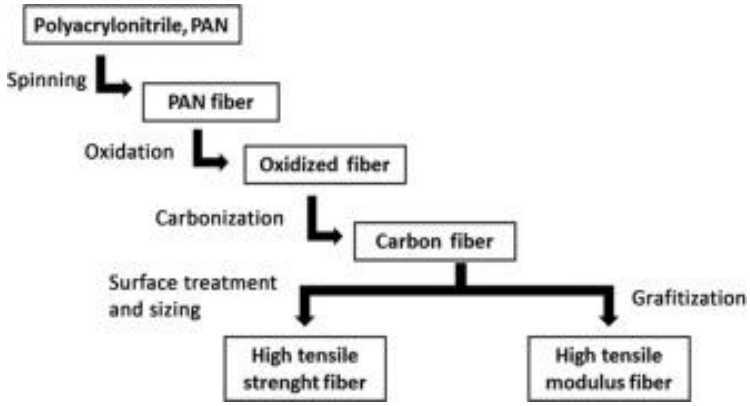


Figure 2: The Manufacturing Process PAN

Image Source: **www.sciencedirect.com**

3.2 Spinning

- Acrylonitrile plastic powder is blended with another plastic, like methyl acrylate or methyl methacrylate, and is reacted with a catalyst in a regular suspension or solution polymerization process to shape a polyacrylonitrile-plastic.
- In other techniques, the plastic mixture is warmed and pumped through tiny jets into a chamber where the solvents dissipate, leaving a solid fiber behind. This is an essential process as the internal atomic structure is formed during the process of spinning.
- After this, the fibers are washed and stretched to their desired width. This process helps in more substantial bond formation as the molecules get aligned within the fiber.

3.3 Stabilizing

Before the carbonization of Carbon Fibers, they are heated in air at around 200-300 C for about an hour so that the fiber can pick up oxygen molecules from the air and form a more thermally stable ladder bonding. This process is achieved with various tools and equipment like heated chambers and hot rollers, to name a few.

3.4 Carbonizing

After the stabilization process, the intermediate is heated at around 100-300 C in the absence of oxygen, and the pressure inside the chamber is kept higher than that of outside. The absence of oxygen prevents the fibers from burning at such high temperatures. As the temperature is increased and the fibers start to heat, they lose the non-carbon atoms in the form of various gases like H20, CO2, and CO. As the non-carbon atoms are lost along with some carbon atoms, the remaining carbon atoms align themselves in tightly formed carbon crystals which is more or less parallel to the axis of the fiber.

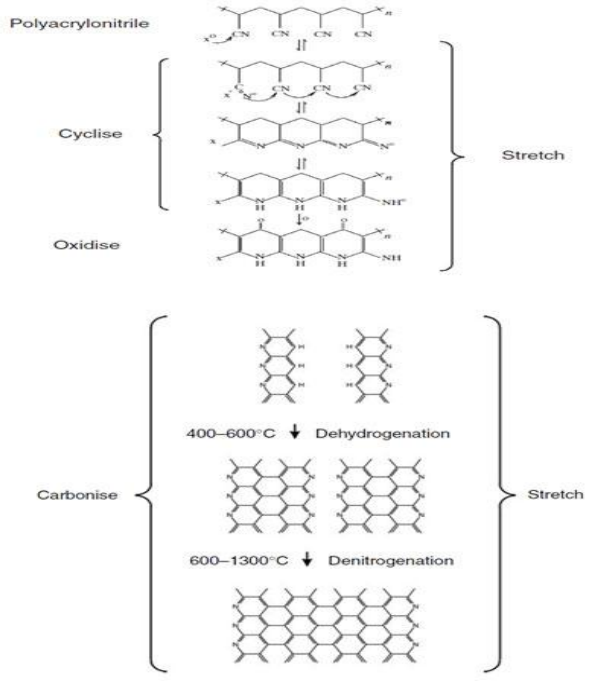


Figure 3: Chemical Reaction that happen during the process

Image Source: www.materialsciencejournal.org

3.5 Treating the surface

The intermediate formed after the previous process does not surfaces that are smooth enough to bond well with the epoxies and other materials used during the utilization of Carbon Fiber. To give it some better bonding properties, they are oxidized a little bit with oxygen. This can be achieved by either immersing the intermediate in gases like air, ozone or by dipping it in liquids such as nitric acid. Surface Coating through electrolysis is also a method that can be used but with great precautions.

3.6 Sizing

- In the process of Sizing, Carbon Fibers are coated with coating materials like epoxy and polyester, which are compatible with the adhesive used in the process of forming composites.
- "The coated fibers are wound onto cylinders called bobbins. The bobbins are loaded into a spinning machine, and the fibers are twisted into yarns of various sizes."[9]

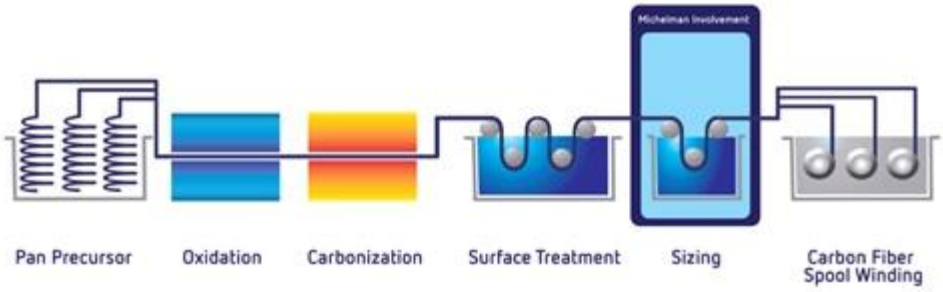


Figure 4: Complete Cycle used for the formation of Carbon Fiber
Image Source: www.azom.com

4. Properties

4.1 Carbon Fiber has High Specific Strength

Carbon Fibers have a good STRENGTH/WEIGHT ratio, making them capable of carrying a large load despite being not heavy without the fiber reaching its breaking point. Other materials that have good strength are Aluminium, High strength steel alloys, Titanium to name a few.

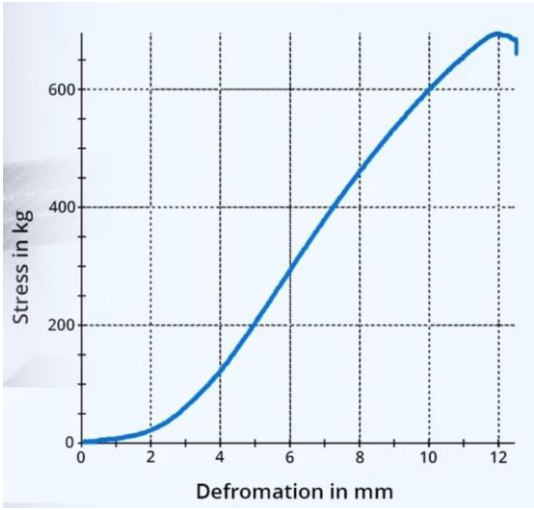


Figure 5: Curve showing the high stress capacity of Carbon Fiber
Image Source: www.easycomposites.uk

4.2 Carbon Fiber is very Rigid

Carbon Fibers have excellent stiffness or Rigidity, which means that they don't reflect much under high stress, which in turn is measured by Young's Modulus. Carbon fiber reinforced plastic is stiffer than Glass-reinforced plastic, pine, and aluminium.

4.3 Carbon fiber is Chemically Stable

Carbon Fibers are Chemically inert and Corrosion resistant; hence do not deteriorate with time, although epoxy and other matrices in which carbon is embedded may be reactive and needs to be protected.

4.4 Carbon fiber is Electrically Conductive

Carbon Fibers are highly conductive, and heat loss during transmission is also comparatively less than other materials, which make it a suitable material to be used in chargers; however, this ability can also be a bane in boat building where like Aluminium, it can facilitate Galvanic corrosion in the fitting;

therefore it has to be taken into account before the installation.

4.5 Carbon Fiber has good Tensile Strength

Carbon Fibers can withstand a large amount of stress while it is being stretched or pulled before its starts necking or failing, i.e., the cross-section of the sample starts to reduce dramatically. It is usually tested by taking carbon Fiber of definite width and then stretching it gradually until the sample changes the shape or breaks. Due to its internal flaws, the carbon Fiber never breaks at constant value but instead fails at minor strains. This property makes them a desirable raw material for various industries.

4.6 Fire Resistance/Non Flammable

Due to their flexible nature, the Carbon Fibers can be pretty soft and can be integrated into the materials of protective clothing made for firefighters. Its property of being chemically inert also helps in using it when the fire is combined with some corrosive agents to protect from both fire and any hazardous chemical reaction. Nickel-coated Carbon Fiber is a famous example of it being used in firefighting equipment.

4.7 Thermal Conductivity of Carbon Fiber

Thermal Conductivity is a measure of how effectively can heat can flow through a material. Due to so many Carbon Fiber variations, different forms can be used to achieve both high and low thermal Conductivity depending upon the requirement.

4.8 Low Coefficient of Thermal Expansion

The coefficient of thermal expansion is a measure of much a material expands or contracts depending upon the temperature variation. Carbon Fibers have a low thermal expansion coefficient, making them suitable for parts where even a tiny movement can be critical, like in a telescope or any other optical type of machinery.

4.9 Non Poisonous, Biologically Inert, X-Ray Permeable

Due to the properties mentioned above, Carbon Fibres have a variety of uses in Medical appliances like implants, tendon repair, X-ray accessories, and various other surgical instruments. Although Carbon Fibers are non-toxic yet their exposure should be limited due to the presence of other polyesters and epoxy in it.

4.10 Fatigue Resistance is good

The resistance to Fatigue ratio in Carbon Fiber Composites is very good than compared to any other materials.

5. Characteristics and Applications of Carbon Fibers

CHARACTERISTICS	APPLICATIONS
Physical strength, light weight , specific toughness	They are used in Aerospace industry, road and marine transportation sector and for making sporting goods.
High dimensional stability, low abrasion, low coefficient of thermal expansion	They are used for the manufacturing of Missiles, aircraft brakes, large telescopes, aerospace antenna and support structure, optical benches.
Good vibration damping ability, strength, and toughness	Useful in Audio equipment, loudspeakers for Hi-fi equipment along with pickup and robot arms
Electrical conductivity	Automobile hoods, novel tooling, casings and bases for electronic equipments, EMI and RF shielding, brushes
Biological inertness and X-ray permeability	Medical applications in prostheses, surgery and x-ray equipment, implants, tendon/ligament repair
Fatigue resistance, self-lubrication, high damping	It is extremely useful in the Textile machinery.
High corrosion resistance , Chemical inertness	It finds various uses in Chemical industry; nuclear field along with valves, seals, and pump components for processing plants
Electromagnetic properties	Carbon Fiber is used in the making of Large generator retaining rings and various radiological equipment

6. Conclusion

For the most part of the century, scientists worldwide have been looking for the next breakthrough element, and Carbon Fiber has proved to be the front runner in that race. With the plethora of applications like in batteries, the aviation industry, and the automobile industry and advantages over the current element being used in these industries, Carbon Fiber is expected to overshadow all other compounds in the near future. Although in the current times, its scope of mass synthesis is limited due to which it is costly, but slowly and steadily, efforts are being made to work on the problems that are stopping Carbon Fiber from booming. In a nutshell, the authors can't emphasize strongly enough that the future is of Carbon Fiber.

Acknowledgements

The Authors would like to convey their sincere gratitude towards Delhi Technological University for its continuous support throughout work

REFERENCES

[1] Frank E., Hermanutz F and Buchmeiser M. R. Carbon Fibers: Precursors, Manufacturing and Properties. Macromol. *Mater. Eng.* 2012;297: 493–501.

[2] Schawaller D., Clauß B and Buchmeiser M. R. Ceramic Filament Fibers – *A Review. Macromol. Mater. Eng*. 2012:297:502–522.

[3] Ho C., Qian K. K and Bismarck H. A. Carbon Fiber: Surface Properties. *Wiley Encyclopedia of Composites*. 2011;1–11.

[4] Huang X. Fabrication and Properties of Carbon Fibers. *Materials.* 2009;*2*:2369-2403.

[5] Bajaj P., Paliwal D. K.,Gupta A. K. Influence of metal ions on structure and properties of acrylic fibers. *J. Appl. Polym. Sci.* 1998;67:1647–1659.

[6] Goodhew P. J., Clarke A. J., Bailey J. E. Review of fabrication and properties of carbon-fibers. *Mater. Sci. Eng.* 1975;17:3–30.

[7] Bhatt P, Goe A. Carbon Fibres: Production, Properties and Potential Use. Mat.Sci.Res.India;14

[8] Shubham S Narwade, Shantanu M Kanadi, Carbon Fibre:The Future Of Building Material. IJSR

[9] Saket Vishal, SWapnil U.Deokar, Properties Of Carbon Fibre Reinforced Polymers. IJIRST

Brayton-Moser Passivity Based Controller for Electric Vehicle Battery Charger

Kumari SHIPRA, Rakesh MAURYA, and Shambhu N. SHARMA

***Abstract*—In this paper Brayton-Moser passivity-based control (BM-PBC) methodology is developed for an on-board battery charger for plug-in electric vehicles(PHEVs). The main features of this electric vehicle (EV) charger include improved power quality, reduced filter size and voltage stress across the switches and fast dynamic response. In this paper, a dynamic model of the three-level (TL) boost power factor correction (PFC) converter is developed using the Brayton-Moser formulation. Then, the Brayton-Moser based control technique is designed by injecting a virtual resistor in series with the input inductor. Further, the stability analysis of the proposed controller is also carried out using energy balance approach. To improve the dynamic performance and reduce the steady state error, a PI controller is integrated with the aforesaid controller. Therefore, the controller comprises of BM-PBC and the PI controller is implemented for the TL boost PFC converter as a battery charger and its performances are investigated under various operating modes with the help of MATLAB/Simulink. Furthermore, power quality of charger is assessed by monitoring source current total harmonic distortion (THD) under different operating conditions. It is also observed that the proposed system provides THD less than 5% in source current which satisfies IEC 61000-3-2 Class C standard. The performance of the aforesaid controller is also compared with the conventional PI controller. In order to validate the proposed controller, a prototype model of same specifications is tested in hardware in loop and obtained test results are also presented.**

***Index Terms*—Brayton-Moser, mathematical modelling, passivity-based control, three-level boost converter.**

Nomenclature

C_1, C_2, C	Capacitor, μF.
$E(t)$	Rectified input voltage, V.
E_f	Energy function.
$\dot{E}_f$	Derivative of E_f.
G_i	Parallel damping injection.
I_d	Desired current, A.
i_L	Inductor current, A.
i_{LD}	Desired average inductor current, A.
I_0	Output current, A.
i_s	Input source current, A.
L	Inductor, mH.
R	Load resistor, Ω.
R_i	Series damping injection.
V_d	Desired output voltage, V.
v_{c1}, v_{c2}, v_c	Capacitor voltage, V.
v_{cd}	Desired average capacitor voltage, V.
V_m	Peak value of input voltage, V.
V_0	Output voltage, V.
v_s	Input voltage, V.
$\tilde{i}_L, \tilde{v}_c$	Error state variables.
ω	Angular frequency.
$\sigma_1, \sigma_2, \sigma$	Duty ratio.
P	Mixed potential function.
G	Content energy.
J	Co-content energy.
P_{Ri}	Series dissipation factor.
P_{Gi}	Parallel dissipation factor.
k_p, k_i	Gain of PI controller.
EL	Euler-Lagrange.
PBC	Passivity-based controller.
BM	Brayton-Moser.
TL	Three-level.
EV	Electric vehicle.

I. Introduction

PLUG-in electric vehicles (PHEVs) are one of the sustainable mean of transportation as compared to internal combustion engine based automobile due to various advantages such as low transportation cost, low maintenance, high efficiency, and environmentally friendly [1], [2]. For successful implementation of PHEVs, the EVs charger is playing key role for charging EVs battery and therefore efficient battery charger along with improved power quality features, are the main requirements in view of large penetration of PHEVs in electrical power grid. According to guidelines prescribed by IEEE STD 519-2014 and IEC 61000-3-2, the power factor should be more than 0.9 and total harmonic distortion (THD)in source current should be less than 5% for any single phase system [3]. In view of aforesaid regulation, several EVs charger as power factor correction (PFC) converters are reported in literature whose operations mainly governed by good design of power circuit

Manuscript received February 29, 2020; revised September 18, 2020; accepted December 23, 2020. Date of publication March 30, 2021; date of current version March 12, 2021.

All authors are with the Department of Electrical Engineering, Sardar Vallabhbhai National Institute of Technology, Surat 395007, India (e-mail: kumarishipra2005@gmail.com; rmaurya@eed.svnit.ac.in; sns@eed.svnit.ac.in).

Digital Object Identifier 10.24295/CPSSTPEA.2021.00004

and its controller. The boost PFC converter is one of converter which has been extensively used as PFC in design of general purpose power supply because of simple circuit, allows low-distorted input currents along with unity power factor [4], [5]. However, it brings large switching stresses and switching losses which yields low efficiency. Hence, a three-level (TL) boost PFC converter topology has been proposed to reduce stress across the switch as well as filter size by adding additional capacitor, diode and switch in existing circuit [6], [7].

Design efficient controller requires accurate mathematical modelling of the system. In the mid-seventies, the Middlebrook and Ćuk was first offered a modelling of the switched regulated DC-DC power converter [8]. The Euler-Lagrange (EL) formulation of a boost, a buck-boost and a fly-back converter are discussed in [9],[10]. In [11], the Lagrangian dynamics approach is used for modelling of a Ćuk, a three-phase boost and a coupled-inductor Ćuk converters. The EL based modelling approach of DC-DC power converter with non-ideal switches including parasitic is presented in [12]. The EL modelling of the multivariable of the boost converter constituted by the cascaded connection is also discussed in [13]. There is drawback of the EL as it is not suitable for non-ideal switches, diodes, transformers, and isolation transformers etc. In [13], the Hamiltonian dynamics approach are discussed for modelling of boost, Ćuk, fly-back for ideal switches and also for non-ideal switches and diodes. A mathematical formulation for switched electrical networks has been discussed by Brayton and Moser (1964) in [9]. In this formulation, the current and voltage are obtained variables, so controller can directly use obtained variables as a measured variable. This is also suitable for non-ideal switches, diodes and transformers etc.

There are various linear and non-linear control methodologies reported in literature which is designed to control the switched electrical system against dynamic variation [14]–[18]. Generally, linear controllers like (P, PI, PID) have been widely used to control the converters as discussed in[14], but these controllers have highly unpredictable tuning problem of controller parameters. So, these controllers are not suitable and user friendly. Recently, many non-linear controllers such as sliding mode control, fuzzy logic control (FLC) etc. are reported in [18], [19]. Several control techniques have been proposed for controlling the switched electrical system to charge EVs battery [20]–[22]. A conventional PI control technique is highly used as it is simple to implement [20]. In [21], the approximate rule based FLC technique has been implemented to achieve battery charging and discharging applications. A non-linear model predictive controller for battery charging application is discussed in [22].

Recently, power shaping based passivity-based controllers (PBCs) have been widely used in several research areas such as robot arms, induction motors, switched power converters etc.[23]. In order to achieve equilibrium stability, a virtual resistor is added as a damping factor in power shaping passivity based controller. So that the system becomes passive (an increase of the storage energy should be less than the rate of flow of energy into the system) against disturbances. The EL and the port controlled Hamiltonian based passivity control technique are very powerful and robust for switched electrical system[13], [24]–[27]. The passivity-based control theory is implemented to bidirectional buck-boost converter for battery charging application in [28]. An adaptive passivity based control design technique for battery hybrid power source has been proposed in [29]. A new concept of passivity-based controller known as Brayton-Moser has overcome the problem of selection of damping factor unlike EL based PBC controller. Additionally, the Brayton-Moser based controller requires controlled variables which are directly expressed in physical measurable variables like current and voltage in switched electrical network while the EL and the Hamiltonian have physical variables charge or flux to control current and voltage of system. The concept of the Brayton-Moser based controller for switched electrical systems is reported in [30]–[32].

In this paper, a Brayton-Moser passivity-based controller has been developed for an on-board battery charger using the TL boost converter to achieve well-regulated voltage and current along with nearly unity power factor operation while charging EV battery. The TL boost converter has an additional switch, diode, capacitor and operation of its switches are shifted by 180° out of phase. Due to these additional elements and its interleaved operation, the switch stress as well as filter size are reduced considerably.

This paper is structured in the following manner. Section I describes the brief introduction of the TL boost converter, PFC and PBCs. The mathematical modelling of the TL boost converter using the Brayton-Moser formulation is carried out in Section II. It also covers brief description of circuit and its operational modes under different switching states. The Brayton-Moser equation for electrical circuit is also review in Section II. Section III narrates the control topology of the Brayton-Moser based PBC for the TL boost converter along with analysis of stability. Section IV presents simulation study and its test results of the proposed converter for battery load constant current (CC) and constant voltage (CV) mode. Concluding remarks are given in Section V.

II. Mathematical Modeling of the TL Boost PFC Converter

A schematic diagram of TL boost PFC converter with the Brayton-Moser passivity-based controller is depicted in Fig. 1. The proposed system is employed for charging the electric vehicle battery under CC and CV modes. The detailed mathematical analysis and operation are carried out in the following subsections.

A. Topology

The power circuit of TL boost PFC converter along with the BM-PBC is presented in Fig. 1. It consists of a single-phase diode bridge rectifier (DBR), TL boost converter, battery load

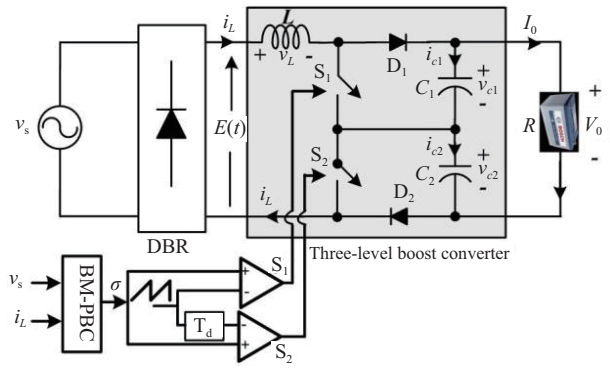


Fig. 1. Schematic diagram of a TL boost PFC converter.

and a PBC as an on-board battery charger for CC/CV charging of EV battery. The proposed TL boost converter is derived from conventional boost DC-DC converter by splitting output capacitor (C) into output capacitors (C_1, C_2), diode (D) into diodes (D_1, D_2) and switch (S) into switches (S_1, S_2) as shown in Fig. 1.

A single-phase AC supply is connected to a DBR to get rectified voltage which act as input voltage source for TL boost converter. The switches are operated in interleaved manner to achieve reduced ripple contents in passive elements. A BM-PBC methodology is applied to control the closed loop operation of converter as shown in Fig. 1.

B. Operational Modes

Prior to steady-state analysis of the proposed converter, few assumptions are made such as ideal passive elements and switching devices. It is also assumed that for a given sampling time (T_s), supply voltage (v_s) remains constant. The duty cycle $\sigma(t)$ varies in each sample either $\sigma(t) < 0.5$ or $\sigma(t) > 0.5$ during each sampling interval as instantaneous value of rectified voltage $E(t)$, keeps on changing. For one sampling time T_s, the switch is on for σT_s and off for $(1-\sigma)T_s$ period as depicted in Fig. 2. Depending upon the switching state of switches (S_1, S_2), four operating modes are identified which repeat several times in one cycle of the line frequency (i.e., 50 Hz). The waveforms of voltage and current of circuit elements are depicted in Fig. 2.

C. The Brayton-Moser Equations for Electrical Circuit

According to the Brayton-Moser [33], the differential equations representing dynamical behaviour of non-linear electrical circuits is given below:

$$\begin{cases} -L\dfrac{\mathrm{d}i_L}{\mathrm{d}t} = \dfrac{\partial P}{\partial i_L}(i_L, v_c) \\ C\dfrac{\mathrm{d}v_c}{\mathrm{d}t} = \dfrac{\partial P}{\partial v_c}(i_L, v_c) \end{cases}, \tag{1}$$

where $i_L \in R''$ is the input inductor currents, $v_c \in R^m$ is the capacitor voltages and P: $R^{n\times m}$ is mixed-potential function.

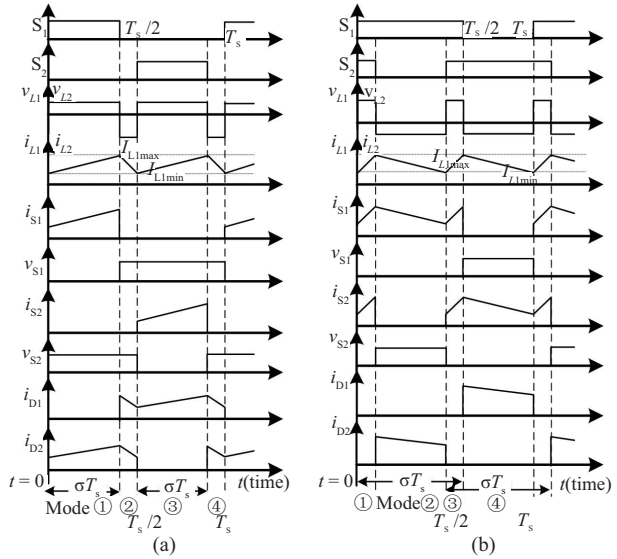


Fig. 2. Switching waveform of the TL boost converter over one switching cycle for (a) and (b) .

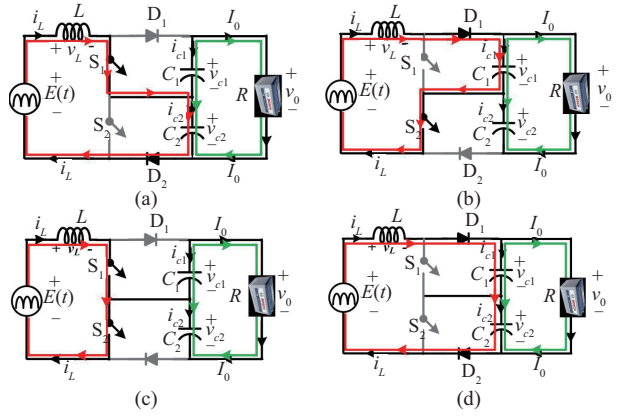


Fig. 3. Equivalent circuit of the TL boost converter under various operating modes: (a) mode 1, (b) mode 2, (c) mode 3 and (d) mode 4.

The matrices $L(i_L) \in R^{n\times n}$ and $C(v_c) \in R^{m\times m}$ are the inductance and capacitance matrices respectively. The mixed-potential function P is: $\mathrm{d}P = \mathrm{d}G - \mathrm{d}J$, where G and J are content and co-content energy function respectively, $\mathrm{d}G = \langle i_L, dv_c \rangle$ and $\mathrm{d}J = \langle v_c, \mathrm{d}i_L \rangle$.

D. Mathematical Modelling of the TL Boost Converter

In this section, the mathematical modelling of the TL boost converter has been carried out using the Brayton-Moser formulation as given in (1). So, first we should obtain the mixed potential function (MPF) for all four possible positions of switches (σ_1, σ_2) such as (1, 0), (0, 1), (1, 1) and (0, 0) as given below:

Mode 1: At switch position (1, 0), means switch (S_1) is conducting and switch (S_2) is off as shown in Fig. 3(a) and correspondingly duty functions are $\sigma_1 = 1$ and $\sigma_2 = 0$. Hence, the MPF for switch position (1, 0) is given below:

$$G(i_L,v_{c1},v_{c2}) = -Ei_L + \frac{1}{2}i_L v_{c2},$$
$$J(i_L,v_{c1},v_{c2}) = \frac{1}{2R}(v_{c1}+v_{c2})^2 - \frac{1}{2}i_L v_{c2},$$
$$P^{(1,0)}(i_L,v_{c1},v_{c2}) = -Ei_L + i_L v_{c2} - \frac{1}{2R}(v_{c1}+v_{c2})^2$$

Mode 2: In this mode, switches are at position (0, 1), when switch (S_1) is off and switch (S_2) is conducting and the diodes (D_1, D_2) are in forward and reverse biased respectively as shown in Fig. 3(b). In this case the duty functions are $\sigma_1 = 0$ and $\sigma_2 = 1$. Hence, MPF for switch position (0, 1) is given below:

$$G(i_L,v_{c1},v_{c2}) = -Ei_L + \frac{1}{2}i_L v_{c1},$$
$$J(i_L,v_{c1},v_{c2}) = \frac{1}{2R}(v_{c1}+v_{c2})^2 - \frac{1}{2}i_L v_{c1},$$
$$P^{(0,1)}(i_L,v_{c1},v_{c2}) = -Ei_L + i_L v_{c1} - \frac{1}{2R}(v_{c1}+v_{c2})^2.$$

Mode 3: In this case, both the switches (S_1, S_2) are conducting and diodes (D_1, D_2) are reverse biased as shown in Fig. 3(c) and correspondingly duty functions are $\sigma_1 = \sigma_2 = 1$. Hence, MPF for switch position (1, 1) is given below:

$$G(i_L,v_{c1},v_{c2}) = -Ei_L,$$
$$J(i_L,v_{c1},v_{c2}) = \frac{1}{2R}(v_{c1}+v_{c2})^2,$$
$$P^{(1,1)}(i_L,v_{c1},v_{c2}) = -Ei_L - \frac{1}{2R}(v_{c1}+v_{c2})^2.$$

Mode 4: In this case, both switches are non-conducting and diodes (D_1, D_2) are forward biased as shown in Fig. 3(d) and correspondingly duty functions are $\sigma_1 = \sigma_2 = 0$. The MPF for switch position (0, 0) is given below:

$$G(i_L,v_{c1},v_{c2}) = -Ei_L + \frac{1}{2}i_L(v_{c1}+v_{c2}),$$
$$J(i_L,v_{c1},v_{c2}) = \frac{1}{2R}(v_{c1}+v_{c2})^2 - \frac{1}{2}i_L(v_{c1}+v_{c2}),$$
$$P^{(0,0)}(i_L,v_{c1},v_{c2}) = -Ei_L + i_L(v_{c1}+v_{c2}) - \frac{1}{2R}(v_{c1}+v_{c2})^2.$$

Now, above written MPFs for all four possible switching positions can be re-written in terms of switching functions (σ_1, σ_2) as given below:

$$P^{(\sigma_1,\sigma_2)}(i_L,v_{c1},v_{c2}) = -Ei_L + (1-\sigma_1)i_L v_{c1} + (1-\sigma_2)i_L v_{c2} - \frac{1}{2R}(v_{c1}+v_{c2})^2, \quad (2)$$

where $P^{(\sigma1,\sigma2)}(i_L, v_{c1}, v_{c2})$ is the MPF for the switch position (σ_1, σ_2). From (1) and (2), partial differentiation of the above MPF is done with respect to each variable and obtain following state equations:

$$-L\frac{di_L}{dt} = \frac{\partial P^{(\sigma_1,\sigma_2)}}{\partial i_L}(i_L,v_{c1},v_{c2}) = -E + (1-\sigma_1)v_{c1} + (1-\sigma_2)v_{c2}, \quad (3)$$

$$C_1\frac{dv_{c1}}{dt} = \frac{\partial P^{(\sigma_1,\sigma_2)}}{\partial v_{c1}}(i_L,v_{c1},v_{c2}) = (1-\sigma_1)i_L - \frac{1}{R}(v_{c1}+v_{c2}), \quad (4)$$

$$C_2\frac{dv_{c2}}{dt} = \frac{\partial P^{(\sigma_1,\sigma_2)}}{\partial v_{c2}}(i_L,v_{c1},v_{c2}) = (1-\sigma_2)i_L - \frac{1}{R}(v_{c1}+v_{c2}). \quad (5)$$

Assuming i_L, v_{c1} and v_{c2} as state variables, (3)–(5) are rearranged in the matrix–form as given below:

$$\begin{bmatrix}\dot{i}_L\\ \dot{v}_{c1}\\ \dot{v}_{c2}\end{bmatrix} = \begin{bmatrix}0 & -\frac{(1-\sigma_1)}{L} & -\frac{(1-\sigma_2)}{L}\\ \frac{(1-\sigma_1)}{C_1} & -\frac{1}{RC_1} & -\frac{1}{RC_2}\\ \frac{(1-\sigma_2)}{C_2} & -\frac{1}{RC_1} & -\frac{1}{RC_2}\end{bmatrix}\begin{bmatrix}i_L\\ v_{c1}\\ v_{c2}\end{bmatrix} + \begin{bmatrix}\frac{1}{L}\\ 0\\ 0\end{bmatrix}E.$$

III. BM Passivity Based Controller

In this section, the BM-PBC based controller is developed for the TL boost PFC converter to charge the EVs battery under CC and CV modes along with improved power factor features at AC supply side.

A. BM Model of the TL Boost PFC Converter

In order to control the TL boost converter, the duty ratio of both switches (S_1, S_2) are same i.e., $\sigma_1 = \sigma_2 = \sigma$, but phase-shifted by 180°. The measured and controlled variables of the proposed system is output voltage V_0 which is average of v_c. So, the new average state variables can be defined as $v_c = v_{c1} + v_{c2}$. Let $C_1 = C_2 = 2C$. The new equivalent state-space equation of the proposed system with new state variables (i_L, v_c) as derived by the Brayton-Moser formulations is given as:

$$\begin{bmatrix}L\frac{di_L}{dt}\\ C\frac{dv_c}{dt}\end{bmatrix} = \begin{bmatrix}0 & -(1-\sigma)\\ (1-\sigma) & -\frac{1}{R}\end{bmatrix}\begin{bmatrix}i_L\\ v_c\end{bmatrix} + \begin{bmatrix}1\\ 0\end{bmatrix}E. \quad (6)$$

The above average state-space equation can be written in the BM form as given below:

$$\begin{cases}-L\frac{di_L}{dt} = \frac{\partial P}{\partial i_L}(i_L,v_c)\\ C\frac{dv_c}{dt} = \frac{\partial P}{\partial v_c}(i_L,v_c)\end{cases}. \quad (7)$$

The desired trajectories of the average state variables are as follow:

$$\begin{cases}-L\frac{di_{Ld}}{dt} = \frac{\partial P}{\partial i_{Ld}}(i_{Ld},v_{cd})\\ C\frac{dv_{cd}}{dt} = \frac{\partial P}{\partial v_{cd}}(i_{Ld},v_{cd})\end{cases}. \quad (8)$$

where i_{Ld}, v_{cd}, are the desired trajectory for the average inductor current and average output voltage for one switching cycle respectively. The error trajectory of the state variables as $\tilde{i}_L = i_L - i_{Ld}$ and $\tilde{v}_c = v_c - v_{cd}$, are obtained from (7) and (8). The error dynamics are described as given below:

$$\begin{cases} -L\dfrac{\mathrm{d}\tilde{i}_L}{\mathrm{d}t} = \dfrac{\partial P}{\partial i_L}(i_L, v_c) - \dfrac{\partial P}{\partial i_{Ld}}(i_{Ld}, v_{cd}) \\ C\dfrac{\mathrm{d}\tilde{v}_c}{\mathrm{d}t} = \dfrac{\partial P}{\partial v_c}(i_L, v_c) - \dfrac{\partial P}{\partial v_{cd}}(i_{Ld}, v_{cd}) \end{cases}. \tag{9}$$

A series and parallel damping dissipation factors are added to the above error dynamics equations which yields below:

$$\begin{cases} -L\dfrac{\mathrm{d}\tilde{i}_L}{\mathrm{d}t} = \dfrac{\partial P}{\partial \tilde{i}_L}(\tilde{i}_L, \tilde{v}_c) + \dfrac{\partial P_{R_i}}{\partial \tilde{i}_L}(\tilde{i}_L, \tilde{v}_c) \\ C\dfrac{\mathrm{d}\tilde{v}_c}{\mathrm{d}t} = \dfrac{\partial P}{\partial \tilde{v}_c}(\tilde{i}_L, \tilde{v}_c) + \dfrac{\partial P_{G_i}}{\partial \tilde{v}_c}(\tilde{i}_L, \tilde{v}_c) \end{cases}. \tag{10}$$

where $P_{R_i} = R_i\tilde{i}_L^2/2$ and $P_{G_i} = G_i\tilde{v}_c^2/2$ are the injected dissipation terms. R_i and G_i are the injected series damping with the inductor and injected parallel damping with the capacitor respectively.

B. Stability Analysis

Let a positive definite energy function E_f of the proposed system is created as:

$$E_f = \frac{1}{2}L\tilde{i}_L^2 + \frac{1}{2}C\tilde{v}_c^2. \tag{11}$$

From (11), $E_f = 0$ only when $\tilde{i}_L = 0$ and $\tilde{v}_c = 0$ or else $E_f > 0$ The derivative of E_f, is given below:

$$\dot{E}_f = L\dot{\tilde{i}}_L \cdot \tilde{i}_L + C\dot{\tilde{v}}_c \cdot \tilde{v}_c. \tag{12}$$

The error dynamic (10), $L\tilde{i}_L$ and $C\tilde{v}_c$ can be represented as:

$$\begin{cases} L\dot{\tilde{i}}_L = -(1-\sigma)\tilde{v}_c - R_i\tilde{i}_L \\ C\dot{\tilde{v}}_c = (1-\sigma)\tilde{i}_L - \dfrac{1}{R}\tilde{v}_c \end{cases}. \tag{13}$$

Then, inserting (13) into (12), we get (14):

$$\dot{E}_f = -R_i\tilde{i}_L^2 - \frac{1}{R}\tilde{v}_c^2 < 0 \tag{14}$$

In practical circuit, R and $R_i > 0$. Then, from (14), derivative of E_f, is always negative unless $\tilde{i}_L = 0$ and $\tilde{v}_c = 0$. Hence from (11) and (14), the system is stable.

From [12], the direct control of output voltage does not achieve stability point. Hence an indirect control can be applied to achieve stability point and correspondingly achieve desired objective. Here, the objective of the control design is to charge the EV battery in CC and CV mode along with improved quality of source current with low THD. To achieve nearly unity power factor, the input inductor current should follow the rectified sinusoidal source voltage of same frequency and same phase i.e., $i_{Ld} = I_d|\sin\omega t|$. I_d is the desired input inductor current and $I_d = V_m/(1-\sigma)^2R$. V_m is the maximum voltage of input source voltage.

C. Controller Design

In PBC strategy, the energy of the system is re-shaped so that a control function can be obtained. The energy re-shaping can be done by injecting virtual damping into the system. A virtual damping factor may be inserted in series with the input inductor or in parallel with the output capacitor. Both methodologies (series and parallel) are discussed here.

1)Series Damping Injection

In this methodology, a virtual damping factor R_i is inserted in series with the input inductor to modify the energy of the proposed system. Putting $P_{G_i} = 0$, in (10), then the obtained closed loop dynamics is:

$$\begin{cases} -L\dfrac{\mathrm{d}\tilde{i}_L}{\mathrm{d}t} = \dfrac{\partial P}{\partial \tilde{i}_L}(\tilde{i}_L, \tilde{v}_c) + \dfrac{\partial P_{R_i}}{\partial \tilde{i}_L}(\tilde{i}_L, \tilde{v}_c) \\ C\dfrac{\mathrm{d}\tilde{v}_c}{\mathrm{d}t} = \dfrac{\partial P}{\partial \tilde{v}_c}(\tilde{i}_L, \tilde{v}_c) \end{cases}. \tag{15}$$

Subtracting (15) from (7), the closed loop dynamics could be written as:

$$\begin{cases} -L\dfrac{\mathrm{d}i_{Ld}}{\mathrm{d}t} = \dfrac{\partial P}{\partial i_{Ld}}(i_{Ld}, v_{cd}) - \dfrac{\partial P_{R_i}}{\partial \tilde{i}_L}(\tilde{i}_L, \tilde{v}_c) \\ C\dfrac{\mathrm{d}v_{cd}}{\mathrm{d}t} = \dfrac{\partial P}{\partial v_{cd}}(i_{Ld}, v_{cd}) \end{cases}. \tag{16}$$

From above (16), the closed loop dynamics can be described as:

$$\begin{cases} -L\dfrac{\mathrm{d}i_{Ld}}{\mathrm{d}t} = -E + (1-\sigma)v_{cd} - R_i\tilde{i}_L \\ C\dfrac{\mathrm{d}v_{cd}}{\mathrm{d}t} = -(1-\sigma)i_{Ld} - \dfrac{1}{R}v_{cd} \end{cases}. \tag{17}$$

Hence, to control the two objectives of the system, let $i_{Ld} = I_d|\sin\omega t|$ and $v_{cd} = V_d$.

After solving (17), we get control law is as given below:

$$\begin{cases} \sigma = 1 - \dfrac{(E - L\dot{\tilde{i}}_{Ld} + R_i \tilde{i}_L)}{v_{cd}} \\ \dot{v}_{cd} = \dfrac{1}{C}\left[-(1-\sigma)i_{Ld} - \dfrac{1}{R}v_{cd}\right] \end{cases}. \tag{18}$$

In this controller, the regulation scheme based on series damping injection. So, the duty ratio is influenced due to series damping factor as the duty ratio depends on the several state variables. From [30], for selection of series damping factor R_i there should be satisfied the following condition:

$$\left\| \frac{1-\sigma}{R_i}\sqrt{\frac{L}{C}} \right\| \leqslant 1-\delta \Leftrightarrow R_i \geqslant \frac{1-\sigma}{1-\delta}\sqrt{\frac{L}{C}}.$$

For all $0 < \delta < 1$.

2)Parallel Damping Injection

In this methodology, to re-shape the energy of the system, parallel damping factor G_i is inserted in parallel with the output capacitor. Putting $P_{R_i} = 0$ in (10), the closed loop dynamics can be rewritten as below:

$$\begin{cases} -L\dfrac{\mathrm{d}\tilde{i}_L}{\mathrm{d}t} = \dfrac{\partial P}{\partial \tilde{i}_L}(\tilde{i}_L, \tilde{v}_c) \\ C\dfrac{\mathrm{d}\tilde{v}_c}{\mathrm{d}t} = \dfrac{\partial P}{\partial \tilde{v}_c}(\tilde{i}_L, \tilde{v}_c) + \dfrac{\partial P_{G_i}}{\partial \tilde{v}_c}(\tilde{i}_L, \tilde{v}_c) \end{cases}. \tag{19}$$

Subtracting (19) from (7), the closed loop dynamics could be written as:

$$\begin{cases} -L\dfrac{\mathrm{d}i_{Ld}}{\mathrm{d}t} = \dfrac{\partial P}{\partial i_{Ld}}(i_{Ld}, v_{cd}) \\ C\dfrac{\mathrm{d}v_{cd}}{\mathrm{d}t} = \dfrac{\partial P}{\partial v_{cd}}(i_{Ld}, v_{cd}) - \dfrac{\partial P_{G_i}}{\partial \tilde{v}}(\tilde{i}_L, \tilde{v}_c) \end{cases}. \tag{20}$$

From above (20), the closed loop dynamics can be described as:

$$\begin{cases} -L\dfrac{\mathrm{d}i_{Ld}}{\mathrm{d}t} = -E + (1-\sigma)v_{cd} \\ C\dfrac{\mathrm{d}v_{cd}}{\mathrm{d}t} = -(1-\sigma)i_{Ld} - Gv_{cd} - G_i\tilde{v}_c \end{cases}. \tag{21}$$

Hence, to control the two objectives of the system, let $i_{Ld} = I_d|\sin\omega t|$ and $v_{cd} = V_d$ After solving (21), we get control law as given below:

$$\begin{cases} \sigma = 1 - \dfrac{(E - L\dot{i}_{Ld})}{v_{cd}} \\ \dot{v}_{cd} = \dfrac{1}{C}\left[-(1-\sigma)i_{Ld} - Gv_{cd} - G_i\tilde{v}_c\right] \end{cases}. \tag{22}$$

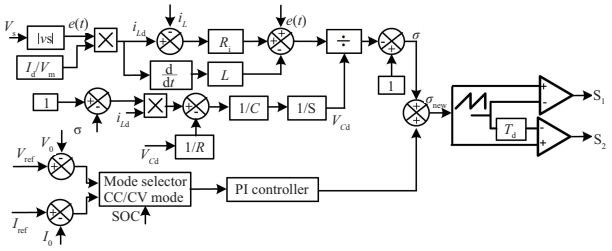


Fig. 4. Schematic block diagram of Brayton-Moser based passivity controller.

In this controller, the regulation scheme depends on parallel damping factor. So, the duty ratio is influenced due to parallel damping factor. From [30], for selection of parallel damping factor G_i there should be satisfied the following condition:

$$\left\| \frac{1-\sigma}{G_i}\sqrt{\frac{C}{L}} \right\| \leqslant 1-\delta \Leftrightarrow G_i \geqslant \frac{1-\sigma}{1-\delta}\sqrt{\frac{C}{L}} - G.$$

For all $0 < \delta < 1$.

In order to obtain desired objective, the controller is designed in which damping factor is injected in series with the input inductor and in parallel with the capacitor. But here, the series damping injected controller is implemented only because a load is already connected in parallel to the capacitor and parallel damping injected controller is not appropriate for PFC.

To improve the dynamic performances of the aforesaid controller, a PI controller is integrated with the conjunction to the aforesaid controller. Due to the addition of the PI controller, the steady-state errors easily eliminated, and the system quickly stabilizes. So, after adding a PI controller in parallel to the aforesaid controller a new control law is obtained, i.e.,

$$\sigma_{new} = \sigma + k_p e(t) + k_i \int e(t)\mathrm{d}t, \tag{23}$$

where k_p , k_i are the gain of the PI controller and $e(t) = V_{ref} - V_0$, the output voltage error in CV mode and $e(t) = I_{ref} - I_0$, the output current error in CC mode.

IV. Results and Discussions

A Simulink model of the proposed BM-PBC controller for the TL boost PFC converter, rated for 2.8 kW is developed using the Simulink and the Sim Power System toolbox of MATLAB. The specifications of system parameter include AC input supply voltage (v_s): 230 V, 50 Hz inductor (L): 6 mH, capacitors (C_1, C_2): 2500 μF with switching frequency of 5 kHz. The proposed controller is used to charge a lithium-ion battery of rated capacity 35 Ah, with nominal voltage of 345 V under CC and CV modes. In CC mode, battery starts charging with state of charge (SOC) of 30% at 7 A current load i.e., C/5 rate and in CV mode, battery is charged with 70% SOC at CV of 400 V. The BM PBC control methodology is same for both modes of battery charging, but in CV mode, the output DC voltage of converter is sensed and compare with constant reference voltage of 400 V and in CC mode, the load current is sensed and compare with constant reference current of 7 A as shown in Fig. 4. Depending upon the SOC of the

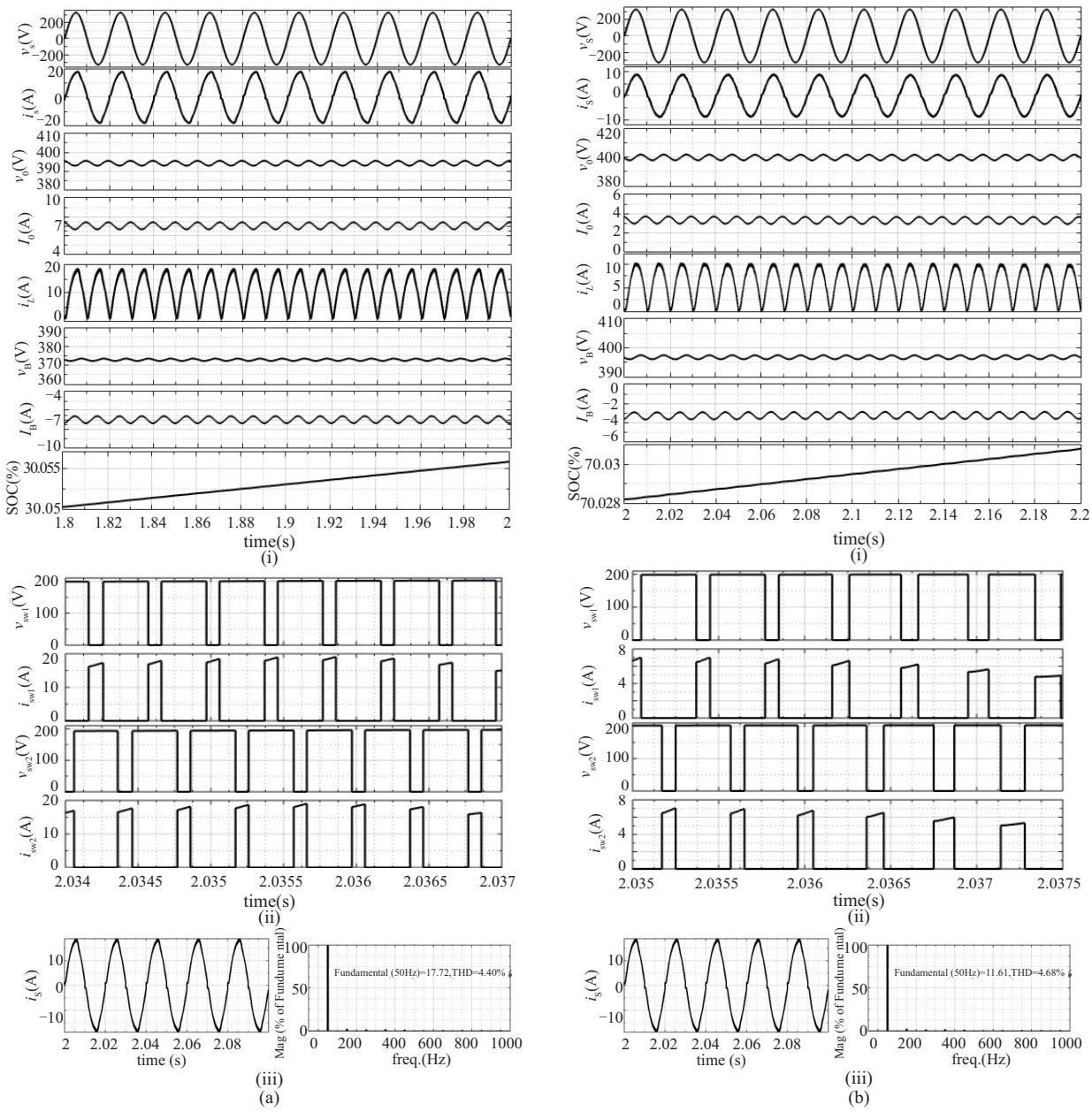


Fig. 5. The steady state waveforms feeding battery load with the BM-PBC (i) v_s, i_s, V_0, I_0, i_L, V_B, I_B, SOC, (ii) v_{swl}, i_{swl}, v_{sw2}, i_{sw2}, (iii) harmonic spectrum of supply current of converter, under (a) CC mode and (b) CV mode.

battery, aforesaid modes are selected by mode selector block. The steady state and dynamic performances of the aforesaid controller are investigated for battery load under both CC and CV mode and obtained results are discussed in Figs. 5 and 6.

A. Simulation Study of the TL Boost Converter

1)Steady State Performance

The simulation study of the BM-PBC based controller for the TL boost PFC converter employed for charging battery (35 Ah) under CC mode or in CV mode is carried out with MATLAB software. As conventional practice, the battery is charged in CC mode with 30% of SOC and after some time when SOC reached to,70%−80%, the converter operation is shifted to CV mode. Fig. 5 depicts steady state response for the BM-PBC of the TL boost PFC converter feeding battery load under CC and CV mode at source voltage (v_s): 230 V, 50 Hz Fig. 5(a)(i) and (b)(i) depicts waveforms of source voltage (v_s), source current (i_s), output voltage (V_o), load current (I_o), input inductor current (i_L), battery voltage (V_B), battery current (I_B) and SOC for CC and CV mode respectively. It has been observed that in CC mode, the battery has been charged with CC of 7 A and in CV mode, CV of 400 V with maintained input PF closed to unity in both modes. Fig. 5(a)(ii) and (b)

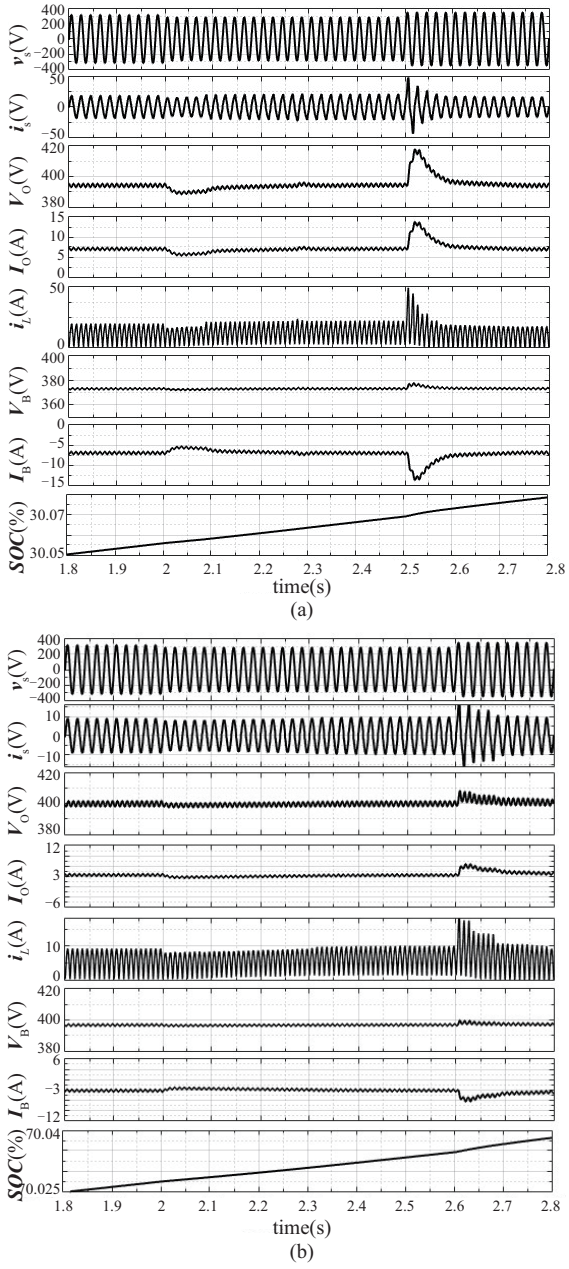


Fig. 6. Dynamic response of converter feeding battery load under step decrease in supply voltage by 10% at time t = 2 s and increase 10% from nominal value (230 V) at time instant t = 2.5 s under (a) CC mode and (b) CV mode.

also depicts waveforms of voltage across switches (v_{SW1}, v_{SW2}) and switch currents (i_{SW1}, i_{SW2}) for CC and CV mode respectively. The fast Fourier transfor (FFT) analysis is also done and depicted in Figs. 5(a)(iii) and (b) which indicates, at steady state total harmonics distortion is 3.17% and 4.55% for CC and CV mode respectively which is below 5%, according to IEEE standard.

2)Dynamic Response Under Input Variations

Fig. 6(a-b) depict dynamic performance against source variation of the TL boost converter in CC and CV charging

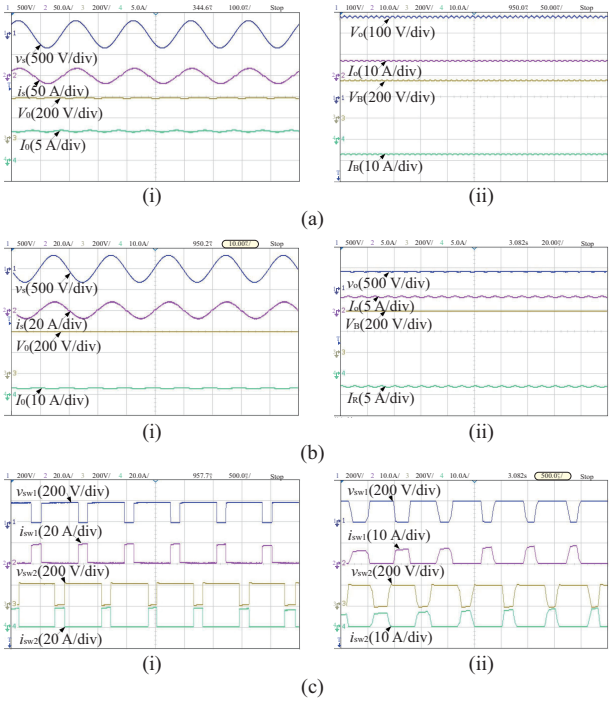


Fig. 7. Steady state performance of converter with battery load under (a) CC mode: (i) v_s , i_s ,V_0 I_0 and (ii) V_0 , I_0 , V_B and I_B. (b) CV mode: (i) v_s , i_s ,V_0, I_0 and (ii) V_0 , I_0 V_B , I_B. (c) Voltage and current of switches S_1, S_2: v_{sw1} , i_{sw1} , v_{sw2}, i_{sw2} : (i) CC mode and (ii) CV mode.

mode. First, the supply voltage decreased by 10% then increased by 10% from its nominal value (230 V) at t = 2s and t = 2.5 s respectively. As observed from the simulation study, the battery has been charged with CC (7A) in CC mode and with CV (400 V) in CV mode against source variation. The input source current is purely sinusoidal and in phase with source voltage which demonstrates the unity power factor as presented in Fig. 6. Fig. 6 also show SOC of the battery gradually increases in both CC and CV modes.

B. Test Results of the TL Boost Converter

The real-time validation of the proposed controller is implemented with TL boost converter with the similar parameters as used in the simulation study and its performances are observed under battery load. The real-time simulator OP-5142 which is FPGA based XC3S5000 series processor, is used to implement Brayton-Moser based PBC controller. A 4-channel Digital Storage Oscilloscope-DSO-X-2004A and a Fluke power quality analyzer (43B) are used for measuring and recording the test results. The several test results are obtained while charging the EV battery under CC and CV mode and few of them are presented in Figs. 7–9. To monitor the power quality features of supply current under different charging modes, are harmonic spectra of source current is recorded using Fluke power quality analyzer (43B) and its results are presented.

1)Steady State Response

The test performance under steady state of the TL boost

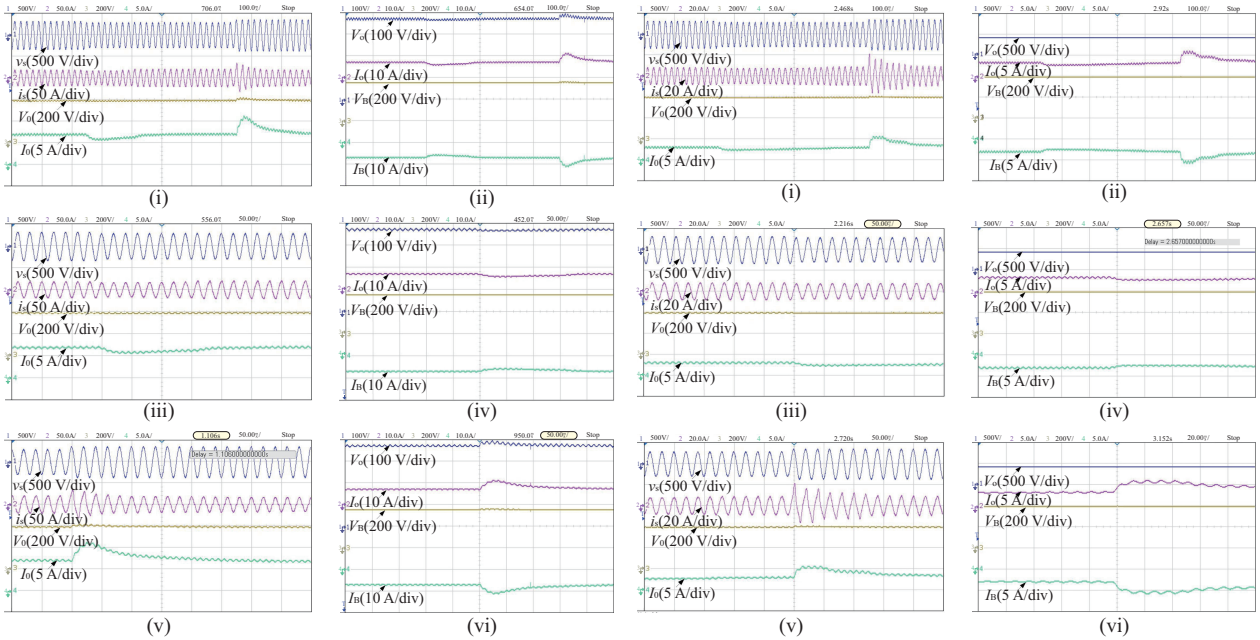


Fig. 8. Dynamic performance of the converter with battery load under (a) CC mode and (b) CV mode. (i) Waveforms of v_s , i_s , V_0 and I_0 against supply voltage variations. (ii) Waveforms of V_0 , I_0 , V_B and I_B against supply voltage variations. (iii) Waveforms of v_s , i_s , V_0 and I_0 during supply voltage decrease 10%. (iv) Waveforms of V_0 , I_0 , V_B and I_B during supply voltage decrease 10%. (v) Waveforms of v_s , i_s , V_0 and I_0 during supply voltage increase 10%. (vi) Waveforms of V_0 , I_0 , V_B and I_B during supply voltage increase 10% .

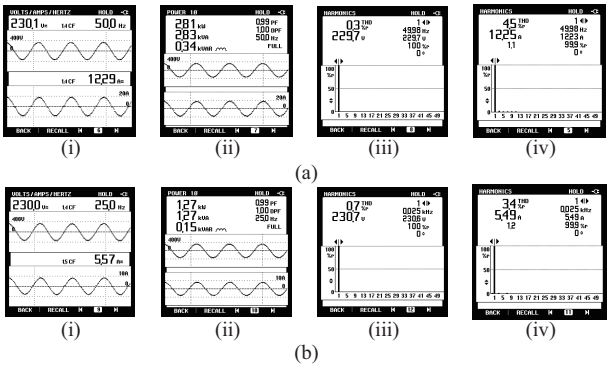


Fig. 9. Power quality assessment of input supply voltage and source current of the system with battery load under (a) CC mode and (b) CV mode. (i) v_s and i_s, (ii) P_s and Q_s, (iii) harmonic spectrum of i_s and (iv) harmonic spectrum of v_s.

PFC for the BM-PBC in CC and CV charging mode at supply voltage (v_s): 230 V, 50 Hz is presented in Fig. 7(a) and (b). Fig. 7(a) shows the battery is charging with CC (7 A) in CC mode along with purely sinusoidal input source current and in same phase with the input supply voltage. Similarly, EV battery is charging with CV of (400 V) in CV mode along with nearly unity power factor as presented in Fig. 7(b). Fig. 7(c)(i–ii) show the switch voltages (v_{SW1}, v_{SW2}) and switch currents (i_{SW1}, i_{SW2}) for both switches, for both modes (CC and CV).

2)Test Results Under Battery Load for Dynamic Response

The dynamic performance of the TL boost converter with BM-PBC in CC and CV charging mode is depicted in Fig. 8. The dynamic performance against input source variation from 230 V to 207 V are displayed in Fig. 8(a)(iii) and (b)(iii) for CC and CV mode respectively and input source variation from 207 V to 253 V is shown in Fig. 8(a)(v) and (b)(v) for CC and CV mode respectively. The waveforms of the battery parameters (V_B , I_B) are also recorded against input variation and observed satisfactory results.

Fig. 9 shows the results of Fluke power analyzer which provides waveforms and THD of source voltage (v_s) and source current (i_s) with their harmonic spectrums for CC and CV modes battery charger at source voltage (v_s): 230 V, 50 Hz Fig. 9(a)(iii) and (iv) shows the THD of 0.3% and 4.5% in source voltage and source current respectively, under CC mode battery charger. In CV mode, when the proposed converter rated at 1.27 kW is supplied at nominal voltage (230 V), the observed THD in source voltage is 0.7% and in source current is 3.4% as displayed in Fig. 9(b)(iii) and (iv). As observed from recorded data, the THD for both loads are less than 5%, which validate the range of the IEEE standard.

The comparative performance of the proposed controller against benchmark conventional PI controller under resistive load is depicted in Fig. 10. The efficiency curve under the variation of input voltage and output power is presented in Fig. 10(a) and (b) and it is observed that the efficiency of the both controllers are lies in 96%–97%. The THD curve of the input current against input voltage and output power is also observed and presented in Fig. 10(c)–(d). The THD of the input current of the proposed controller is less than 10% against dynamics conditions. The comparative study of the proposed controller

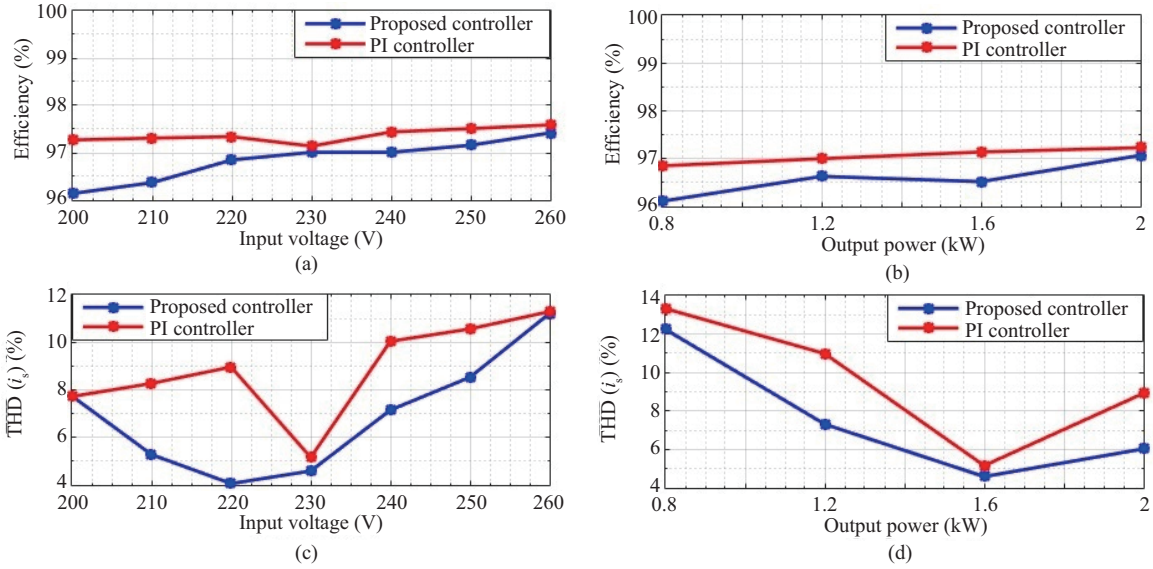


Fig. 10. Comparative performance curve: (a) efficiency vs. input voltage, (b) efficiency vs. output power, (c) source current THD vs. input voltage and (d) source current THD vs. output power.

TABLE I
COMPARATIVE STUDY AGAINST INPUT VOLTAGE VARIATION

Input voltage /V	Proposed controller				PI controller			
	Peak overshoot/%	Rise time/s	Peak time/s	Settling time/s	Peak overshoot/%	Rise time/s	Peak time/s	Settling time/s
207	31.50	0.022	0.032	0.36	31.75	0.028	0.036	0.38
230	42.50	0.019	0.031	0.39	43.75	0.024	0.038	0.41
254	49.00	0.018	0.029	0.38	49.25	0.024	0.036	0.41

TABLE II
COMPARATIVE STUDY AGAINST OUTPUT POWER VARIATION

Output power /kW	Proposed controller				PI controller			
	Peak overshoot/%	Rise time/s	Peak time/s	Settling time/s	Peak overshoot /%	Rise time/s	Peak time/s	Settling time /s
1.6	42.50	0.019	0.031	0.39	43.75	0.024	0.038	0.41
2.0	41.25	0.020	0.031	0.37	42.35	0.026	0.037	0.39
2.4	40.25	0.020	0.031	0.38	41.50	0.026	0.037	0.39

against benchmark conventional PI controller is also done in terms of control parameters like peak overshoot, rise time, peak time and settling time against input source and output power variations as presented in Tables I and II respectively. It is observed that the proposed controller BM-PBC achieve stability point faster than conventional PI controller. The performance curve of the aforesaid system with proposed controller is presented in Fig. 11 for both charging mode (CC and CV). Fig. 11(a) and (b) shows the efficiency curve against input voltage variation for CC and CV mode respectively. It is observed that efficiency is varying around 96%–97% against large variation of input voltage in CC mode and for CV mode, it is 94%–96%. The plot of THD against input variation is shown in Fig. 11(c) and (d) for both CC and CV mode. It is observed that the THD is less than 10% for both modes (CC and CV) against input variation.

V. CONCLUSION

In this paper, the mathematical model of the TL boost PFC converter is developed using the Brayton-Moser formulation and then BM-PBC based controller is established. The proposed controller is based on energy modification in which its control of energy is done through suitable damping injection by incorporating a virtual resistor in series with the inductor or in parallel to the capacitor. The stability analysis is also carried out with the help of Lyapunov stability. To improve the dynamic performance and to reduce the steady state

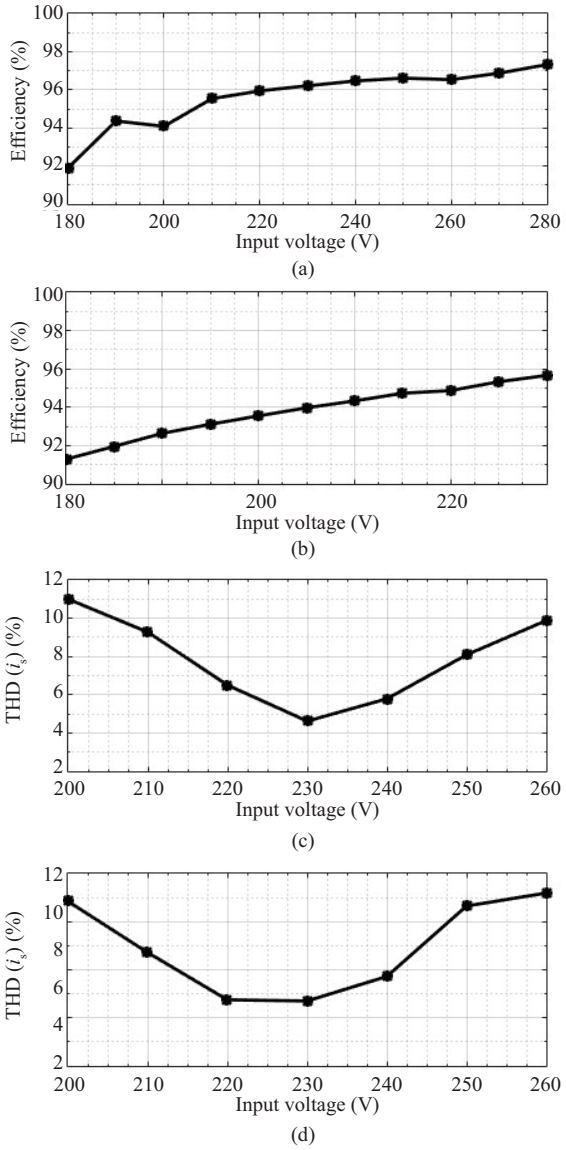


Fig. 11. Performance curve: efficiency vs. input source voltage under (a) CC mode and (b) CV mode. The source current THD vs. input source voltage under and (c) CC mode and (d) CV mode.

errors, a PI controller is integrated in conjunction with the proposed controller. The Brayton-Moser based PBC controller for TL boost converter is simulated using Matlab software and validated through hardware in loop (HIL) using FPGA based XC3S5000 series processor of real-time simulator OP-5142. The proposed system is employed as an on-board battery charger to charge the EV battery under CC and CV modes. It has satisfactory output response under steady-state and dynamic response and it is observed that the aforesaid passivity-based controller control strategies for the TL boost exhibits good power qualities under different operating conditions. The total harmonic distortion of input source current is also observed for both charging mode (CC and CV) and it is found below 5% which fulfilled the criteria of IEEE standard and IEC 61000-3-2. It is also observed that design of passivity-based control is simple and robust against dynamic variations.

References

[1] S. S. Williamson, *Energy Management Dtrategies for Electric and Plug-In Hybrid Electric Vehicles*. NewYork, NY, USA: Springer, 2013.

[2] J. Larminie and J. Lowry, *Electric Vehicle Technology Explained*. Swindon, UK.: John Wiley & Sons, 2012.

[3] *IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems*, IEEE standard, 519-1992.

[4] J. Linares-Flores, H. Sira-Ramirez, J. Reger, and S. Hernandez-Marcial, "A boost unity power factor pre-compensator, " in *2008 IEEE Power Electronics Specialists Conference*, Rhodes, Greece, 2008, pp. 3623–3627.

[5] R. W. Erickson and D. Maksimovic, *Fundamentals of Power Electronics* (2nd ed). New York, NY, USA: Springer,2007.

[6] X. Ruan, B. Li, Q. Chen, S.-C. Tan, and K. T. Chi,"Fundamental considerations of three-level DC–DC converters: Topologies, analyses, and control," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 11, pp. 3733–3743, Dec. 2008.

[7] W.-Y. Choi and S.-J. Lee,"Three-level SEPIC with improved efficiency and balanced capacitor voltages," in *Journal of Power Electronics*, vol. 16, pp. 447–454, 2016.

[8] R. D. Middlebrook and S. Cuk,"A general unified approach to modelling switching-converter power stages," in *1976 IEEE Power Electronics Specialists Conference*, Cleveland, OH, USA, 1976, pp. 18–34.

[9] D. Jeltsema, "Modeling and control of nonlinear networks: a power-based perspective," Netherlands, UK, 2005.

[10] H. A. Yildiz and L. Goren-Sumer, "Lagrangian modeling of DC-DC buck-boost and flyback converters," in *2009 European Conference on Circuit Theory and Design, Antalya*, Turkey, 2009, pp. 245–248.

[11] J. M. Scherpen, D. Jeltsema, and J. B. Klaassens, "Lagrangian modeling of switching electrical networks," in *Systems & Control Letters*, vol. 48, no. 5, pp. 365–374, Apr. 2003.

[12] R. Ortega, J. A. L. Perez, P. J. Nicklasson, and H. J. Sira-Ramirez, *Passivity-Based Control of Euler-Lagrange Systems: Mechanical, Electrical and Electromechanical Applications*(1st edition). UK.:Springer-Verlag London, 1998.

[13] G. Escobar, A. J. van der Schaft, and R. Ortega, "A Hamiltonian viewpoint in the modeling of switching power converters," in *Automatica*, vol. 35, no. 3, pp. 445–452, Mar. 1999.

[14] H. Sira-Ramirez, "Nonlinear P-I controller design for switchmode DC-to-DC power converters," in *IEEE Transactions on Circuits and Systems*, vol. 38, no. 4, pp. 410–417, Apr. 1991.

[15] B. J. Patella, A. Prodic, A. Zirger, and D. Maksimovic, "High-frequency digital PWM controller IC for DC-DC converters," in *IEEE Transactions on Power Electronics*, vol. 18, no. 1, pp. 438–446, Jan. 2003.

[16] P. Mattavelli, L. Rossetto, G. Spiazzi, and P. Tenti, "General-purpose fuzzy controller for DC-DC converters," in *IEEE Transactions on Power Electronics*, vol. 12, no. 1, pp. 79–86, Jan. 1997.

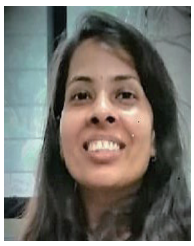
[17] W. Stefanutti, P. Mattavelli, S. Saggini, and M. Ghioni, "Autotuning of digitally controlled DC–DC converters based on relay feedback," in *IEEE Transactions on Power Electronics*, vol. 22, no. 1, pp. 199–207, Jan. 2007.

[18] S. Baev, Y. Shtessel, H. Biglari, and R. Adhami, "Sliding mode control of a unity power factor AC-to-DC boost converter," in *2007 46th IEEE Conference on Decision and Control, New Orleans*, LA, USA, 2007, pp. 2005–2010.

[19] H. S. H. Chung, E. P. W. Tam, and S. Y. R. Hui, "Development of a fuzzy logic controller for boost rectifier with active power factor correction," in *30th Annual IEEE Power Electronics Specialists Conference Record (Cat. No.99CH36321)*, Charleston, SC, USA, 1999, pp. 149–154 vol.1.

[20] M. Saleh, Y. Esa, Y. Mhandi, W. Brandauer, and A. Mohamed, "Design and implementation of CCNY DC microgrid testbed," in *2016 IEEE Industry Applications Society Annual Meeting*, Portland, OR, USA, 2016,

pp. 1–7.
[21] T. F. Wu, C. H. Chang, and Y. H. Chen, "A fuzzy-logic-controlled single-stage converter for PV-powered lighting system applications," in *IEEE Transactions on Industrial Electronics*, vol. 47, no. 2, pp. 287–296, Apr. 2000.
[22] M. Abedi, B. Song, and R. Kim, "Nonlinear-model predictive control based bidirectional converter for V2G battery charger applications," in *2011 IEEE Vehicle Power and Propulsion Conference*, Chicago, IL, USA, 2011, pp. 1–6.
[23] R. Ortega and M. W. Spong, "Adaptive motion control of rigid robots: a tutorial," in *Proceedings of the 27th IEEE Conference on Decision and Control*, Austin, TX, USA, 1988, pp. 1575–1584, vol.2.
[24] R. Ortega, A. van der Schaft, B. Maschke, and G. Escobar, "Interconnection and damping assignment passivity-based control of port-controlled Hamiltonian systems," in *Automatica*, vol. 38, no. 4, pp. 585–596, Apr. 2002.
[25] H. Sira-Ramirez and M. D. deNieto, "A Lagrangian approach to average modeling of pulsewidth-modulation controlled DC-to-DC power converters," in *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 43, no. 5, pp. 427, May 1996.
[26] H. Sira-Ramirez, R. Ortega, and G. Escobar, "Lagrangian modeling of switch regulated DC-to-DC power converters," in *Proceedings of 35th IEEE Conference on Decision and Control*, Kobe, Japan, 1996, pp. 4492–4497, vol.4.
[27] J. M. Scherpen, D. Jeltsema, and J. B. Klaassens, "Lagrangian modeling and control of switching networks with integrated coupled magnetics," in *Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No. 00CH37187)*, Sydney, NSW, Australia, 2000, pp. 4054–4059, vol.4.
[28] O. D. M. Giraldo, A. G. Ruiz, I. O. Velázquez, and G. R. E. Pérez, "Passivity-based control for battery charging/discharging applications by using a buck-boost DC-DC converter," in *IEEE Green Technologies Conference (GreenTech)*, Austin, TX, 2018, pp. 89–94.
[29] M. R. Mojallizadeh and M. A. Badamchizadeh, "Adaptive passivity-based control of a photovoltaic/battery hybrid power source via algebraic parameter identification," in *IEEE Journal of Photovoltaics*, vol. 6, no. 2, pp. 532–539, Mar. 2016.
[30] D. Jeltsema and J. M. A. Scherpen, "Tuning of passivity-preserving controllers for switched-mode power converters," in *IEEE Transactions on Automatic Control*, vol. 49, no. 8, pp. 1333–1344, Aug. 2004.
[31] M. M. J. de Vries, M. J. Kransse, M. Liserre, V. G. Monopoli, and J. M. A. Scherpen, "Passivity-based harmonic control through series/parallel damping of an H-bridge rectifier," in *2007 IEEE International Symposium on Industrial Electronics*, Vigo, Spain, 2007, pp. 3385–3390.
[32] H. Zhou, A. M. Khambadkone, and X. Kong, "Passivity-based control for an interleaved current-fed full-bridge converter with a wide operating range using the Brayton-Moser form," in *IEEE Transactions on Power Electronics*, vol. 24, no. 9, pp. 2047–2056, Sep. 2009.
[33] R. K. Brayton and J. K. Moser, "A theory of nonlinear networks. I," in *Quarterly of Applied Mathematics*, vol. 22, no. 1, pp. 1–33, Apr. 1964.



Kumari Shipra received her B.S. degree in engineering in electrical engineering from the Muzaffurpur Institute of Technology, Muzaffurpur, India, in 2003, and M.E. degree in electrical engineering with a specialization in control and instrumentation from the Delhi College of Engineering (now Delhi Technological University), Delhi, India in 2012. She is currently working toward her Ph.D. degree in electrical engineering at the Sardar Vallabhbhai National Institute of Technology, Surat, India. From 2008 to 2014, she was working in Inderprastha Engineering College, Ghaziabad, India as an Assistant Professor. Her current research interests include the passivity-based controllers, high power factor AC/DC converter and battery charger.

Rakesh Maurya received his B.Tech. degree in electrical engineering from the Kamla Nehru Institute of Technology, Sultanpur, India, in 1998, M.T. degree in power electronics and electric drive and Ph.D. degree in electrical engineering from the Indian Institute of Technology Roorkee, Roorkee, India, in 2002 and 2014 respectively. He is currently serving as an Associate Professor in the Department of Electrical Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, India. In last five years, he has published more than 50 research papers in journals of international repute particularly IEEE Transactions, IETs-UK and Elsevier, Taylor & Francis and many conference papers. His current research interests include design of switching power converters, high power factor AC/DC converters, hybrid output converters, improved power quality converters for battery charging applications, power quality problems, advanced electric drives and applications of real-time simulator for the control of power converters. He is a member of IEEE and life member of System Society of India.

Shambhu N. Sharma received his B.E. degree from the Government College of Engineering, Rewa (M.P.), India in 1994, M.Tech. degree from the Banaras Hindu University (now IIT BHU), UP, India in 2000, and Ph.D. degree from the Delhi University in 2007. Currently, he is working as a Professor in Electrical Engineering Department of the National Institute of Technology, Surat, India. He specialises in the areas of stochastic systems, control theory, stochastic differential equations with applications to electrical and electronic networks. One of his works is known as a pioneering work in stochastic systems. A stochastic system bears his name, the Sharma Parthasarathy stochastic two-body problems. He has published over 20 journal papers in reputed journal, which include International Journal of Control, Automatica, Royal Society Proceedings, Non-linear Dynamics, Journal of the Franklin Institute, AMC etc. He has also published about 12 papers in international conferences and symposia.

Archives of Microbiology
https://doi.org/10.1007/s00203-021-02297-4

ORIGINAL PAPER

Catabolic enzyme activities during biodegradation of three-ring PAHs by novel DTU-1Y and DTU-7P strains isolated from petroleum-contaminated soil

Sakshi[1] · S. K. Singh[1] · A. K. Haritash[1]

Received: 26 November 2020 / Revised: 17 March 2021 / Accepted: 19 March 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous pollutants having health hazards. PAH-utilizing bacterial strains were isolated from petroleum-contaminated soil from siding area, Bijwasan supply location of BPCL, Delhi, India. Bacterial strains with different morphology were isolated and acclimatized to a mixture of low molecular weight PAH compounds in the concentration range of 50–10,000 mg/L. Two bacterial strains surviving at 10,000 mg/L PAH concentration were identified as *Kocuria flava* and *Rhodococcus pyridinivorans,* based on 16S rRNA gene sequencing and phylogenetic analysis over MEGA X, are reported for the first time for PAH degradation. The strain *K. flava* could degrade phenanthrene, anthracene, and fluorene with efficiency of 55.13%, 59.01%, and 63.46%, whereas *R. pyridinivorans* exhibited 62.03%, 64.99%, and 66.79% degradation for respective PAHs at initial PAH concentration of 10 mg/L. Slightly lower degradation of phenanthrene could be attributed to its more stable chemical structure. The consortium of both the strains degraded 61.32%, 64.72%, and 66.64%, of 10 mg/L of phenanthrene, anthracene, and fluorene, respectively, in 15 days of incubation period indicating no synergistic or antagonistic effect towards degradation. Catechol 2,3-dioxygenase (C23O), dehydrogenase and peroxidase enzyme activities during PAH degradation coincided with degradation of PAHs, thus highlighting the role of these enzymes in catabolising three-ring PAHs. This is the first investigation confirming the participation of C23O, dehydrogenase and peroxidases enzyme profiles throughout the period of degradation. The study concludes that these strains can play significant role in microbial remediation of PAH-contaminated environment.

Keywords Bioremediation · Catechol 2,3-dioxygenase · Dehydrogenase · Enzyme activity · Peroxidase · PAHs

Introduction

The contamination due to polycyclic aromatic hydrocarbons (PAHs) has become a serious environmental threat due to their toxic, carcinogenic and mutagenic nature. PAHs are composed of benzene rings fused together, arranged in linear or branched structural arrangement. The sources of PAHs are natural (Abdel-shafy and Mansour 2016; Sakshi et al. 2019) as well as anthropogenic (Sharma and Jain 2019); and these compounds have been reported ubiquitously in air water and soil around the globe (Haritash and Kaushik, 2009b; Kaushik et al. 2012). Since these compounds have a tendency to deposit over suspended dust, these ultimately settled down over the soil thereby resulting in accumulation/contamination (Wild and Jones 1995). Bio-chemical transformation of soil PAHs by microorganisms present in soil has been reported in several studies (Sakshi and Haritash 2020). Since physico-chemical methods of PAH remediation are cost/energy prohibitive (Kuppusamy et al. 2016; Sakshi et al. 2019), microbial degradation has emerged as an eco-friendly, cost effective, and sustainable technique for removal of hydrocarbons from soil or water (Haritash and Kaushik 2009a; Megharaj et al. 2011; Bhattacharya et al. 2015; Parthipan et al. 2017; Chaudhary et al. 2019; Vaishnavi et al. 2020). The microorganisms have been reported to metabolise PAHs into water and carbon dioxide, thus, highlighting their role in decontamination of soils (Reddy and Adams 2015; Dhanya and Kalia 2020).

Communicated by Erko Stackebrandt.

✉ A. K. Haritash
akharitash@dce.ac.in

[1] Department of Environmental Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Delhi, India

Published online: 02 April 2021

The degradation capacity of microbes can be stimulated by exposing them to high PAH concentration which results in genetic adaptation for higher degradation efficiency. A number of bacterial species such as *Janibacter, Microbacterium, Gordonia, Rhodococcus, Pseudomonas, and Bacillus* (Zhang et al. 2009; Lara-Severino et al. 2016; Qi et al. 2017; Subashchandrabose et al. 2019; Rabodonirina et al. 2019) and white rot fungi such as *Phanerochaete chrysosporium*, *and Pleurotus ostreatus* (Bishnoi et al. 2007; Pozdnyakova et al. 2018) have been investigated for proficient PAH biodegradation. The degradation of PAHs by microbial species essentially depends on the catalytic action of enzymes produced by them. Mono-oxygenases, dioxygenases (catechol 1,2-dioxygenase), dehydrogenases, peroxidases and lipases are predominant enzymes responsible for the degradation of PAHs (Cerniglia 1992; Lyu et al. 2014; Pandey et al. 2012). Since catabolic enzymes are involved in PAH removal, so, enzymatic assay during degradation may provide better understanding of metabolic pathway and can help in design of an efficient remediation technique for PAH-contaminated sites. The novelty of present study lies in isolation and characterization of bacterial strains from contaminated soil to degrade low molecular weight (LMW) PAHs (phenanthrene, anthracene and fluorene), and to study the enzyme profile of catechol 2,3-dioxygenase, dehydrogenase and peroxidase in biodegradation of three-ring PAHs to propose the probable mechanism of degradation. Apart from it, the variation in degradation based on chemical structure of LMW PAHs has also been investigated in the present study.

Materials and method

Chemicals and sample collection

Naphthalene (NAP), phenanthrene (PHE), anthracene (ANT), fluorene (FLU), and pyrene (PYR) (≥ 98% purity), and HPLC-grade Acetonitrile used as mobile phase in HPLC were procured from Sigma-Aldrich, India and SRL chemicals, India**.** All the chemicals used for Luria–Bertani (LB) medium and mineral salt medium (MSM) were of analytical grade and purchased from SRL chemicals, India and CDH, India. n-Hexane, dichloromethane (DCM), silica (100–200 mesh size), sodium sulphate used for PAH extraction, and acetone used for preparation of PAH-stock solutions (NAP, PHE, ANT, FLU and PYR each having concentration 5000 mg/L) were purchased from Thermo Fisher Scientific, India and RANKEM, Avantor Performance materials, India. Triphenyl tetra-zolium chloride (TTC), triphenyl formazan (TF) dye (to establish a standard curve for absorbance (OD_{484}) vs. TF concentration) and glucose used to measure the dehydorogenase activity were procured from CDH, India, Alfa Aesar and Sigma. A TTC-glucose reagent was prepared by dissolving 1-g glucose and 2-g TTC in 100-mL distilled water and stored at 4 °C in the dark until used. Methanol used to extract the TF from bacterial cells to study dehydrogenase activity was purchased from SRL chemicals. Petroleum-contaminated soil was collected from siding area, Bijwasan supply location—Bharat Petroleum Corporation Limited (BPCL) in Delhi, India.

Media

LB medium with pH value of 7.0–7.2 was prepared using 10.0-g NaCl, 5.0-g yeast extract, and 10.0-g tryptone per litre. 2% agar was added in 1.0 L medium for preparing solid LB plates. The MSM with pH value of 7.0–7.2 was prepared using 3.0-g KNO_3, 0.5-g KH_2PO_4, 0.5-g $K_2HPO_4{\cdot}3H_2O$, 2.0-g $(NH_4)_2SO_4$, 1.0-g NaCl, amount of 0.008-g $MgSO_4.7H_2O$, 0.002-g $CuSO_4.7H_2O$, and 0.002-g $FeSO_4{\cdot}7\ H_2O$ per litre. pH value of media was adjusted with 1 N NaOH.

Isolation and acclimatization of bacteria

Bacterial strains used in this study were isolated from petroleum contaminated soil. The enrichment and isolation of PAH-degrading microbes from petroleum contaminated soil was done using LB medium by traditional methods based on serial dilution. Soil suspension (1:10 w/v) prepared by dissolving soil in normal saline (0.9% NaCl) was kept in shaking incubator at 30 °C for 48 h. Supernatant of soil suspension was collected and serially diluted to 10^{-4} order, which was used (1 mL) for inoculation on LB petri plates, incubated at 30 °C for 3 days for bacterial growth. Bacterial strains based on different colony morphology with best growth on the plates were repeatedly streaked over new LB-agar plates at regular interval of 7 days; incubated at 30 °C every time to obtain pure cultures. Screening of PAH-degrading bacterial strains was done using spray plate technique (Kiyohara et al. 1982). The obtained pure bacterial stains were exposed to a dose of 50 mg/L of mixed PAHs by inoculating them on different LB-agar plates which were uniformly sprayed with a mixture of PAHs solution (50 mg/L) containing NAP, PHE, ANT, FLU and PYR. The solvent, *i.e.* acetone, was allowed to evaporate leaving a fine coating of mixed PAHs on LB-agar surface, later inoculated with obtained pure bacterial stains and incubated at 30 °C for 7 days with close observation each day. Bacterial strains which survived on LB plates sprayed with mixed PAHs (50 mg/L) were selected and exposed to high doses of mixed PAHs by sub-culturing them with steadily decreasing concentration of yeast extract (5–0 g), and progressively increasing concentration of mixed PAHs (50–10,000 mg/L). The pure cultures were initially subjected to a dose of 50 mg/L with 0.5% yeast extract for 7 days; 200 mg/L with 0.25%

yeast extract for next 7 days; similarly, with 500 mg/L with 0.25% yeast extract, 1000 mg/L with 0.25% yeast extract, 1000 mg/L without yeast extract for 7 days each time. Colonies obtained at this stage were further exposed to higher PAHs concentrations without yeast extract *i.e.* 1500 mg/L, 2000 mg/L, 3000 mg/L, 4000 mg/L, 5000 mg/L, 7000 mg/L and 10,000 mg/L for 7 days of incubation each time. In such a way, isolated bacterial strains were exposed to high doses of PAHs under static incubation conditions at 30 °C and well adapted for PAH degradation as these bacterial strains used PAHs as sole carbon and energy source when no yeast extract was present in the medium. All isolated bacterial strains that were able to grow on medium containing mixed PAHs (10,000 mg/L), were selected for further study. The selected potent bacterial strains i.e. DTU-1Y and DTU-7P were identified by color, morphology and Gram staining. The molecular identification was done by 16S rRNA sequencing at CSIR- Institute of Microbial Technology, Chandigarh (IMTECH), India. The bacterial isolates were streaked on LB-agar slants and stored at 4 °C for preservation. Regular sub-culturing was done after every 20–25 days.

Inoculum preparation

The inoculum was prepared using pure culture of each selected bacterial isolate, which had showed viable growth at high dose (10,000 mg/L) of mixed PAHs when no yeast extract was present in growth medium. The selected bacterial strains considered as potent isolates were grown in 50 ml MSM broth to attain the exponential growth phase at 30 °C. When cell density of the suspension was found to be OD = 0.5–0.6 at 600 nm, 1% (v/v) of MSM broth containing bacterial strain was used as inoculum for further PAH-degradation analysis.

Microbial degradation assay

Assessment of PHE, ANT and FLU degradation was carried out separately in MSM broth in laboratory ensuring sterile and controlled conditions during 15 days. Each degradation study was carried out using sterilized 250 mL Erlenmeyer's flasks containing 100 ml MSM broth supplemented with respective PAH (dissolved in acetone) to attain a final concentration of 10 mg/L and inoculated aseptically with 1% (v/v) of MSM broth containing bacterial monocultures of DTU-1Y and DTU-7P (OD = 0.5–0.6 at 600 nm) individually and in consortia (containing both strains). Abiotic controls *i.e.* blank were prepared in the same way with no bacterial inoculums. All flasks containing monoculture, consortia and control were incubated at a temperature of 30 °C with rotation of 120 rpm under dark conditions for 15 days in an orbital shaking incubator as reported in earlier studies (Rabodonirina et al. 2019). All the experiments were performed in triplicates for biodegradation study. A volume of 5 ml was extracted from each experiment at regular time intervals of 0, 3, 6, 9,12 and 15 days to analyze the residual PAH concentration in medium using high-performance liquid chromatography (HPLC) (Bishnoi et al. 2007; Haritash and Kaushik, 2016). The percentage of PAHs degradation was calculated using the following relation:

$$\text{Percentage degradation} = \frac{\left(C_i - C_f\right)_s - \left(C_i - C_f\right)_b \times 100,}{\left(C_i\right)_s}$$

where C_i is the initial concentration of PAH compound in the blank and C_f is the final concentration of PAH compound in the medium, on a particular day.

Biomass growth of each strain in monoculture and in consortium was monitored during degradation study at regular intervals by measuring the increase of absorbance at OD_{600} over a spectrophotometer (Labtronics, model LT-290).

HPLC analysis: analytical method

A volume of 5 mL of sample from each flask was taken for analysis by HPLC (Shimadzu, LC-20AD) equipped with degassing unit (DGU-$20A_{3R}$), binary pump, UV/VIS detector (SPD-20A), auto sampler (SIL-$20A_{HT}$), communications bus module (CBM-20A), analytical column (Inertsil® ODS-3 V, 5 µm 4.6 × 250 mm), column oven (CTO-$10AS_{VP}$) and a computer system. Standard chromatograms for PHE, ANT and FLU using HPLC were prepared. The stock solutions of each PAH were diluted to attain the concentrations of 2, 4, 6, 8 and 10 mg/L and were used to prepare standard chromatograms. Residual PAHs were extracted with dichloromethane (DCM) in three cycles (5, 3, 2 mL) by vigorous shaking for 5 min in each cycle using separating funnel to ensure complete PAH extraction, and the organic phase extracts were pooled. The extracts (10 mL organic phase) were dried using column prepared by pre-heated (120–140 °C, 3–4 h) anhydrous silica and sodium sulphate (80:20) to remove water content/aqueous phase. The extracts were reduced in volume and concentrated by placing them under dark conditions for evaporating DCM. Finally extracted residual PAHs were dissolved in 1-ml acetonitrile and quantified by HPLC. Gradient mixture of acetonitrile and water (80:20) was used as mobile phase at a flow rate of 1 mL/min in binary mode with 10-µL sample injection volume, and UV/VIS-Detector wavelength was set at 254 nm for chromatographic analysis of PAHs.

Enzyme assay

For extraction of crude enzyme after every 3-day intervals, 10 mL of bacterial cell suspension from each flask were subjected to cell disruption using sonicator (Labman

scientific instruments LMUC–9, with frequency 40 kHz) maintaining the temperature of 10 °C for 10 min (short impulses of 30–40 s with pauses 10–20 s to maintain low temperature). Sonicated samples were centrifuged at 12,000 RCF at 4 °C for 15 min (REMI, CPR-30 PLUS), the supernatant of each sample carrying enzymes were stored at 4 °C and used for further enzyme activity assay while cell debris remained as sediments were removed.

The enzymatic activities in supernatant were monitored using a double beam UV–Vis spectrophotometer (Lab India make UV 3092 model). Catechol 1,2-Dioxygenase (C12O) and catechol 2,3-dioxygenase (C23O) activities were assayed using catechol as a substrate, by recording increase in absorbance at 260 nm and 375 nm, respectively. Assay mixture for enzyme activity contained 2.8 ml of 0.1 M potassium phosphate buffer of pH 7.5 (12.814-g K_2HPO_4 and 3.598-g KH_2PO_4 in 1L), 0.1 ml of 0.01 M catechol and 0.1 ml crude enzyme in a quartz cuvette with 1-cm light path. The increase in absorbance at 260 nm due to formation of cis-cis muconic acid ($\varepsilon_M = 1.6 \times 10^4$ $mol^{-1}lcm^{-1}$); and at 375 nm due to formation of 2-hydroxymuconic semialdehyde ($\varepsilon_M = 4.4 \times 10^4$ $mol^{-1}lcm^{-1}$) by enzymatic reaction were monitored (Nozaki 1970; Nakazawa and Nakazawa 1970; Fernandez–Lafuente et al. 2000). The peroxidase activity was monitored in 3-mL reaction mixture containing crude enzyme, 0.1 M potassium phosphate buffer of pH 7.0 (4.672 g K_2HPO_4 and 3.154 g KH_2PO_4 in 500 ml), and 5-mM guaiacol, and 0.6-mM hydrogen peroxide as the substrates. The increase in absorbance at 470 nm due to formation of tetraguaiacol (2.66×10^4 $mol^{-1}lcm^{-1}$) due to enzymatic reaction was monitored (Whitaker 1972). The least count of the increase in absorbance measurement was 0.001 and one unit of activity was defined as the amount of enzyme that oxidized 1 µmol of substrate per minute.

The dehydrogenase activity was examined using 2,3,5-triphenyltetrazoilum chloride (TTC, as the artificial electron acceptor), which was reduced to red-coloured triphenyl formazan (TPF). For dehydrogenase enzyme assay, 1 mL of bacterial cell suspension from each flask was transferred into test tubes containing tris buffer (2.5 mL) and 1 mL TTC-glucose solution, after every 3-day intervals. The tubes were lightly swirled to mix the contents and incubated at 30 °C for 1 h in a temperature controlled oven (Yamato, DKM400). After 1-h incubation the reaction product (TPF) was extracted from bacterial cells with methanol and TPF concentration was measured using TPF standard curve at OD_{484} using UV–Vis spectrophotometer (Lab India make UV 3092 model) (Ghaly and Mahmoud 2006; Burdock et al. 2011).

Result and discussion

Identification of PAH-degrading microorganisms

Long-term contaminated soil may have indigenous microbial species capable of metabolically utilizing contaminant as its carbon and energy source. The PAH-degrading bacterial strains from petroleum contaminated soil were isolated by enrichment culture technique. Different bacterial colonies were seen on LB-agar plates inoculated with petroleum contaminated soil and these colonies were first purified by repeated streaks on LB agar. By visualization, total seven different bacterial colonies having varied colony shape and color were isolated from petroleum contaminated soil. These seven bacterial strains were exposed to increasing concentration of mixed PAHs (50–10,000 mg/L containing PHE, ANT, FLU and PYR) gradually on LB plates by spray plate method to assay their growth. Among these seven isolates, two potent Gram-positive strains DTU-1Y (forming yellow, aerobic, smooth, non-motile, coccoid wet colonies on LB agar) and DTU-7P (forming antique pink, aerobic, smooth, non-motile, coccoid wet colonies on LB agar) were able to grow on LB plate without yeast extract and with high PAHs concentration. The 16S rRNA sequencing of both strains was done using Sanger sequencing method. Based on 16S rRNA sequence analysis of pure bacterial strains by nucleotide BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) search DTU-1Y was showing 99.58% similarity with *Kocuria flava* and isolate DTU-7P was showing 98.94% similarity with *Rhodococcus pyridinivorans*. The position of identified bacterial strains in phylogeny was confirmed by constructing phylogenetic tree for both strains using MEGA X (Molecular Evolutionary Genetics Analysis) (Kumar et al. 2018) by neighbor-joining algorithm (Saitou and Nei 1987) using a maximum composite likelihood method and 1000-bootstrap replications. Phylogenetic tree indicated that DTU-1Y strain and *Kocuria flava* shared one cluster, and DTU-7P and *Rhodococcus pyridinivorans* shared one cluster (Fig. 1).

The 16 s rRNA sequences of these strains obtained in this study were deposited to NCBI (National Center for Biotechnology Information), *Kocuria flava* DTU-1Y with sequence accession number: MN841976 and *Rhodococcus pyridinivorans* DTU-7P with sequence accession number: MN841977, reported elsewhere (Sakshi et al. 2020).

Degradation of PAH

Biodegradation experiments demonstrated that the two bacterial isolates (*Kocuria flava* DTU-1Y and

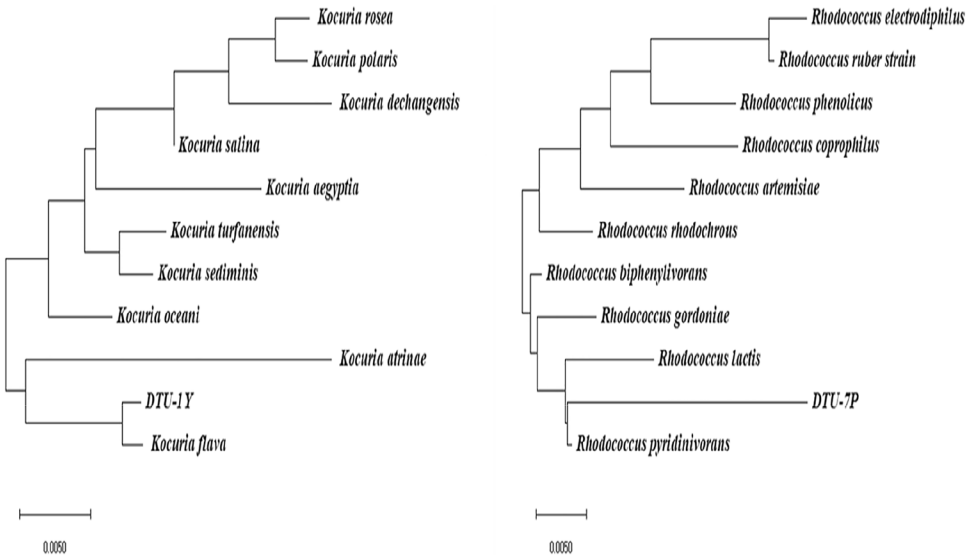


Fig. 1 Phylogenetic analysis of 16S rRNA gene sequence of isolates DTU-1Y and DTU-7P with most closely related species found by BLAST search (Sakshi et al. 2020) were conducted in MEGA X (Kumar et al. 2018) by Neighbor-Joining method (Saitou and Nei 1987). The tree was generated with 1000 replicates (bootstrap test). The scale bar corresponds to approximately 0.0050 changes per nucleotide position

Rhodococcus pyridinivorans DTU-7P) were able to efficiently degrade three ring PAHs with efficiency ranging from 55 to 66% after 15 days of incubation at 30 °C. As far available data, it is the first study where *R. pyridinivorans* DTU-7P is reported to use PHE, ANT and FLU as carbon and energy source and *K. flava* DTU-1Y to have ability to use ANT and FLU. Both strains exhibited potential in utilising all selected PAHs as carbon source as indicated by their growth on the substrate, illustrated in Fig. 2. Degradation of PHE, ANTH and FLU by monocultures and consortium was recorded from 3rd day, and maximum degradation was achieved within first 9 days of incubation. It was observed that, within 15 days, strain *K. flava*-DTU-1Y degraded 55.13%, 59.01%, 63.46%; strain *R. pyridinivorans*-DTU-7P degraded 62.03%, 64.99% 66.79%; and consortium of both strains degraded 61.32%, 64.72%, 66.64% of 10 mg/l of PHE, ANT and FLU, respectively, in liquid MSM cultures. There is no significant difference in degradation efficiencies of *R. pyridinivorans*-DTU-7P and consortium indicating that there is no synergistic or inhibitive effect of both isolated strains over each other. Monitoring of bacterial growth curve during PAH degradation has notable significance in understanding biodegradation capabilities of bacterial strain. The biomass of both bacterial strains gradually increased during 3–6th day of incubation and after that maximum biomass accumulation was achieved. The maximum growth of both strains was observed on ANTH and FLU over PHE. Maximum degradation rate was observed in logarithmic growth phase of bacterial isolates; more then 40% degradation was achieved within 9 days. The degradation rate for PHE was found to be slightly low for both the bacterial strains as monocultures and as consortium.

It was observed that the degradation efficiency of monoculture *R. pyridinivorans*-DTU-7P for all three PAHs is more as compared to monoculture *K. flava* DTU-1Y and consortium (Fig. 2), however insignificant difference was observed in the degradation efficiencies of *R. pyridinivorans*-DTU-7P and consortium. *K. flava* DTU-1Y, *R. pyridinivorans*-DTU-7P and consortium had maximum degradation efficiency for FLU 63.46%, 66.79% and 66.64% as compared to other two PAHs. Minimal degradation was observed for PHE (55.13%) by *K. flava.* Upon comparing the degradation among isolated strains, minimum degradation for PHE was observed for both strains and for consortium. The minimum degradation for PHE might be due to its stable chemical structure as biodegradability primarily depends on the complexity of chemical structures as well (Bamforth and Singleton 2005). The benzene rings in ANTH are fused linearly having four resonance structures whereas, in PHE, these rings are fused at an angle having five resonance structures, this extra resonance in PHE makes it more stable than ANTH (Poater et al. 2018). The degradation profile of PAHs based on initial and final concentration determined over HPLC confirmed that complete degradation of PAHs has been observed, and no formation major secondary metabolites was noted (Fig. 3).

Whereas, other bacterial species belonging to the genus *Rhodococcus* were found to have PAH-degradation ability (Song et al. 2011; Goswami et al. 2017), *R. pyridinivorans* has not been investigated for PAH degradation. *Rhodococcus sp*. P14, isolated from crude oil contaminated sediment degraded 43% PHE in 30 days when supplemented at concentration of 50 mg/L (Song et al. 2011); and *R. opacus* degraded ANTH within 7 days (Goswami et al. 2017). In the present study, *R. pyridinivorans* DTU-7P strain has potential in utilising three ring PAHs as sole source of

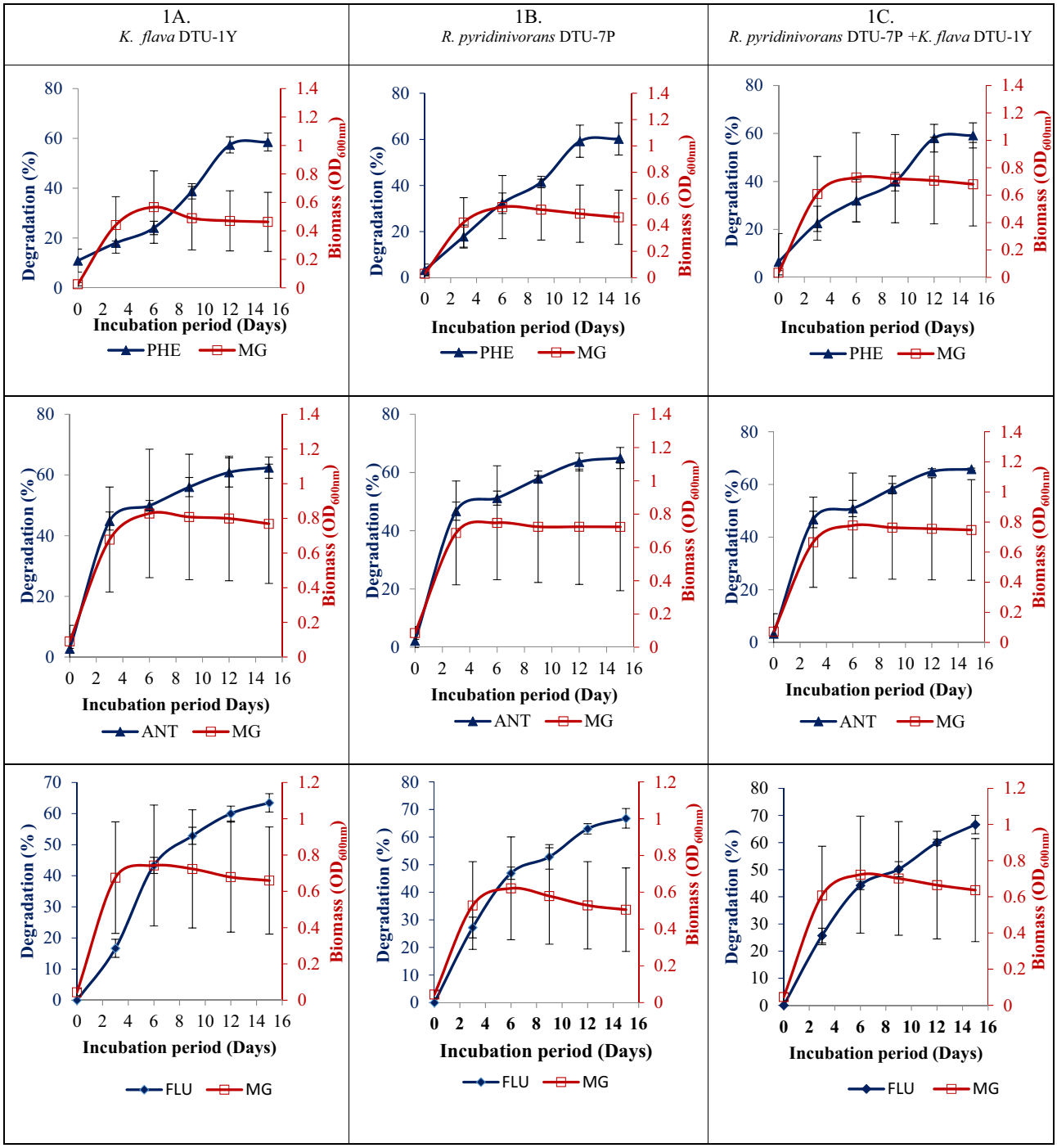


Fig. 2 Dynamics of microbial growth during biodegradation of phenanthrene (PHE), anthracene (ANT) and fluorene (FLU) by bacterial strains (1A.) *K. flava* DTU-1Y, (1B.) *R. pyridinivorans* DTU-7P, and (1C.) consortium of *R. pyridinivorans* DTU-7P and *K. flava* DTU-1Y in 15 days. (Solid symbol is about degradation; hollow symbol is about the biomass *i.e.* absorbance at OD_{600nm} which indicate bacterial growth and error bars represents standard deviation of replicates)

carbon and energy with 62.02%, 64.99% and 66.79% degradation efficiency for PHE, ANTH and FLU, respectively, after 15 days of incubation. Similarly, *K. flava* has not been reported for three-ring PAH degradation till date, but *K. flava* CMG2028 has exhibited growth in the presence of PHE (Ahmed et al. 2010). *K. rosea* has been reported to degrade 54.6% PHE (1 mg/L) in 15 days (Haritash and Kaushik, 2016), and 46–54% of ANTH (0.27 mg/L) within 2 days (Lara-Severino et al. 2016). The strains used in the present study represent slightly higher degradation

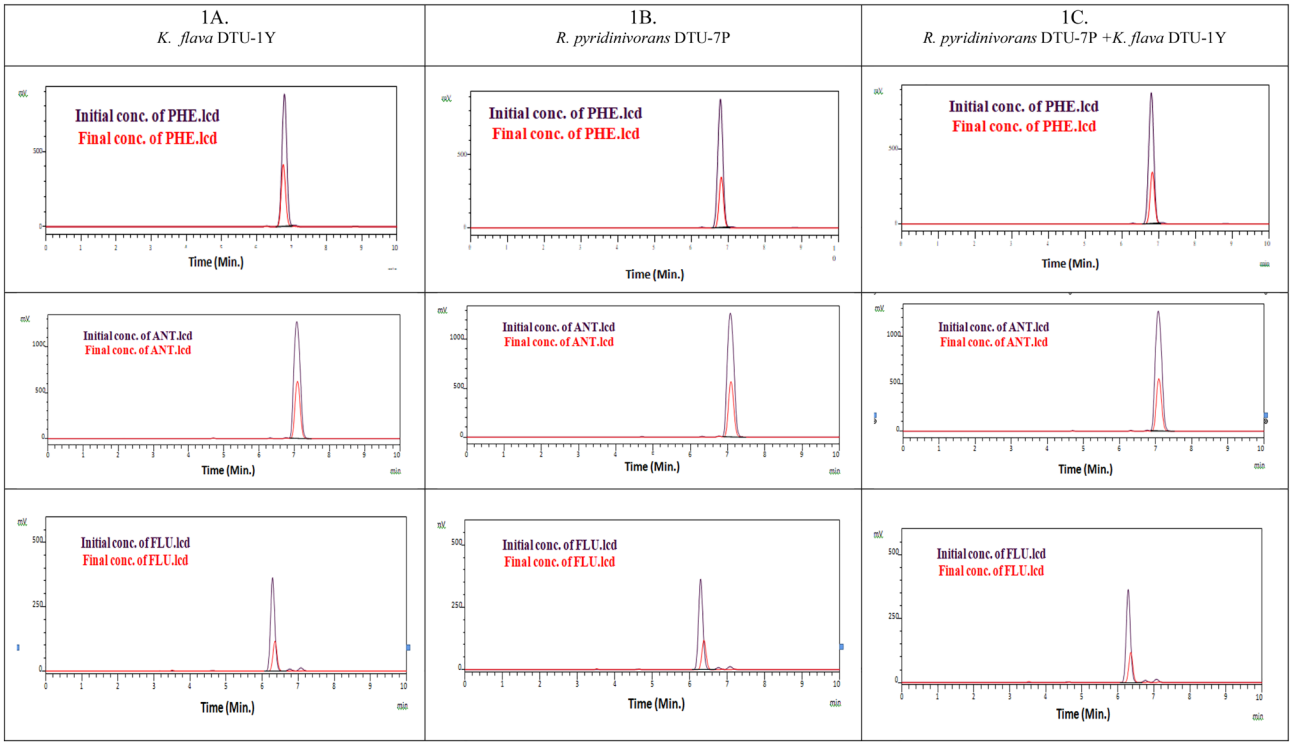


Fig. 3 Initial and final HPLC chromatograms of PHE, ANT and FLU during degradation by *K. flava*, *R. pyridinivorans* and consortium

efficiency compared to similar strains reported in other studies.

Enzyme activity

In the present study, catabolic enzymes such as C12O, C23O, dehydrogenase and peroxidase assay during PAH degradation were analysed at 0, 3, 6, 9, 12 and 15 days to understand the role of catabolic enzymes during degradation. None of the strain exhibited C12O activity in the present study. The levels of C23O, dehydrogenase and peroxidase activity were observed to increase during initial phase of degradation, and slightly decrease at 12th and 15th day of incubation (Fig. 4). C23O activity for *K. flava* was maximum during 6th, 9th and 12th days of culturing for PHE (5.7×10^{-4} µmoles/mL/min), ANTH (5.7×10^{-4} µmoles/mL/min) and FLU (2.6×10^{-4} µmoles/mL/min), respectively, whereas for *R. pyridinivorans,* C23O showed its highest levels during 6–9th, 12th and 6th days of culturing for PHE (3.8 × 10–4 µmoles/mL/min), ANTH (4.0 × 10–4 µmoles/mL/min) and FLU (3.4×10^{-4} µmoles/mL/min), respectively. Peroxidase activity for *K. flava* as well as for *R. pyridinivorans* was maximum during 9th day of incubation for all three PAHs. Moreover, for consortium C23O and peroxidase activity was maximum at 9th day for PHE (5.5 × 10–4 µmoles/mL/min, 17.4 × 10–4 µmoles/mL/min), 12th and 6th day for ANTH (2.1×10^{-4} µmoles/mL/min, 32.89×10^{-4} µmoles/mL/min), as well as for FLU (2.5×10^{-4} µmoles/mL/min, 14.0×10^{-4} µmoles/mL/min), respectively. Dehydrogenase activity for both strains was maximum during 6th and 9th day of incubation for all three PAHs and almost similar dehydrogenase activity was observed during initial phase of degradation for all three PAHs. It was observed that during 12th and 15th day of incubation dehydrogenase activity decreased.

Dioxygenases and peroxidase have been reported to play an important role in oxidation of PAHs and the results of this study demonstrated that activity of C23O and peroxidase enzymes in isolated strains show the opportunity to use these strains in potential in-situ biodegradation. Based on the activity of enzymes, C23O may be responsible for initiation of hydrolysis through meta-cleavage of benzene ring, whereas dehydrogenases and peroxidases might catalyze further oxidation of PAHs into simpler non-toxic forms. The dehydrogenase enzyme has been reported for its role in transformation of PAHs to dihydrodiols which are relatively less stable then PAHs (van der Waarde et al. 1995). The chemically reactive PAH dihydrodiols are easily oxidised by peroxidases into secondary metabolites broken down in Kreb's cycle (Cerniglia 1993). As C23O activity was observed during degradation, meta-cleavage by C23O might be common for PAH degradation (Meyer et al. 1999). C23O and peroxidase both are known to degrade PAHs but peroxidase activity has not been reported for bacterial species during PAHs degradation. To our knowledge, this is the

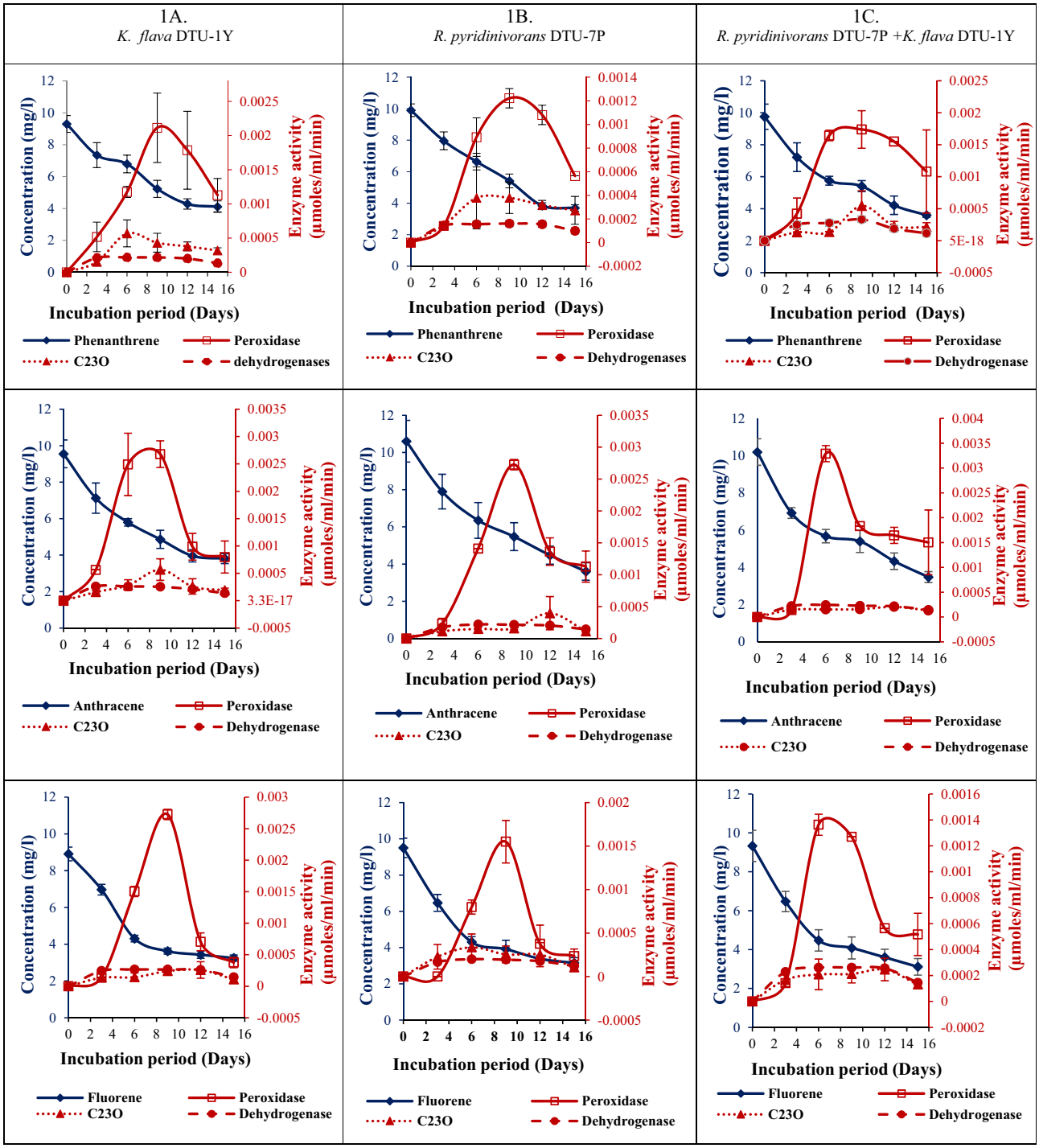


Fig. 4 Catabolic enzyme activities with average concentration of residual PAHs during biodegradation with *K. flava* DTU-1Y, *R. pyridinivorans* DTU-7P, and consortium in 15 days

first study to demonstrate the dynamic patterns of these catabolic enzymes, i.e. C23O, dehydrogenase and peroxidase simultaneously from two different bacterial isolates. The role of C23O in meta-cleavage of PHE has also been reported by *Pseudomonas* and *Sphingomonas* in earlier literature (Meyer et al. 1999), thus, confirming the meta-cleavage of benzene ring by C23O. In this study, C23O activity was found during degradation of each PAH which points towards involvement of C23O in initial ring cleavage of PAHs. The findings of this study indicate that *K. flava* DTU-1Y and *R. pyridinivorans* possess the ability to degrade PHE, ANTH and FLU; and have a potential for use in bioremediation processes.

Based on the results, it is suggested that PAH degradation mechanism in *K. flava* DTU-1Y and *R. pyridinivorans* could be mediated by dioxygenase enzyme (C23O), which may possibly play a role in the initial degradation of three-ring PAHs, subsequently, dehydrogenase and peroxidase enzyme could contribute to further degradation of remaining intermediates of PAHs.

Conclusion

The results reported in this study showed that PAH degradation is common in species previously adapted to the PAH-contaminated conditions. Results indicate that *R. pyridinivorans*-DTU-7P has better degradation potential for three-ring PAHs as compared to *K. flava* DTU-1Y and the degradation efficiency of PHE might be relatively less owing to its stable chemical structure. These strains were found to show high tolerance levels towards exposure of mixed PAHs and could be used for bioremediation of PAHs contaminated sites without any inhibitory/synergistic effects. Catabolic enzyme activity of dioxygenase (C23O), dehydrogenase, and peroxidase enzymes were found to play a crucial role in metabolism of PHE, ANTH and FLU. The PAH-degradation capabilities of *R. pyridinivorans*-DTU-7P and *K. flava* DTU-1Y with screening of catabolic enzymes can be exploited further for developing efficient bioremediation technology for cleanup of PAHs from a contaminated environmental matrix.

Acknowledgements The authors acknowledge the help of Bharat Petroleum Corporation Ltd. (BPCL), Bijwasan Delhi for providing the samples of contaminated soil; and CSIR- Institute of Microbial Technology, Chandigarh (IMTECH), India, for helping us in sequencing.

Declarations

Conflict of interest The authors have no 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.

References

Abdel-Shafy HI, Mansour MSM (2016) A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation. Egypt J Pet 25:107–123

Ahmed RZ, Ahmed N, Gadd GM (2010) Isolation of two *Kocuria* species capable of growing on various polycyclic aromatic hydrocarbons. Afr J Biotechnol 9(24):3611–3617

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bamforth SM, Singleton I (2005) Bioremediation of polycyclic aromatic hydrocarbons: current knowledge and future directions. J Chem Technol Biotechnol 80:723–736. https://doi.org/10.1002/jctb.1276

Bhattacharya A, Gupta A, Kaur A, Malik D (2015) Simultaneous bioremediation of phenol and Cr (VI) from tannery wastewater using bacterial consortium. Int J Appl Sci Biotechnol 3(1):50–55. https://doi.org/10.3126/ijasbt.v3i1.11889

Bishnoi K, Kumar R, Bishnoi NR (2007) Biodegradation of polycyclic aromatic hydrocarbons by white rot fungi *Phanerochaete chrysosporium* in sterile and unsterile soil. J Sci Ind Res 67:538–542

Burdock T, Brooks M, Ghaly A, Dave D (2011) Effect of assay conditions on the measurement of dehydrogenase activity of *Streptomyces venezuelae* using triphenyl tetrazolium chloride. Adv Biosci Biotechnol 2:214–225. https://doi.org/10.4236/abb.2011.24032

Cerniglia CE (1992) Biodegradation of polycyclic aromatic hydrocarbons. Biodegradation 3:351–368. https://doi.org/10.1007/BF00129093

Cerniglia CE (1993) Biodegradation of polycyclic aromatic hydrocarbons. Curr Opn Biotechnol 4:331–338

Chaudhary P, Beniwal V, Kaur R, Mehra R, Kumar A, Chhokar V (2019) Efficacy of *Aspergillus fumigatus* MCC 1175 for bioremediation of tannery wastewater. Clean-Soil Air Water 47(6):1900131

Dhanya MS, Kalia A (2020) Bioremediation: an eco-friendly cleanup strategy for polyaromatic hydrocarbons from petroleum industry waste. In: Saxena G, Bharagava R (eds) Bioremediation of industrial waste for environmental safety. Springer, Singapore. https://doi.org/10.1007/978-981-13-1891-7_18

Fernandez-Lafuente R, Guisan JM, Ali S, Cowan D (2000) Immobilization of functionally unstable catechol-2,3-dioxygenase greatly improves operational stability. Enzyme Microb Technol 26:568–573

Ghaly AE, Mahmoud NS (2006) Optimum conditions for measuring dehydrogenase activity of *Aspergillus niger* using TTC. Am J Biochem Biotechnol 2:186–194

Goswami L, Kumar RV, Manikandan NA, Pakshirajan K, Pugazhenthi G (2017) Anthracene biodegradation by oleaginous *Rhodococcus opacus*for biodiesel production and its characterization. Polycycl Aromat Comp. https://doi.org/10.1080/10406638.2017.1302971

Haritash AK, Kaushik CP (2009a) Biodegradation aspects of polycyclic aromatic hydrocarbons: a review. J Hazard Mater 169:1–15

Haritash AK, Kaushik CP (2009b) Seasonal and spatial occurrence and distribution of respirable particulate-bound atmospheric polycyclic aromatic hydrocarbons in Hisar City (India) and their potential health-risks. Asian J Water Environ Pollut 8(1):73–80

Haritash AK, Kaushik CP (2016) Degradation of low molecular weight polycyclic aromatic hydrocarbons by microorganisms isolated from contaminated soil. Int J Environ Sci 6(5):646–656

Kaushik CP, Sangwan P, Haritash AK (2012) Association of polycyclic aromatic hydrocarbons (PAHs) with different sizes of atmospheric particulate in Hisar City and its health aspects. Polycycl Aromat Comp 32(5):626–642

Kiyohara H, Nagao K, Yana K (1982) Rapid screen for bacteria degrading water-insoluble, solid hydrocarbons on agar plates. Appl Environ Microbiol 43(2):454–457

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549

Kuppusamy S, Thavamani P, Venkateswarlu K, Lee YB, Naidu R, Megharaj M (2016) Remediation approaches for polycyclic aromatic hydrocarbons (PAHs) contaminated soils: technological constraints, emerging trends and future directions. Chemosphere 168:944–968

Lara-Severino RC, Camacho-López MA, Casanova-González E, Gómez-Oliván LM, Sandoval-Trujillo ÁH, Isaac-Olivé K, Ramírez-Durán N (2016) Halo alkali tolerant actinobacteria with capacity for anthracene degradation isolated from soils close to

areas with oil activity in the State of Veracruz, Mexico. Int Microbiol 19:15–26
Lyu Y, Zheng W, Zheng T, Tian Y (2014) Biodegradation of polycyclic aromatic hydrocarbons by Novosphingobium pentaromativorans US6–1. PLoS ONE 9(7):e101438
Megharaj M, Ramakrishnan B, Venkateswarlu K, Sethunathan N, Naidu R (2011) Bioremediation approaches for organic pollutants: a critical perspective. Environ Int 37:1362–1375
Meyer S, Moser R, Neef A, Stahl U, Kampfer P (1999) Differential detection of key enzymes of polyaromatic-hydrocarbon-degrading bacteria using PCR and gene probes. Microbiology 145:1731–1741
Nakazawa T, Nakazawa A (1970) Pyrocatechase (*pseudomonas*). In: Taber H, Tabor CW (eds) Methods in enzymology. Academic Press, New York, pp 522–525
Nozaki M (1970) Metapyrocatechase of pseudomonas. In: Taber H, Tabor CW (eds) Methods in enzymology. Academic Press, New York, pp 522–525
Pandey AK, Chaudhary P, Singh SB, Arora A, Kumar K, Chaudhry S, Nain L (2012) Deciphering the traits associated with PAH degradation by a novel *Serratia marcesencs* L-11 strain. J Environ Sci Health 47(5):755–765
Parthipan P, Preetham E, Machuca LL, Rahman PK, Murugan K, Rajasekar A (2017) Biosurfactant and degradative enzymes mediated crude oil degradation by bacterium *Bacillus subtilis* A1. Front Microbiol 8:193. https://doi.org/10.3389/fmicb.2017.00193
Poater J, Duran M, Solà M (2018) Aromaticity determines the relative stability of kinked vs. straight topologies in polycyclic aromatic hydrocarbons. Front Chem 6:561. https://doi.org/10.3389/fchem.2018.00561
Pozdnyakova N, Dubrovskaya E, Chernyshova M, Makarov O, Golubev S, Balandina S, Turkovskaya O (2018) The degradation of three-ringed polycyclic aromatic hydrocarbons by wood-inhabiting fungus *Pleurotus ostreatus* and soil-inhabiting fungus *Agaricus bisporus*. Fungal Biol 122(5):363–372
Qi YB, Wang CY, Lv CY, Lun ZM, Zheng CG (2017) Removal capacities of polycyclic aromatic hydrocarbons (PAHs) by a newly isolated strain from oil field produced water. Int J Environ Res Public Health 14:215
Rabodonirina S, Rasolomampianina R, Krier F, Drider D, Merhaby D, Net S, Ouddane B (2019) Degradation of fluorene and phenanthrene in PAHs-contaminated soil using *Pseudomonas* and *Bacillus* strains isolated from oil spill sites. J Environ Manage 232:1–7. https://doi.org/10.1016/j.jenvman.2018.11.005
Reddy KR, Adams JA (2015) Sustainable remediation of contaminated sites. Momentum Press, New York
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
Sakshi, Haritash AK (2020) A comprehensive review of metabolic and genomic aspects of PAH-degradation. Arch Microbiol. https://doi.org/10.1007/s00203-020-01929-5
Sakshi, Singh SK, Haritash AK (2019) Polycyclic aromatic hydrocarbons: soil pollution and remediation. Int J Environ Sci Technol. https://doi.org/10.1007/s13762-019-02414-3
Sakshi, Singh SK, Haritash AK (2020) Evolutionary relationship of polycyclic aromatic hydrocarbons degrading bacteria with strains isolated from petroleum contaminated soil based on 16S rRNA diversity. Polycycl Aromat Compd. https://doi.org/10.1080/10406638.2020.1825003
Sharma D, Jain S (2019) Impact of intervention of biomass cookstove technologies and kitchen characteristics on indoor air quality and human exposure in rural settings of India. Environ Int 123:240–255
Song X, Xu Y, Li G, Zhang Y, Huang T, Hu Z (2011) Isolation, characterization of Rhodococcus sp. P14 capable of degrading high-molecular-weight polycyclic aromatic hydrocarbons and aliphatic hydrocarbons. Mar Pollut Bull 62:2122–2128
Subashchandrabose SR, Venkateswarlu K, Naidu R, Megharaj M (2019) Biodegradation of high-molecular weight PAHs by *Rhodococcus wratislaviensis* strain 9: Overexpression of amidohydrolase induced by pyrene and BaP. Sci Total Environ 651:813–821
Vaishnavi J, Devanesan S, AlSalhi MS, Rajasekar A, Selvi A, Srinivasan P, Govarthanan M (2020) Biosurfactant mediated bioelectrokinetic remediation of diesel contaminated environment. Chemosphere 264:128377. https://doi.org/10.1016/j.chemosphere.2020.128377
van der Waarde JJ, Dijkhuis EJ, Hessen MJC, Keuning S (1995) Enzyme assay as indicators for biodegradation. In: van der Brick WJ, Bosman R, Arendt F (eds) Contaminant soil. Springer, Berlin, pp 1377–1378
Whitaker JR (1972) Principles of enzymology for food sciences. Marcel Dekker, New York
Wild SR, Jones KC (1995) Polynuclear aromatic hydrocarbons in the United Kingdom environment: a preliminary source inventory and budget. Environ Pollut 88(1):91–108
Zhang GY, Ling JY, Sun HB, Luo J, Fan YY, Cui ZJ (2009) Isolation and characterization of a newly isolated polycyclic aromatic hydrocarbons-degrading *Janibacter anophelis* strain JY11. J Hazard Mater 172(2–3):580–586

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

Clickbait in Social Media : *Detection and Analysis of the Bait*

Mini Jain*, Peya Mowar*, Ruchika Goel*, and Dinesh K. Vishwakarma*
Department of Information Technology
Delhi Technological University, Delhi, India
Email: mini.j2799@gmail.com, peyajm29@gmail.com, ruchika2ar@gmail.com, dinesh@dtu.ac.in
**All authors have contributed equally*

***Abstract*—Taking advantage of visual-centric social media's rising popularity, content creators have started using enticing images to lure users into clicking on bothersome clickbaits, in place of previously used text-based baits. In addition, the development of a single model to detect clickbait on multiple image-centric social media platforms is largely an unexplored problem. Therefore, we introduce a novel model that can detect visual clickbaits on both Instagram and Twitter posts. The proposed model consists of a stacking classifier framework composed of six base models (K-Nearest Neighbors, Support Vector Machine, XGBoost, Naive Bayes, Logistic Regression, and Multilayer Perceptron) and a meta-classifier (Random Forest). The developed classifier achieved an accuracy of 88.5% for Instagram posts and 85% for Twitter posts, which is an improvement over previous separate state-of-the-art models for both platforms. Additionally, the stated classifier does not use meta-features (e.g., the number of likes or followers) for classification, which helps to detect potential clickbaits right away, enhancing its applicability in real-time clickbait detection use cases. Furthermore, based on our analysis, we have drawn essential conclusions about the telling characteristics of clickbaits.**

Keywords—clickbait detection, ensemble learning, stacking classifier, visual, instagram, twitter

I. INTRODUCTION

Clickbait in social media refers to misleading and exaggerated headlines and hashtags, often accompanied by specious images whose primary objective is to entice the user to click on them. The headlines and images used in clickbait take advantage of the 'curiosity gap' by providing barely enough information in the form of a teaser message to make the crowd on social media interested, however insufficient to fulfil their interest without navigating to the connected link or post, often disregarding the content's target audience. Fig. 1(a) shows an instance of a clickbait post on Twitter that exploits this curiosity gap. Besides, clickbait can refer to instances wherein searching via hashtags shows posts and tweets containing unrelated images to the searched hashtag. In such cases, the content creators aim to reach a larger audience by using popular hashtags, such as luxurious brands that are likely to get clicked on, thus deceiving them into clicking on irrelevant and unrelated content. Fig. 1(b) shows an instance of a clickbait post on Instagram using an irrelevant brand name hashtag, "#hollister" to lure the user. In recent times, the most popular model of generating revenue online is through advertising since creators earn income for each ad view. Therefore, online content creators use clickbaits to reach a larger audience, leading to increased page views and readership, thus increasing their ad revenue.

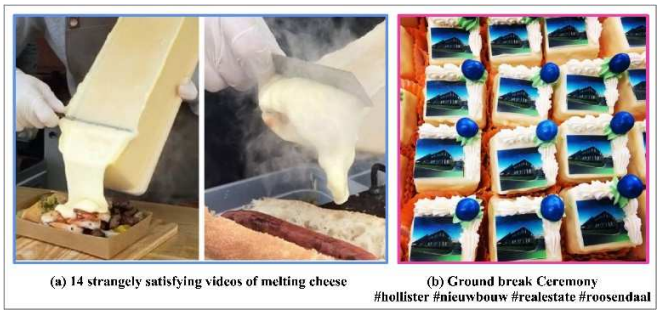


Fig. 1. Clickbait instances for (a) Twitter and (b) Instagram

Clickbaits are a nuisance for users and brands alike. When users click on a hashtag or link, users hope to be directed to relevant content; however, in the case of clickbaits, they are often driven towards low quality or unrelated content. This leads to frustration and a poor experience for the user. On the other hand, it damages the brand's reputation and future engagement prospects. Therefore, in light of an enormous increase in clickbait over the years, as shown by [1], we propose a novel approach to detect clickbaits on both Twitter and Instagram. We have chosen Instagram and Twitter for our study's scope because of their increasing popularity amongst users and brands in contrast to other social media platforms. This claim is supported by [2], which mentions that 33% of B2B businesses, 71% of businesses, and 98% of fashion brands use Instagram. Furthermore, according to [3], 92% of companies tweet more than once a day, and twitter's site referral traffic is up 6%, year-over-year. This research's main intention is to create a classification model that can predict the presence of potential clickbaits on Twitter or Instagram, by using only teaser text and image features of posts, with improved accuracy over existing methods or baseline models.

Previous work in the field of clickbait detection has majorly aimed to create models that detect clickbaits in only one type of social media at a time [4]; however, we propose a model that can catch clickbaits on both Twitter and Instagram posts, taking both the text and associated image into account. Furthermore, not using metadata features can facilitate the detection of clickbaits in real-time and improve users' social media feed and online experience by eliminating or marking such content as potential clickbait for users' reference.

This paper is organized as follows. Section II reviews the related work on clickbait detection, research gaps, and future scope. Section III presents an overview of our proposed methodology and the architecture of the employed stacking

2021 55th Annual Conference on Information Sciences and Systems (CISS) | 978-1-6654-1268-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISS50987.2021.9400293

978-1-6654-1268-1/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:49:54 UTC from IEEE Xplore. Restrictions apply.

classifier. Section IV discusses the results of clickbait detection on Twitter and Instagram with the proposed scheme. Section V outlines a few findings about the defining characteristics of clickbaits. Finally, Section VI concludes the paper and explains the future scope for the underlying research.

II. Literature Review

Some research has been done in the field of clickbait detection. With time, the definition of clickbait has evolved and broadened to encompass a broader class of deceiving posts, and various clickbait detection models have been developed. In this section, we review the major work done in this domain.

A. Early Work

Prior research and studies have focused on clickbaits primarily in news media wherein an article is marked clickbait if there is a disagreement between the news headline and the link's content. Biyani, Tsioutsiouliklis, and Blackmer [5] developed a clickbait detection model for news streams based on the headline's content and the linked webpage. They engineered several features based on the headline and web page content, the similarity between them, and the web page content's informality. They achieved a high performance (0.749 F1-score) in detecting clickbait articles. A similar definition of clickbaits based on incongruence between headline and content has been explored[6].

B. Redefining Clickbaits

However, there was a shift over time in the definition of the clickbait from the disparity between the headline of the post and linked content to one where content creators aimed to take advantage of the 'curiosity gap' of the user (i.e., by revealing some information to raise their curiosity but not enough to satisfy it unless they clicked on the link). In light of this new definition, Potthast et al. [7] created a corpus of 2992 tweets. They annotated the tweets only based on the teaser headline and accompanying image. Further, they developed a Random Forest classifier to detect clickbait posts on Twitter and achieved 0.79 ROC-AUC. However, the features designed in the dataset were primarily text-based using the information extracted from the headline and the linked webpage in addition to meta-features such as the presence of media in the tweet or if the tweet had been retweeted.

Furthermore, Chakraborty et al. [8] developed a clickbait detection and blocking mechanism based only on the various linguistic features extracted from the headline content. They achieved an accuracy of 93% in the detection and 89% in blocking of clickbaits, showing that clickbait can be successfully detected based only on the user's information to manipulate him into clicking without using the actual contents of the link.

C. Clickbait in Visual Media

The rising acclaim of visual-based social networking platforms (e.g., Facebook, Pinterest) presented online content creators with an opportunity to deceive users by using clickbaits and increasing their page views. Therefore, moving toward detecting clickbaits in majorly visual-centric social media, Ha et al. [4] created a clickbait detection mechanism for Instagram. Previously, most clickbait detection methods had been primarily developed using text-based features and were thus, not applicable to visual social media. Using a manually annotated dataset of Instagram posts, they achieved an accuracy of 0.864. They showed via analysis that content-based features (like text and image in post) are much more critical than meta-features (like the number of followers of the content creator). Furthermore, they defined clickbaits on Instagram as incorrectly labelled posts with misleading hashtags causing poor-quality information retrieval when searched via hashtags. They argued that such posts are intended to reach a larger audience and manipulate them into clicking by labelling them with popular hashtags irrelevant to the post. However, clickbait detection in visual social media remains an unexplored domain to a great extent with the further scope of research and analysis on Instagram and other platforms like Pinterest, Vine, etc.

D. Research Gaps and Future Work

Since clickbait detection in social media is a relatively uncharted field, there is enormous scope to improve such models' accuracy and efficacy by using advanced Computer Vision and NLP techniques. Research has been conducted to make clickbait detection mechanisms independent of social media platforms, but primarily in text-based social media or using only text-based features [9]. However, these approaches have failed to use copious information present in images that accompany such posts. Therefore, there is scope to develop a robust cross-platform clickbait detection model that takes advantage of image-based features in addition to text-based features.

III. Methodology

The proposed methodology for our framework for clickbait identification is illustrated in Fig. 2. We suggest a robust pipeline developed using Python that incorporates feature engineering and ensemble learning for classification.

A. Feature Engineering

Feature engineering includes extracting text and image features from the post text and the associated image. We propose textual feature extraction using BERT-Base [12] and image feature extraction using ResNet50 [13].

We extracted features from tweet's text using BERT-Base. BERT stands for Bidirectional Encoder Representations from Transformers [12]. It is devised to pre-train deep bidirectional representations from an unsupervised text by simultaneously conditioning on both left and right context. Consequently, the BERT-Base model can be fine-tuned by adding merely one supplementary output layer to create classifiers for a broad array of NLP assignments such as Sentiment Analysis and Next Sentence Prediction Tasks. Accordingly, the BERT-Base model can extract conceptualized word embeddings for the text from the post to get the pooled output of size 768 units as the feature vector.

Image features from tweet's image were extracted using ResNet50. ResNet50 is a 50-layer deep convolutional neural network. A pre-trained neural network model is accessible, trained on more than a million images from the ImageNet database [14], which can classify photos into a thousand target categories. Consequently, the network has acquired deep

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:49:54 UTC from IEEE Xplore. Restrictions apply.

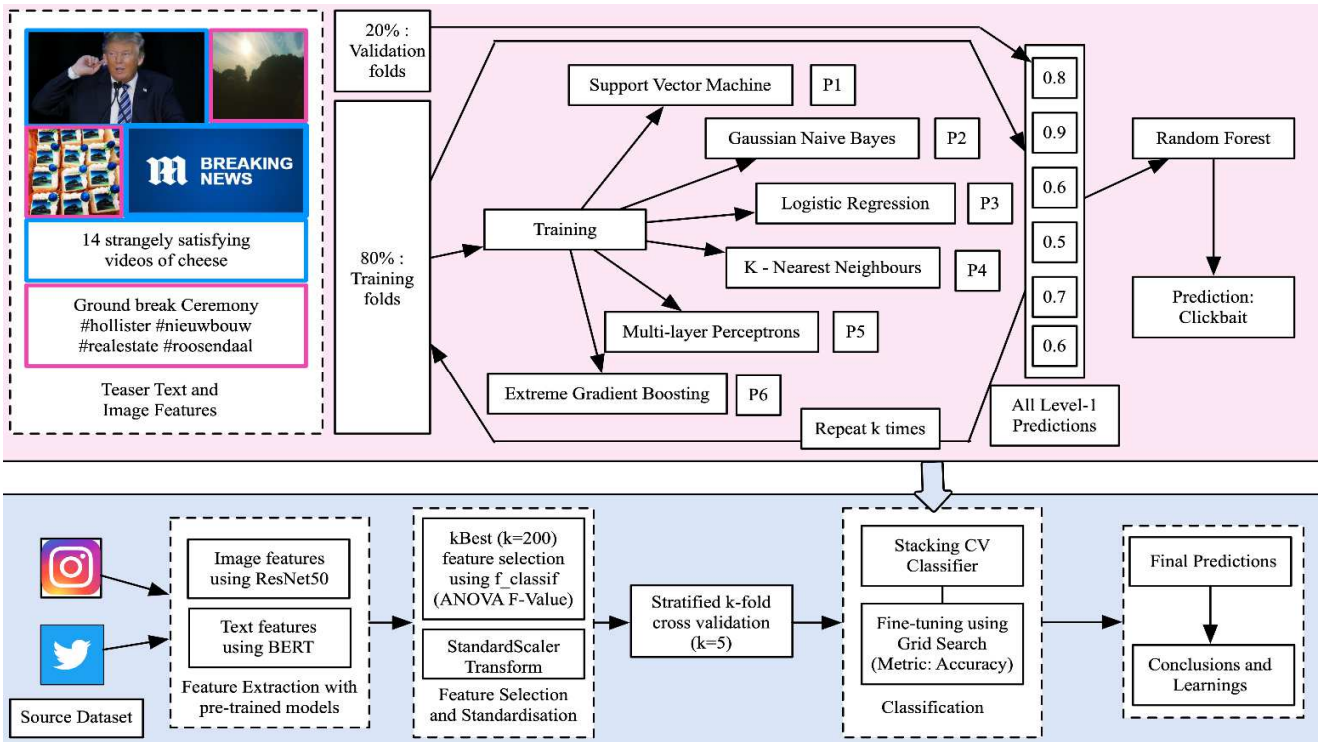


Fig. 2. Pipeline of the proposed methodology

characteristic representations for an extensive collection of images and, in most cases, can be used without any fine-tuning.

To reduce overfitting, improve accuracy, and decrease the training time, we suggest using the ANOVA F-value (f_classif) statistical test to pick the best attributes. This statistical selects attributes having the most robust association with the dependent variable. After selection, the normalization of the features should be carried out per fold to obtain normalizing factors using the training data and applying them to the validation data such that the population has an average value of 0 and a standard deviation of 1.

B. Classification

In order to classify the instances as clickbait and non-clickbait, we recommend ensemble learning techniques to model a clickbait detection classifier. Ensemble learning models are meta-learning programs that blend different deep learning methods into an individual prediction model. The benefits of using ensemble learning are that they aid in decreasing variance (using bagging techniques), reducing bias (using boosting techniques), and improving predictions (by using stacking).

1) ***Bagging****:* Bagging is an acronym for bootstrap aggregation. It helps to decrease the variance of an independent attribute by averaging multiple estimates from models trained on several sets of training data picked randomly with replacement. It employs majority voting for classification and equalizing for regression to aggregate the yields of the base learners. In our research, we use the following models as our base learners: K-Nearest Neighbours (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) and XGBoost (XGB). The output from each of these models, i.e., the binary classification of a post being clickbait or not, is used to perform majority voting and ultimately, classify a post as clickbait or not. Furthermore, we compare the performance of bagging and stacking techniques. For the latter, we use the output probability of a post being clickbait from all the base learners as features for training the stacking classifier.

2) ***Boosting****:* Boosting belongs to a class of algorithms that are applied to transform weak learners into strong learners wherein learners are trained sequentially, with first learners fitting uncomplicated models to the data and examine data for inaccuracies. Continuous trees are fit, and at each level, the aim is to increase the efficiency from the former tree. If a hypothesis incorrectly assigns a label, its weight is decreased so that the next assumption is more prone to classify it correctly. We use XGB[15], short for 'eXtreme Gradient Boosting,' which implements gradient boosted trees, to perform boosting in our research.

3) ***Stacking:*** Stacking is an ensemble learning method that can be applied for achieving classification via a meta-classifier. The base-level or level-0 classifiers are trained on the complete training set, and then the meta-classifier is trained on the predictions of the level-0 classifier-like features. The base-level model often consists of different machine learning techniques, and therefore stacking ensembles are usually wide-ranging and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:49:54 UTC from IEEE Xplore. Restrictions apply.

non-homogeneous. We use the previously mentioned classifiers as base-level models and feed their output, i.e., probability of a post being clickbait as features into the meta-classifier, which is a Random Forest (RF) classifier.

We suggest hyperparameters for all the models be tuned using Grid Search wherein a grid of hyperparameter values is created the model is trained for each combination, and the set of hyperparameters resulting in highest accuracy is selected.

IV. Experimental Results

For experimentation, we selected instances from the Webis Clickbait Corpus 2017 dataset that contained images and extracted features from it using pre-trained ResNet50 and BERT models. We combined this dataset with a publicly available clickbait classification dataset for Instagram and used as our research dataset. Subsequently, we applied and compared various classification techniques with the proposed classification model and evaluated the results. We also gauged the feature importance by testing for different feature combinations.

A. Source and Derived Datasets

The Clickbaits Labeling Data on Instagram (CLDI) dataset by Harvard [10] contains information about fashion-pertaining posts collected from Instagram. The clickbait posts in this dataset exploit "visual cues" and use unrelated but popular hashtags to lure the user into visiting their post. In the Webis Clickbait Corpus 2017 (WCC) [11] for Twitter, each instance in the dataset has been labeled as clickbait or non-clickbait only based on the information presented in the teaser message and the accompanying image. A post is tagged clickbait if it aims to take advantage of the user's "curiosity gap" to mold him into clicking the link. A more detailed description of the source datasets has been provided in Table I.

TABLE I. Source Dataset Characteristics

Characteristics	Social Media Platform	
	Instagram	Twitter
Dataset source	CLDI	WCC
Total instances	7769	19538
Clickbait instances	4260	4761
Non-clickbait instances	3509	14777
Number of attributes	2080	14
Labelling characteristics	Binary	Graded
Attribute characteristics	Integers	Strings

Since our study focuses on clickbaits in visual social media, we selected the tweets from the WCC dataset that contained an accompanying image, which reduced the number of instances from 19,538 to 10,306. Image and text feature extraction was performed for the WCC dataset.

Feature extraction was already done for the CLDI dataset by the dataset curators; therefore, we used it as it was available without performing any additional feature extraction. The derived dataset for WCC had 2215 clickbait instances, while the derived CLDI dataset had 4260 clickbait instances, as shown in Fig. 3. A detailed description of the derived datasets is also provided in Table II.

TABLE II. Derived Dataset Characteristics

Characteristics	Dataset	
	CLDI	WCC
Total selected instances	7769	10306
Clickbait instances	4260	2215
Non-clickbait instances	3509	8091
Source dataset features	2070	14
Extracted features	0	2804
Total features	2070	2818
Selected features	200	200
Labelling characteristics	Binary	Binary
Attribute characteristics	Integers	Decimals

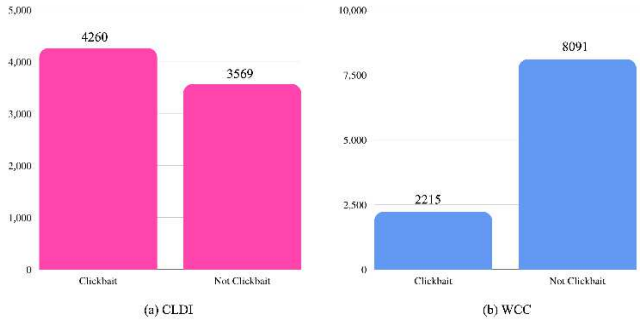


Fig. 3. Distribution of clickbait and non-clickbait instances in (a) CLDI and (b) WCC

B. Classification

We tried six different classification models; namely, KNN, LR, SVM, GNB, XGB, MLP, and the results that we observed are summarized in Table III.

TABLE III. Derived Dataset Characteristics

Model	Dataset	Accuracy	Precision	Recall	F1	AUC
KNN	CLDI	0.761	0.729	0.747	0.739	0.760
	WCC	0.723	0.811	0.843	0.827	0.563
LR	CLDI	0.600	0.159	0.129	0.157	0.579
	WCC	0.773	0.868	0.838	0.813	0.756
GNB	CLDI	0.518	0.515	0.896	0.649	0.568
	WCC	0.721	0.860	0.769	0.813	0.692
XGB	CLDI	**0.877**	0.865	0.864	0.863	0.949
	WCC	**0.838**	0.859	0.949	0.902	0.848
MLP	CLDI	0.844	0.809	0.854	0.831	0.925
	WCC	0.815	0.862	0.909	0.885	0.807
SVM	CLDI	0.807	0.788	0.785	0.786	0.890
	WCC	0.765	0.866	0.830	0.847	0.750

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:49:54 UTC from IEEE Xplore. Restrictions apply.

All the metrics are the average of the individual metrics over the 5-fold cross-validation run of the classifier. We then proceeded to try a Majority Voting Classifier that did majority voting on the result of those six models. We compared this with a Stacking CV Classifier, which performed ensemble learning and used two levels of classifications to predict the output. We split each dataset into training and test sets (8:2) and carried out 5-fold stratified cross-validation for evaluating the model. We adjusted the parameters to use the top 200 best performing attributes of the complete input data for classification.

Table IV shows the accuracy, f1-score, precision, recall, and ROC-AUC value for both the datasets using our Generalised Clickbait Classifier as well as a Majority Classifier. All the metrics are the average of the individual metrics over the 5-fold cross-validation run of the classifier.

TABLE IV. MAJORITY AND STACKING CLASSIFIER RESULTS

Model	Dataset	Accuracy	Precision	Recall	F1	AUC
Majority classifier	CLDI	0.865	0.861	0.834	0.847	0.862
	WCC	0.834	0.851	0.954	0.9	0.682
Stacking classifier	CLDI	**0.885**	0.870	0.877	0.874	0.930
	WCC	**0.850**	0.865	0.960	0.904	0.814

*1) **Model Performance:*** We chose the Stacking Classifier as our clickbait classification model. It performed best with an accuracy of 85% for the WCC dataset and an accuracy of 88.5% for the CLDI dataset as shown in Table V. Our model showed accurate and robust classification performance (with the accuracies exceeding 84%) when both image and text features are included. For the CLDI dataset, the image features, in particular, yielded a high accuracy (86%), which was comparatively more significant than the accuracy with the text features (75.1%). This was backed by the fact that the top 200 attributes extracted for this dataset are mostly image-based, as demonstrated in Fig. 4.

TABLE V. RESULTS FOR DIFFERENT FEATURE COMBINATIONS

Features	Dataset	Accuracy	Precision	Recall	F1	AUC
Image	CLDI	**0.860**	0.857	0.843	0.843	0.910
	WCC	0.791	0.795	0.991	0.882	0.646
Text	CLDI	0.751	0.720	0.737	0.729	0.750
	WCC	**0.849**	0.865	0.960	0.904	0.814
Image + Text	CLDI	**0.885**	0.870	0.877	0.874	0.930
	WCC	**0.850**	0.865	0.960	0.904	0.814

For the WCC dataset, the text features alone yielded the highest accuracy (84.9%), which was equivalent to the accuracy achieved by combining the text and image features. This was confirmed by the fact that the top 200 features extracted for this dataset were utterly text-based, as shown in Fig 4.

*2) **Baseline performance:*** For the CLDI dataset, we successfully improved the accuracy by 2.1% from the baseline (RF model in [4]). For the WCC dataset, we used a modified version of the dataset and hence considered the RF model trained on this dataset as the baseline that gave an accuracy of 80.5%. We were successful in improving the classification accuracy by 4.5%.

V. FINDINGS AND LEARNINGS

While creating the clickbait classification model, we studied how clickbaits entice users and what sets them apart from regular non-clickbait posts to draw insights by carefully studying the data and conducting rigorous research. These learnings or "reasons" for the bait in clickbait are explained below:

*1) **Relevance of Image and Text Features on different kinds of clickbaits:*** We discovered that clickbaits could majorly be classified into one of two types, one that exploited the "curiosity gap" and one that lured users by using visual cues. Clickbaits exploiting curiosity gaps contained enticing text and descriptions that nudged the user to know more, whereas clickbaits based on visual or directional cues subtly drew user attention. In the datasets that we used, the WCC dataset contained the former kind of clickbaits, while the CLDI dataset contained the latter in abundance. To analyze which features were more relevant to which type of clickbaits, we visualized the number of text and image features in the 200 best features that were retained after feature selection, as shown in Fig. 4. As was expected, the text features were much more relevant to the former kind of clickbaits, and the image features were given more importance in the latter.

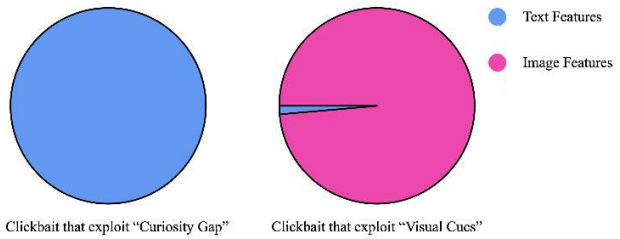


Fig. 4. Two types of clickbaits

*2) **Brand mentioning by clickbaiters to attract attention from the brand-conscious generation:*** Social-media statistics suggested that Instagram is host to most fashion and lifestyle brands with raging popularity, with brands such as Nike having up to 90 million followers. In this era of a brand-conscious generation, clickbaiters found it possible to take advantage of sought-after brands' popularity into gaining attention from novice users. Often, clickbaiters mentioned these brands in the form of hashtags or tags in their posts. In the CLDI dataset, the posts have been collected by searching for hashtag mentions of desired brands in the fashion industry. The traction of this brand mentioning trend among clickbait posts is proved by the percentage of clickbait posts collected by this method. We observed that of the 7,769 posts compiled by the searched tag, 4260 posts were labeled as clickbaits. Moreover, the average

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:49:54 UTC from IEEE Xplore. Restrictions apply.

clickbait post mentioned up to 3 of the 100 most popular brands in this data. We additionally observed that 80% of the top 10 brands classified as clickbaits were luxury fashion brands. This is because the desirability of these brands was much higher. Hence clickbaiters targeted such brands to increase their reach more often.

3) ***Presence of an additional link in clickbait posts:*** Very often, the motivation for posting clickbaits was to get the user to a redirected webpage, which may be ranging from an advertisement to a user's profile desiring more followers to a phishing website. We observed that up to 2.4% of clickbait posts in the CLDI dataset contained additional URLs in the post, and up to 18.1% contained user mentions. All the tweets in the WCC dataset also included linked articles. This suggested that the presence of an additional link in a caption was an indicative factor of clickbait.

4) ***Presence of an additional image in clickbait tweets:*** We observed that for the source dataset of WCC, 53.4% of the clickbait posts had images associated with them as images increase the post's attractiveness factor. This was supported by the claim that tweets with images receive 18% more click-throughs, 89% more likes, and 150% more retweets [16].

5) ***Less text in clickbait posts:*** In our analysis, we found that clickbait posts had minimal and emphasized text providing incomplete information to grab the viewers' attention and take advantage of the curiosity gap to lure the user. We discovered that the average length of a clickbait tweet was 130 characters, in contrast to 145 character length of an average benign tweet.

VI. Conclusions And Future Scope

Clickbait detection and subsequent filtration are essential for a quality user experience across social media. This paper presents a generalized clickbait classifier for detecting clickbait in social media and draws important conclusions about the nature of clickbaits and trends prevalent in clickbaits across social media. Our model is flexible and is expected to perform well across different social media and various forms of clickbait and takes both the post text and the associated image into account for classification. It is also independent of the topic of clickbait as it performs equally well for a fashion-focused CLDI dataset and a news-focused WCC dataset. As the proposed classifier is not dependent on meta-features, it is possible to deploy the model on posts from the moment they are posted. This would decrease the number of clickbait posts circulating by filtering them out from the recommendation feed early, thus improving user experience and engagement.

Since our goal comprised of analysing the importance of text-based and image-based features to gain an understanding of the types of cues employed by content-creators to allure users into clicking on clickbaits, we performed feature selection prior to cross-validation. However, a more conservative approach can be to perform feature-selection per cross validation round considering only the training data. There is also scope to broaden testing of the framework across more visual-centric social-media platforms such as Pinterest, Facebook, and Reddit. Furthermore, the model can be extended to incorporate clickbait detection for videos on Twitter and Instagram developing on previous works done for video-streaming platforms such as YouTube [17]. Deployment of the model as a web and mobile extension for Twitter and Instagram can be another possible scope for exploration in the future.

References

[1] M. M. Uddin Rony, N. Hassan, and M. Yousuf, "Diving deep into clickbaits: Who use them to what extents in which topics with what effects?," Proc. 2017 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2017, no. July, pp. 232–239, 2017, doi: 10.1145/3110025.3110054.

[2] "Social Media Demographics to Inform your Strategy in 2020 | Sprout Social." https://sproutsocial.com/insights/new-social-media-demographics/ (accessed Dec. 19, 2020).

[3] "309 Social Media Statistics You Must Know In 2020." https://www.socialpilot.co/blog/social-media-statistics/ (accessed Dec. 19, 2020).

[4] Y. I. Ha, J. Kim, D. Won, M. Cha, and J. Joo, "Characterizing clickbaits on instagram," 12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018, no. Icwsm, pp. 92–101, 2018.

[5] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, "Detecting Clickbaits in News Streams Using Article Informality," Thirtieth AAAI Conf. Artif. Intell., pp. 94–100, 2016.

[6] S. Chesney, M. Liakata, M. Poesio, and M. Purver, "Incongruent Headlines: Yet Another Way to Mislead Your Readers," no. 1, pp. 56–61, 2018, doi: 10.18653/v1/w17-4210.

[7] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait Detection," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9626, no. 1, pp. 810–817, 2016, doi: 10.1007/978-3-319-30671-1_72.

[8] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016, pp. 9–16, 2016, doi: 10.1109/ASONAM.2016.7752207.

[9] A. Agrawal, "Clickbait detection using deep learning," Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016, no. October, pp. 268–272, 2017, doi: 10.1109/NGCT.2016.7877426.

[10] "Clickbaits Labeling Data on Instagram - Harvard Dataverse." https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DEZMRA (accessed Dec. 19, 2020).

[11] M. Potthast et al., "Crowdsourcing a Large Corpus of Clickbait on Twitter," Proc. 27th Int. Conf. Comput. Linguist., pp. 1498–1507, 2018.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," no. May 2014, pp. 248–255, 2010, doi: 10.1109/cvpr.2009.5206848.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[16] "(65) Tweets With Images Get 18% More Clicks, 89% More Favorites And 150% More Retweets | LinkedIn." https://www.linkedin.com/pulse/tweets-images-get-18-more-clicks-89-favorites-150-retweets-mansell/ (accessed Dec. 19, 2020).

[17] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Appl. Intell.*, 2021, doi: 10.1007/s10489-020-02057-9.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:49:54 UTC from IEEE Xplore. Restrictions apply.

arXiv:2104.05321v1 [cs.CL] 12 Apr 2021

Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of COVID-19 fake tweets

Rachit Bansal[1⋆] **William Scott Paka**[2⋆] **Nidhi**[3]
Shubhashis Sengupta[3] **Tanmoy Chakraborty**[2]

[1] Delhi Technological University, [2] IIIT-Delhi, [3] Accenture Labs, India
rachitbansal_2k18ee152@dtu.ac.in
{william18026, tanmoy}@iiitd.ac.in
{nidhi.sultan, shubhashis.sengupta}@accenture.com

Abstract. Fake tweets are observed to be ever-increasing, demanding immediate countermeasures to combat their spread. During COVID-19, tweets with misinformation should be flagged and neutralised in their early stages to mitigate the damages. Most of the existing methods for early detection of fake news assume to have *enough propagation information* for *large labelled tweets* – which may not be an ideal setting for cases like COVID-19 where both aspects are largely absent. In this work, we present ENDEMIC, a novel early detection model which leverages exogenous and endogenous signals related to tweets, while learning on limited labelled data. We first develop a novel dataset, called ECTF for early COVID-19 Twitter fake news, with additional behavioural test-sets to validate early detection. We build a heterogeneous graph with follower-followee, user-tweet, and tweet-retweet connections and train a graph embedding model to aggregate propagation information. Graph embeddings and contextual features constitute endogenous, while time-relative web-scraped information constitutes exogenous signals. ENDEMIC is trained in a semi-supervised fashion, overcoming the challenge of limited labelled data. We propose a co-attention mechanism to fuse signal representations optimally. Experimental results on ECTF, *PolitiFact*, and *GossipCop* show that ENDEMIC is highly reliable in detecting early fake tweets, outperforming nine state-of-the-art methods significantly.

1 Introduction

Over the past couple of years, several social networking platforms have seen drastic increase in the number of users and their activities online [1]. Due to the lockdown situations and work from home conditions during COVID-19 pandemic, the screen time on social media platforms is at an all time high. Twitter is one such micro-blogging platform where users share opinions and even rely on news updates. Twitter users exposed to unverified information and opinions of others often get influenced and become contributors to further spreading. The characteristics of fake news spreading faster farther and deeper than genuine news is well-studied [12]. Fake news on health during COVID-19 might endanger people's lives as a few of them call for action. Although our

⋆ Equal Contribution. The work was done when Rachit was an intern at IIIT-Delhi.

proposed method is highly generalised, we choose to specifically focus on the ongoing pandemic situation of COVID-19 as it is timely and needs scaleable solutions.

State-of-the-art fake news detection models for Twitter have proved useful when trained on a sufficient amounts of labelled data. Any emerging fake news in its initial stages could not be detected by such models due to the lack of corresponding labelled data. Moreover, these models tend to fail when the fake news is not represented largely in the training set. By the time the fake news is detected, it has spread and caused damages to a wide range of users. Detecting a fake news in an early stage of its spread gives us the advantage of flagging it early. A few state-of-the-art models for early fake news detection use propagation based methods [9], i.e., they are based on how a particular news event (fake/genuine) spreads (both wider and deeper); retweet chain in Twitter is one such example. These models work with the propagation chains and require sufficient historical data (retweet/reply cascade) of each tweet, which are very hard to collect within limited time.

In this work, we design ENDEMIC, a novel approach to detect fake news in its early stage. To this end, we developed an Early COVID-19 Twitter Fake news dataset, called ECTF with additional test set (early-test) for early detection. We collected vast amount of COVID-19 tweets and label them using trusted information, transformer models and human annotation. We finally curate a standard training dataset, composed of numerous rumour clusters, while simulating a scenario of emerging fake news through early-test set. Next, we extract exogenous signals in form of most suitable stances from relevant external web domains, relative to the time the tweet is posted. Unlike existing studies which extract propagation paths per tweet, we create a massive heterogeneous graph with follower-followee, tweet-retweet and tweet-user connections and obtain the node representations in an unsupervised manner. The graph embeddings (both users and tweets) constitute the major component of the endogenous signals (within Twitter). The time-variant contextual tweet and user features are used to provide additional context, and a few of them are masked to validate the model for early detection. Lastly, to overcome the challenge of limited labelled dataset, we setup the whole model in a semi-supervised fashion, learning in an adversarial setting to utilise the vast unlabelled data.

In Summary, fake news on Twitter is an inevitable threat especially during COVID-19 pandemic, where inaccurate or deviating medical information could be harmful. As a result, a timely model which can detect fake news in its early stages is important due to the current conditions of emerging fake news. We propose ENDEMIC, a semi-supervised co-attention network which utilises both exogenous and endogenous signals. Experimental results show that ENDEMIC outperforms nine state-of-the-art models in the task of early fake tweet detection. ENDEMIC produces 93.7% accuracy, on ECTF for fake tweet detection and 91.8% accuracy for early detection, outperforming baselines significantly. We also show the generalisation of ENDEMIC by showing its efficacy on two publicly available fake news datasets, *PolitiFact* and *GossipCop*; ENDEMIC achieves 91.2% and 84.7& accuracy on the two datasets, respectively.

In particular, our major contributions are as follows:

- We introduce ECTF, an Early COVID-19 Twitter Fake news dataset with additional early-test set to evaluate the performance of fake tweet detection models for early detection.
- As propagation paths for tweets might not be available in all scenarios, we build connections of follower-followee, tweet-retweet and user-tweet network and extract representations upon learning the graph, constituting of our endogenous signals.
- We use exogenous signals which informs the model on the realities of information available on the web at tweet time, and helps in learning from weak external signals.
- Adding an effort towards early detection, time variant features are masked at test time for further evaluation on early detection.
- We further show the generalisation of ENDEMIC by presenting its superior performance on two other general fake news datasets .

Reproducibility: The code and the datasets are public at: https://github.com/LCS2-IIITD/ENDEMIC.

2 Related Work

Our work focuses majorly on early detection of fake tweets. As our model involves techniques such as graphs and semi-supervised learning, we present related studies pertaining to our model.

Fake news detection. Proliferation of fake news over the Internet has given rise to many research, and hence there exist an abundance of literature. Early studies on fake news relied on linguistic features of texts to detect if the content is fake [3]. Wang et al. [27] used textual embeddings and weak labels with reinforcement learning to detect fake news. They leveraged an annotator component to obtain fresh and high-quality labelled samples to overcome the challenge of limited labelled data. Recent approaches have started exploring directed graph and Weisfeiler-Lehman Kernel based model for social media dataset that use similarity between different graph kernels [21]. A recent survey [31] shows that there are four approaches to detect fake news: (i) knowledge-based methods, where the content is verified with known facts, (ii) style-based methods, by analysing the writing style of the content, (iii) propagation methods, based on how a particular news event spreads, and (iv) source-based methods, by verification of credibility of sources. As studies show, each of these methods used individually is not enough to build an efficient classifier [16].

Semi-supervised learning. Semi-supervised models have been used often in the past to leverage vast unlabelled datasets in various fields. Helmstetter et al. [8] explored weakly supervised learning for fake news detection which automatically collects large scale noisy dataset to aid the classification task . Yu et al. [29] used constrained semi-supervised learning for social media spammer detection. Tensor embeddings are used to design a semi-supervised model for content based fake news detection [6]. A few studies leveraged variational auto-encoders in the form of sequence-to-sequence modelling on text classification and sequential labelling [7]. Nigam et al. [18] classified the text using a combination of Naive Bayes and Expectation Maximisation algorithms and demonstrated substantial performance improvements. Miyato et al. [15] utilised adversarial and virtual adversarial training to the text domain by applying perturbations

to the word embeddings. Chen et al. [4] introduced MixText that combines labelled, unlabelled and augmented data for the task of text classification.

Early detection. Early fake news detection methods detect fake news at an initial stage where the news has not yet been popularised. There exist very limited studies on early detection. Liu et al. [9] built an early detection model using propagation paths of news spreading as a multivariate time series and then training a recurrent and convolution classifier on user features. Rosenfeld et al. used graph kernels on Twitter cascades to capture intricate details of the data-set for fake news detection without feeding user identity and time for an early detection model [21]. Shu et al. used content engagement and cleaned labelled dataset for early detection with deep neural network [24].

3 Our Proposed Dataset: ECTF

As there is no publicly available COVID-19 Twitter dataset particularly for *early detection*, we attempt to develop our own dataset, ECTF, specifically crafting it for early detection. We expand on CTF, a general COVID-19 Twitter fake news dataset, proposed by Paka et al. [19]. CTF was formed using multiple sources- unlabelled Twitter datasets that are publicly released [2, 22, 26], hydration of tweets using predefined hashtags, and governmental health organisations and fact checking websites for verified news. They considered statements released by the latter to be true, and applied Sentence-BERT [20] and RoBERTa [11] to convert these tweets and verified news into contextual embeddings, pairwise cosine similarity is then computed to assign a label 'fake' or 'genuine. This way, CTF composes of $72,578$ labelled and $2,59,469$ unlabelled tweets, partially verified manually.

We took a sample of labelled and unlabelled tweets from CTF, forming our train set. A **'general-test'** set is created by randomly sampling from the remaining labelled and unlabelled tweets. The training dataset, which contains a wide variety of rumour clusters, is kept constant. We identified small rumour clusters in the labelled dataset and use those to form additional test set, called **'early-test'** set, to perform behavioural analysis of our algorithm for early detection. These small rumour clusters contain the fake news that are not popularised yet (having chance of getting popular), simulating early stages of fake news events. We extracted more tweets belonging to these rumour clusters, while keeping an upper limit on the time since they were posted. This ensures that the tweets belonging to this early-test set are in their early stages.

We refer to the complete dataset composed of the train and test sets as **ECTF**, **E**arly **C**OVID-19 **T**witter **F**ake news dataset.

4 Our Proposed Methodology: ENDEMIC

Here we present our model, called **ENDEMIC** (**E**xogenous and e**ND**ogenous signals with s**EMI**-supervised **C**o-attention network). Figure 1 represents the model architecture. Here we show various input features, each of which passes through separate modules before being concatenated. Our approach towards early detection relies on simulating the early stages of the news event testing with time invariant features.

4.1 Exogenous signals

Traditional supervised fake news detection methods are bounded by the knowledge present in the data they are trained on [30]. This makes them incapable of classifying

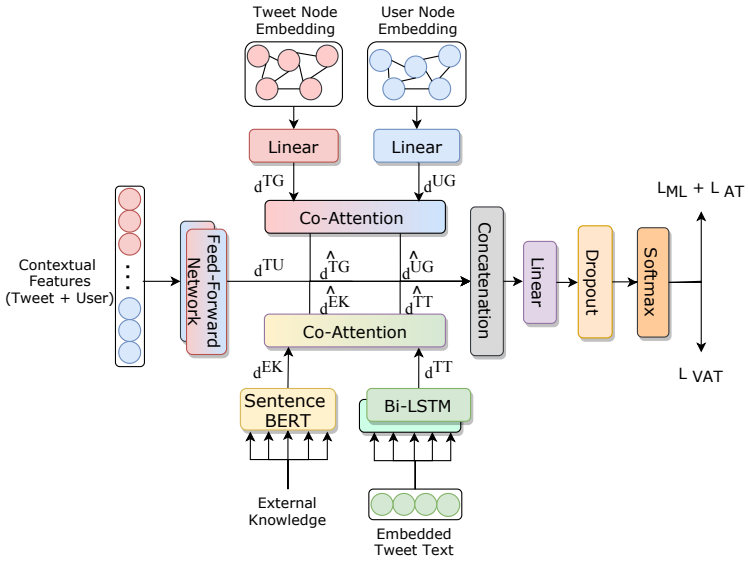


Fig. 1: A schematic architecture of ENDEMIC. The encoded interpolations of external knowledge (d^{EK}), tweet text (d^{TT}), contextual features (d^{TU}), and tweet (d^{TG}) and user node embeddings (d^{UG}) are shown. $\hat{\cdot}$ represents an output from co-attention.

tweets on information the model is not trained on. This is particularly challenging for domains where large amount of new theories and their consequent studies arrive rapidly within small duration. Very often, the general stance of such news also changes radically over time. Just as a human expert, classification models too need to be well-aware of current information in a domain in order to be reliable and efficient. To address this problem, we make use of exogenous signals through *external knowledge*.

We curate exogenous content using web-scraped articles from various sources. In order to simulate the early stages of detection, this knowledge base is collected differently for training and testing instances. We build the external knowledge for training as a static knowledge base composed of various sources for each input. We hypothesise that collecting this external knowledge in accordance with the time the post was made simulates an early detection scenario, consequently making the model adapt to such scenarios better where the information about an emerging fake news will be limited. Using this, the model learns the weak signals that are present to perform a classification. In case of testing, the external knowledge is scraped for every test instance at the time of testing itself. The dynamic nature of building the external knowledge base during testing ensures that the model makes use of latest sources to make the prediction.

We perform a web scraping with tweets as queries using Google Web API. Since a large amount of information in form of web pages are available per query, and the content per page could also be excessively large, selecting the right content is vital. We do so by firstly tokenizing each web page $x^{EK,i}$ into sentences. The j^{th} sentence of $x^{EK,i}$, denoted by $x_j^{EK,i}$, is encoded using Sentence-BERT [20] to obtain its contextual representation, $d_j^{EK,i} \in \mathbb{R}^K$, where K is the dimension of the encoding. This represen-

tation is then compared with the representation of the tweet text, $x^{TT} \rightarrow d_x^{TT} \in \mathbb{R}^K$, encoded using the same Sentence-BERT model. This comparison is made using cosine similarity. If the cosine similarity between these two encodings, $cos(d_j^{EK,i}||d_x^{TT})$, is greater than a threshold ϵ (set as 0.8), then the sentence $x_j^{EK,i}$ is added to the set of input external knowledge for that particular tweet. The same representation, $d_j^{EK,i}$ is used as the corresponding input to the model for all sentences belonging to the tweet. This process is done for the entire set of input queries, until we obtain 50 such sentences for each, with the amount of phrases per web-source being limited to 10. Thus, the net external knowledge input to `ENDEMIC` during training is obtained by concatenating the encoding for each input in 2D fashion and is given by $d^{EK} \in \mathbb{R}^{n \times 50 \times K}$, where n is the input size. For most of our experiments, we keep $K = 512$.

4.2 Endogenous signals

Input tweet embedding. The original tweet text x_i^{TT} of sequence length N (say) is first represented by a one-hot vector using a vocabulary of size V. A word embedding look-up table transforms these one-hot vectors into a dense tensor. This embedding vector is further encoded using a Bidirectional LSTM. A final state output $d \in \mathbb{R}^{K/2}$ is obtained at both the forward and backward layers, which are then concatenated to give a 2D vector corresponding to each text input. The final representation of the tweet text input to the model is, thus, $d^{TT} \in \mathbb{R}^{n \times N \times K}$.

Graph embeddings. Extracting the propagation chain of retweets has been proven effective by the existing early detection models [9, 14]. However, these methods limit the training as for each training or test sample, the entire set of retweet chains needs to be extracted which is often computationally expensive. Here we build a heterogeneous graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ as follows: the set of nodes V can be users or tweets, and edges E are formed in the following ways: two users are connected via follower-followee link, a user is connected to her posted tweet, two tweets are connected if one is a retweet of another. We keep $\mathcal{G}$ as undirected intentionally ignoring the direction of some edge types (follower-followee) in order to maintain uniformity. The formed heterogeneous graph contains around $51M$ nodes and $70M$ edges. Such huge connections, when added into one graph, form many disconnected clusters. In our graph, we observe one giant cluster with almost millions of nodes and edges, which stands dominating compared to other small (and disconnected) clusters.

We obtain the embedding of the graph using GraphSAGE [17] in an unsupervised fashion due to its capability to scale and learn large graphs easily. We label each node with its characteristics such as parent, tweet, retweet, user, fake tweet. The generalisability of GraphSAGE helps in extracting the embeddings of unseen users and tweets. We use a teleportation probability of 0.3 for the graph to randomly jump across nodes, which helps with clusters having less number of connections or being disconnected from the rest. In this work, we show that using the embeddings of both users and tweets in combination with co-attention leads to better supervision and learning of the model.

We represent the tweet and user graph embeddings as d^{TG} and $d^{UG} \in \mathbb{R}^{n \times G}$ respectively, where G represents the embedding dimension, and is kept as $G = 768$, for most of our experiments.

Contextual features. Social media platforms like Twitter offer a variety of additional features that can play a crucial role in identifying the general sentiment and stance on a post by the users. Moreover, some features of user could also be used as indicative measures of their social role and responsibility, in general. Therefore, we use a variety of such tweet and user features to provide additional context regarding the input tweet and the corresponding users who interact with it. Some of the tweet features used are *number of favourites, number of retweets, PageRank reliability score of domains mentioned in the tweet*, *text harmonic sentiment*[1], etc. Some of the user features include *follower and followee counts*, *verified status*, and *number of tweets made by the user*.

Note that majority of these features continually change over time and can be regarded as time-variant and accordingly, the inferences drawn also would change over time. Therefore, for early detection, it is vital not to rely too heavily on such features. For instance, the number of likes and retweets for a tweet changes over time. Similarly, such additional features for a new user cannot be expected to give a proper indicative measure of a user's tendency of believing, spreading, or curbing misinformation. And masking the time-variance during evaluation is better explained in Section 5.3.

Throughout this study, we represent these contextual tweet and user features as $x^{TF} \in \mathbb{R}^{n \times N_{TF}}$ and $x^{TF} \in \mathbb{R}^{n \times N_{UF}}$, respectively, where N_{TF} and N_{UF} indicate the number of such features. As shown in Fig. 1, these input features are concatenated and passed across a common feed-forward network (FFN), which interpolates $x_i^{TF} \oplus x_i^{UF} \in \mathbb{R}^{1 \times (N_{TF}+N_{UF})}$ to $d_i^{TU} \in \mathbb{R}^C$, where C is the output dimension of FFN.

4.3 Connecting components and training

Co-attention. In order to jointly attend and reason about various interpolated inputs, we use the parallel co-attention mechanism [13] (Figure 2). As shown in Figure 1, this is done at two places of ENDEMIC, namely, to co-attend between external knowledge $d_i^{EK} \in \mathbb{R}^{50 \times K}$ and tweet text $d^{TT} \in \mathbb{R}^{N \times K}$ in 2D, and tweet $d^{TG} \in \mathbb{R}^{1 \times G}$ and user $d^{UG} \in \mathbb{R}^{1 \times G}$ graph embeddings in 1D. The same process is followed for two; therefore, to unify the notation, we use $d^A \in \mathbb{R}^{X \times Z}$ and $d^B \in \mathbb{R}^{Y \times Z}$ to explain the mechanism used.

Firstly, an affinity matrix $C \in \mathbb{R}^{Y \times X}$ is obtained as, $C = tanh(d^A W_b {d^B}^\top)$. Here $W_b \in \mathbb{R}^{Z,Z}$ represents the learnable weight matrix. Further the corresponding attention maps between A and B are calculated as follows:

$$H^A = tanh(W_A {d^A}^\top + (W_B {d^B}^\top)C), \quad H^B = tanh(W_B {d^B}^\top + (W_A {d^A}^\top)C^\top)$$

where, W_A, $W_B \in \mathbb{R}^{k \times Z}$ again represent the learnable weight matrices. Further, to compute the attention probabilities of each element in A with each element of B, we use,

$$a^A = Softmax({w_{hA}}^\top H^A), \quad a^B = Softmax({w_{hB}}^\top H^B)$$

where, $w_{hA}, w_{hB} \in \mathbb{R}^k$ represent the weight parameters, while $a^A \in \mathbb{R}^X$ and $a^B \in \mathbb{R}^Y$ represent the resultant attention probabilities. Finally, the net attention vectors between A and B are computed as a weighted sum of the corresponding features, i.e.,

[1] obtained from the textblob tool in Python.

$$\widehat{A} = \sum_{i=1}^{X} a_i^A d_i^A, \quad \widehat{B} = \sum_{i=1}^{Y} a_i^B d_i^B \tag{1}$$

$\widehat{A}$ and $\widehat{B} \in \mathbb{R}^{1\times Z}$ are the learned feature vectors obtained through co-attention. This, for instance, represents how each representation in the tweet graph embeddings attends to the user graph embeddings, when A represents tweet graph embeddings (TG), and B represents user graph embeddings (UG).

The interpolations of the various inputs, obtained through the separate smaller pipelines are combined using a unified head pipeline in the model architecture. Considering a single input, firstly, the representations from the two co-attention layers, $\widehat{d^{EK}}$ and $\widehat{d^{TT}} \in \mathbb{R}^{1\times K}$, and $\widehat{d^{TG}}$ & $\widehat{d^{UG}} \in \mathbb{R}^{1\times G}$, are concatenated along with $d^{TU} \in \mathbb{R}^{1\times C}$.

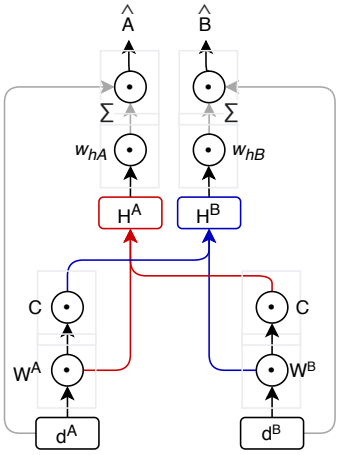


Fig. 2: The co-attention mechanism.

$$d = \widehat{d^{EK}} \oplus \widehat{d^{TT}} \oplus \widehat{d^{TG}} \oplus \widehat{d^{UG}} \oplus d^{TU}$$

This net representation, $d \in \mathbb{R}^{2K+2G+C}$, then passes onto a dropout ($p_{drop} = 0.2$) regularised feed-forward layer with an output size of 2, and finally a Softmax function, to give the probability per output class.

Training. In order to overcome the limitations produced by scarce labelled data, we use the Virtual Adversarial Loss (VAT) across both labelled and unlabelled data. VAT introduces a small perturbation in the input embedding and computes the loss using weak labels. Maximum Likelihood (ML) loss and Adversarial Training (AT) losses are further used to train ENDEMIC on the labelled data [15]. These additional losses allow ENDEMIC to be more robust, and the abundant unlabelled data being used this way allow it to form better understanding and representation of the domain-specific texts.

5 Experimental Setup and Results

Here we present our comparative analysis by empirically comparing ENDEMIC with state-of-the-art models, including techniques for general fake news detection, early detection, and text classification.

5.1 Baseline methods

We consider nine baselines as follows. Liu and Wu [9] introduced a GRU and CNN-based model which relies on the propagation path (PP) of a news source to classify its veracity, and is termed as **PPC** (Propagation Path Classification). **FNED** [10] makes use of a PU-learning framework to perform weak classification, and is claimed to be suitable in data scarce settings with large amount of unlabelled data, as is the scenario we deal with in this study. Further, Shu et al. [25] proposed to learn from Multiple-sources of weak Social Supervision (**MWSS**), on top of a classifier such as CNN and

Model	Feature used				Performance			
	TG	UG	EK	UL	Accuracy	Precision	Recall	F1 Score
HAN					0.865	0.685	0.865	0.762
MixText				✓	0.875	0.835	0.843	0.847
GCAN		✓			0.852	0.817	0.820	0.813
CSI	✓		✓		0.873	0.805	0.915	0.854
dEFEND	✓		✓		0.890	0.830	0.891	0.862
CNN-MWSS				✓	0.805	0.785	0.810	0.790
RoBERTa-MWSS				✓	0.825	0.815	0.852	0.840
FNED				✓	0.911	0.907	0.920	0.915
PPC	✓				0.861	0.860	0.832	0.845
ENDEMIC	✓	✓	✓	✓	**0.937**	**0.922**	**0.943**	**0.932**

Table 1: Features used and performance comparison on **general-test** dataset of ECTF (TG: Tweet Graph, UG: User Graph, EK: External Knowledge, UL: Unlabelled Data).

RoBERTa [11]. They too relied on weak labelling, but constructed them through social engagements. Both **dEFEND** and **CSI** deployed an RNN-based framework for classifying news; the former makes use of user comments with co-attention, while the latter relies on a three-step process of *capturing, scoring* and *integrating* the general user response. Furthermore, **GCAN** [14] presents a dual co-attention mechanism over source and retweet tweet representations, relying on interaction and propagation for detecting fake news. Finally, we also make some interesting observations by employing text classification models, namely **MixText** [4] and **HAN** [28], as additional baselines.

5.2 Evaluating on general-test set

Table 1 shows the performance comparison of ENDEMIC compared to the baselines on the **general-test** set of ECTF. We observe that all the features (graph based, external knowledge and unlabelled data) play a major role in determining the corresponding performance of detecting fake tweets. While general fake news detection models like dEFEND and CSI strive to integrate more such features, early detection models like FNED and PPC tend to rely more on the time-invariant features like text alone. The effect produced by the absence of one over the other is apparent from Table 1. In general, ENDEMIC shows a benchmark performance of 0.937 accuracy, outperforming all the baselines across all the evaluation measures significantly.

5.3 Evaluating on early detection

Early-test. Table 2 shows the comparative results on the specially curated evaluation set (**early-test**) for early detection of fake news. Even though the change in accuracy (ΔAcc) of ENDEMIC as compared to its performance on general fake news (as shown in Table 1) is not least among the rest of the models, it can be seen that it still comfortably outperforms all other baselines. Interestingly, the general purpose text classifiers, which are not particularly designed for fake news detection, show a relatively lesser ΔAcc, while dEFEND [5] suffers from the largest difference of 9%. We attribute this to the heavy reliance of these models on time-variant features which provide critically less context in early stages. At the same time, as shown in Table 1, only text-based and closely related non-time-variant features are not enough to reliably detect fake news. Thus, the set of input features used by ENDEMIC optimises the trade-off between reliability and time-variance.

Model	Performance				
	Acc.	Prec.	Rec.	F1	ΔAcc
HAN	0.850	0.692	0.841	0.751	1.5%
MixText	0.865	0.832	0.851	0.844	1.0%
GCAN	0.809	0.800	0.785	0.795	4.3%
CSI	0.808	0.821	0.795	0.815	6.5%
dEFEND	0.800	0.790	0.815	0.790	9.0%
MWSS	0.798	0.800	0.812	0.805	2.7%
FNED	0.901	0.887	0.913	0.895	1.0%
PPC	0.831	0.805	0.797	0.810	3%
ENDEMIC	**0.918**	**0.910**	**0.923**	**0.920**	1.9%

Table 2: Comparing model performance on **early-test** set of ECTF.

Model	Time-variant Features	Performance				ΔAcc
		Acc.	Prec.	Rec.	F1	
HAN	–	0.858	0.705	0.850	0.755	1.2%
MixText	–	0.860	0.815	0.833	0.820	1.0&
GCAN	Tweet Propagation	0.792	0.812	0.775	0.805	6.0%
CSI	User Response	0.783	0.701	0.745	0.725	9.0%
dEFEND	User Comments	0.768	0.750	0.804	0.774	12.2%
MWSS	–	0.810	0.800	0.835	0.844	1.5%
FNED	–	0.892	0.890	0.925	0.905	1.9%
PPC	Propagation Paths	0.855	0.810	0.790	0.805	0.6%
ENDEMIC	Contextual Features	**0.928**	**0.920**	**0.930**	**0.927**	0.9%

Table 3: Performance comparison for **mask-detect** on ECTF. The masked features and corresponding metric scores are shown.

Masking time variance. To further verify our model's prowess in detecting early fake news, we introduce a unique masking approach to be applied upon the early-test evaluation set. For this technique, the tweets belonging to the smaller rumour clusters (as defined in Section 3) are further simulated as actual early tweets for a model, by masking all time-variant features used by the model. We call this approach **'mask-detect'**.

Most of the existing techniques for fake news detection make use of some particular input features which are time-variant. In case of ENDEMIC, these are the additional contextual tweet features and user features, which rapidly change over time. When such features are masked, there is effectively no way to distinguish an old tweet from a new one. Therefore, we perform mask-detect by replacing the numerical values of the relevant time-variant features with a common masking token. These features are different for each model. Therefore, we first identify such features and then perform the masking. Table 3 shows the features masked for the various fake news detection models and their corresponding performance.

5.4 Evaluating on general-domain fake news

Although we tested our model on COVID-19, ENDEMIC is highly generalised to other domains. And to prove the generalisability, we also evaluate ENDEMIC on the FakeNewsNet [23] datasets – *PolitiFact* and *GossipCop*. Both of these datasets are obtained from the respective sources and contain labelled textual news content along with the social context. Evaluating on these datasets allows us to validate ENDEMIC's generalisation across domains outside of COVID-19 and beyond ECTF.

Table 4 shows that ENDEMIC outperforms all baselines with more than 0.5% accuracy on *PolitiFact* and 1.5% on *GossipCop*. The performance of ENDEMIC is highly comparable compared to other baselines w.r.t. other evaluation metrics. These results clearly establish that ENDEMIC generalises well on any fake news domain and dataset.

6 Conclusion

In this work, we introduced the task of early detection of COVID-19 fake tweets. We developed a COVID-19 dataset with additional test set to evaluate models for early

Model	PolitiFact				GossipCop			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
HAN	0.863	0.730	0.855	0.790	0.800	0.685	0.775	0.770
MixText	0.870	0.840	0.860	0.853	0.785	0.805	0.795	0.813
GCAN	0.867	0.855	0.880	0.875	0.790	0.805	0.795	0.803
CSI	0.827	0.847	0.897	0.871	0.772	0.732	0.638	0.682
dEFEND	0.904	0.902	**0.956**	**0.928**	0.808	0.729	0.782	0.755
CNN-MWSS	0.823	0.811	0.830	0.820	0.770	0.793	0.815	0.795
RoBERTa-MWSS	0.825	0.833	0.795	0.805	0.803	0.815	0.810	0.807
FNED	0.907	**0.910**	0.922	0.916	0.832	0.830	0.825	0.830
PPC	0.885	0.887	0.870	0.880	0.791	0.805	0.832	0.811
ENDEMIC	**0.912**	**0.910**	0.904	0.915	**0.847**	**0.835**	**0.822**	**0.840**

Table 4: Performance comparison on two general-domain datasets to check the generalisability of the models.

detection. We took measures to simulate the early stages of the fake news events by extracting the external knowledge relative to the time the tweet is posted. Our proposed model, ENDEMIC overcomes the challenges of limited dataset in a semi-supervised fashion. We adhere to co-attention as a better information fusion tested on time invariant features. ENDEMIC outperformed nine baselines in both the tasks of general and early-stage fake news detection. Experimental results compared to nine state-of-the-art models show that ENDEMIC is highly capable for early detection task. We also showed the generalisability of ENDEMIC on two other publicly available fake news dataset which are not specific to COVID-19.

Acknowledgements

The work was partially supported by Accenture Labs, SPARC (MHRD) and CAI, IIIT-Delhi. T. Chakraborty would like to thank the support of the Ramanujan Fellowship.

References

1. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013)
2. Carlson: Coronavirus Tweets, Tweets (json) for Coronavirus on Kaggle. https://www.kaggle.com/carlsonhoo/coronavirus-tweets (2020), online; accessed 2020
3. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: WWW. pp. 675–684 (2011)
4. Chen, J., Yang, Z., Yang, D.: Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. arXiv preprint arXiv:2004.12239 (2020)
5. Cui, L., Shu, K., Wang, S., Lee, D., Liu, H.: defend: A system for explainable fake news detection. In: CIKM. pp. 2961–2964 (2019)
6. Guacho, G.B., Abdali, S., Shah, N., Papalexakis, E.E.: Semi-supervised content-based detection of misinformation via tensor embeddings. In: ASONAM. pp. 322–325. IEEE (2018)
7. Gururangan, S., Dang, T., Card, D., Smith, N.A.: Variational pretraining for semi-supervised text classification. arXiv preprint arXiv:1906.02242 (2019)
8. Helmstetter, S., Paulheim, H.: Weakly supervised learning for fake news detection on twitter. In: ASONAM. pp. 274–277. IEEE (2018)
9. Liu, Y., Wu, Y.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: AAAI. pp. 354–361 (2018)

10. Liu, Y., Wu, Y.F.B.: Fned: A deep network for fake news early detection on social media. ACM Trans. Inf. Syst. **38**(3) (May 2020)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
12. Lohr, S.: It's true: False news spreads faster and wider. and humans are to blame. The New York Times **8** (2018)
13. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) NIPS. vol. 29, pp. 289–297. Curran Associates, Inc. (2016)
14. Lu, Y.J., Li, C.T.: GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In: ACL. pp. 505–514. Online (Jul 2020)
15. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725 (2016)
16. Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M.: Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673 (2019)
17. Niepert, M., Ahmed, M., Kutzkov, K.: Learning convolutional neural networks for graphs. In: ICML. pp. 2014–2023 (2016)
18. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Machine learning **39**(2-3), 103–134 (2000)
19. Paka, W.S., Bansal, R., Kaushik, A., Sengupta, S., Chakraborty, T.: Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. arXiv preprint arXiv:2102.08924 (2021)
20. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: EMNLP-IJCNLP. pp. 3982–3992. Hong Kong, China (2019)
21. Rosenfeld, N., Szanto, A., Parkes, D.C.: A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In: The WebConf. pp. 1018–1028 (2020)
22. Shane Smith: Coronavirus (covid19) Tweets - early April. `https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april` (2020), online; accessed 2020
23. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 **8** (2018)
24. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. In: MIPR. pp. 430–435. IEEE (2018)
25. Shu, K., Zheng, G., Li, Y., Mukherjee, S., Awadallah, A.H., Ruston, S., Liu, H.: Leveraging multi-source weak social supervision for early detection of fake news. arXiv preprint arXiv:2004.01732 (2020)
26. Sven Celin: COVID-19 tweets afternoon 31.03.2020. `https://www.kaggle.com/svencelin/covid19-tweets-afternoon-31032020` (2020), online; accessed 2020
27. Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., Gao, J.: Weak supervision for fake news detection via reinforcement learning. In: AAAI. vol. 34, pp. 516–523 (2020)
28. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: NAACL. pp. 1480–1489 (2016)
29. Yu, D., Chen, N., Jiang, F., Fu, B., Qin, A.: Constrained nmf-based semi-supervised learning for social media spammer detection. Knowledge-Based Systems **125**, 64–73 (2017)
30. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: A theory-driven model. Digital Threats: Research and Practice **1**(2) (Jun 2020)
31. Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM CSUR **53**(5), 1–40 (2020)

VIA MEDICA

LETTER TO THE EDITOR

Int Marit Health
2021; 72, 1: 82–83
10.5603/IMH.2021.0014
www.intmarhealth.pl
Copyright © 2021 PSMTTM
ISSN 1641-9251
eISSN 2081-3252

Consequences of COVID-19 pandemic on global maritime trade industry

Shailender Kumar, Aman Jolly

Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

The coronavirus disease 2019 (COVID-19) pandemic caused by the recently discovered novel coronavirus has caused confusion and hysteria around the world, leading to daily activities such as work, walking outside, or leaving home in some countries, to cease. The maritime sector is also one of the areas affected by the pandemic. This crisis has forced the shipping and marine industries to go through the toughest times in the present day, as the work in these fields has been shut down for the protection and prevention of the COVID-19 outbreak.

Though COVID-19 has placed the world in a massive humanitarian crisis, the pillars of our globalised economy have also been destroyed by it. Perhaps any country, small or large, has been infected by this virus. This pandemic has led to profound disruptions in supply and demand, thereby upsetting world commerce and causing serious economic pain. Depending on how the situation unfolds, analysts at the World Trade Organisation plan world merchandise trade volumes to decline by 13% to 32% (the decline was 12% at the height of the 2008 crisis) [1].

This global pandemic has provided a major negative impact for not just the ports that are located in China (where the outbreak is said to arise) but also the ports for the shipping and maritime industry at the global level are equally affected.

Some of the maritime and shipping industry's problems include border restrictions for airlines and port shutdown, reduced demand for freight, disputes in laytime arbitration, disputes between owners and charters of such vessels due to lack of time and resources, bankruptcy due to reduced demand and the failure to control the company's finances during this era of decreased demand for shipping and freight [2].

In pandemic, the quarantine conditions and prohibitions, for workers crossing borders in certain countries that led to crew shifts and repatriations for mariners, especially those operating on cruise ships and yachts, are critical issues that have a negative impact on the global maritime industry [3]. The study showed a large majority who preferred shore leave previously has decreased dramatically during the pandemic. The effect on the job and well-being of seaman was exposed when only one-fifth of them voluntarily negotiated to renew the deal. In the outbreak with many parameters, this analysis reported occurrence rates around 6 months: overdue resettlements (21.44%) including crew with contract extensions (12.48%), the crew with a completed contract pending repatriation (8.96%) and crew who had crossed 12 months of continuous operation (0.82%). Compensation, if given, is insufficient and has had the most effect on ratings. The survey also revealed the deprivation of medical assistance [4].

For the global maritime shipping, the pandemic has sent shockwaves and laid the ground for a changing industry and related supply chains management. United Nations Conference on Trade and Development (UNCTAD) predicts the momentum of economic growth to return in 2021.

But because of the enormous damage triggered by COVID-19 outbreak, the overall maritime trade will fall by 4.1% in 2020 worldwide. This prediction is done by UNCTAD in its Maritime Transport Assessment 2020 [5]. Figure 1 explains all the COVID-19 pandemic factors that are interconnected with each other that have resulted in the decline of maritime trade [6].

The study warns that new pandemic waves may cause a steeper downturn, further affecting supply chains and economies. The report claims that the pandemic caused a massive hindrance across supply chains, transport networks and ports, which led to a decrease in cargo volumes and stifled the prospects for expansion [7].

"The short-term prospect for maritime trade is poor", according to the report [5]. Moreover, the long-term consequences of a pandemic along with the timing and extent of the industry's recovery are difficult to predict [5].

Mr. Aman Jolly, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India, e-mail: aman.jolly1994@gmail.com

This article is available in open access under Creative Common Attribution-Non-Commercial-No Derivatives 4.0 International (CC BY-NC-ND 4.0) license, allowing to download articles and share them with others as long as they credit the authors and the publisher, but without permission to change them in any way or use them commercially.

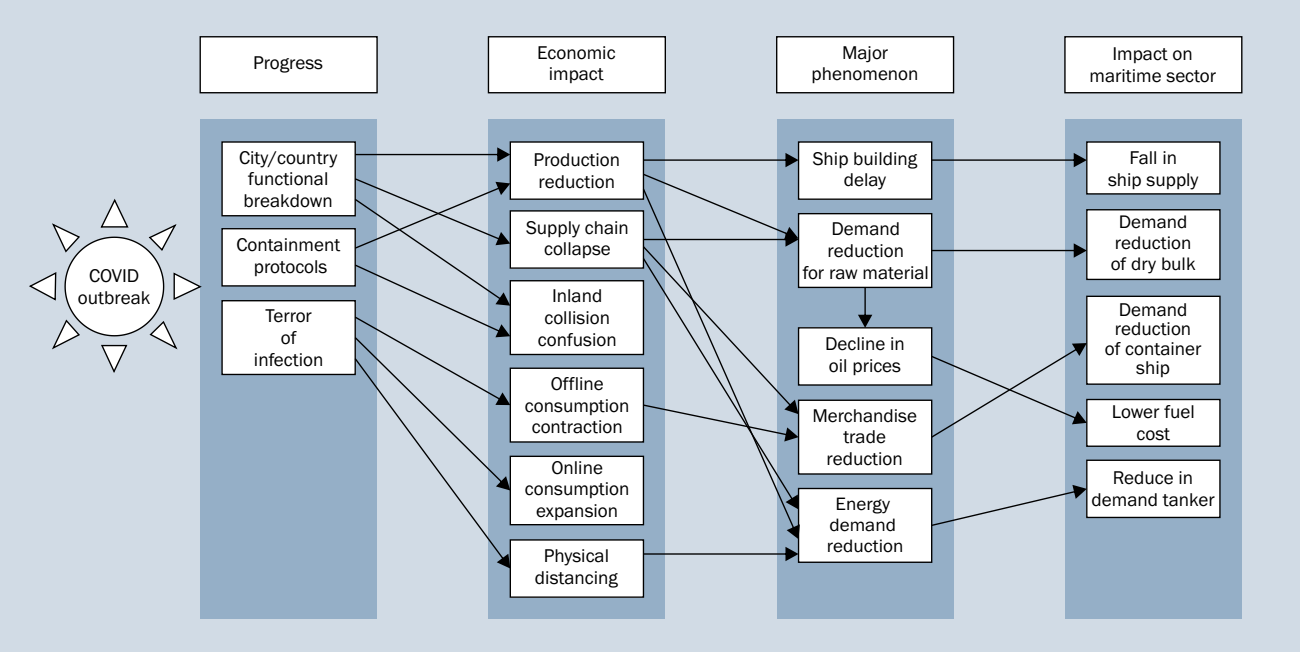


Figure 1. COVID-19 factors for decline in maritime trade [6]

"As a crucial enabler of the smooth running of international supply chains, the global shipping sector will be at the forefront of efforts for a sustainable recovery" [6]. UNCTAD Secretary-Kituyi said "The industry must be a key stakeholder to help adjust the logistics of 'just-in-time productivity' to 'just-in-case' preparedness", UNCTAD has made the prediction that maritime trade development will recover its losses and increase by 4.8% in 2021 by assuming that global economic performance is recovering at a steady pace. However, they stress that there is a strong need to brace for transition in the maritime transport industry and be well-positioned for a changed post-pandemic environment [5].

In the near future as well, it is expected that COVID-19 will have a major impact not only in the maritime and shipping industries but also in the domain of the cruise ship industry, therefore reflecting its repercussions on world trade and tourism. It has been catastrophic in its effects on seaman worldwide, be it in the marine industries or port industries. The boundaries of collaboration have been demonstrated by government actions focused on their concerns about the welfare of people. By citing health and safety concerns, many countries have closed their naval facilities for seaman; however, seafarers are the forgotten heroes of the pandemic. The major takeaway from this scenario is that the seafarers, who are not in sight or mind of anyone except their relatives, have to be better recognised as we are highly dependent on them for goods which are needed every day. They are the ones that make world trade possible.

REFERENCES

1. Doumbia-Henry C. Shipping and COVID-19: protecting seafarers as frontline workers. WMU Journal of Maritime Affairs. 2020; 19(3): 279–293, doi: 10.1007/s13437-020-00217-9.
2. Hebbar AA, Mukesh N. COVID-19 and seafarers' rights to shore leave, repatriation and medical assistance: a pilot study. Int Marit Health. 2020; 71(4): 217–228, doi: 10.5603/IMH.2020.0040, indexed in Pubmed: 33394486.
3. STA Law Firm. Impact Of COVID-19 On The Shipping And Maritime Industry. June 2020. https://www.mondaq.com/marine-shipping/958770/impact-of-covid-19-on-the- shipping-and-maritime-industry (02/25/2020).
4. WTO. Trade set to plunge as COVID-19 pandemic upends global economy. Apr. 2020. https://www.wto.org/english/news_e/pres20_e/pr855_e.htm (02/25/2021).
5. UNCTAD. Review of Maritime Transport 2020 (Executive summary). Nov. 2020.
6. UNESCAP. The Impact and Policy Responses for COVID-19 in Asia and the Pacific. Mar.2020.
7. Yazir D, Şahin B, Yip TL, et al. Effects of COVID-19 on maritime industry: a review. Int Marit Health. 2020; 71(4): 253–264, doi: 10.5603/IMH.2020.0044, indexed in Pubmed: 33394490.

Electrical Engineering
https://doi.org/10.1007/s00202-021-01273-x

ORIGINAL PAPER

Control of Grid Integrated Photovoltaic system using new Variable Step Size Least Mean Square adaptive filter

Avdhesh kumar[1] · **Rachana Garg**[1] · **Priya Mahajan**[1]

Received: 2 August 2020 / Accepted: 18 March 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
In this paper, a new variable step size least mean square adaptive filter-based control algorithm is proposed to generate reference current for voltage source converter of a grid-integrated PV system. The proposed control generates reference current by extracting the fundamental active and reactive component from load current with fast convergence. The proposed control offers better initial transients and dynamic performance than conventional LMS and synchronous reference frame-based control. Grid-integrated PV system is modelled in MATLAB/Simulink and is tested on different loading condition viz. nonlinear load, unbalanced load and variable load. Further, to validate the efficacy of the proposed control, it is also tested in real time on prototype hardware in the laboratory. It has been observed that proposed control provides compensation of harmonic and reactive power of the local load. Distortion at grid side of the system is within the limits as per IEEE-519 standard.

Keywords Variable step size least mean square (VSSLMS) · Power quality (PQ) · Voltage source converter (VSC)

1 Introduction

The nonlinear loads, viz. arc furnaces and variable frequency drives, etc., at consumer side have deteriorated the quality of power in utility grid. As a consequence, distribution system suffers from low power factor, unbalance current, poor voltage regulation, increase in neutral current, transients, harmonic distortion and flicker in distribution grid. To improve these power quality (PQ) issues, photovoltaic (PV) system has been integrated to a distribution grid. The main features of a PV system integrated to grid are: maintaining grid current sinusoidal by compensating the harmonics, neutralizing the effect of unbalance loads and maintaining power factor of the grid at unity [1–5]. To ensure the effectiveness of the proposed control, IEEE-519 standards are taken in to the consideration [6, 7]. As per IEEE-519, less than 5% total harmonic distortion (THD) in grid current and less than ± 5% voltage fluctuations are allowable at PCC [7, 8]. Therefore, to maintain PQ within reasonable limits of IEEE requirements, it is necessary to establish an effective control of grid-integrated PV system with improved performance. PV system operated to get maximum efficiency using various algorithms [9–11]. Various conventional control algorithms of MPPT such as incremental conductance (INC), adaptive step size MPPT with single sensor and perturb & observe (P&O) are presented.

Many conventional control techniques viz. synchronous reference frame (SRF), unit template, least mean square (LMS) and least mean fourth (LMF) have been implemented [4, 7–17]. A phase locked loop (PLL) is required to transform three phases to synchronous frame in SRF-based control, but sometimes it shows computational delay and also it has to be tuned preceding to its operation. In the literature [18], PLL performance has been found to be easily influenced by distortions and noise and to overcome these drawbacks SOGI PLL has been proposed, but it also suffers

from poor dynamic response and uses generalized integrator thus requiring tuning of gains. Reference current is estimated using the voltage and current of the PCCC in the implementation of IRPT; therefore, fluctuation in voltage will reflect in the reference current. LMS and LMF suffer from fluctuations and less accuracy in estimating mean square error (MSE). In the past few decades, adaptive control algorithms have been proposed for many applications. Among them, the least mean square (LMS) algorithm is a distinguished one, whose stability and optimality are fully discussed in many literatures [13–20]. Due to its simplicity,

✉ Avdhesh kumar
iesavd@gmail.com

[1] Electrical Engineering Department, Delhi Technological University, Delhi, India

Published online: 15 April 2021

conventional LMS adaptive control has been implemented widely, but the performance is often unsatisfactory because of poor dynamic performance, due to compromise between tracking capability and accuracy in fixed step size.

In the conventional LMS adaptive control algorithm, estimation of mean square of error (MSE) is accomplished by fixed step size. Fixed step size (μ) is a positive scalar constant that governs the rate of convergence and stability of the algorithm. Step size (μ) value must be small and positive to maintain the stability of the adaptive filter. A trade-off point between the convergence rate and SME is an issue while choosing the value of the step size. The step size, used in most of these adaptive algorithms, is the function of a filtering error signal that is not equal to zero when the filter converges [21, 22, 23]. However, the characteristics of the error are responsible for deciding their step size. For further improvement, an improved new variable step size LMS (VSSLMS) control is proposed and implemented for grid integration of solar PV system in MATLAB/Simulink.

In this paper, a new variable step size LMS (VSSLMS) adaptive filter is derived by regulating the step size. The step size (μ) in proposed control merges estimation error, small positive constants α and β, respectively. Step size (μ) helps to reduce instantaneous estimation errors thus accelerating the convergence rate of the algorithm [21]. The purpose of the grid-integrated PV system using proposed control is to accomplish the load necessity, then after filling the load necessity, remaining power is supplied to the grid. Whenever, generated power is not enough and then load necessity is fulfilled by receiving extra essential power from the grid. The proposed algorithm successfully improves the initial transients and dynamic performance of the system with the faster convergence rate; hence, it decreases the error faster compared to the LMS and SRF. LMS and SRF controls show poor/slower convergence rate which reduces dynamic performance of the system. These challenges can be improved by proposed control. The proposed VSSLMS control algorithm is also implemented and tested in real time on prototype hardware in the laboratory.

The prime contributions of the paper are given below:

1. Proposed VSSLMS algorithm extracts active and reactive fundamental component from load current with a faster rate of convergence; hence, dynamic performance get improved.
2. PF of the grid has been improved by compensating the reactive power.
3. Maintains grid current to be sinusoidal and balanced during the nonlinear and unbalanced load, respectively, thus improves PQ.
4. It is more efficient than conventional LMS and SRF control due to faster tracking abilities during dynamic loading condition.
5. The proposed control is efficient under adverse scenario of the grid viz. unbalance and distorted grid voltage.

2 System description

Figure 1 demonstrates the proposed grid-integrated PV system. Proposed system consists of PV array, DC-DC boost converter, MPPT, VSC, interfacing inductors and various loads. To get the maximum efficiency from PV array, P&O MPPT is employed. Grid integration of PV system is accomplished using proposed VSSLMS adaptive control algorithm of VSC. Efficacy of the proposed algorithm is validated in MATLAB/Simulink and experimentally on a prototype hardware in the laboratory.

3 VSC Control technique

The proposed control algorithm is depicted in Fig. 2. Proposed control algorithm extracts the fundamental components from load current to estimate reference current. It improves the PQ of the grid by compensating harmonics. Performance of the system is analyzed in power factor correction (PFC) mode of operation. Grid-integrated PV system is modelled in MATLAB/Simulink2016 (a), and various system parameters are given in the appendix.

3.1 VSC switching pulse generation

3.1.1 Estimation of in-phase and orthogonal unit templates of voltages

The peak amplitude of voltage ($\mathcal{V}_t$) of grid voltage, can be calculated by:

$$\mathcal{V}_t = \sqrt{\frac{2}{3}\left(\mathcal{V}_{sa}^2 + \mathcal{V}_{sb}^2 + \mathcal{V}_{sc}^2\right)} \tag{1}$$

The in-phase unit template voltages ($\mathcal{U}_{pa}$,$\mathcal{U}_{pb}$, $\mathcal{U}_{pc}$) can be calculated from phase voltages ($\mathcal{V}_{sa}, \mathcal{V}_{sb}, \mathcal{V}_{sc}$) and peak amplitude of voltages ($\mathcal{V}_t$) [14] as follows:

$$\mathcal{U}_{pa} = \frac{\mathcal{V}_{sa}}{\mathcal{V}_t}, \mathcal{U}_{pb} = \frac{\mathcal{V}_{sb}}{\mathcal{V}_t}, \mathcal{U}_{pc} = \frac{\mathcal{V}_{sc}}{\mathcal{V}_t} \tag{2}$$

The reactive/orthogonal/quadrature unit templates ($\mathcal{U}_{pa}$,$\mathcal{U}_{pb}$, $\mathcal{U}_{pc}$) can be obtained from in phase unit template voltages as follows:

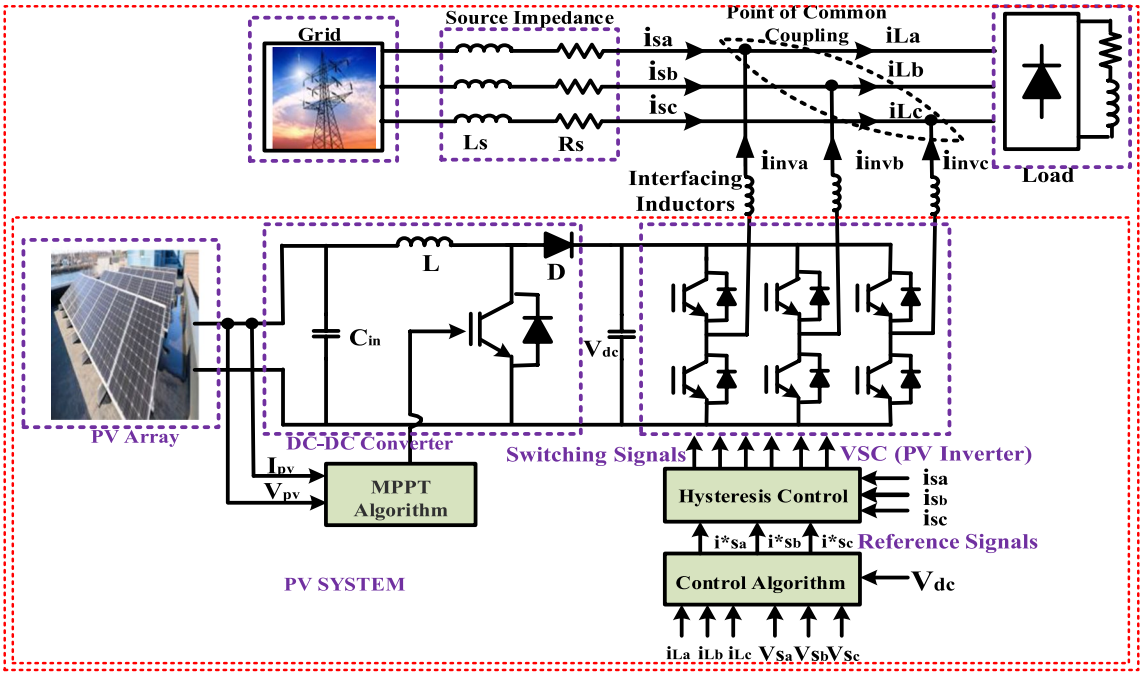


Fig.1 Proposed grid-integrated PV system

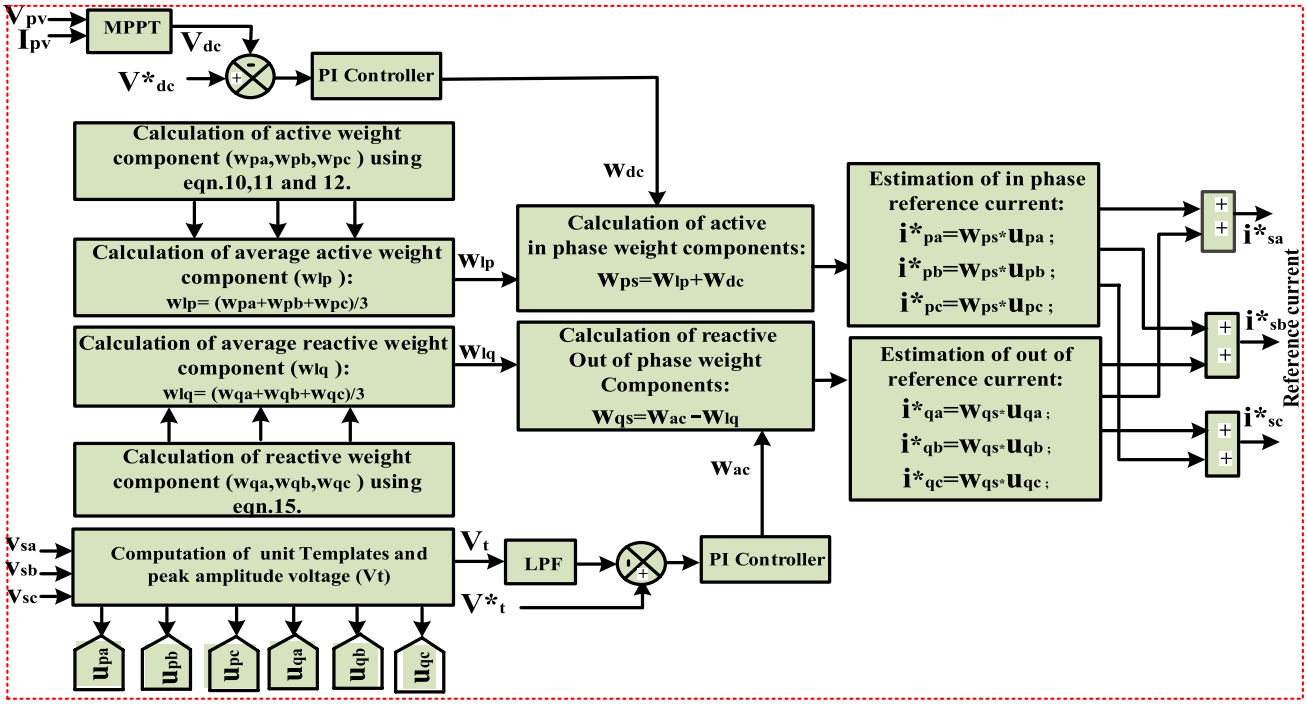


Fig.2 Proposed VSSLMS adaptive control for reference current generation

$$u_{qa} = -\frac{u_{pb}}{\sqrt{3}} + \frac{u_{pc}}{\sqrt{3}}; u_{qb} = \frac{\sqrt{3}u_{pa}}{2} + \frac{(u_{pb} - u_{pc})}{2\sqrt{3}};$$
$$u_{qc} = -\frac{\sqrt{3}u_{pa}}{2} + \frac{(u_{pb} - u_{pc})}{2\sqrt{3}} \quad (3)$$

3.1.2 Estimation of active and reactive weight components

Instantaneous estimation error (e_p) of active component in each phase at nth instant is estimated from in-phase unit template (u_p) and load current (i_L) for any phase "m" as follows:

$$e_{pm}(n) = i_{Lm}(n) - u_{pm}(n) * w_{pm}(n-1) \quad (4)$$

where w_p is fundamental active weights and "m" represents the phases "a", "b" and "c".

New variable step size (μ) for proposed control algorithm can be realized by mathematical equations [21, 23] as given below:

$$\mu(N) = \frac{\beta}{\{[1 + \exp(-\alpha|e(n)e(n-1)|)] - 0.5\}} \quad (5)$$

where $e\,(n)$ represents the estimation error at nth instant.

Estimation error (e_p) of different phases e_{pa}, e_{pb}, e_{pc} is estimated using in phase unit template (u_p) and load current (i_L) at nth instant as follows:

$$e_{pa}(n) = i_{La}(n) - u_{pa}(n) * w_{pa}(n) \quad (6)$$

$$e_{pb}(n) = i_{Lb}(n) - u_{pb}(n) * w_{pb}(n) \quad (7)$$

$$e_{pc}(n) = i_{Lc}(n) - u_{pc}(n) * w_{pc}(n) \quad (8)$$

Fundamental active weights w_{pa}, w_{pb}, w_{pc} at $(n+1)$ instant for phase any phase "m" can be computed as:

$$w_{pm}(n+1) = w_{pm}(n) + \mu(n) * e_{pm}(n) \quad (9)$$

where $e_{pm}(n)$ is the priori estimation error and $\mu(n)$ is step size for minimum error. Updated weight vector $w_{pm}(n+1)$ is calculated from weight $w_{pm}(n)$ for overall error minimization. Updated weight vector for different phases given below:

$$w_{pa}(n+1) = w_{pa}(n) + \mu(n) * e_{pa}(n) \quad (10)$$

$$w_{pb}(n+1) = w_{pb}(n) + \mu(n) * e_{pb}(n) \quad (11)$$

$$w_{pc}(n+1) = w_{pc}(n) + \mu(n) * e_{pc}(n) \quad (12)$$

The fundamental active weight component of load (w_{lp}) is estimated by taking average of updated weight vector.

$$w_{\mathrm{lp}} = \frac{(w_{pa} + w_{pb} + w_{pc})}{3} \quad (13)$$

The purpose of proposed control technique is to minimize the error. Similarly, reactive component of error (e_q) and fundamental reactive weight components of load (w_{lq}) can be computed as follows:

$$e_{qm}(n) = i_{Lm}(n) - u_{qm} * w_{qm} \quad (14)$$

$$w_{\mathrm{lq}} = \frac{(w_{qa} + w_{qb} + w_{qc})}{3} \quad (15)$$

The error in sensed actual dc-link voltage (V_{dc}) as compared to (V^*_{dc}) reference voltage is compensated using PI controller. Output is dc loss weight (w_{dc}) of PI controller is given as:

$$w_{\mathrm{dc}}(n+1) = w_{\mathrm{dc}}(n) + K_{pd}\{v_{\mathrm{dc}(e)}(n+1) - v_{\mathrm{dc}(e)}(n)\} + K_{id}v_{\mathrm{dc}(e)}(n+1) \quad (16)$$

where K_{pd}, K_{id} are gains of dc bus PI controller. $v_{\mathrm{dc}(e)}(n+1)$ is the error of the sensed V_{dc} and V^*_{dc} dc bus voltage at $(n+1)$ th sampling time.

Total active weight (w_{ps}) component in reference current can be calculated as:

$$w_{\mathrm{ps}} = w_{\mathrm{lp}} + w_{\mathrm{dc}} \quad (17)$$

The active in-phase reference current components can be evaluated as:

$$i^*_{pa} = w_{ps} * \mathcal{U}_{pa};\ i^*_{pb} = w_{ps} * \mathcal{U}_{pb};\ i^*_{pc} = w_{ps} * \mathcal{U}_{pc} \quad (18)$$

The sensed peak magnitude of supply voltage is compared with set reference peak magnitude and generated error is compensated by PI controller. Output of the controller is ac loss weight (w_{ac}) is given as:

$$w_{\mathrm{ac}}(n+1) = w_{\mathrm{ac}}(n) + K_{pa}\{v_{te}(n+1) - v_{\mathrm{te}}(n)\} + K_{ia}v_{\mathrm{te}}(n+1) \quad (19)$$

where K_{pa}, K_{ia} are gains of ac voltage PI controller and $v_{\mathrm{te}}(n+1)$ is the error of the sensed ac voltage and reference ac bus voltage at $(n+1)$th sampling time.

Total reactive weight component (w_{qs}) of the reference current as shown below:

$$w_{\mathrm{qs}} = w_{\mathrm{ac}} - w_{\mathrm{lq}} \quad (20)$$

Reactive reference current components can be evaluated as:

$$i^*_{qa} = w_{qs} * \mathcal{U}_{qa};\ i^*_{qb} = w_{qs} * \mathcal{U}_{qb};\ i^*_{qc} = w_{qs} * \mathcal{U}_{qc} \quad (21)$$

3.1.3 Estimation of reference current and switching signal for VSC

The summations of reference active and reactive $\left(i^*_{pa},\ i^*_{pb},\ i^*_{pc}\right)$,$(i^*_{qa},\ i^*_{qb},\ i^*_{qc})$, respectively, are used to generate reference current $(i^*_{sa},\ i^*_{sb},\ i^*_{sc})$ which is given by:

$$i^*_{sa} = i^*_{pa} + i^*_{qa},\ i^*_{sb} = i^*_{pb} + i^*_{qb},\ i^*_{sc} = i^*_{pc} + i^*_{qc} \tag{22}$$

Further, reference current is compared with sensed grid current in hysteresis current controller for generation of gating signal for six switches of voltage source converter.

4 Simulation results and discussions

In the present section, simulation results have been presented. Performance of proposed new VSSLMS-based control of grid-integrated PV system is analyzed and compared with conventional SRF and LMS adaptive control. Proposed control maintains the current and voltage harmonics within limit of IEEE standard and reduces the reactive power burden on the grid.

Proposed control is modelled in MATLAB/Simulink environment and studied in PFC mode of operation. Simulation and experimental parameters of the system are given in the appendix.

4.1 Performance of grid-integrated PV system

4.1.1 Dynamic and Steady state behaviour under linear, nonlinear and unbalanced load at STC input

The performance of proposed new VSSLMS adaptive control under linear and nonlinear load depicted in Figs. 3 and 4, respectively. Various parameters of the system viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, V_{dc} and power balance between grid, load and inverter are also analyzed.

It can be seen from Fig.3 that till 0.15 s PV inverter supplies 4kW active power and 3.0 kVAR of reactive demand of the load, Vgrid and Igrid are 180° out of phase (as excess 6.25 kW power is supplied to the grid), V_{dc} is maintained at 750V. The inverter alone supplies the load active and reactive power demand, which decreases the reactive power

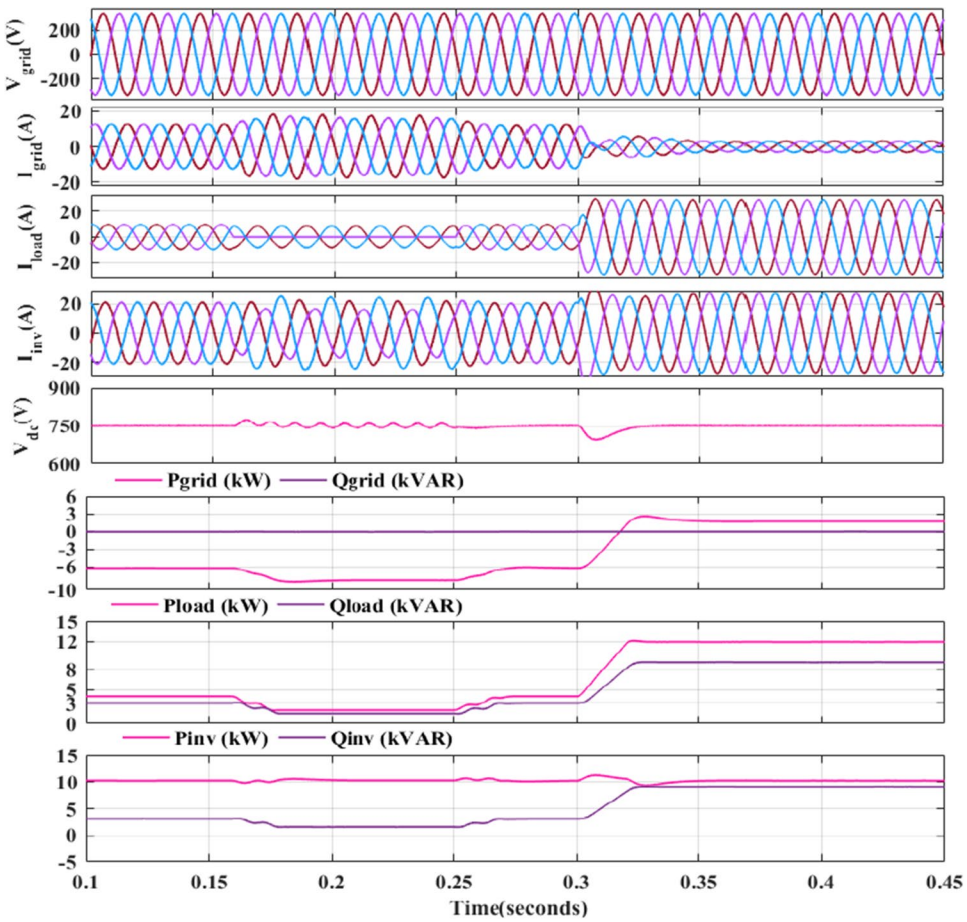


Fig. 3 Simulated performance of the system under linear (balance/unbalance) and variable load at STC

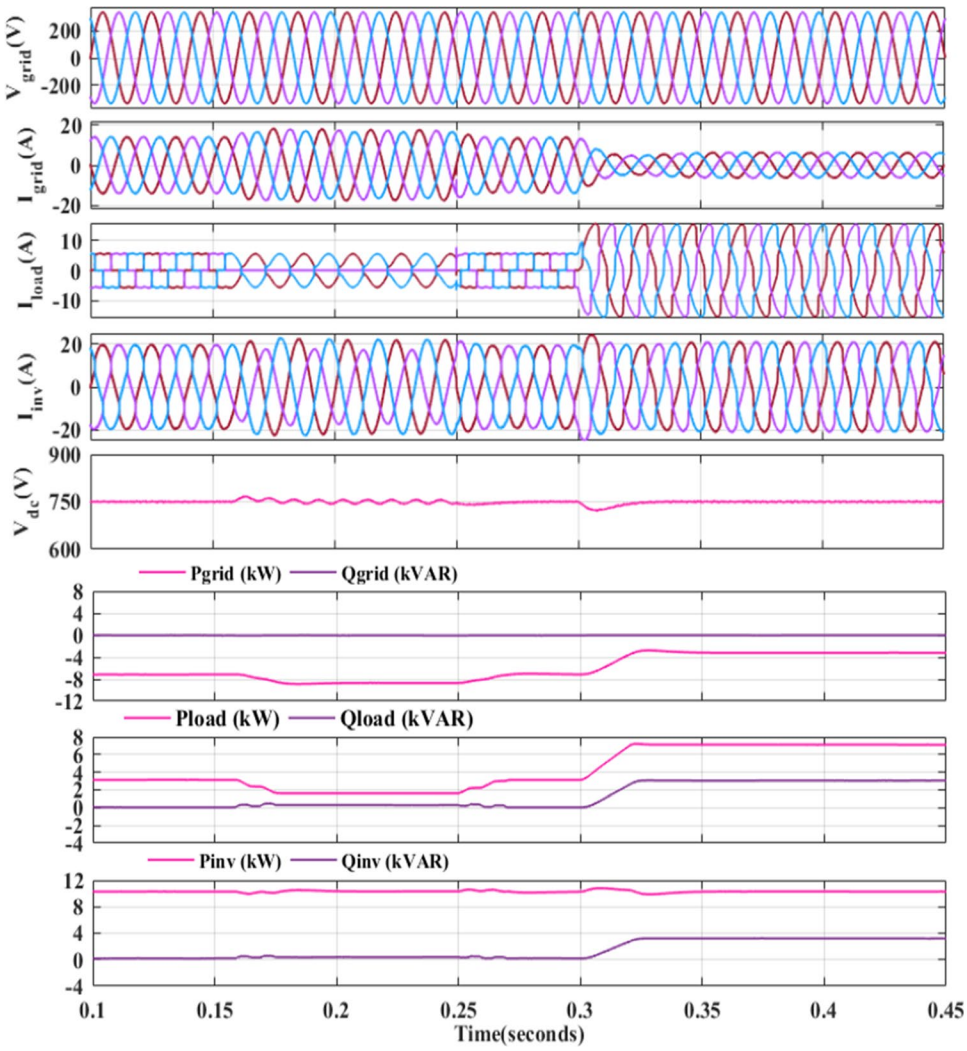


Fig. 4 Simulated performance of system under nonlinear (balance/unbalance) and load varying at STC input

drawn from the grid to zero, demonstrating that the proposed control is effective in maintaining the grid at UPF.

Further, unbalanced condition and load varying condition occur at 0.15 s and 0.30 s, respectively. Performance of the proposed control under linear unbalanced load condition by keeping one phase (phase *b*) disconnected between 0.15 and 0.25 s is studied and various parameters viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, P_{grid}, Q_{grid}, P_{inv}, Q_{inv}, P_{load}, Q_{load}, and V_{dc} are analyzed and depicted in Fig. 3. During unbalance period, inverter current compensates, effect of load unbalance and maintains grid current to be balance. It also supplies the load requirement for reactive power and thus holds the grid at UPF along with the balance grid current. For the load varying condition between 0.3 and 0.45 s, an extra linear load of 10kVA, 0.8 pf lagging is added at 0.3 s in the system. After 0.3 s, load demand is 15kVA and 0.8 pf lagging (12 kW and 9 kVAR). Power supplied by PV inverter is 10.25 kW which is not enough to fulfill the load requirement, therefore, to meet the load demand extra required power 1.75 kW is taken from distribution grid.

It has been realised that during load variation and single phasing also that I_{grid} is sinusoidal and V_{dc} is maintained at 750 V. THD in I_{grid} is also maintained at 3.37% as depicted in Fig. 7a which is as per IEEE standard 519–2014, well within the limit.

Further, to test the proposed control under nonlinear load, it has been considered and results are shown in Fig. 4. It is observed that till 0.15 s PV system supplies its nominal power of 10.25 kW. It is observed from Fig. 4 that till 0.15 s V_{grid} and I_{grid} are 180° out of phase, maintains V_{dc} at 750 V, PV inverter supplies 3.1 kW active power demand of the load. Further, remaining 7.15 kW power of the PV system is supplied to the grid.

Further, proposed control under nonlinear unbalanced and load varying condition is also tested. Under nonlinear unbalanced load condition by one phase '*b*' of load disconnected between 0.15 and 0.25 s and various parameters viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, P_{grid}, Q_{grid}, P_{inv}, Q_{inv}, P_{load}, Q_{load}, and V_{dc} are analyzed and presented in Fig. 4. During unbalance period, inverter current compensates, effect of

load unbalance and maintains balanced grid current. It also supplies the load requirement for reactive power and thus holds the grid at UPF along with the balance grid current.

For the load varying conditions between 0.3 and 0.45 s, an extra linear varying load of 5kVA, 0.8 pf lagging is added at 0.3 s in the system of existing nonlinear load. After 0.3 s, load demand is 7.1 kW and 3 kVAR, out of which 7.1 kW is supplied by PV inverter, as its generated power is 10.25 kW, after satisfying the load requirement, remaining power 3.15 kW is supplied to grid. It has been realised that during load variation and single phasing also grid current is sinusoidal and V_{dc} is maintained at 750 V. THD in I_{grid} is also maintained at 4.12% as depicted in Fig. 7b which is as per IEEE standard 519–2014, well within the limit.

4.1.2 Dynamic and steady state behavior under linear and nonlinear load at varying insolation

Further, the proposed new VSSLMS control is also tested at varying insolation under linear and nonlinear load. Solar irradiance is decreased at 0.15 s from 1000 to 700 W/m^2 and again increased to 1000 W/m^2 at 0.25 s; thus, solar PV current decreased; thus, PV output power is reduced and vice versa.

A 20kVA, 0.8 pf lag linear load is considered, simulated results and corresponding various parameters viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, P_{grid}, Q_{grid}, P_{inv}, Q_{inv}, P_{load}, Q_{load}, and V_{dc} of the PV inverter are shown in Fig. 5. It is observed from Fig. 5, that both V_{grid} and I_{grid} are in phase (as power is supplied from grid to load), V_{dc} is maintained at 750 V, PV inverter supplies its generated power 10.25 kW which is not enough to fulfill the load demand of 20kVA, 0.8pf lag(16 kW and 12kVAR); therefore, to meet the load demand, extra power 5.75 kW is taken from distribution grid, while 12 kVAR

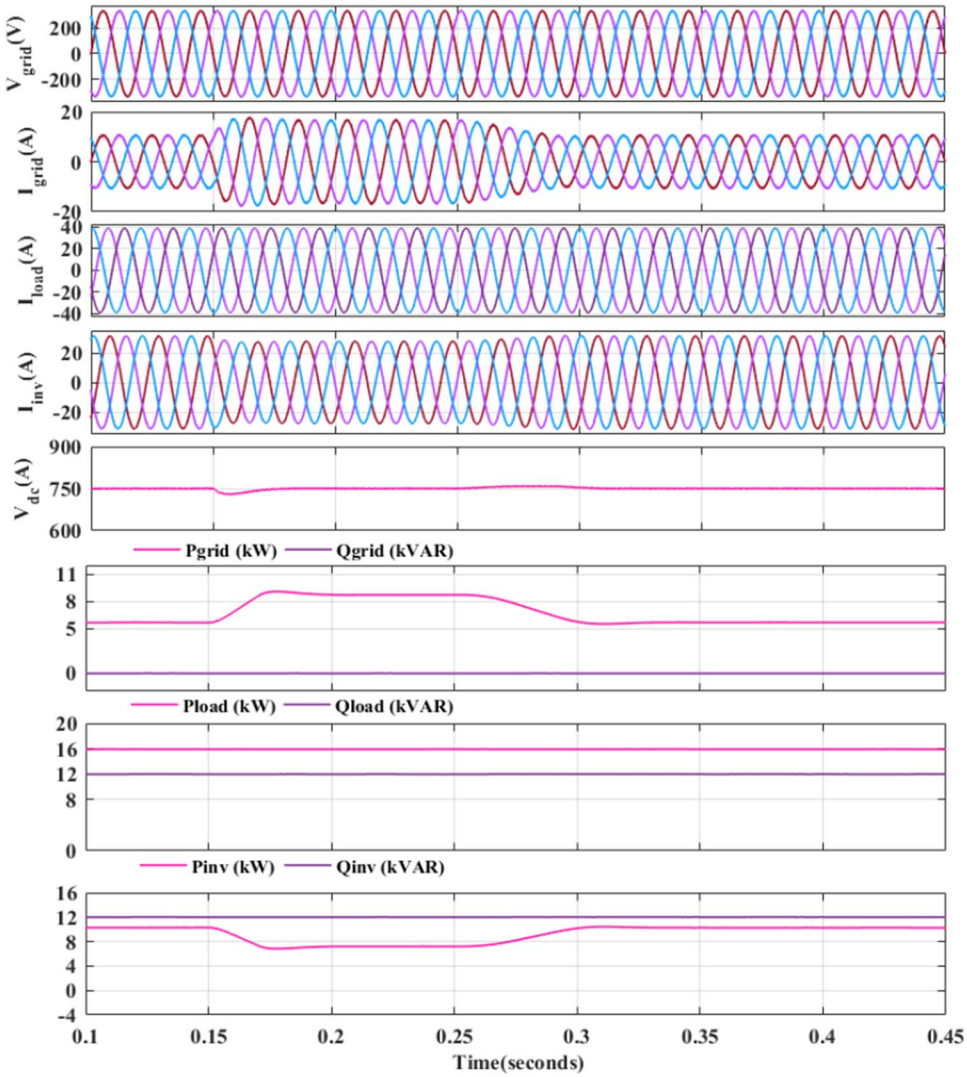


Fig. 5 Simulated performance of system under linear load at varying insolation

of reactive demand of the load is supplied by PV inverter alone, showing proposed control is efficient in maintaining the grid at UPF.

Further, the proposed algorithm on linear load is tested under varying insolation from 0.15 to 0.25 s. The performance of system using proposed algorithm under linear load condition under varying insolation between 0.15 and 0.25 s is analysed and various parameter viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, P_{grid}, Q_{grid}, P_{inv}, Q_{inv}, P_{load}, Q_{load}, and V_{dc} are presented in Fig. 5. Due to decrease in solar insolation, P_{inv} is

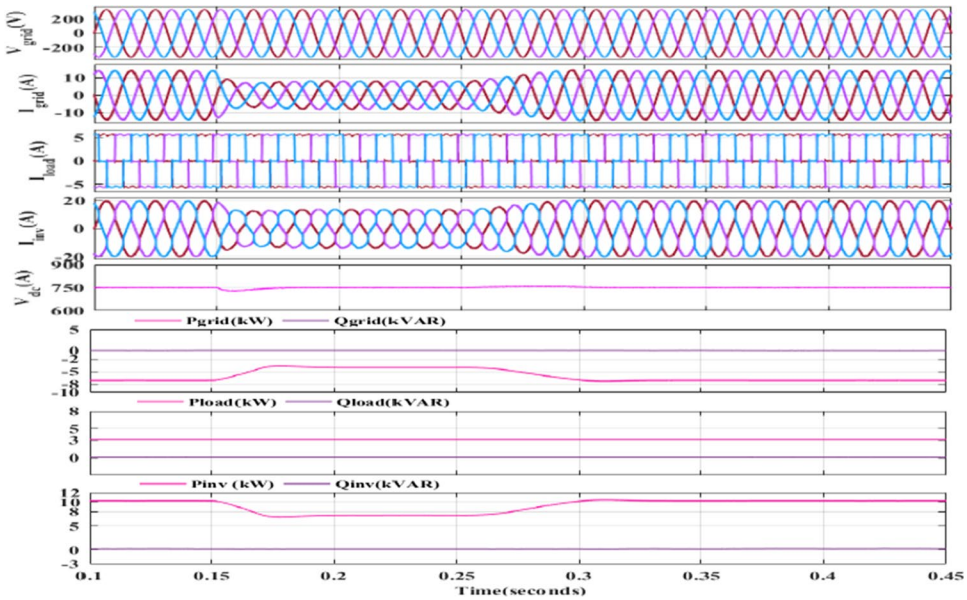


Fig. 6 Simulated performance of system under nonlinear load at varying insolation

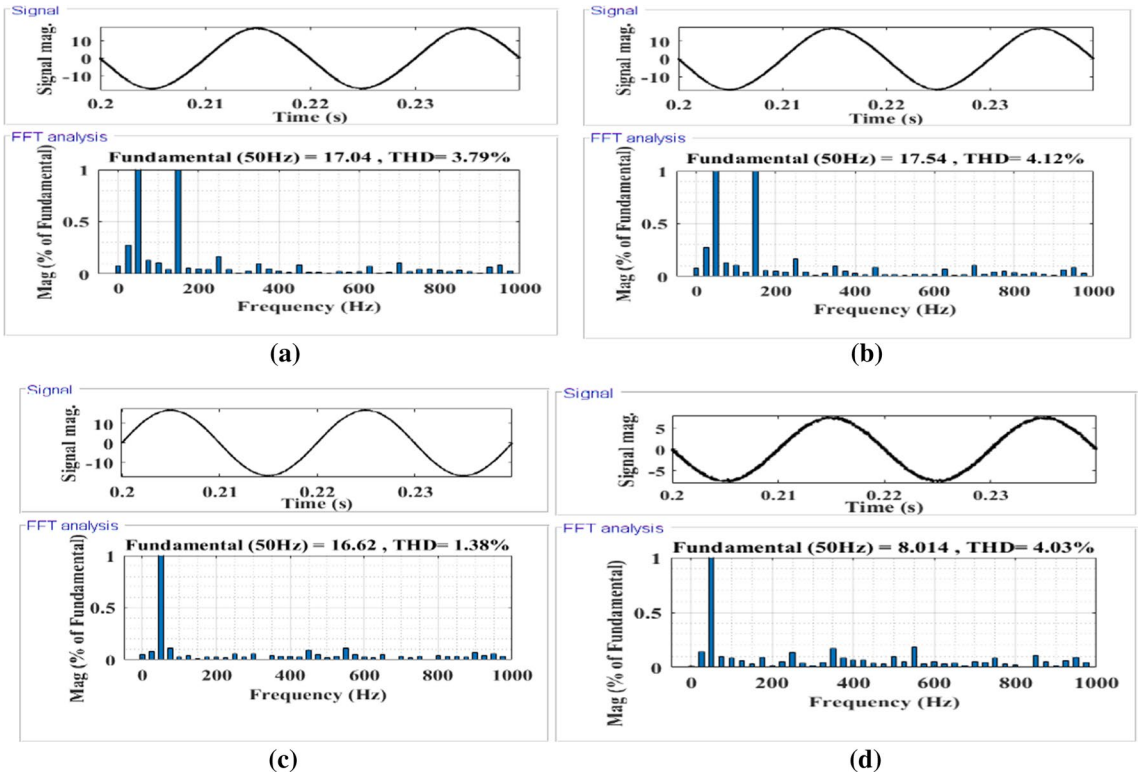


Fig. 7 **a**, **b** Compensated grid current waveform and THD under linear and nonlinear load, respectively, at STC; **c** and **d** compensated grid current waveform and THD under linear and nonlinear load, respectively, at varying insolation

decreased to 7.75 kW at 0.15 s causing the grid to supply 2.5 kW more power (total 8.25 kW) to the load as depicted in Fig. 5, while total reactive power demand of load is still supplied by inverter alone hence keeping grid at UPF during the adverse scenario of solar insolation. It has been observed that during solar insolation variation, sharing of active and reactive power between load (P_{load}, Q_{load}), grid (P_{grid}, Q_{grid}) and inverter (P_{inv}, Q_{inv}) is maintained, grid current is sinusoidal and maintains V_{dc} at 750 V. THD in I_{grid} is also maintained at 1.38% during the varying insolation also as shown in Fig. 7c which is as per IEEE standard 519–2014, well within the limit.

Further for a nonlinear load at varying irradiance, simulated results and corresponding various parameters viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, P_{grid}, Q_{grid}, P_{inv}, Q_{inv}, P_{load}, Q_{load}, and V_{dc} of the PV inverter are shown in Fig. 6. V_{grid} and I_{grid} are out of phase by 180^0 (as power is supplied to grid), V_{dc} is maintained at 750 V, PV inverter supplies its generated power 10.25 kW after satisfying the load demand of 3.1 kW, rest power 7.15 kW is fed back to grid, while PV inverter alone supplying reactive power demand of the load, showing proposed control is efficient in maintaining grid at UPF.

After decrease in solar insolation, PV inverter power is decreased from 10.25 to 7.75 kW at 0.15 s leading to decrease in the power fed to the grid by 2.5 kW as shown in Fig. 6. While reactive power demand of load is supplied by PV inverter (VSC) alone hence keeping grid at UPF during the adverse scenario of solar insolation. After 0.25 s, insolation and PV power both have been increased and system remains in steady-state condition.

It has been observed that during solar insolation variation, active and reactive power is balanced between inverter (P_{inv},

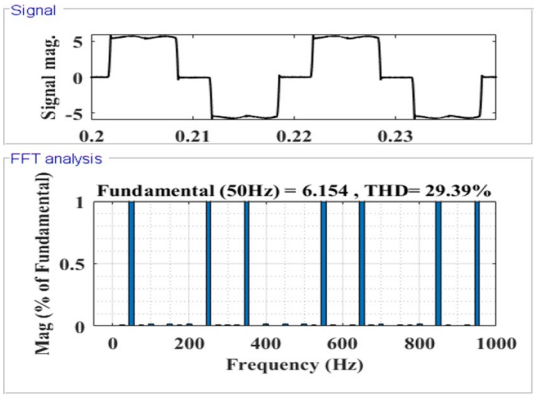


Fig. 8 Uncompensated nonlinear load current waveform and THD

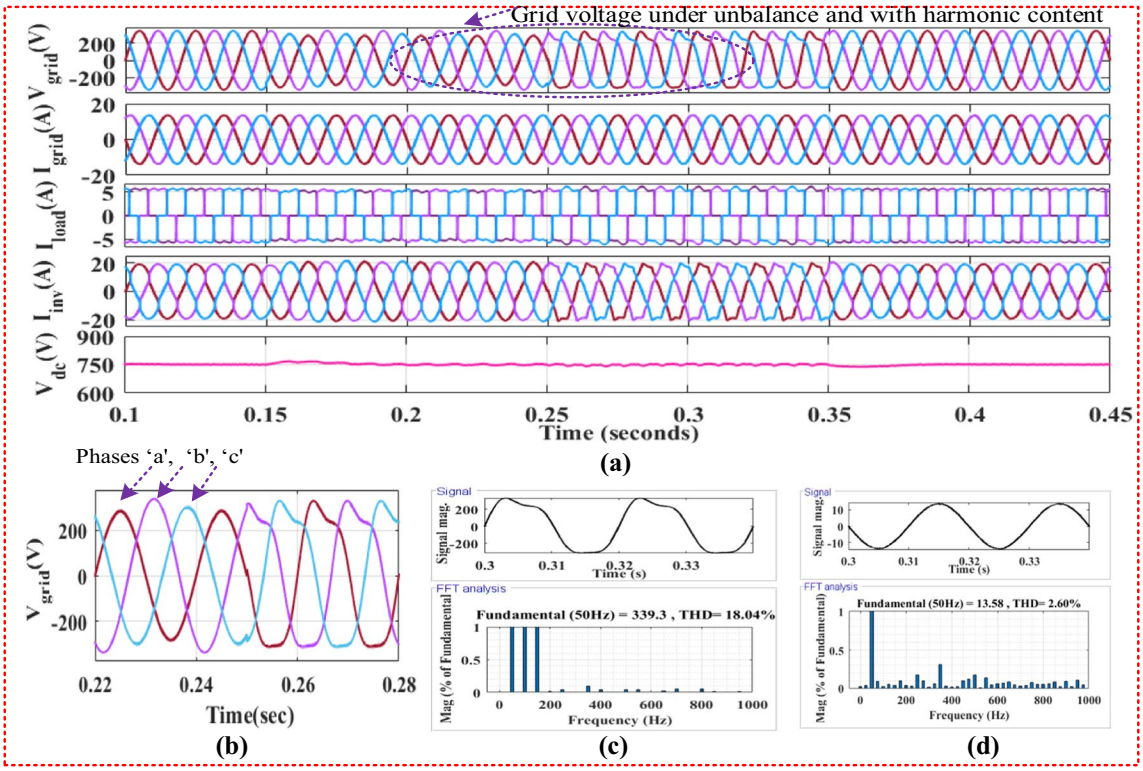


Fig. 9 **a** Performance with grid voltage under unbalance and grid voltage with harmonic content/distortion; **b** zoomed area of grid voltage waveform under unbalance and with harmonic content; **c** THD in grid current under grid voltage with harmonic content

Q_{inv}) load (P_{load}, Q_{load}) and grid (P_{grid}, Q_{grid}) is maintained, I_{grid} is sinusoidal and V_{dc} is maintained at 750 V. THD in I_{grid} is also maintained at 4.03% during insolation varying as shown in Fig. 7d, while load current THD is 29.39% as shown in Fig. 8.

4.1.3 Performance under unbalance grid voltage and grid voltage with harmonic contents

Further, proposed control is also tested under unbalanced grid voltage and grid voltage with harmonic content for nonlinear load. Figure 9 depicts the response of the system. Various parameters viz. V_{grid}, I_{grid}, I_{load}, I_{inv}, V_{dc} and THDs in grid voltage and current are analyzed and presented.

It can be seen from Fig. 9a that from 0.15 to 0.25 s, grid voltage is under unbalance and suffers from distortion between 0.25 and 0.35 s, respectively. During unbalance voltage of grid which is decreased by 16% and 12% of phases "*a*" and "*c*", respectively, as can be seen from Fig. 9a, b. Under unbalanced grid voltage condition, the proposed algorithm maintains grid current to be balanced and free from harmonics irrespective of voltage unbalance and load harmonics. It is also maintaining grid at UPF along with the balance grid current.

Further, the behaviour of the system is observed when grid voltage with harmonic content up to 18.4% THD as shown in Fig. 9c. During the adverse scenario of grid voltage with harmonic content, VSC maintains the THD in grid current as 2.6% as shown in Fig. 9d, which is within the IEEE standards. Hence, proposed control is efficient in maintaining grid current to be sinusoidal viz. free from harmonics and balanced irrespective of grid voltage harmonics and load harmonics as depicted in Fig. 9.

4.1.4 Comparison of proposed VSSLMS control and conventional SRF and LMS

Figure 10 shows comparison of proposed VSSLMS control with conventional SRF and LMS adaptive control. Nonlinear load viz. bridge rectifier with RL load ($R = 100\ \Omega$, 100 mH) is considered and further to investigate the transient response

Table 1 Comparison of proposed control with SRF and LMS control

Control algorithm	Settling time (sec)	Overshoot in active component (W_{ps})	Undershoot in active component (W_{ps})
SRF	0.06	160	− 58
LMS	0.05	140	− 50
Proposed	0.03	75	− 25

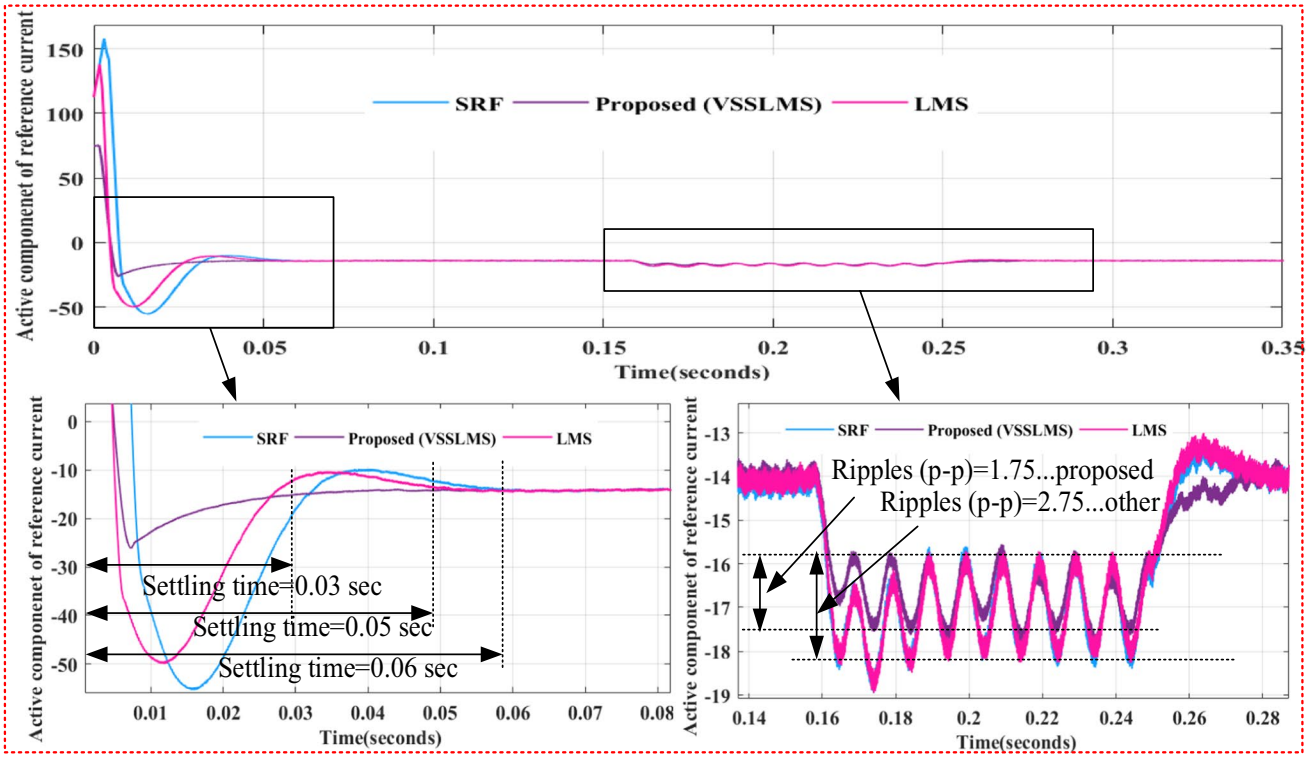


Fig. 10 Comparative performance of proposed control with SRF and LMS control

during load perturbation one phase of load is kept removed from 0.15 to 0.25 s.

It can be seen from Fig. 4 that till 0.30 s, PV system supplies its rated active power. It is observed that till 0.30 s V_{grid} and I_{grid} are 180° out of phase (as power is fed to the grid); therefore, active component of reference supply current is negative as it can be seen from Fig. 10.

Figure 10 shows the comparison of transient and dynamic performance under considered load. It can be seen from the Fig. 10 that during grid integration, proposed control gives best initial transient response in terms of settling time, overshoot and undershoot as given in Table 1 and has less oscillations. Further during load perturbation, it can be seen that proposed control has lower ripples as compared to that of the conventional SRF and LMS control, settling time using proposed VSSLMS control is 30 ms, whereas it's 50 ms and 60 ms, respectively, in case of conventional SRF and LMS. Hence, the proposed control is more efficient than that of other conventional SRF and LMS adaptive control.

5 Experimental Results

Due to limitation of the lab, developed prototype is made to operate as DSTATCOM by taking reference DC link voltage of 200V. Proposed control is tested to conduct VSC as a DSTATCOM. Figure 11 shows the prototype hardware setup developed in the laboratory.

Developed hardware has: 1. Grid, 2. interfacing indctors, 3. loads 4. power analyzer (Fluke/HIOKI-198), 5. inverter 6. DC supply, 7. DSO (Agilent made DSO7014A): Four-channel, 8. DSP-dSPACE-1202 MicroLabBox, 9.control desktop. Hall-effect-based sensors: voltage (LV 25-P) and current (LA 55-P).

5.1 Performance of proposed control under various load and grid conditions

5.1.1 Linear load

Performance of proposed control, under linear load of 0.48 kVA, 0.93 pf lagging has been considered. Various experimental parameters of the system viz. V_{grid}, I_{grid}, I_{load}, current supplied by inverter (I_{inv}) of one phase "a", before and after compensation are shown in Fig. 12a. Power factor of the load is 0.93 pf lagging, before compensation, it can be seen from Fig. 12b, while after compensation reactive power supplied by grid is approximately zero, showing UPF operation of proposed control as shown in Fig. 12c. After compensation, load reactive demand is fulfill by the VSC alone. Figure 12b and c show the power factor of the grid before and after compensation, respectively. Power factor of grid is maintained UPF after compensation as it can be seen from Fig. 12c as reactive power fed by grid zero.

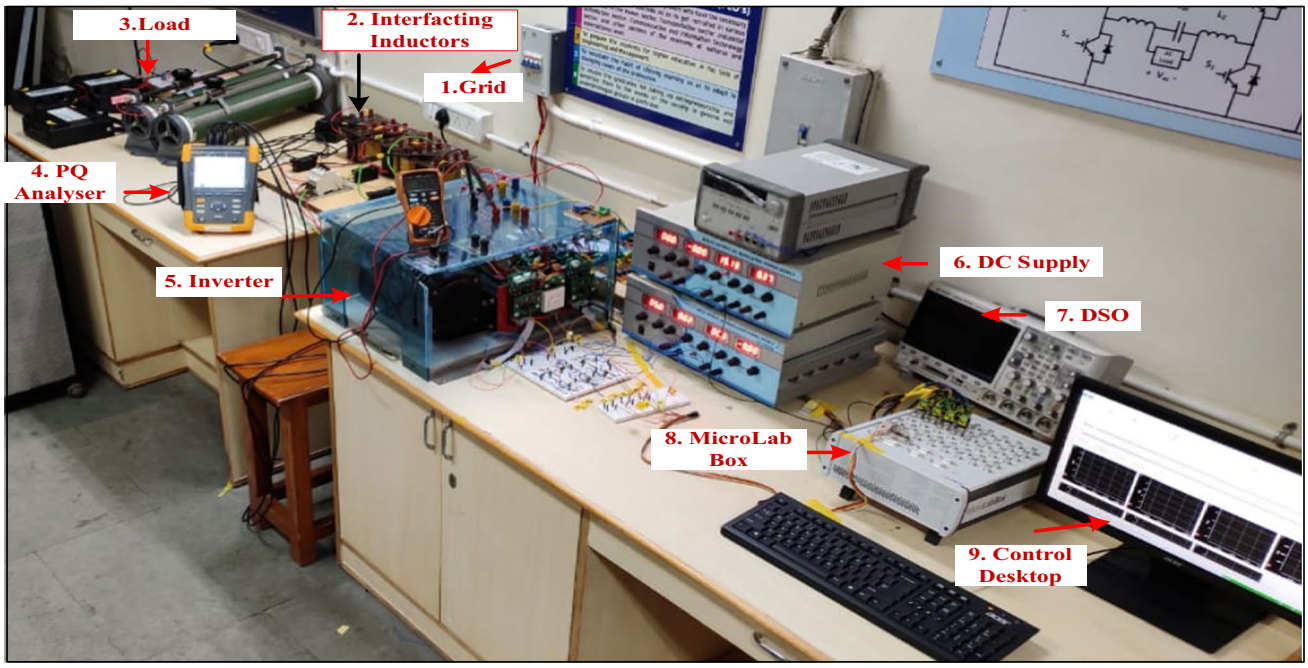


Fig. 11 Hardware prototype of DSTATCOM

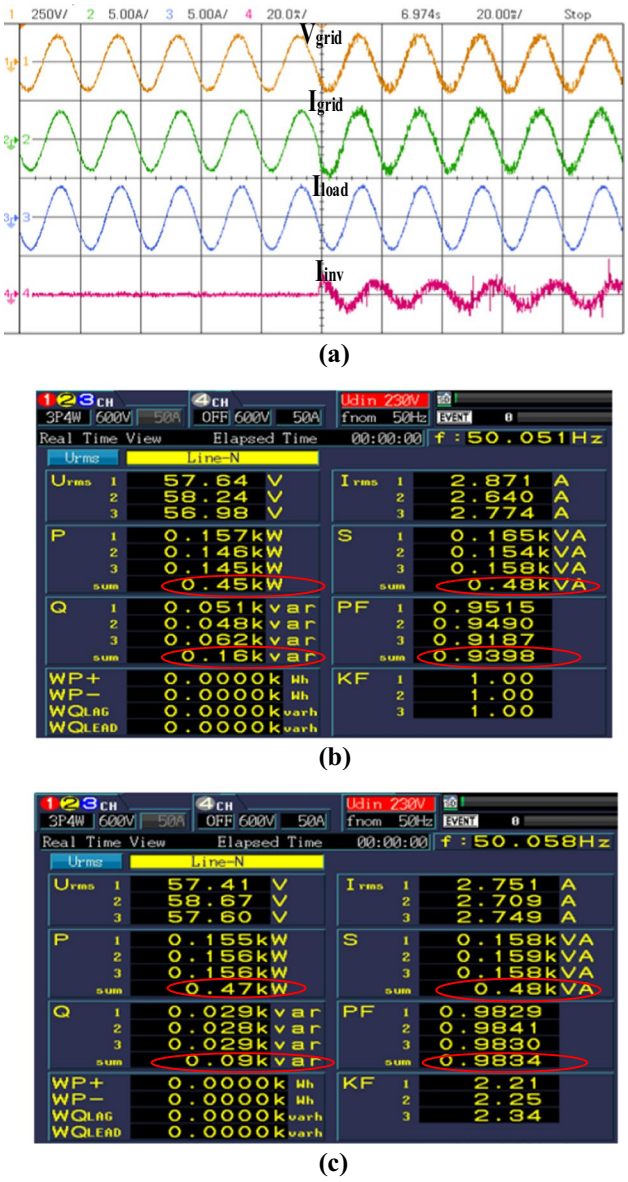


Fig. 12 Experimental result using proposed control under: **a** waveform; **b** and **c** load parameter and pf before and after compensation respectively

5.1.2 Nonlinear load

Performance of proposed control is tested on hardware prototype under nonlinear load by connecting a diode bridge rectifier with R load ($R = 50\ \Omega$). Figure 13a shows the V_{grid}, I_{grid}, I_{load}, current supplied by inverter (I_{inv}) respectively of one phase ('a') and Fig. 13b shows the load variation from 50 to 40 Ω. Inverter act as a harmonic compensator by compensating the harmonics of local loads. THD in grid current is less than 1%, while nonlinear load current THD is approximately 27% as shown Fig. 13c, d, respectively, which is well within IEEE standard 519–2014. It is analyzed that using

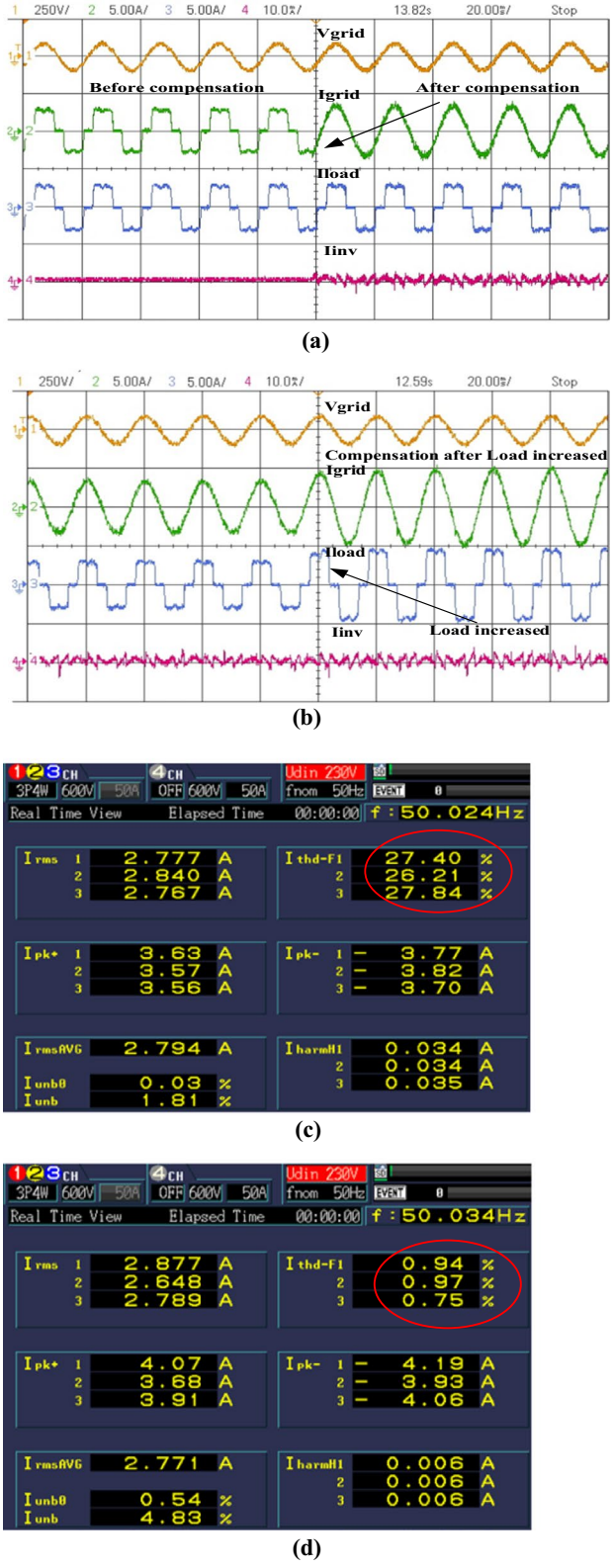


Fig. 13 Parameter under: **a** performance under nonlinear load; **b** performance under nonlinear load during load variation(increased); **c** THD in grid current before compensation under nonlinear **d** THD in grid current after compensation under nonlinear load

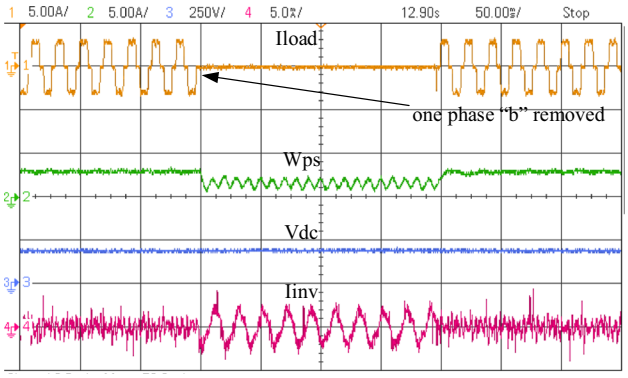


Fig. 14 Experimental result under nonlinear unbalance load

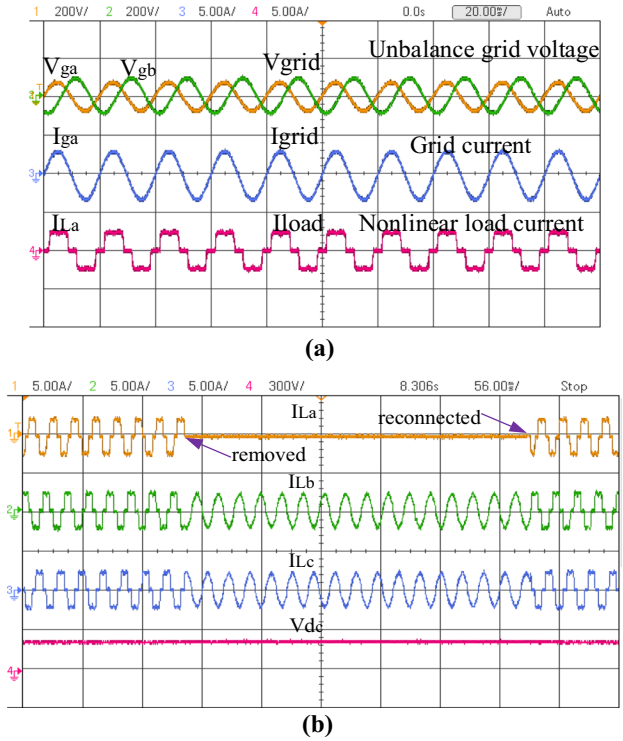


Fig. 15 Response of the system under unbalance grid voltage: **a** load under unbalance voltage of grid; **b** load unbalance

proposed control, grid current is improved significantly in terms of distortion by removing the harmonics.

5.1.3 Nonlinear unbalance load

To study the performance of proposed control under nonlinear unbalanced load, existing diode bridge rectifier with R load ($R = 50\ \Omega$) is taken. Unbalanced is created by removing one phase for few cycles and again connected back. Various experimental parameters of the system viz. I_{load}, fundamental active component (W_{ps}), V_{dc} and I_{inv} are shown

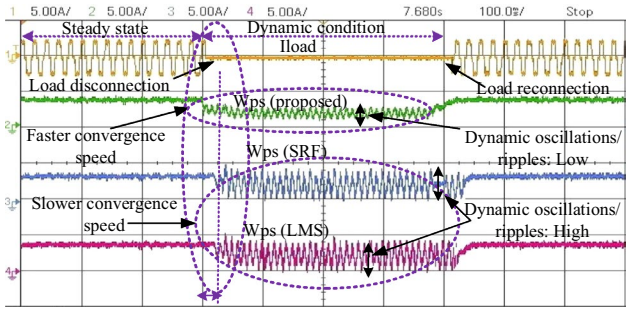


Fig. 16 Comparison of proposed and conventional SRF- and LMS-based control

Table 2 Comparison of proposed algorithm, conventional SRF and LMS

S.N	Criteria	SRF	LMS	Proposed
1	Accuracy	Good	Moderate	Better
2	Convergence	Slow	Moderate	Good
3	Complexity	High	Low	Low
4	Step size	NA	Constant	Variable
5	dSPACE-1202/sampling time	50 µS	40µS	40 µs
6	Filtering type	Time domain	Adaptive	Adaptive
7	Static error	High	High	Low
8	Dynamic oscillations in active component	High	High	Low

in the Fig. 14. From the experimental result, performance is efficient.

5.1.4 Performance under unbalance grid Voltage

Performance of proposed control is also tested under unbalance grid voltage at nonlinear and unbalanced load both. From Fig. 15a, it can be seen that, during the unbalance of voltage of the grid which is decreased by 17% (approx.) in phase '*a*', proposed control is efficient in maintaining the grid current to be balance and free from harmonics. Furthermore, during the unbalance of the load by means of removing one phase of load, different phases of load current are depicted in Fig. 15b. It has been found that performance using proposed control is efficient under unbalance scenario of the grid voltage also.

5.1.5 Comparison of proposed and conventional SRF and LMS based control

Performance of proposed VSSLMS control has been compared with conventional LMS and SRF control and is depicted in Fig. 16. It has been observed that proposed control's performance is efficient over other conventional

control and provide comparatively 40 ms faster convergence speed with minimum oscillations during dynamic condition viz. unbalanced load. Moreover, proposed control does not require any complex block like phase lock loop (PLL) for the synchronisation, which reduces the additional computational burden and provides faster response in terms of convergence speed compared to the SRF and LMS algorithm to estimate the reference current. Table 2 shows the comparative analysis of proposed control with other conventional control.

6 Conclusion

Solar PV system has been integrated to grid and has been studied using proposed new VSSLMS adaptive control algorithm of VSC. Proposed algorithm performance is tested at STC and varying insolation both under various balanced/unbalanced loads. Under various conditions, UPF operation of inverter is maintained and the THD in grid current is within the limits as per IEEE standards 519–2014. Proposed control is compared with conventional SRF and LMS adaptive control and it is found that proposed algorithm depicts better transient response under PV grid integration and dynamic load (unbalance) also. Proposed algorithm is validated experimentally in the laboratory on prototype hardware. Simulation and experimental investigations justify the efficacy of the proposed algorithm. Moreover, solar-PV system using proposed new VSSLMS adaptive control maintains the PQ of the system along with the PV power generation.

Appendix

Simulation parameters grid voltages (V_{sabc}): 415 V, frequency: 50 Hz, line impedances (R_s, L_s): 0.01 Ω, 0.1mH, interfacing inductor (L_f): 7mH, DC link capacitor (C_{dc}): 1000 µF, $V_{dc(ref)}$: 750 V; $\alpha = 20$; $\beta = 0.01$; sampling time (T_s) = 5.5 µs; source impedance:($R_s = 0.01$ Ω, $Ls = 0.1$ mH) per phase;

Solar PV parameters: nominal power:10.25 kW, nominal voltage: 410 V, nominal current: 25 A;

Boost converter parameters: L: 0.5 mH, switching frequency (f_s): 10 kHz, capacitor (C): 1000 µF, duty cycle (D): 0.43;

Various loads: linear load (balanced/unbalanced): 5 kVA, 0.8 pf lag (unbalanced between 0.15 and 0.25 sec); nonlinear load: bridge rectifier $R = 100$ Ω, $L = 100$mH (unbalanced between 0.15 and 0.25 sec); linear varying load: 5 kVA, 0.8 pf lag + 10 kVA, 0.8 pf lag (between 0.3 and 0.45 sec.); nonlinear varying load: (bridge rectifier $R = 100$ Ω, $L = 100$ mH) + 5 kVA, 0.8 pf lag (between 0.3 and 0.45 sec);

Experimental parameters of the prototype hardware: grid voltage ($V_{L\text{-}L}$) = 100 V(58 V/phase); reference dc link voltage ($V_{dc(ref)}$) = 200 V; VSC rating = 25 kVA; interfacing inductor (L_f) = 5 mH; nonlinear load ($R = 50$ Ω, 30 Ω); linear load: 0.48 kVA (0.93 pf lag); $\alpha = 20$; $\beta = 0.01$; gain of dc PI controller: $k_p = 0.50$, $k_i = 0.1$; $T_s = 40$ µs.

References

1. Verma P, Garg R, Mahajan P (2020) Asymmetrical interval type-2 fuzzy logic control based MPPT tuning for PV system under partial shading condition. ISA Trans 100:251–263. https://doi.org/10.1016/j.isatra.2020.01.009
2. Gupta N, Garg R (2017) Tuning of asymmetrical fuzzy logic control algorithm for SPV system connected to grid. Int J Hydrog Energy 42(26):16375–16385. https://doi.org/10.1016/j.ijhydene.2017.05.103
3. IEEE Std 1547:2003, IEEE standard for interconnecting distributed resources with electric power systems, IEEE Std 1547–2003, vol. 2014, pp. 1–16, 2003. https://doi.org/10.1109/IEEESTD.2003.94285
4. Choi W, Lee W, Han D, Sarlioglu B (2018) New configuration of multifunctional grid-connected inverter to improve both current-based and voltage-based power quality. IEEE Trans Ind Appl 54(6):6374–6382. https://doi.org/10.1109/TIA.2018.2861737
5. Trishan E, Patrick LC (2007) Comparison of photovoltaic array maximum power point tracking techniques. IEEE Trans Energy Convers 22(2):439449. https://doi.org/10.1109/TEC.2006.874230
6. Jiang Y, Qahouq JAA, Haskew TA (2012) Adaptive step size with adaptive-perturbation-frequency digital MPPT controller for a single-sensor photovoltaic solar system. IEEE Trans Power Electron 28(7):3195–3205. https://doi.org/10.1109/TPEL.2012.2220158
7. Shukl P, Singh B (2019) Grid integration of three-phase single-stage PV system using adaptive Laguerre filter based control algorithm under nonideal distribution system. IEEE Trans Ind Appl. https://doi.org/10.1109/tia.2019.2931504
8. Beniwal N, Hussain I, Singh B (2019) Second-order volterra-filter-based control of a solar PV-DSTATCOM system to achieve lyapunov's stability. IEEE Trans Ind Appl 55(1):670–679. https://doi.org/10.1109/TIA.2018.2867324
9. Liang X, Andalib-Bin-Karim C (2018) Harmonics and mitigation techniques through advanced control in grid-connected renewable energy sources: a review. IEEE Trans Ind Appl 54(4):3100–3111. https://doi.org/10.1109/TIA.2018.2823680
10. Beniwal N, Hussain I, Singh B (2018) Implementation of the DSTATCOM with an i-PNLMS-based control algorithm under abnormal grid conditions. IEEE. https://doi.org/10.1109/TIA.2018.2846739
11. Yang Bo, Li W, Zhao Yi, He X (2009) Design and analysis of a grid-connected photovoltaic power system. IEEE Trans Power Electron 25(4):992–1000. https://doi.org/10.1109/tpel.2009.2036432
12. Kanavaros D, Oriti G, Julian AL (2020) Novel implementation and comparison of active and reactive power flow control methods in a single phase grid-connected microgrid. IEEE Trans Ind Appl 56(2):1631–1639. https://doi.org/10.1109/tia.2020.2965080
13. Kazmierkowski MP (2015) Power quality: problems and mitigation techniques. IEEE Ind Electron Mag 9(2):62

14. Singh B, Chandra A, Al-haddad K Power quality problems and mitigation techniques. Wiley Online Library, ISBN: 978-1-118-92205-7
15. Barik PK, Shankar G, Sahoo PK (2019) Power quality assessment of microgrid using fuzzy controller aided modified SRF based designed SAPF. Int Trans Electr Energy Syst. https://doi.org/10.1002/2050-7038.12289
16. Singh B, Jayaprakash P, Kothari DP, Chandra A, Al Haddad K (2014) Comprehensive study of dstatcom configurations. IEEE Trans Ind Inform 10(2):854–870. https://doi.org/10.1109/TII.2014.2308437
17. Singh B, Shahani DT, Verma AK (2012) Power balance theory based control of grid interfaced solar photovoltaic power generating system with improved power quality. In: PEDES 2012—IEEE international conference on power electronics, drives and energy systems, pp. 1–7. https://doi.org/10.1109/PEDES.2012.6484359
18. Yada HK, Murthy MSR (2017) An improved control algorithm for DSTATCOM based on single-phase SOGI-PLL under varying load conditions and adverse grid conditions. In: IEEE international conference on power electronics, drives and energy systems. PEDES 2016, vol. 2016, pp. 1–6. https://doi.org/10.1109/PEDES.2016.7914564
19. Zeb K, Uddin W, Khan MA, Ali Z, Ali MU, Christofides N, Kim HJ (2018) A comprehensive review on inverter topologies and control strategies for grid connected photovoltaic system. Renew Sustain Energy Rev 94:1120–1141. https://doi.org/10.1016/j.rser.2018.06.053
20. Patel SK, Arya SR, Maurya R (2017) Harmonic mitigation technique for DSTATCOM using continuous time LMS adaptive filter. In:IEEE Uttar Pradesh section international conference on electrical, computer and electronics engineering. UPCON 2016, pp. 19–24. https://doi.org/10.1109/UPCON.2016.7894617
21. Huang HC, Lee J (2012) A new variable step-size NLMS algorithm and its performance analysis. IEEE Trans Signal Process 60(4):2055–2060. https://doi.org/10.1109/TSP.2011.2181505
22. Li Z, Li D, Xu X, Zhang J (2019) New normalized LMS adaptive filter with a variable regularization factor. J Syst Eng Electron 30(2):259–269. https://doi.org/10.21629/JSEE.2019.02.05
23. Yuan Z, Songtao X (2017) New LMS adaptive filtering algorithm with variable step size. In: Proceedings of the international conference on vision, image and signal processing. ICVISP 2017, pp. 1–4. https://doi.org/10.1109/ICVISP.2017.11

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Covid Detection from X-RAY and CT Scans using Transfer Learning – A Study

Shailender Kumar
Department Of Computer Engineering,
Delhi Technological University,
NCT of Delhi, India
shailenderkumar@dce.ac.in

Naman
Department Of Computer Engineering
Delhi Technological University,
NCT of Delhi, India
yadavnaman1999@gmail.com

Mohd Talib,
Department Of Computer Engineering
Delhi Technological University, NCT of Delhi, India
talibansari96487@gmail.com

Praduman Verma
Department Of Computer Engineering
Delhi Technological University,
NCT of Delhi, India
verma.praduman007@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395784

***Abstract*: CoronaVirus cases have reached 55.6 million (as of 19th Nov 2020) out of which about 9 million cases are in India alone. Out of these 55.6 million people, nearly 1.34 million people have unfortunately lost their life which is about 3 % of the total cases. The disease has become quite a risky infection that has spread quickly across the world and has been announced as a Pandemic by the World Health Organization (WHO). As the greater part of the Covid patients show respiratory manifestations (basically Pneumonia), Chest X-Ray Images can turn into an elective answer to help analyze COVID-19 positive patients from suspected partners. Anyway because of the unexpected episode of the COVID-19 pandemic, radiologists are currently confronting extraordinary weight in analyzing gigantic amounts of X-Ray Images. Other than being tedious and careful, the assessment cycle can become blunder inclined thinking about the huge measure of picture information and the outstanding task at hand. This article focuses on an examination of various procedures identified with the discovery of Coronavirus or pneumonia disease from the X-Ray Images or CT-Scan pictures. It additionally gives an investigation and examination of move learning calculations that were utilized to execute these Coronavirus identification methods.**

***Keywords* - COVID, Transfer learning, Ensemble, VGG16, Resnet, Inception, Image Recognition, Deep Learning.**

I. INTRODUCTION

Coronavirus Disease 2019 (COVID-19) is alarmingly spread worldwide from the time it was first noted and it has been declared as a Pandemic across the world. Timely Detection of COVID-19 is essential for each disease control and patient care.

Table 1: COVID Cases Status (till 19th Nov' 2020)

Country Name	Total Cases	Total Deaths
United States	11.6 Million	250 K
India	8.96 Million	132 K
Brazil	5.95 Million	167 K
France	2.07 Million	46,698
Russia	2.02 Million	34,850

COVID-19 is an acutely treated and highly contagious disease. In severe cases, it may lead to acute respiratory problems or multi-organ failure. The statistics regarding corona cases in the world are shown in Table 1. The laboratory testing technique used for clinical diagnosis of the Corona Virus infection is RT-PCR testing which stands for Reverse Transcription Polymerase Chain Reaction. The COVID-19 RT-PCR test is done for the detection of nucleic acid from the SARS-CoV-2 in lower and upper respiratory compiled samples, using a table compiling strategy. Negative results from the compiled test shouldn't be considered surety. If a

© IEEE 2021. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

patient's physical actions and symptoms are not matched with a negative result then the patient should be accounted for individual testing. Examples with low popular burdens may not be distinguished in example pools because of the diminished affectability of pooled testing. More rapid testing technique is also being used named Rapid Antigen testing. But due to pandemic situations; there is a worldwide scarcity of RT-PCR testing equipments. Also, confirming COVID patients using RT-PCR tests may have high false-negative rates which may put obstacles for the patients to be detected and treated early [1]. Also, the Rapid Antigen testing is not much accurate and hence its reliability is questionable. Therefore, an Automated Detection sample that can constantly distinguish COVID-19 from X-Ray and computed tomography (CT) images can be of great clinical value in fighting the pandemic. Some of the methodologies include Ensemble Convolution Networks, Transfer Learning Algorithms like VGG16, ResNet, Inception. The main purpose behind writing this review paper is analysis and comparison of various methodologies related to the detection of covid infection from X-Ray and CT images. A comparison of transfer learning algorithms like VGG16 and ResNet were discussed. This article is partitioned into various sections: Section II is about related work. Section III is about transfer learning, Section IV includes a comparison of VGG16 and ResNet algorithms. Section V is about Ensemble Convolutional Neural Network (CNN) technique. Section VI includes conclusion, and Section VII consists of future scope.

II. RELATED WORK

In 2020, Hu et al. [1] proposed a method that utilizes the learning-dependent CT radiological scans to classify COVID cases from community-acquired pneumonia (CAP) and non-pneumonia (NP) utilizing the deep neural networks. Community-acquired pneumonia is perhaps the most widely recognized irresistible sicknesses and is a significant reason for mortality and dreariness around the world. Streptococcus pneumonia, Haemophilus influenza, and Moraxella catarrhalis are common bacterial microorganisms that cause CAP. In any case, with the coming of novel demonstrative advances, viral respiratory microbes are progressively being recognized as incessant etiologies of CAP. The most well-known viral microorganisms recuperated from hospitalized patients conceded with CAP incorporate human rhinovirus and flu. Non-Pneumonia (NP) based scans have an observable change image that could be noticed and hence classification can be done and can establish NP-based research alongside CAP.

A. DUAL TRACK RANKING

In the very year of 2020, Yuexiang Li et al. [2] proposed a dual-track ranking approach to deal with unequivocally measuring the estimation of negative examples for network preparation in a disconnected way. The presented approach positions the non-covid examples in the two terms of trouble and variety. Likewise, to remove highlights from the COVID-19 and non COVID CT volumes, a novel administered learning strategy for example Rubik's 3D square star was proposed.

Again in 2020, Jose Ranieri et al.[3] proposed a Multi-view ensemble convolutional neural network to improve characterization of Pneumonia in low contrast pictures. This work assessed the presentation of deep learning technique VGG16 in the discovery of Pneumonia and to recognize bacterial and viral pneumonia.

B. ITERATIVE PRUNING MODEL

Rajaraman et al.[4] (2020) proposed an iterative pruning model for the discovery of COVID from X-Ray pictures to diminish the unpredictability and improve the memory effect of the best-performing profound learning models. Adjusted models were iteratively pruned using several iterations to locate the ideal number of neurons in the convolutional layers to decrease the multifaceted nature of the model with no decrease in its presentation. The best performing pruned models were then chosen to develop a gathering to improve the expectation execution and speculation when contrasted with any individual constituent model. The output systems utilized were max casting a ballot, averaging, and weighted averaging. The algorithms utilized in this model were VGG16, VGG19, Inception V3, Inception ResNet-V2, etc.

III. TRANSFER LEARNING

Transfer Learning is an optimization technique that acts as an easy way to save computational time or get superior performance or accuracy. In this type of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

learning, the models already trained for a task to be reused on another task are called pre-trained models. A pre-trained model is a model that has been previously trained on a dataset and is being used on a different task with the already trained parameters. This saves the training time of the model because pre-trained model is used to train the model again as the already fitted parameters are used. In image classification tasks, commonly used pre-trained models are CNN models like VGG16 or VGG19 or VGG32, GoogleNet, residual networks like Resnet50, etc. Popular examples of pre-trained models are given in Figure 2.

Transfer learning using CNN models can prove to be beneficial in improving the accuracy of image classification tasks. CNN stands for Convolutional Neural Networks. These are the deep neural networks that are used to extract complex features from images using a combination of convolution layers which consist of filters that slide over the image matrix. In detail, the convolutional layers closer to the input layer extract low-level features, then the convolutional layers in the middle extract complex abstract features, and finally, the layers close to the output layer interpret the extracted features in the context of the image classification task.

Pre-trained models utilized for image recognition tasks are the convolutional neural networks previously trained on the ImageNet dataset. ImageNet is an enormous virtual dataset of human clarified pictures that comprises more than 14 million pictures at present and is utilized for research purposes identified with visual article acknowledgment area. A competition named ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is held yearly that utilizes subsets from the ImageNet dataset and is intended to support the advancement and benchmarking of cutting-edge calculations. The basic steps involved in an Image Classification task are shown in Fig1.

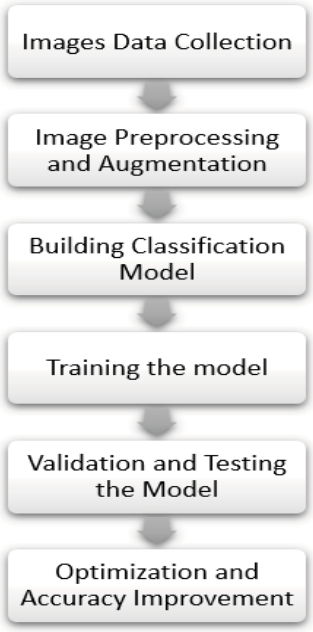


Figure 1: Basic Steps involved in Image Recognition task

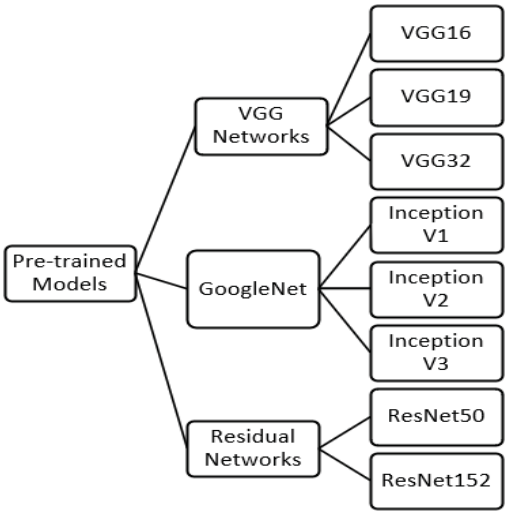


Figure 2: Some Popular Examples of Pre-trained models.

IV. COMPARISON BETWEEN VGG16 AND RESNET

(i) VGG16

It is a computer vision model designed specifically for Image Recognition tasks. It is a convolution neural network (CNN) design that was developed by Simonyan and Zisserman during the ImageNet

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

challenge of 2014 in which it came as a runner-up and 16 in VGG16 mean it has 16 layers that have weights. VGG16 has convolution layers of 3X3 configuration of filters with stride=1 and padding="same". The activation function applied for these convolutional levels is Rectified Linear Unit (ReLu) so that no negative value is passed onto the next layers. It has pooling layers with the Max-Pool function used and the filter size is 2X2 with stride=2. Finally, it has two fully connected layers followed by a Softmax layer for output. So, in total, the network has 13 Convolutional layers, 5 Max pooling layers, 2 Fully connected layers, and 1 SoftMax layer. This network is significantly deep and has about 138 million (approx.) parameters [7].

(ii) **Residual Networks (ResNet)**

A residual network comprises a combination of residual blocks or units which have skip connections. Residual networks are deeper than the VGG networks because it has more layers in the network which makes the network more robust for image recognition tasks. The remaining has become considerably popular for image-related classification tasks because of their ability to conquer the issues of disappearing and detonating inclinations [16] [17]. The issues of exploding and vanishing gradient, skip connections play an important role because by using a skip connection, the context of the previous layers is preserved which prevents the output from previous layers from getting vanished or exploding. Resnet50 is one of the residual networks used frequently for image-related tasks. It consists of 50 layers that have weights and like VGG16, it also uses 3X3 sized filters for convolutional layers.

The performance of deep learning neural networks increases with an increase in-depth of the model. But with increased depth, the problem of vanishing and exploding gradient also arises. ResNet helps to overcome this problem by using a combination of residual blocks. Thus, there are deep networks required for better performance, and ResNet must be used because of their capability to overcome the gradient problems. VGG16 is an accurate model but it is computationally expensive because it is time-consuming and computational resources due to the presence of large numbers of training parameters. Because of VGG16's high accuracy, it is desirable to use it in research where high power computational resources are available and high reliability is required. But for general use, where time is an important and deciding factor, it is not advisable to be utilized [8].

V. ENSEMBLE CONVOLUTIONAL NEURAL NETWORK TECHNIQUE FOR DETECTING PNEUMONIA

Patterns can be automatically detected in low resolution or contrast images using deep neural networks which make beneficial to use when dealing with low contrast X-Ray images for detecting infections like pneumonia and COVID. Image pre-processing techniques also play an important role in such cases. To enhance detection of pneumonic infection, the CNN network with four training ways can prove beneficial and can improve the performance of the model, especially accuracy. To classify these X-Ray images, the CNN model of VGG16 with its fully connected layers replaced by a customized multilayer perceptron layer proved significantly efficient. The image pre-processing techniques that were used were image cropping using U-Net CNN and histogram equalization using the CLAHE method [3].

(i) Image Cropping

Image cropping involves removing the irrelevant regions from the chest X-Ray. This cropping can enhance the learning abilities of the CNN and thus improve the detection ability. U-Net CNN is a pre-trained convolutional network that is often used for cropping images. In the ensemble CNN technique, U-Net CNN was used for cropping the chest cavity from the radiographic chest X-Ray images. The steps of the algorithm are given in Fig 3.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

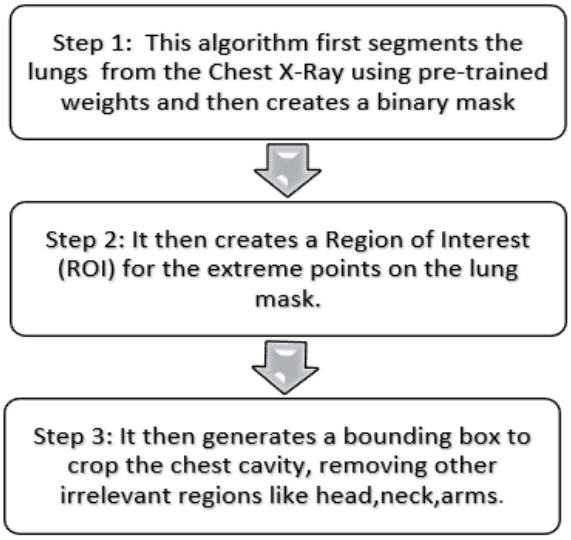


Figure 3: Steps of U-Net CNN Algorithm.

So, the steps shown in Fig3 are the steps involved in the procedure of the U-Net CNN algorithm. In Step-1, lungs are first cropped from the X-Ray image with the help of pre-trained parameters and then a binary mask is created. Then, a Region of Interest is marked using the corner points of the mask. Finally, the chest cavity is cropped using the Region of Interest by eliminating the body parts that are not relevant to our model like head, shoulders, etc. In case of large opacities in the lungs area, U-Net may not be able to detect one of the two lungs. To detect and handle these cases, the condition used was: When the breadth of chest cavity in Binary mask (wref) is not greater than w/2 (w= X-Ray image width), wref is considered miscalculated and recalculated as wref=w-2a, where a is the smaller distance between the image's an adjacent border and extreme points of the region of interest (ROI)[10][11].

(ii) Histogram Equalization

Histogram equalization is a process used to improve the contrast of the image so that the image looks more balanced and stable. A histogram is the plot of Frequency versus Intensity. The frequency corresponding to an intensity value means the number of pixels of the current image having that intensity value. High Contrast images are the images in which most of the image pixels are having high-intensity values due to which the histogram is mostly concentrated in high intensity regions. And, Low Contrast images are the images that have most of the image pixels in the low-intensity range. So during this equalization procedure, the contrast of the image pixels is adjusted so that the spread of intensity values i.e. the Histogram is balanced. Adaptive Histogram Equalization (AHE) is different from simple histogram equalization in the way that AHE computes several histograms of the image with each histogram corresponding to a particular section of the image. AHE then uses these histograms to adjust and redistribute the contrast values of the image [13-15].

To increase the contrast of the projections of low contrast tissues, the Ensemble CNN approach used Histogram equalization by Contrast Limited Adaptive Histogram Equalization (CLAHE) technique. CLAHE method is an advanced method to improve the balance of an image by limiting contrast enhancement at a pre-decided value (0.01) on the histogram. The problem with ordinary AHE is that it over amplifies the contrast in near-constant regions that is the regions in which the intensity values are nearly constant. Because the histogram in these sections is highly concentrated, it can lead to amplification of noise in those areas. CLAHE is a type of Adaptive equalization in which amplification of contrast is limited to avoid amplification of noise. [12].

With this Ensemble CNN architecture, four different training strategies were evaluated: (1) original chest X-Ray image, (2) Cropped image (with chest cavity) (M), (3) CLAHE equalized cropped image (N). (4) The last strategy was implemented based on an ensemble between strategies M and N. In this ensemble strategy, the strategies M and N, making a multi-view were frozen and their outputs combined in the first multi-layer perceptron layer. The training of the network and the extra multi-layer perceptron layer of ensemble strategy was done with 30 and 10 iterations respectively. These four training strategies, including the ensemble, were evaluated by the technique of stochastic gradient descent in 16 image batches per iteration. The optimizer used was RMSprop optimizer with 0.0001 as a learning rate.

The metrics or parameters used to compare the performance of the four strategies were:- area under

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

the Receiver Operating Characteristic curve (ROC curve) Accuracy and F1-score. The metrics values are depicted in Table 2.

Table 2: Comparison of Accuracy and F1-Score

Author	Accuracy	F1 Score
Kermany[13]	0.928	-
Liang[14]	0.905	0.927
This Work (Jose [3])	**0.974**	**0.979**

VI. CONCLUSION

The automated detection of COVID infection or pneumonia from chest X-Ray Images or CT-Scans will prove beneficial in coming months and years because a vaccine for COVID is yet to be discovered and given the rapid spread of the infection, soon the automated systems will become a need of the hour for reducing the workload of health care workers. This article discussed different existing approaches and techniques to detect infection from chest X-Ray Images or CT-scans in detail. It also illustrated the technique of transfer learning and analyzed the performances of commonly used pre-trained models- VGG16 and ResNet. The datasets used in the approaches and the comparison of the metrics of these approaches are displayed.

The comparative study of the existing approaches and methods of infection detection domain shows that using the ensemble CNN approach yields the maximum accuracy because of the evaluation of four different training strategies.

The weakly supervised deep learning approach [1] effectively-recognized COVID-19 cases from CAP and NP cases precisely. Yet, one of the limitations of this system is that it isn't discriminative enough with regards to isolating Community-acquired pneumonia from COVID-19. This is because of the restricted limit of the spine Convolutional Network (CNN). Accordingly, this issue can be resolved by utilizing further developed designs like Inception or Resnet.

The Dual Track Ranking approach's exploratory outcomes indicated that in examination with vanilla organization preparing plan utilizing all accessible negative examples, this methodology had the option to accomplish unrivaled execution utilizing just 50% of the negative examples for preparing the COVID-19 order organization, considerably diminishing the preparation season of the model. But two principle challenges that were experienced while building up this PC-supported determination framework for Covid-19 utilizing CT checks were: the amount of COVID-19 CT volumes is inadequate, contrasted with the accessible negative examples, experiencing the class irregularity issue. Because of the Corona Pandemic, there is restricted time for the preparation and tuning of network parameters.

In the Iterative Pruning Model, it was seen that for the two-fold grouping task, all the models were completely precise with VGG16 having the least number of boundaries while in multi-class characterization, Inception V3 was more exact.

From the comparison of VGG16 and ResNet algorithms, a conclusion can be derived that when there is no restriction on the training time and computational resources, VGG16 performs exceptionally well in terms of accuracy, and in the cases when the depth of the neural network is increased, ResNet algorithms are the more reliable networks given their ability to prevent the problems of vanishing and exploding gradient.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

Table 3: Datasets Used

Author	Approach	Dataset Used
Hu[1]	Weakly supervised Learning	450 Chest CT images (150 each of COVID-19, CAP and NP patients).
Li[2]	Dual Track Ranking	305 Chest CT images of COVID-19 positive patients and 2,370 CT images for CAP and NP patients.
Rajaman [4]	Iterative Pruning model	Pediatric CXR Dataset, RSNA CXR dataset, Twitter COVID-19 CXR dataset, Montreal COVID-19 CXR dataset
Jose[3]	Ensemble CNN	Guangzhou Pneumonia Chest X-ray dataset

Table 4: Comparison of Approaches for Infection Detection.

Author	Approach	Accuracy	F1-Score
Li[2]	Dual Track Ranking	0.84	0.848
Rajaraman (Inception V3) [4]	Iterative Pruning Model	0.9683	0.968
Hu[1]	Weakly supervised Learning	0.81	-
Jose[3]	Ensemble CNN	0.974	0.979

VII. FUTURE SCOPE

There are numerous unexplored areas in the infection detection domain, as well as the transfer learning domain, that need to be explored in order to address the challenges of a lack of COVID-19 or pneumonia CT scans compared to available non-infected samples, as well as the availability of limited time to train and tune the model. The futuristic approach must also include the exploration of more ensemble strategies, as research has shown that ensemble approaches outperform conventional approaches in terms of accuracy. The iterative Pruning technique also needs to become an important topic of research to develop more simplified and more memory efficient deep learning models [4]. Exploring more image pre-processing techniques and time-wise evolution of techniques like U-Net CNN and the CLAHE method is also desirable. Therefore, this concludes that currently infection detection domain is an unfamiliar domain to an extent and comparatively less explored domain. Thus, there is enormous scope for research and development of new more accurate and reliable techniques. Also, there is great scope for advancement and improvement of the existing techniques. In the future, this can be beneficial and instrumental to those researchers who have been researching in the field of infection detection domain.

VIII. REFERENCES

[1] Hu, S., Gao, Y., Niu, Z., Jiang, Y., Li, L., Xiao, X., … Yang, G. (2020). Weakly Supervised Deep Learning for COVID Infection Detection and Classification from CT Images.
[2] Yuexiang Li*, Don Wei*, Jiawe Chen*, Shili Cao, Hongyu Zhou, Yanchun Zhu, Jianrong Wu, Lan Lan, Wenbo Sun, Tianyi Qian, Kai Ma, Haibo Xu, Yefeng Zhng.(Dual Track Ranking).
[3] Jose Raniery Ferrira Junior, Diego Armndo Cardona Cardens, Ramon Alfredo Moreno, Marina de Fatima de S´a Rebelo ´ , Jose Eduardo Krieger ´ 1 , Marco Antonio Gutierrez1. (Ensemble CNN)
[4] Rajaraman, S., Siegelman, J., Alderson, P. O., Folio, L. S., Folio, L. R., & Antani, S. K. (2020). (Iterative Pruning).
[5] 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE-2020), 07-08 February 2020, (IEEE Conference Record # 48199)
[6] S.J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, pp. 1345– 1359, 2010.
[7] Karen Simonyan, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." International Conference on Learning Representations (2015).
[8] Agarwal, T., & Mittal, H. (2019). Performance Comparison of Deep Neural Networks on Image Datasets. 2019 Twelfth International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2019.8844924
[9] Wei Zhao, Zheng Zhong, Xingzhi Xie, Qizhi Yu, and Jun Liu. "Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study." American Journal of Roentgenology (2020): 1-6.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2015, pp. 234–241
[11] I. Pazhitnykh and V. Pets Uk, "Lung segmentation (2D)," https://github.com/imlab-uiip/lung-segmentation-2d, 2017. (U-NET Part 2)
[12] K. Zuiderveld, Graphic Gems IV. Academic Press Professional, Inc., Cambridge, Massachusetts, USA, 1994. (CLAHE)
[13] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122–1131, 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

[14] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," Comput Methods Programs Biomed, 2019, DOI:10.1016/j.cmpb.2019.06.023.
[15]Chen, Jun, Lianlian Wu, Jun Zhang, Liang Zhang, Dexin Gong, Yilin Zhao, Shan Hu et al. "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study." medRxiv (2020).
[16] Dhaya, R. "Deep net model for detection of covid-19 using radiographs based on roc analysis." Journal of Innovative Image Processing (JIIP) 2, no. 03 (2020): 135-140.
[17] Kumar, T. Senthil. "Video based Traffic Forecasting using Convolution Neural Network Model and Transfer Learning Techniques." Journal of Innovative ImageProcessing (JIIP) 2, no. 03 (2020): 128-134.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:25:59 UTC from IEEE Xplore. Restrictions apply.

VIA MEDICA

LETTER TO THE EDITOR

Int Marit Health
2021; 72, 1: 76–77
10.5603/IMH.2021.0010
www.intmarhealth.pl
Copyright © 2021 PSMTTM
ISSN 1641-9251
eISSN 2081-3252

COVID-19 vaccine on board ships: current and future implications of seafarers

Gopi Battineni[1], Shailender Kumar[2], Mamta Mittal[3], Francesco Amenta[1, 4]

[1]Telemedicine and Telepharmacy Centre, School of Medicinal and Health Products Sciences, University of Camerino, Italy
[2]Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India
[3]Department of Computer Science and Engineering, G.B. Pant Government Engineering College, Okhla, New Delhi, India
[4]Research Department, International Radio Medical Centre (C.I.R.M.), Rome, Italy

The impact and effect of novel coronavirus disease 2019 (COVID-19) are expanding all over the world without boundaries. When it began in China in late 2019, it was seen to be not as scary and devastating as it has become today. However, the hope for overcoming the pandemic lies in finding an effective treatment and inventing a vaccine. If we look at the history of the spread of this contagion, within a short span of seven months it has spread all over the world. Statistics are saying that more than 104 million individuals got infected and 2.2 million deaths were currently occurred [1].

In the last few months, there has been a massive revelation in the medical domain regarding the outbreak of COVID-19 and its vaccination [2]. It has been truly amazing to see scientists and doctors come together across all over the world to share their expertise to develop vaccination in the past few months. Thanks to the collaboration between researchers and scientists are always way ahead to take up challenges and bear the torch of a bright and prosperous future as far as the COVID-19 struggle is concerned. To channelise their potential and encourage them to develop solutions by joining hand together in this pandemic time, the government has also been supportive of various initiatives and funding. Medical doctors sharing their experiences and the patients who have recovered from this near-fatal disease largely help in vaccine development. This brought a whole new perspective on how to handle the current situation the world is facing with the COVID-19 pandemic.

Seafarers are unsung heroes of this pandemic because shipping plays an important role in this serious outbreak. Unfortunately, because of the global emergency that happened in the last year, seafarers went through the depressive symptoms and faced various neuropsychological and psycho-affective alterations [3]. To slow down the rate of onboard infection, various shipping companies provided some serious guidelines such as social distancing at working places, self-hygiene, and room quarantine for any suspicious cases. For some seafarers, this situation is unbearable. Because of the novel pandemic, tens of thousands of sailors were stuck on board for more than a year now. Despite the fact of crew changes have been made conceivable in certain nations, the implementation stays exceptionally challenging. These situations demand immediate access to the vaccine for seafarers.

The speakers in the CrewConnect Global Virtual Event 2020 highlighted that, with some vaccine trials now showing promising results, sailors need to be treated as fundamental workers and be among the first to get vaccine [4]. “I think all the work we’ve done in lobbying governments now needs to be even more concerted effort about the point that seafarers are essential workers... we need to find a way to find a way to access vaccines” said Mr Stephen Cotton, general secretary of International Transport Workers Federation [4].

Some countries, such as Singapore, already started vaccination for seafarers and it becomes one of the first nations to give importance for COVID-19 vaccine to frontline workers of maritime. According to the Maritime and Port Authority of Singapore, seafarers need to undergo fewer testing before getting vaccinated. It is scheduled that at least 10,000 seafarers can be vaccinated by the end of January 2021 under air vaccination exercise [5].

COVID-19 is considered a type of pneumonia which can be effectively treated by vaccination. All national bodies have to step forward and take Singapore as an example in the provision of vaccine for frontline workers such as seafarers. On other hand, the International Maritime Or-

Dr. Gopi Battineni, Telemedicine and Telepharmacy Centre, School of Medicinal and Health Products Sciences, University of Camerino, Italy, e-mail: gopi.battineni@unicam.it

This article is available in open access under Creative Common Attribution-Non-Commercial-No Derivatives 4.0 International (CC BY-NC-ND 4.0) license, allowing to download articles and share them with others as long as they credit the authors and the publisher, but without permission to change them in any way or use them commercially.

ganization also urges that maritime workers and seafarers have to obtain COVID-19 vaccination on a priority basis to allow them to work freely and sustain dynamic global supply chain.

ACKNOWLEDGEMENTS

This paper was supported in part by the ITF Trust grant No. 1508/2020 to Centro Internazionale Radio Medico (C.I.R.M.).

REFERENCES

1. Home — Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/ (Feb 02, 2021).
2. Li Q, Lu H. Latest updates on COVID-19 vaccines. Biosci Trends. 2021; 14(6): 463–466, doi: 10.5582/bst.2020.03445, indexed in Pubmed: 33390384.
3. Battineni G, Sagaro GG, Chintalapudi N, et al. Assessment of awareness and knowledge on novel coronavirus (COVID-19) pandemic among seafarers. Healthcare (Basel). 2021; 9(2), doi: 10.3390/healthcare9020120, indexed in Pubmed: 33503921.
4. CrewConnect, Shipping needs to ensure seafarers prioritised for vaccines. https://www.seatrade-maritime.com/regulation/shipping-needs-ensure-seafarers-are-prioritised-covid-19-vaccines (Feb 02, 2021).
5. Singapore to prioritise maritime workers in Covid-19 vaccination effort | TradeWinds. https://www.tradewindsnews.com/people/singapore-to-prioritise-maritime-workers-in-covid-19-vaccination-effort/2-1-940231 (Feb 02, 2021).

The Visual Computer
https://doi.org/10.1007/s00371-021-02100-x

ORIGINAL ARTICLE

Crowd anomaly detection with LSTMs using optical features and domain knowledge for improved inferring

Mohammad Sabih[1] · **Dinesh Kumar Vishwakarma**[1]

Accepted: 23 February 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
With the increasing population, the probability of occurrence of different kinds of crowd anomalies gets frequent. Blockage on roads, the lighting condition, and the uneven movement of humans and vehicles makes it a tough and challenging problem. The paper proposes the combined use of a convolutional neural network and bidirectional LSTM to solve the task. CNN helps extract frame-level features of the optical flow over the video evaluated by the Lucas Kanade algorithm. A novel approach of improving the predicted class with the domain knowledge of datasets is also performed. The proposed methodology is tested on the crowd anomaly dataset's benchmark datasets, namely UCSD Ped-1 and UCSD Ped-2, and it outperforms various other existing state-of-the-art methods.

Keywords Optical flow · Bidirectional LSTM · Regularizers · CNN · Domain knowledge

1 Introduction

Anomaly detection is one of the most interesting and diverse fields of research in the present era. It has helped the generation capture better the pattern flow of videos. It is a binary classification problem and extracts anomalous instances within a video series; they are the ones that are not normal for an observer. The anomaly can be detected on busy roads, in-hospital reports, bank cheques, LIC policies etc. In general, the occurrence of anomaly is very rare so we progressively try to study the behaviour and pattern of normal instances, and any input which surpasses a threshold of mismatch we proclaim to have found an anomaly. A set of anomalous features obtained for one sector cannot be generalized for all. For example, driving a car over some speed limit on a street road may be anomalous, but it'd be totally normal for a highway. With the increasing population and abstract and rare road rage, it is important to lay focus and improve the anomaly detection methods over the crowded streets to ensure the mass's safety. The branch of computer vision comes in handy, providing fruitful information using machine learning algorithms much needed. The information obtained is likely to have a huge data imbalance, which makes it tough to design strong working neural networks. Second lighting condition plays a significant role too, and also the idea of the presence of anomaly is not crystal clear and is very vague. The main contribution of the work is as follows:

- An improved end-to-end experiment-driven supervised learning approach using a combination of LSTM and CNN to solve crowd anomaly detection by extracting frame-level features of the optical flow over the video evaluated by Lucas Kanade.
- Novel post-processing based inferring method incorporating the distribution of optical flow features and domain knowledge after the raw predictions of LSTM to reduce false predictions is developed, which gives improved accuracy.
- Comparison between varieties of LSTM networks using different regularizers and batch size.
- It is reducing the effect of class-imbalance in the supervised method by using oversampling methods, adding noise to LSTM states, and tweaking loss function.

✉ Dinesh Kumar Vishwakarma
dinesh@dtu.ac.in

Mohammad Sabih
thegenius411@gmail.com

[1] Biometric Research Laboratory, Department Information Technology, Delhi Technological University, Delhi 110042, India

Published online: 02 April 2021

2 Related work

Throughout the last decade, many methods have been proposed for anomaly detection over a publicly available dataset. Researchers mostly rely on unsupervised learning that is extrapolating an anomaly after studying just the features of non-anomalous (normal) occurrence, any incoming anomaly will have different features, and various existing classifiers can be used to predict it. But the problem of crowd anomaly usually needs temporal information, and it gets tough constructing a temporal context classification network. Initially, the focus was laid on crowd density analysis [1] and object trackers [2, 3] to detect an anomaly. The tracklet obtained can be used as a feature vector, and any occurrence which deviates much from the normally averaged tracklet can be counted as an anomaly. But these methods fail if the density is too much or there's a blockage in the crowd [4]. There was a need for high-level features; hence, the technique of optical flow was effectively used. Optical flow helps extract the spatial features over the video samples. It can be calculated using Lucas Kanade [5] or SIFT [6] algorithm and various other methods. With the unique flow features of video frames, different design feature vectors can be constructed, for example, the histogram of orientation. Optical flow is calculated with a combination of frames so in a way it even helps in reducing the dimension of the dataset.

Convolutional neural networks are known to show unique and exceptional results in different computer vision fields. COV-nets can also be used to extract features for anomaly detection. But since the problem of anomaly detection needs temporal information, the idea of solving the problem just with COV-nets gets much more complex [7]. COV-nets can be combined with LSTM (long short term memory) to obtain a much stronger temporal information extractor over the dataset. LSTM is an artificial neural network used in the field of deep learning; it has the capability of processing an entire input sequence altogether. LSTM networks can be combined with convolution neural networks, and very high-level features can be learned over the video samples. Zhuang et al. [8] uses LSTM combined with a CNN over the raw dataset frames. Optical flow features can also be used instead of raw images for better classification. Autoencoders [9] or PCA based unsupervised learning is used to build a generative model which classifies an anomalous frame on the basis of its reconstruction error for the frame's optical flow features.

For the proposed methodology, initially, experiments were conducted, and Inceptionv3 was used to extract features of data frames, but the accuracy was not good and temporal information was much needed. A dummy model was trained for the classification of optical flow features alone, this was much better than the inceptionv3 classification, but still, it wasn't good enough. A sequential model combining Inceptionv3 and LSTM was also built, whose results are discussed in Sect. 4 of this paper. A variety of LSTM networks were tested, and conclusions were made regarding the batch size and regularizers. Section 3 of this paper discusses the novel approach of inferring over the predicted samples. It helps reduce the false positive rate of predictions and hence making the classification network better.

3 Proposed methodology

A novel inferring approach over the predictions obtained from bidirectional LSTM combined with a convolutional neural network is proposed. It is motivated by the shortcomings in the related work, specifically addressing false negatives and the lack of an extensive experiment-driven architecture. The proposed methodology makes the use of optical flow features obtained from the well-known Lucas Kanade algorithm to extract frame-level spatial information. The bidirectional LSTM captures the temporal information over the dataset. The main constituents of the proposed architecture are (1) formation of optical feature matrix, (2) CNN combined with bidirectional LSTM, (3) novel inferring. Figure 1 shows the workflow of the proposed methodology. The following subsections describe each step in detail.

3.1 Optical flow

In the branch of computer vision, Lucas Kanade is extensively used for computing optical flow. The method assumes that the disturbance caused in the coordinates of image contents between two nearby instances is negligible around a point. Thus, the equation of flow must satisfy all pixels around the periphery of that point. The velocity vectors V_x, V_y obtained must satisfy Eq. (1):

$$I_x(q_n)V_x + I_y(q_n)V_y = -I_t(q_n), \tag{1}$$

where q_n is the nth pixel inside the window and I_x, I_y and I_t are the partial derivatives of x, y, t of image I calculated at point q_n. The equation can be written in the form of a matrix, as shown in Eq. (2):

$$AV = B, \tag{2}$$

$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \vdots \\ I_x(q_n) & I_y(q_n) \end{bmatrix}, \tag{2.1}$$

$$B = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(q_n) \end{bmatrix}, \tag{2.2}$$

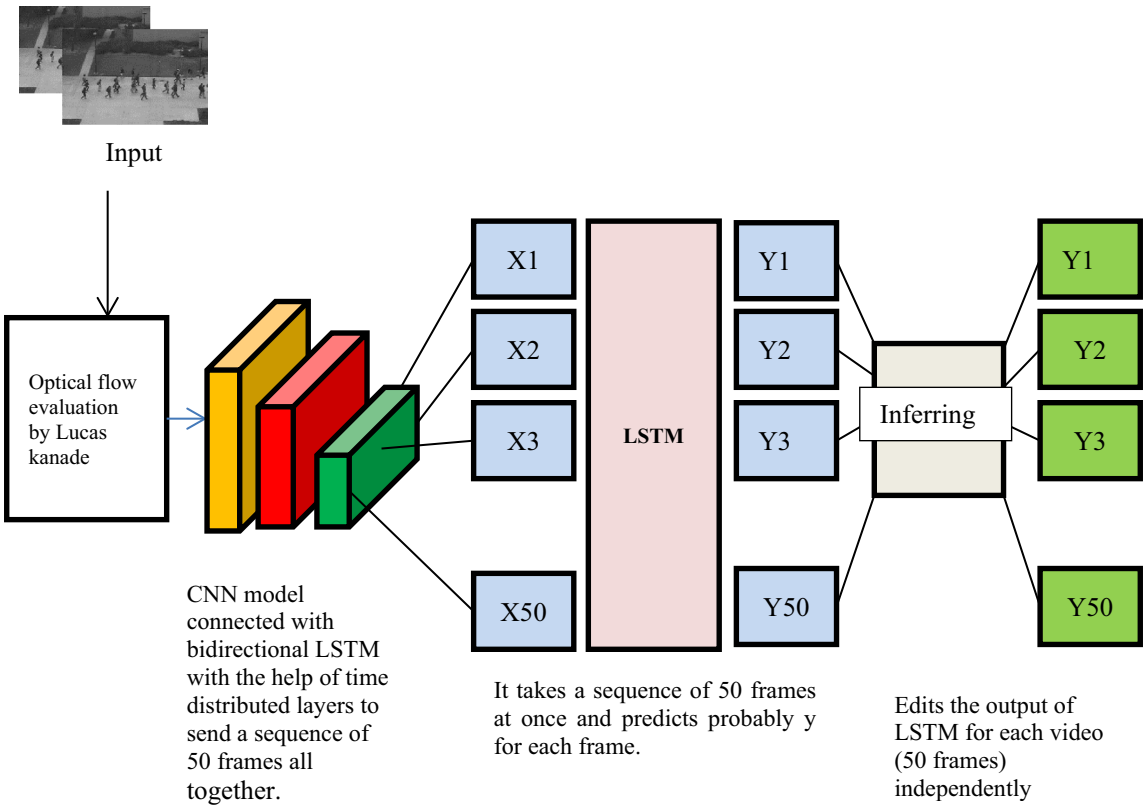


Fig.1 The work flow of the proposed methodology

$$V = \begin{bmatrix} V_x \\ V_y \end{bmatrix}. \tag{2.3}$$

The Lucas–Kanade uses least squares principle to obtain the solution. It solves the following equation,

$$A^{\mathrm{T}}AV = A^{\mathrm{T}}B, \tag{3}$$

where A^{T} is the transpose of matrix A. From this equation V, therefore, V_x, V_y are obtained.

3.1.1 Data preprocessing

For the optical flow feature matrix evaluation, only those velocity vectors are considered whose magnitude is greater than 0.3 obtained from Lucas Kanade algorithm, which is working on every two consecutive frames of the dataset. A Histogram of orientation and magnitude is made using these vectors. The image frame is divided into 9 patches (image regions) and a separate histogram is made for each patch. Over all the patches, 9 different orientations, which are $(0 \text{ to } 20)°, (20 \text{ to } 40)°, (40 \text{ to } 60)°, (60 \text{ to } 80)°((80 \text{ to } 90)°$ and $((0) \text{ to } (-10))°$, $((-10) \text{ to } (-30))°, ((-30) \text{ to } (-50))°, ((-50) \text{ to } (-70))°$, and $((-70) \text{ to } (-90))°$ are considered for evaluation. A counter is made for each orientation in each batch and it is constantly incremented with the magnitude of velocity which is in its bounded orientation. Therefore, a matrix of dimension $[9 \times 9]$ is made for every two consecutive frames in the dataset.

3.2 CNN

The CNN model is used to capture frame level features. Figure 2 shows the architecture of CNN Model. Conv2D-1 has 32 filters with a kernel size of (5×5) and Conv2D-2 has 16 filters with a kernel size of (3×3). Since the input to the CNN network is the optical flow which itself has condensed information down to 9×9 size, pooling and more Conv2D layers are avoided. Both convolutional layers have 'relu' activation. Flatten layer is added followed by a batch normalization layer to increase the speed, stability and performance of CNN.

3.3 LSTM

Long short term memory recurrent neural networks (LSTM) successfully solves the problem of vanishing gradients [10] for RNNs. The architecture of LSTM contains blocks which

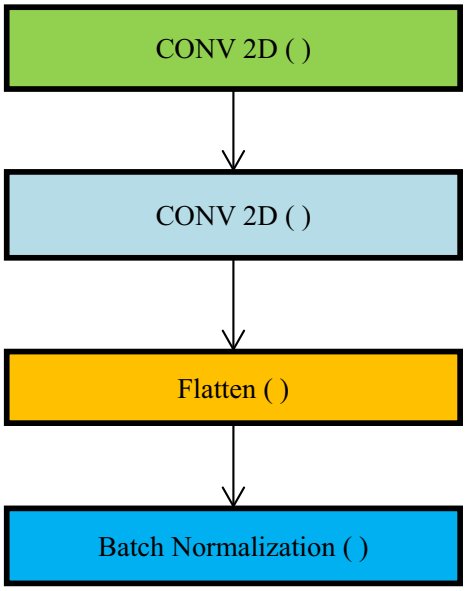


Fig.2 Architecture of CNN model

store memory. There are self-connected memory cells within each memory block and three multiplicative units, which are the input, output and forget gates. The architecture of RNN and LSTM is almost similar just the units in the hidden layers of RNN are replaced by memory blocks. The hidden layer function H in LSTM is implemented as per the Eqs. (4), (5), (6), (7) and (8):

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right), \tag{4}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right), \tag{5}$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + W_{hc}h_{t-1} + b_c\right), \tag{6}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right), \tag{7}$$

$$h_t = o_t \tanh(c_t). \tag{8}$$

The paper proposes to initialize the cell state, i.e. c_0 from Eq. (6) in a noisy way as shown in Eq. (9):

$$c_0 = N(0, 1). \tag{9}$$

This represents sampling c_0 from a Normal distribution with $\mu = 0$, $\sigma = 1$, so as to reduce over fitting and class-dominance. This is motivated by the fact that if c_0 is initialized to 0 as is done normally, then the high loss observed in initial time steps for any given sample (during training) would force the model to "compensate" for this by appropriately modifying the corresponding weights to reduce this high initial time-step loss, which would in-turn reduce the importance of the latter time-steps inputs on the loss-function. This can be alleviated if the model is forced to not over-fit on the initial time steps this way by adding noise to the initial cell state as Eq. (9) suggests.

The weights are initialized by sampling from an normal distribution and the bias terms are initialized to 0. The paper additionally "carefully" initialized the forget gate bias, i.e. b_f to a vector of 1 s and not 0 s so that forget gate stays active right from the initial epochs of training, thereby reducing training time.

i,f and o are the input, forward and output gates, respectively. c is the vector of activation for the cell and σ is the logistic sigmoid function. W represents the weight matrix and b the bias terms. Bidirectional RNNs have the ability to make predictions by extracting information from past as well as future context during training time [11]. The information extracted from both directions is fed forward to evaluate the output; therefore there are two hidden layer sequence (h_f) and backward (h_b). The time-dependent evaluation of the output sequence is shown in Eq. (10), (11) and (12):

$$h_{f_t} = H\left(W_{xh_f}x_t + W_{h_f h_f}h_{f_{t+1}} + b_{h_f}\right), \tag{10}$$

$$h_{b_t} = H\left(W_{xh_b}x_t + W_{h_b h_b}h_{b_{t+1}} + b_{h_b}\right), \tag{11}$$

$$y_t = W_{h_f y}h_{f_t} + W_{h_b y}h_{b_t} + b_y. \tag{12}$$

Bidirectional LSTM is the combination of BRNN and LSTM, and henceforth they are able to make predictions using context from past as well as future samples during training time, hence boosting learning during training time. The paper uses a sequence of 50 frames as an input to the bidirectional LSTM. The LSTM models the probability distribution, $p(yi|x_1, x_2, x_3 \ldots x_{50})$. Figure 3 shows the architecture of bidirectional LSTM. "Sigmoid function", was used as the activation layer.

3.4 Training

For training the complete network is trained end-to-end, i.e. the weights of the LSTM network are optimized using backpropagation through time (BPTT) algorithm (gradient descent) frequently used in RNNs. Additionally, in the same backward pass, the weights of the CNN network are optimized in tandem using simple gradient descent. This strategy ensures that the features extracted by the CNN are directly relevant for sequence classification by the LSTM and vice-versa.

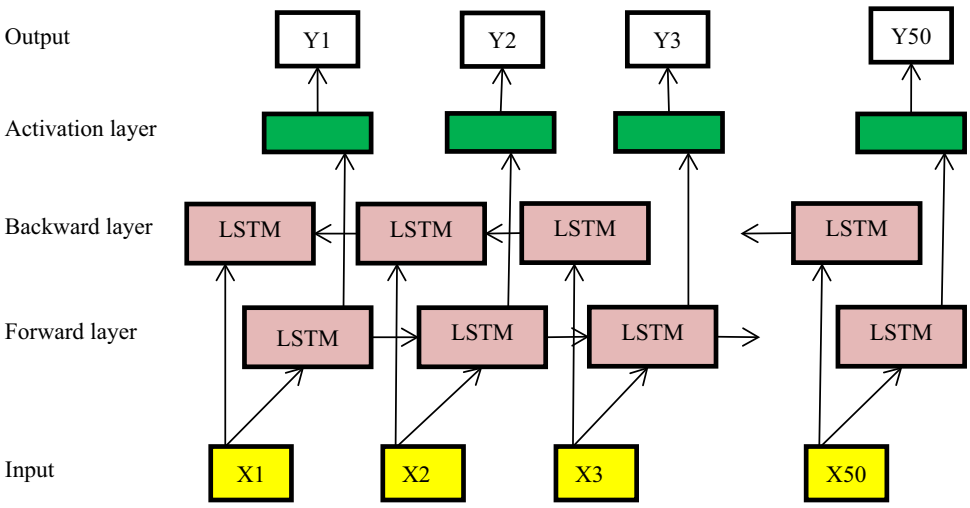


Fig. 3 The architecture of bidirectional LSTM

For loss optimization, stochastic gradient descent (SGD) optimizer with a learning rate of 0.01 and momentum of 0.9 is used. Training sample parameters are updated as per the Eq. (13):

$$\theta = \theta - \eta \cdot \nabla_{\theta} J\left(\theta; x_i; y_i\right), \tag{13}$$

where η is the learning rate and J is the loss function, the loss function taken for training is, "binary cross entropy" it calculates the loss of input by computing the following average loss value given by Eq. (14):

$$\text{Loss} = -\frac{\sum_{i=1}^{50} \left\{ y_i \log\left(\hat{y_i}\right) + \left(1 - y_i\right) \cdot \log\left(1 - \hat{y_i}\right) \right\}}{50}, \tag{14}$$

where $\hat{y_i}$ and y_i are the predicted scalar and probable class value of x_i, respectively. L2 regularizer is also used hence a new regularization term is added in the Loss function. The new loss function is given by Eq. (15):

$$\text{Newloss} = \text{Loss} + \frac{\lambda}{2.50} \times \sum ||w||^2. \tag{15}$$

Here, λ is the regularization parameter. Apart from these two values v and u are also calculated where v is the average value of the optical feature vector of anomalous frames and u of the normal frame. They are given by Eqs. (16) and (17):

$$v = \frac{\sum_{j=1}^{n} v_j}{n}, \tag{16}$$

$$v_i = \frac{\sum_{i=1}^{n} x_i}{81}, \tag{16.1}$$

$$u = \frac{\sum_{j=1}^{m} u_j}{m}, \tag{17}$$

$$u_i = \frac{\sum_{i=1}^{m} x_i}{81}, \tag{17.1}$$

where x_i and x_j is the flattened form of optical flow feature matrix obtained from Lucas Kanade, therefore having dimensions $[81 \times 1]$ and n and m are the total number of anomalous and normal frames, respectively, over the training dataset. These values are later used in novel inferring steps.

3.5 Novel inferring

For a given input sequence of 50 frames, i.e. a video, the LSTM produces a probability of that frame belonging to the anomalous class (after the sigmoid activation). At each time step, the hidden state is used as a context to produce the next probability of the next frame. This iterative process is realized through a simple forward pass through the time of the bidirectional LSTM for the given collection of 50 frames. After the raw inference, the paper proposes a novel post-processing step to reduce false positive and false negative introduced in the predicted. In the dataset domain it is observed between every two anomalies there are at least 4 normal frames; similarly, it is observed that every anomaly at least lasts for 4 frames. The novel inferring algorithm produces two arrays namely y_1 and y_2. The y_1 array handles the occurrence of negative class ($y = 0$) in between the positive class and the y_2 array manages the false predictions of positive class ($y = 1$) in between the negative class. For a given continuous sequence of "1's" in a video, there shouldn't be 3, 2 or 1 predicted "0's" in between. Also, there shouldn't be 3 or less predicted "1's" in the middle of a continuous sequence of "0's" in a video. Figure 4 shows the edits made by an array y_1 and y_2 over the predictions y done by the proposed methodology. To further

Fig. 4 Edits made by array y_1 and y_2 over the predicted frames sequence, y

reduce false predictions, a vector v_x is obtained given by Eq. (18):

$$v_x = \frac{\sum_{i=1}^{81} x_i}{81}. \tag{18}$$

Here x_i is the flattened form of the optical flow matrix obtained from Lucas Kanade method. The value of v_x is compared with the value of v and u calculated before in Sect. 3.4. If y_1 of x_i is "0" and $\boldsymbol{v_x}$ magnitude is greater than some thresh(decided by v and u) then y_1 corresponding to this x_i is made "1", i.e. a possible anomaly sample which was misclassified by the LSTM as a false negative is correctly converted to a true positive sample, a similar process is followed for entries of y_2 and false positives are converted to true negatives. Table 1 shows the dimensional map of the proposed methodology.

Table 1 Dimensional flow of the proposed methodology

Layer	Output dimension
Input	$N \times 9 \times 9 \times 1$
CNN (CONV2D 1)	$N \times 50 \times 50 \times 5 \times 5 \times 32$
CNN (CONV2D 2)	$N \times 50 \times 50 \times 3 \times 3 \times 16$
CNN (FLATTEN)	$N \times 50 \times 144$
Bi-LSTM	$N \times 50 \times 100$
Bi-LSTM batch norm	$N \times 50 \times 100$
Dense	$N \times 50 \times 1$

4 Experiments and results

In order to test the performance of the proposed methodology, an experiment is conducted on publicly available datasets, namely UCSD (PED-I) and UCSD (PED-2). The UCSD-PED dataset [14] was built by a stationary camera

mounted at some height over a pedestrian walkway. The crowd density in both of these datasets is constantly varying with time. Anomaly is introduced whenever a forbidden (non-pedestrian) entity like car, bike etc. is observed in the crowd or with a sudden change in direction and speed of the pedestrians. Attached with each clip in the dataset is a binary ground truth annotation flag which is 1 if the frame is anomalous and 0 otherwise. A comprehensive analysis of the proposed method in comparison with other state of art methods is presented.

4.1 Dataset

The proposed methodology was tested on UCSD Ped-1 which has pedestrians moving towards and away from the camera plane; it contains 34 training videos composed of all normal (non-anomalous) frames and 36 testing video samples which contain the anomaly. Next, there's UCSD Ped-2 dataset which has pedestrians moving parallel to the camera plane, it contains 16 training video samples and 12 testing video all available in the form of frames just like in UCSD Ped-1. Table 2 shows the division of the dataset. The sample images of the dataset are as shown in Fig. 5.

4.1.1 Training–testing division

For the train-test split, the criteria of (75–25)% were considered overall available videos from each dataset. For training 27 and 12 non-anomalous videos and 29 and 10 anomalous videos were taken from UCSD Ped-1 and Ped-2 dataset, respectively. Similarly for testing 7 and 4 non-anomalous videos and 7 and 2 videos containing anomaly were taken from the UCSD Ped-1 and Ped-2 dataset, respectively.

4.2 Feature dimensions

There will be N optical flow matrices for a video having $2N$ frames. A matrix of dimension $[N \times 9 \times 9 \times 1]$ is formed for every two consecutive frames. A sequence of 50 such frames is taken as 1 video in Bi-Directional LSTM. Therefore a matrix of dimensions $[M \times 50 \times 9 \times 9 \times 1]$ is obtained, where $N = M \times 50$.

Table 2 Division of dataset

Dataset	Normal video	Anomalous video	Frame dimension
UCSD Ped-1	34	36	238×158
UCSD Ped-2	16	12	360×240

4.3 Parameter setting

It was observed that with the increase in batch size the accuracy decreases across all architecture types of LSTMs. We interpret this due to class imbalance in the data. In particular, large batch-sizes causes unstable gradient updates during training of the network (e.g. there would be many batches containing samples of only the dominant class. This could be alleviated if the training set is shuffled and the "shift" of weights toward the dominant class is compensated by reducing batch size). Hence imbalanced datasets should be trained with the smallest possible batch size. Figure 6 shows the AUC curve obtained for the UCSD Ped-1 dataset trained till 15 epochs over an LSTM network.

It can be seen from Fig. 6 with the increase in batch size the AUC curve suffers from a downfall, hence it is better to use the smallest possible batch size.

L2 regularizers of different values were also tested on LSTM architecture. When UCSD Ped-1 dataset is trained till 30 epochs with different regularizes, it is observed that in the absence of regularizer the model tends towards over fitting and with a regularizer of 0.1 it performs good and with regularizer of 0.06 it shows much greater accuracy compared to what it displayed with regularizer 0.1, hence regularizer 0.06 turns out to be the best for UCSD Ped-1 dataset.

The LSTM network was also tested over different lengths of sequence and the best results were obtained on a 50 frame video sequence. Therefore, the proposed methodology takes a sequence of 50 frames as one input in the bidirectional LSTM network.

4.4 Results

In initial experiments, the raw images of UCSD Ped-1 dataset were tested on a variety of LSTM networks after extracting features from inceptionV3. A sequence of 50 frame video was passed as one input in the LSTM. Thus, a feature vector of dimensions $[108 \times 50 \times 2048]$ was produced for training and $[28 \times 50 \times 2048]$ for testing. Here 2048 represents the feature matrix given by Inceptionv3. Table 3 shows the results obtained with different LSTM networks.

Fig. 5 Sample images of UCSD Ped-1 and UCSD Ped-2 dataset

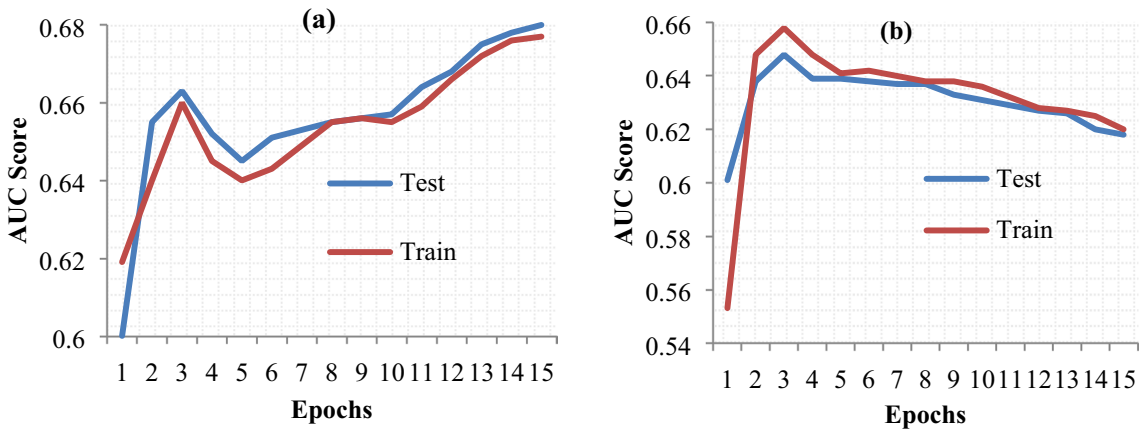


Fig. 6 AUC curve of batch size 4 (**a**), batch size 32 (**b**), of a network trained over the UCSD Ped-1 for 15 epochs

Table 3 Results obtained with different LSTM networks

Network	Two stacked LSTM'S	Single bidirectional LSTM	Two stacked bidirectional LSTM
Validation accuracy (%)	~76	~77	~75

It is clear from the above observation that the model needed some high-level temporal information to show better results; hence optical flow features are used. Figure 7 shows the histogram of orientation and magnitude as described in Sect. 3.1.1, for a normal and anomalous frame. It can be observed from Fig. 7 that with an incoming anomaly the histogram of the patch region in which the anomaly is present shows a marked difference in magnitude.

The results of the proposed methodology before and after the novel inferring are discussed in this section. Table 4 shows the results obtained on the proposed methodology before inferring. It is seen that for UCSD Ped-1 regularizer with of 0.06 gives the best results, and for UCSD Ped-2 dataset regularizer of 0.01 is needed for better results.

4.5 Improving class-imbalance problem

It was observed that adding class weight reduced the false positive and increased the recall by 10% but simultaneously increased the false negative to some extent thereby decreasing the accuracy; the same thing was observed when oversampling was done over the dataset. For over sampling the optical flow features of anomalous videos were stacked in the reverse order in the dataset, therefore, increasing the anomalous frames and balancing the dataset to some extent. We also experimented with changing the loss function to directly optimize the objective, i.e. the AUC score. In particular, a new proxy loss function was developed to calculate the F1 score given by Eq. (19):

$$\text{new_loss} = \frac{2}{1/\text{prob}_{\text{precision}} + 1/\text{prob}_{\text{recall}}}, \tag{19}$$

$$\text{prob}_{\text{precision}} = \frac{(\sum_{i=1}^{\text{tp}} P(y_i))}{(\sum_{i=1}^{\text{EE}} P(y_i)) + (\sum_{j=1}^{\text{fp}} P(y_j))}, \tag{19.1}$$

Similarly,

$$\text{prob}_{\text{recall}} = \frac{(\sum_{i=1}^{\text{tp}} P(y_i))}{(\sum_{i=1}^{\text{tp}} P(y_i)) + (\sum_{j=1}^{\text{fn}} P(y_j))}, \tag{19.2}$$

$P(y_i)$ is the probability of the sample being positive as predicted by the final activation layer (Sigmoid) (in the case of true or false positives) or the probability of the sample being negative as predicted by the network (in the case of false negatives or true negatives).

tp = #true positive samples (i.e. in a 50 frame sequence, how many samples were truly positive).

fp = #false positive samples (i.e. in a 50 frame sequence, how many samples were false positive as predicted by the model).

fn = #false negative samples (i.e. in a 50 frame sequence, how many samples were falsely predicted as negative by the model).

Care was taken to use probabilities instead of the raw indicator values, i.e. instead of using, $\sum_{i=1}^{\text{tp}} 1\{P(y_i > 0.5\}$, (where $1\{<\text{Condition}>\}$ is the usual indicator function which takes the value 1 when $<\text{Condition}>$ is true, 0 otherwise), it was replaced with probabilities so that the overall loss function is differentiable.

Finally, all these techniques qualitatively seemed to improve the number of false positives and negatives in the

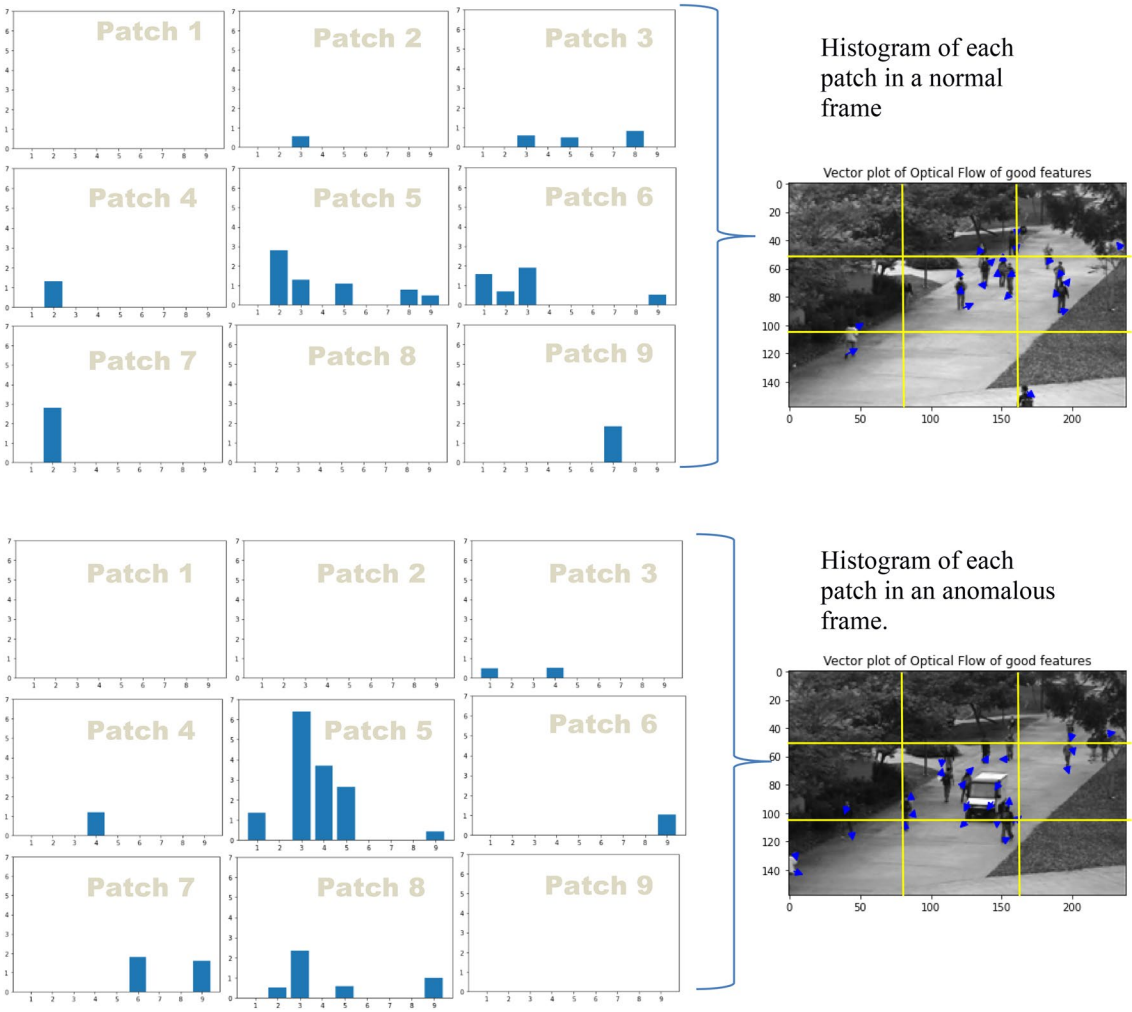


Fig. 7 The histogram of orientation and magnitude for a normal and anomalous frame. The x axis represents the nine orientations, and y axis represents the optical flow magnitude

Table 4 Accuracy, AUC score, training time with different regularizers

Dataset	Regularizer	Training accuracy	Testing accuracy	AUC score	Training time
UCSD Ped-1	–	0.8341	0.7785	0.8274	618.29 s
	0.01	0.8474	0.8407	0.9087	597.97 s
	0.06	**0.8081**	**0.8564**	**0.9483**	**598.90 s**
UCSD Ped-2	–	0.8268	0.8644	0.9325	155.23 s
	0.01	**0.8131**	**0.9222**	**0.9650**	**160.74 s**
	0.06	0.8342	0.9020	0.9450	161.39 s

The bold value shows the best combination of parameters

final predictions, yet the overall accuracy wasn't hampered much. The results obtained over the dataset after doing the novel inferring are shown in Table 5. And Table 6 shows the improvement in the confusion matrix.

It is clear from Table 6 that the novel method of inferring has helped improve the results obtained by the proposed methodology. We can clearly see from the confusion

Table 5 Results obtained after novel inferring

Dataset	Testing accuracy	AUC score
UCSD Ped-1	0.87	0.9483
UCSD Ped-2	0.9378	0.9650

Table 6 Changes in the confusion matrix before and after novel inferring

Dataset	Novel inferring	True(+ve) frames	True(−ve) frames	False(−ve) frames	True(+ve) frames
UCSD Ped-1		968	39	162	231
		952	55	127	266
UCSD Ped-2		291	24	11	124
		287	28	0	135

matrix that our proposed method increases the number of true positives across both datasets by reducing the number of false negatives which is more important than false positives. Figure 8 shows instances in UCSD Ped-1 and Ped-2 where proposed novel inferring helped improve the prediction. It is observed from Fig. 8 that sometimes the anomalous entity (bicycle) gets behind some object because of which there's not much change in the pixel value it constitutes, thereby diminishing its optical flow vector magnitude and hence getting predicted as a normal frame, it is noticed the novel inferring helped cover up these cases.

In general, the positive samples have higher values of v score (as defined Sect. 3.4) and hence this aids in making the predictions better. Although the number of false positives has increased, this is due to the fact that for those samples, both v and u scores are the same as well as the model probabilities are also the same. We have analyzed and deduced that the model misclassified these negative samples falsely as positive because of the fact that the boundary for differentiation is loose and it is difficult for the human eye too to classify them as anomalous. Although, this proposed novel method uses insights from domain data, however, the method of making automatic edits to the predicted vector (*as described in* Sect. 3.5) is very general since in real life scenario it is very unlikely for an anomaly to have a discontinuity of few frames as it continues for atleast some frames. Similarly, it is pretty uncommon in a normal situation (non-anomalous) for an anomaly to pop in between for a few frames. Table 7 shows the comparison of the proposed methodology with the existing state of the art methods, taking AUC score as the criteria. Revathi et al. [17] stands for deep learning based anomaly detection; it uses 4 modules to solve the task, one generates a unique background that is a high level image frame, next object tracking module is used, followed by a feature extractor module, followed by a feature extractor module and finally classification is done with

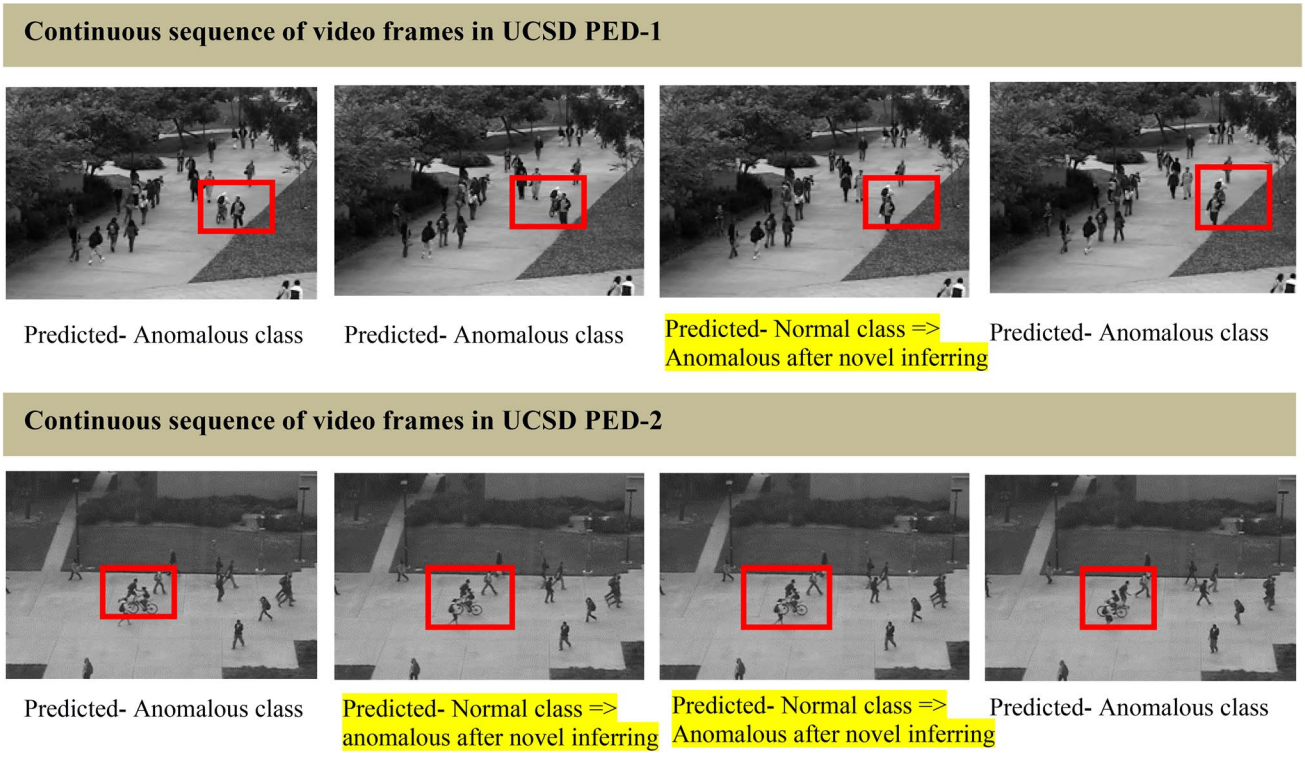


Fig. 8 Instances in UCSD Ped-1 and Ped-2 where novel inferring helped

Table 7 Comparison of AUC score with the existing state of art methods

Method	UCSD Ped-1	UCSD Ped-2
Feng et al. [12]	0.925	–
AMDN [13]	0.921	–
MDT [14]	0.818	–
Xu et al. [15]	0.854	0.882
Cong et al. [16]	0.471	0.868
DLAD [17]	0.554	0.726
AnamolyNet [18]	0.835	0.949
SCG-SF [19]	0.909	0.902
Khan et al. [20]	0.8108	0.9375
AOE [7]	0.946	0.959
Proposed	0.9483	0.9650

deep learning. Khan et al. [20] is a non-deep learning based methodology where super pixel are evaluated over the image frames. Feature of each super pixel is calculated which even depends on the surrounding super-pixels if they both share the same directional characteristics, classification of these features is done with a Gaussian based k-means algorithm. From the Table 7, it is clear the proposed methodology has performed good in comparison with the existing state of the art methods.

5 Conclusion

The paper has proposed and implemented a novel supervised methodology for frame level anomaly detection which uses bidirectional LSTM combined with convolution neural network over the optical features and then the domain knowledge for improved inferring. The architecture of the methodology is simple as compared to a lot of other existing methods moreover it provides good accuracy by reducing the number of false predictions which makes it more reliable for real time applications. Although the proposed method is not strictly real-time as the paper currently makes use of 50 frames to be known in-order to detect anomalous events in these frames. This corresponds to a delay of roughly 1.7 s (assuming real-time FPS is 24). However, as the first batch of 50 frames have arrived; afterwards immediate decisions can be made on the future frames (by sliding the window of 50 frames, one position in time). The method is supervised so if an extraordinary anomaly drops in which it has not seen before, it won't be able to predict this occurrence anomalous which is the backdrop of the proposed method. To further improve, the accuracy under such scenarios, deep learning based optical flow estimation method can be implemented. This will also help in aiding the raw optical flow calculation under some scenarios such as occlusion of objects in a frame, under which deep learning based optical flow calculation will not falter.

References

1. Hettiarachchi, A.L., Thathsarani, H.O., Wickramasinghe, P.U., Wickramasuriya, D.S., Rodrigo, R.: Extensible video surveillance software with simultaneous event detection for low and high density crowd analysis. In: 7th International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, pp. 1–6 (2014). https://doi.org/10.1109/ICIAFS.2014.7069590
2. Piciarelli, C., Micheloni, C., Foresti, G.L.: Trajectory-based anomalous event detection. IEEE Trans. Circuits Syst. Video Technol. (2008). https://doi.org/10.1109/TCSVT.2008.2005599
3. Tung, F., Zelek, J.S., Clausi, D.A.: Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. Image Vis. Comput. (2011). https://doi.org/10.1016/j.imavis.2010.11.003
4. Sabokrou, M., Fayyaz, M., Fathy, M., Klette, R.: Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans. Image Process. (2017). https://doi.org/10.1109/TIP.2017.2670780
5. Xie, S., Zhang, X., Cai, J.: Video crowd detection and abnormal behavior model detection based on machine learning method. Neural Comput. Appl. (2019). https://doi.org/10.1007/s00521-018-3692-x
6. Ojha, N., Vaish, A.: Spatiooral anomaly detection in crowd movement using SIFT. In: 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, pp. 646–654 (2018). https://doi.org/10.1109/ICISC.2018.8398878
7. Singh, K., Rajora, S., Vishwakarma, D.K., Tripathi, G., Kumar, S., Walia, G.S.: Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets. Neurocomputing **371**, 188–198 (2020). https://doi.org/10.1016/j.neucom.2019.08.059
8. Zhuang, N., Ye, J., Hua, K.A.: Convolutional DLSTM for crowd scene understanding. In: Proc.—2017 IEEE Int. Symp. Multimedia, ISM 2017, vol. 2017-Janua, pp. 61–68. https://doi.org/10.1109/ISM.2017.19 (2017)
9. Dhole, H., Sutaone, M., Vyas, V.: Anomaly detection using convolutional spatiotemporal autoencoder. In: 2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019, pp. 1–5. https://doi.org/10.1109/ICCCNT45670.2019.8944523 (2019)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997). https://doi.org/10.1162/neco.1997.9.8.1735
11. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. (1997). https://doi.org/10.1109/78.650093
12. Feng, Y., Yuan, Y., Lu, X.: Learning deep event models for crowd anomaly detection. Neurocomputing (2017). https://doi.org/10.1016/j.neucom.2016.09.063
13. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. In: British Machine Vision Conference (BMVC), pp 8.1–8.12 (2015). https://doi.org/10.5244/c.29.8
14. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. (2014). https://doi.org/10.1109/TPAMI.2013.111
15. Xu, D., Song, R., Wu, X., Li, N., Feng, W., Qian, H.: Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing **143**, 144–152 (2014). https://doi.org/10.1016/j.neucom.2014.06.011

16. Cong, Y., Yuan, J., Tang, Y.: Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE Trans. Inf. Forensics Secur. **8**(10), 1590–1599 (2013). https://doi.org/10.1109/TIFS.2013.2272243
17. Revathi, A.R., Kumar, D.: An efficient system for anomaly detection using deep learning classifier. Signal Image Video Process. **11**(2), 291–299 (2017). https://doi.org/10.1007/s11760-016-0935-0
18. Zhou, J.T., Du, J., Zhu, H., Peng, X., Liu, Y., Goh, R.S.M.: AnomalyNet: an anomaly detection network for video surveillance. IEEE Trans. Inf. Forensics Secur. (2019). https://doi.org/10.1109/TIFS.2019.2900907
19. Chu, W., Xue, H., Yao, C., Cai, D.: Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos. IEEE Trans. Multimed. **21**(1), 246–255 (2019). https://doi.org/10.1109/TMM.2018.2846411
20. Khan, M.U.K., Park, H.S., Kyung, C.M.: Rejecting motion outliers for efficient crowd anomaly detection. IEEE Trans. Inf. Forensics Secur. **14**(2), 541–556 (2018). https://doi.org/10.1109/TIFS.2018.2856189

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

Mohammad Sabih is under graduate student of B.Tech. Electrical Engineering. He is working as Research Intern at Biometric Research Laboratory, Delhi Technological University, Delhi. His research interest includes Machine Learning, Deep Learning and Neural Networks, Image Analysis, End-to-End System Development.

Dinesh Kumar Vishwakarma (M'16, SM'19) received Doctor of Philosophy (Ph.D.) degree in the field of Computer Vision from Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India, in 2016. He is currently a Professor with the Department of Information Technology, Delhi Technological University, New Delhi. His current research interests include Computer Vision, Deep Learning, Sentiment Analysis, Fake News Detection, Crowd Analysis, Human Action and Activity Recognition. He is a reviewer of various Journals/Transactions of IEEE, Elsevier, and Springer. He has been awarded "Premium Research Award" by Delhi Technological University, Delhi, India in 2018.)

Design and Analysis of LFSR based Content Addressable Memory

Vyom Garg
Department of Electronics and Communication
Delhi Technological University
New Delhi, India
vyom250999@gmail.com

Parangat Mittal
Department of Electronics and Communication
Delhi Technological University
New Delhi, India
parangatmittal1@gmail.com

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397071

***Abstract*— This paper presents a Linear Feedback Shift Register (LFSR) based Content Addressable Memory (CAM) that focuses on reducing the area compared to a combinational design of the memory. A simple sequential counter-based memory architecture can also be used for the same, but, an LFSR based CAM architecture increases the randomness and thus liable to acquire better access time within the memory. The design majorly consists of a Reconfigurable LFSR and an encoder module coupled with a ROM to configure the LFSR. Further, a counter is used for writing data sequentially into the memory array. The proposed architecture is simulated on Icarus Verilog to validate the hypothesis and achieved the hypothesised improvements in memory access times.**

Keywords*— *Content Addressable Memory, Linear Feedback Shift Register, Counter.

I. Introduction

As the volume of data increases in the system, the need for large memories also arises. With large quantities of data in the memory, the speed of operation usually gets compromised. A Content Addressable Memory (or Associative Memory) [1] is the approach to achieve a greater speed. For a supplied input data word, it is the one that returns the address/es where the input word was located in the memory array. It is called "Content-Addressable" since it addresses its data by contents. As compared to conventional methods of addressing, this provides a relatively higher searching speed and system performance. While a standard Random-Access Memory (RAM) returns a data located at a particular address, a CAM returns the address containing a specific data, allowing it to search through its entire contents in a single clock cycle. Being a completely hardware-based implementation of associative arrays, CAMs are massively parallel search engines allowing them to match the memory contents with the searched pattern simultaneously in multiple chunks of data.

For the same reason, CAMs are often used in search-intensive applications high-speed applications like network routers [2], lookup tables [3]-[4], pattern matching, data compression [5], parametric curve extraction [6], Hough transformation [7], Huffman coding/decoding [8],[9], image coding [10], and digital signal processors. These are also used widely as single-cycle associative search accelerators. Many architectures [11],[12] have been proposed to implement an associative memory. The most common architecture is having dedicated combinational matching circuitry [13] for memory cells. This significantly increases the hardware overhead. Another architecture in use is the one with a sequential counter. The counter generates input address patterns sequentially, and at each address, the data located is compared with the data searched. This paper presents an alternate associative memory architecture employing a Linear Feedback Shift Register (LFSR). [14],[15]

A Linear Feedback Shift Register is a sequential circuit consisting of D Flip Flops and exclusive OR gates, arranged linearly. For an N-bit LFSR, N refers to the degree of the characteristic polynomial or the output length. The characteristic polynomial of the LFSR defines the arrangement of the exclusive OR gates in the network. On each clock edge, the circuit generates an N-bit value based on the value generated on the previous clock edge. The initial value supplied to the LFSR is called the seed, and it controls the further values being generated as per the characteristic equation. LFSR is widely used in Automatic Test Pattern Generation (ATPG) [16] and Built-in Self-Test (BIST) [17] applications in the field of testing of digital circuits. This property of generation of random outputs by an LFSR is used in the proposed associative memory architecture. This LFSR is used to generate random addresses for the memory locations, data at which is compared with the searched data. The randomness brought in by the LFSR is proposed to increase the search speed as opposed to a sequential counter which would check for the searched data at each of the memory locations consecutively.

This paper presents an architecture for a parameterised Linear Feedback Shift Register with configurable width to generate random patterns of different lengths, and a sequential associative memory architecture utilising the LFSR mentioned above to increase the search speed of the memory. The paper is organised into five sections with the inclusion of this introduction. In Section II, the architecture for a Reconfigurable LFSR has been discussed. This is followed by the proposed CAM architecture in Section III. Thereafter, Section IV contains the results of the experiment performed. The paper is finally concluded in Section V.

II. Previous Works

A. Linear Feedback Shift Register

LFSRs are widely used in several applications, which is why many hardware and software-based implementations of the Linear Feedback Shift Register logic are available. A number of studies [18] have been carried out, proposing and evaluating

ISBN: 978-1-7281-8529-3/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:10:13 UTC from IEEE Xplore. Restrictions apply.

the various architectures for optimised physical parameters like utilisation and power analysis.

In [19], the reversible LFSR architecture has been proposed to reduce the number of reversible gates and constant inputs, and generate the least garbage outputs. The study has presented architectures for SISO, PISO, SIPO and PIPO shift registers. [20] has focused on the implementation of a configurable LFSR. The proposed LFSR is configurable in terms of the number of bits, by configuring the tap positions in the shift register and seed value of LFSR.

In [21], a high-speed pipelined architecture for an LFSR has been proposed. The design was found to have a higher throughput rate and a reduced critical path. [22] has proposed a novel parallel architecture based on transposed LFSR wherein the output value is determined not from the input value and register value, but the feedback value. The proposed architecture has a significant improvement in the Area-Time product.

B. Content Addressable Memory

Many architectures [23] for Content-Addressable Memory or Associative Memory have been put forward in various studies to counter the main design challenges of area, speed, reliability and noise margin. Application-specific architectures catering to a particular use case with modifications in the search and update [24] algorithm have also been proposed.

Most of the CAMs are based on unit storage bit cells. These implementations typically result in an increased footprint of the architecture. This problem has been addressed in [25],[26]. The speed of a CAM is limited due to two main delay elements - matching algorithm delay and priority encoder delay. Various proposals have been put forward to minimise these delays, such as in [27] and [28] respectively. The reliability of memory is governed by its underlying technology. CAM architectures based on the new-age technologies like CNTFETs [29] and FinFETs [30] are in prominence.

III. LFSR with Configurable Width

As mentioned in the introduction, an LFSR based CAM has been proposed to induce randomisation in the searching process. However, unlike sequential counter-based searching, the LFSR will search in the unfilled locations because of random searching of all filled and unfilled locations in the memory. Therefore, as shown in Fig. 1 an adaptable LFSR using additional AND logic gates has been developed whose size varies corresponding to the number of bits required to address the filled locations in the memory.

To access all the locations within the memory, the feedback polynomial for LFSR must be a characteristic polynomial. Table.1 presents the characteristic polynomials for all possible degrees required for a 16-bit addressable memory has been presented. A decoder logic was designed to enable and disable the feedback paths corresponding to the characteristic equations, fulfilling the requirement for searching all the filled.

TABLE.1 CHARACTERISTIC POLYNOMIAL FOR 1-16-BIT LFSR

Degree(n)	Characteristic Polynomial
1	$X + 1$
2	$X^2 + X + 1$
3	$X^3 + X + 1$
4	$X^4 + X + 1$
5	$X^5 + X^2 + 1$
6	$X^6 + X + 1$
7	$X^7 + X + 1$
8	$X^8 + X^6 + X^5 + X + 1$
9	$X^9 + X^4 + 1$
10	$X^{10} + X^3 + 1$
11	$X^{11} + X^1 + 1$
12	$X^{12} + X^7 + X^4 + X^3 + 1$
13	$X^{13} + X^4 + X^3 + X + 1$
14	$X^{14} + X^{12} + X^{11} + X + 1$
15	$X^{15} + X + 1$
16	$X^{16} + X^5 + X^3 + X^2 + 1$

*DEGREE=NUMBER OF BITS FOR ADDRESSING FILLED LOCATIONS OF MEMORY

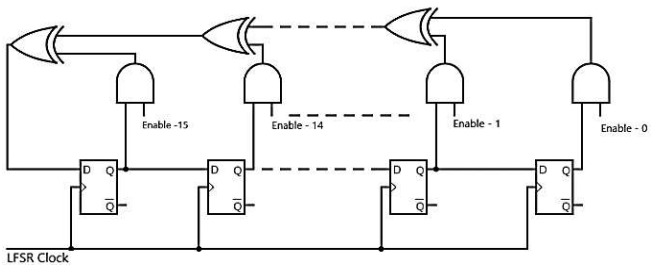


Fig. 1. LFSR with Taps in Feedback

locations within the memory. The width of the LFSR is truncated from lesser significant bits since while shortening from more significant bits requires bi-directional feedback elements. Finally, the output was shifted appropriately using a barrel shifter to align with the address lines of the memory.

IV. Proposed Memory Architecture

The LFSR Module described in the previous section is used with the associative memory module to generate addresses to search in the memory for the given input data. The proposed memory performs its operations as follows:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:10:13 UTC from IEEE Xplore. Restrictions apply.

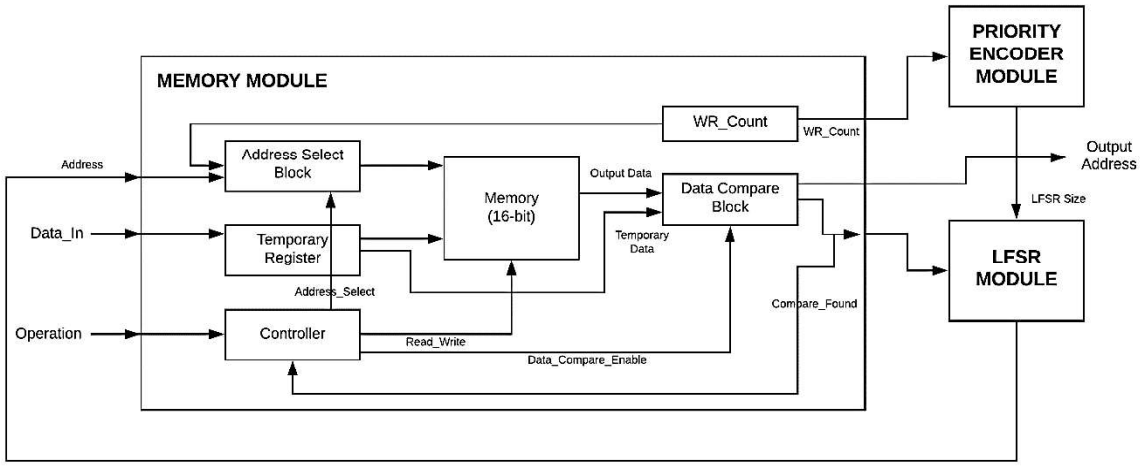


Fig. 2. Block Diagram for the Proposed Architecture

A. Writing Data to Memory:

The associative memory has an address size of 16-bits, i.e., can store $2^{\wedge}16$ locations. For the write operation, the memory module is provided with input data, and that data is stored at the sequentially next available location. However, at any given instant, not the whole of the memory would be populated by data, and generating addresses beyond the last-filled location would be useless. Hence, a signal indicating the terminal address is also generated during the write operation. The terminal address is also indicative of the bit-length that the LFSR should generate addresses of. The terminal count is fed to a Priority Encoder, which decides the length of the LFSR polynomial and correspondingly enables the necessary feedback taps.

B. Reading Data from Memory:

The queried data is latched into an internal register. Soon enough, the LFSR operation is triggered. The length of the address the LFSR must generate is supplied by the Priority encoded terminal count of the last write operation. The values generated are used as input addresses for the memory module. The memory reads the data at the generated addresses and compares it with the latched data. The LFSR operation is suspended as soon as the two data match. The address at which the data found in memory and the latched input data matches is the final output and is retained in the LFSR as seed for the next search operation.

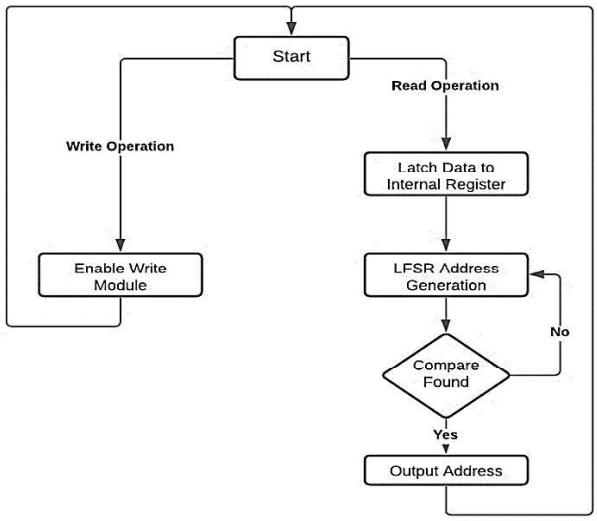
C. Controller Block:

To achieve the primary purpose of speeding up the data search, as mentioned in previous sections, the LFSR is operated at a much higher frequency than the whole driver system. This gives the LFSR sufficient time to generate all the possible addresses before any other operation is performed on the memory.

The whole memory module as shown in Fig. 2 is governed by a Controller block, which conditionally enables or disables the different subsidiary blocks present in the module such as Address Select Block and Data Compare Block, based on the operation being performed on the memory. The Controller consists of a Finite State Machine of 5 states - Start, Write Data, Latch Data, LFSR Address Generation and Output Address Generation, and 3 control signals - Address Select, Data_Compare_Enable, and Compare_Found. The state transition diagram for the same is given in Fig. 3.

V. Results

The proposed architecture was simulated on Icarus Verilog and GTKWave for verification of the design. For a write operation, the Write Count is updated, and if required, the number of bits needed to address the filled locations is also updated as in Fig 4(a). The Encoded value for bits required causes the LFSR to reconfigure the taps in its feedback path for the prospective search operation. Fig. 4(b) and Fig. 4(c) illustrate the search operation of the proposed LFSR memory. The clock for the FSM was taken at 10 MHz while the LFSR generates addresses at 1 GHz. The synchronous external read signal is applied, as shown in Fig 4(b), and latches the data to be searched in the memory.

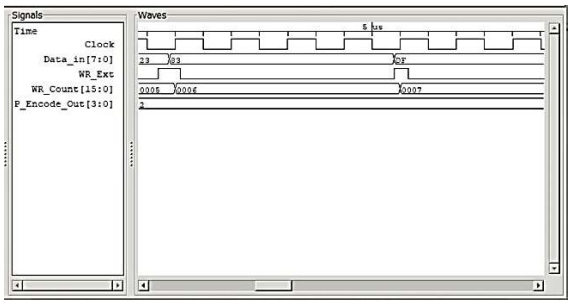


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:10:13 UTC from IEEE Xplore. Restrictions apply.

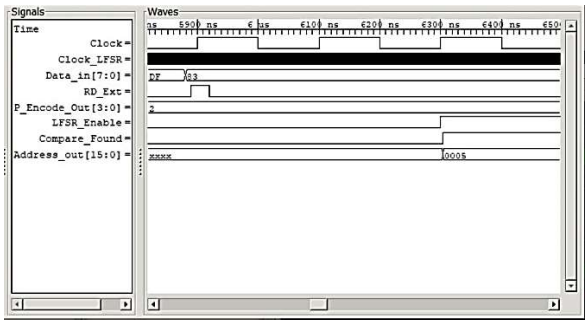
Fig. 3. FSM for the Memory Operation

The Fig. 4(b) and Fig. 4(c) show the duration of the search, i.e., time taken from the read signal's detection until the generation of the 'Compare Found' signal hence triggering latching of the LFSR output to the output of the memory unit. The basic operation of this memory was tested by partially populating it, and the comparison with counter-based content addressable memory was made by performing a large number of iterations on a completely populated memory.

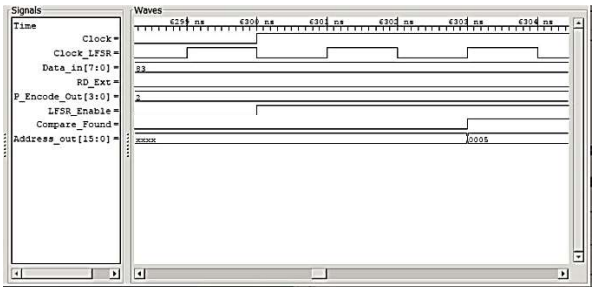
The LFSR based CAM was proposed to reduce the access times when compared to synchronous counter-based design and to prove the result approximately 1,370,000 iterations were performed on both the memory designs. Out of the 1.3 Million iterations that were performed, the LFSR based architecture performed better in roughly 50.2% of the total cases, and 0.4% cases the time taken by both the setups was nearly the same. Therefore, the LFSR based CAM outperformed the Counter based design approximately 10,500 times. The maximum advantage that the LFSR based design had over the counter-based design amounted to 508 nanoseconds.



(a)



(b)



(c)

Fig. 4. (a) Simulation Window for Write Operation (b) Simulation Window Read operation from Read Signal to output Address (c) Simulation Window showing Time lag between LFSR Enable and Compare Found

VI. Conclusions

As evident from the statistical results, the proposed Design performs better on the ground of memory access times in the majority of the situations though narrowly beating the sequential design based on the synchronous counter. However, it can be speculated that the application of other LFSR, as discussed in the introduction can further improve the results. A lot of work has been done on LFSR structures to obtain greater randomisation, and if the standard LFSR employed in the proposed design is replaced with one with greater randomisation, there can be a substantial improvement in the performance parameters.

References

[1] A. M. S. Abdelhadi and G. G. F. Lemieux, "Modular SRAM-Based Binary Content-Addressable Memories," 2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines, Vancouver, BC, 2015, pp. 207-214, doi: 10.1109/FCCM.2015.69.

[2] Nen-Fu Huang, Whai-En Chen, Jiau-Yu Luo and Jun-Min Chen, "Design of multi-field IPv6 packet classifiers using ternary CAMs," GLOBECOM'01. IEEE Global Telecommunications Conference (Cat. No.01CH37270), San Antonio, TX, 2001, pp. 1877-1881 vol.3, doi: 10.1109/GLOCOM.2001.965900.

[3] T. -. Pei and C. Zukowski, "VLSI implementation of routing tables: tries and CAMs," IEEE INFCOM '91. The conference on Computer Communications. Tenth Annual Joint Conference of the IEEE Computer and Communications Societies Proceedings, Bal Harbour, FL, USA, 1991, pp. 515-524 vol.2, doi: 10.1109/INFCOM.1991.147548.

[4] A. J. McAuley and P. Francis, "Fast routing table lookup using CAMs," IEEE INFOCOM '93 The Conference on Computer Communications, Proceedings, San Francisco, CA, USA, 1993, pp. 1382-1391 vol.3, doi: 10.1109/INFCOM.1993.253403.

[5] C. -. Lee and R. -. Yang, "High-throughput data compressor designs using content addressable memory," in IEE Proceedings - Circuits, Devices and Systems, vol. 142, no. 1, pp. 69-73, Feb. 1995, doi: 10.1049/ip-cds:19951443.

[6] M. Meribout, T. Ogura and M. Nakanishi, "On using the CAM concept for parametric curve extraction," in IEEE Transactions on Image Processing, vol. 9, no. 12, pp. 2126-2130, Dec. 2000, doi: 10.1109/83.887981.

[7] M. Nakanishi and T. Ogura, "A real-time CAM-based Hough transform algorithm and its performance evaluation," Proceedings of 13th International Conference on Pattern Recognition, Vienna, Austria, 1996, pp. 516-521 vol.2, doi: 10.1109/ICPR.1996.546878.

[8] Komoto, Homma and Nakamura, "A high-speed and compact-size JPEG Huffman decoder using CAM," Symposium 1993 on VLSI Circuits, Kyoto, Japan, 1993, pp. 37-38, doi: 10.1109/VLSIC.1993.920528.

[9] Liang-Ying Liu, Jhing-Fa Wang, Ruey-Jen Wang and Jau-Yien Lee, "CAM-based VLSI architectures for dynamic Huffman coding," in IEEE Transactions on Consumer Electronics, vol. 40, no. 3, pp. 282-289, Aug. 1994, doi: 10.1109/30.320807.

[10] S. Panchanathan and M. Goldberg, "A content-addressable memory architecture for image coding using vector quantisation," in IEEE Transactions on Signal Processing, vol. 39, no. 9, pp. 2066-2078, Sept. 1991, doi: 10.1109/78.134438.

[11] V. R. Datti and P. V. Sridevi, "Performance Evaluation of Content Addressable Memories," 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2018, pp. 596-598, doi: 10.1109/ICRITO.2018.8748808.

[12] S. Ruan, C. Wu and J. Hsieh, "Low Power Design of Precomputation-Based Content-Addressable Memory," in IEEE Transactions on Very

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:10:13 UTC from IEEE Xplore. Restrictions apply.

Large Scale Integration (VLSI) Systems, vol. 16, no. 3, pp. 331-335, March 2008, doi: 10.1109/TVLSI.2007.915514.

[13] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," in IEEE Journal of Solid-State Circuits, vol. 41, no. 3, pp. 712-727, March 2006, doi: 10.1109/JSSC.2005.864128.

[14] R. Oommen, M. K. George and S. Joseph, "Study and Analysis of Various LFSR Architectures," 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), Kottayam, India, 2018, pp. 1-6, doi: 10.1109/ICCSDET.2018.8821227.

[15] P. Mittal and D. Shah, "Linear Feedback Shift Register-Based Test Pattern Generators: A Comparative Study," in International Journal of Advanced Research in Electronics and Communication Engineering, vol. 9, no. 8, pp. 60-62, August 2020.

[16] S. L. Babu and N. S. M. Sarma, "Automatic test pattern generation in VLSI — A survey," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 3012-3017, doi: 10.1109/ICECDS.2017.8390009.

[17] B. Singh, A. Khosla and S. Bindra, "Power Optimisation of Linear Feedback Shift Register (LFSR) for Low Power BIST," 2009 IEEE International Advance Computing Conference, Patiala, 2009, pp. 311-314, doi: 10.1109/IADCC.2009.4809027.

[18] R. Mahajan, K. Devi and D. Bagai, "Review on Standard and Multi-bit Linear Feedback Shift Register," in Journal of Emerging Technologies and Innovative Research, vol. 5, no. 12, pp. 701-705, December 2018.

[19] P. Ghose, M. N. Rahman, M. M. Alam Polash and U. K. Acharjee, "Design of Reversible Shift Registers Minimising Number of Gates, Constant Inputs and Garbage Outputs," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 752-758, doi: 10.1109/ICACCI.2018.8554454.

[20] S. Mishra, R. R. Tripathi and D. K. Tripathi, "Implementation of configurable linear feedback shift register in VHDL," 2016 International Conference on Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES), Sultanpur, 2016, pp. 342-346, doi: 10.1109/ICETEESES.2016.7581406.

[21] J. Jung, H. Yoo, Y. Lee and I. Park, "Efficient Parallel Architecture for Linear Feedback Shift Registers," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 62, no. 11, pp. 1068-1072, Nov. 2015, doi: 10.1109/TCSII.2015.2456294.

[22] G. Hu, J. Sha and Z. Wang, "High-Speed Parallel LFSR Architectures Based on Improved State-Space Transformations," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 25, no. 3, pp. 1159-1163, March 2017, doi: 10.1109/TVLSI.2016.2608921.

[23] R. Karam, R. Puri, S. Ghosh and S. Bhunia, "Emerging Trends in Design and Applications of Memory-Based Computing and Content-Addressable Memories," in Proceedings of the IEEE, vol. 103, no. 8, pp. 1311-1330, Aug. 2015, doi: 10.1109/JPROC.2015.2434888.

[24] A. T. Do, C. Yin, K. Velayudhan, Z. C. Lee, K. S. Yeo and T. T. Kim, "0.77 fJ/bit/search Content Addressable Memory Using Small Match Line Swing and Automated Background Checking Scheme for Variation Tolerance," in IEEE Journal of Solid-State Circuits, vol. 49, no. 7, pp. 1487-1498, July 2014, doi: 10.1109/JSSC.2014.2316241.

[25] Y. Zhang, W. Zhao, J. Klein, D. Ravelsona and C. Chappert, "Ultra-High Density Content Addressable Memory Based on Current Induced Domain Wall Motion in Magnetic Track," in IEEE Transactions on Magnetics, vol. 48, no. 11, pp. 3219-3222, Nov. 2012, doi: 10.1109/TMAG.2012.2198876.

[26] I. Bayram and Y. Chen, "NV-TCAM: Alternative interests and practices in NVM designs," 2014 IEEE Non-Volatile Memory Systems and Applications Symposium (NVMSA), Chongqing, 2014, pp. 1-6, doi: 10.1109/NVMSA.2014.6927206.

[27] I. Arsovski, T. Chandler and A. Sheikholeslami, "A ternary content-addressable memory (TCAM) based on 4T static storage and including a current-race sensing scheme," in IEEE Journal of Solid-State Circuits, vol. 38, no. 1, pp. 155-158, Jan. 2003, doi: 10.1109/JSSC.2002.806264.

[28] C. Wang, J. Wang and C. Yeh, "High-Speed and Low-Power Design Techniques for TCAM Macros," in IEEE Journal of Solid-State Circuits, vol. 43, no. 2, pp. 530-540, Feb. 2008, doi: 10.1109/JSSC.2007.914330.

[29] S. L. Murotiya and A. Gupta, "CNTFET based design of content addressable memory cells," 2013 4th International Conference on Computer and Communication Technology (ICCCT), Allahabad, 2013, pp. 1-4, doi: 10.1109/ICCCT.2013.6749593.

[30] D. Bhattacharya, A. N. Bhoj and N. K. Jha, "Design of Efficient Content Addressable Memories in High-Performance FinFET Technology," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, no. 5, pp. 963-967, May 2015, doi: 10.1109/TVLSI.2014.2319192.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:10:13 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/350408876

Developments in Machine Tool Spindle Technology

Research · March 2021

CITATIONS
0

READS
65

2 authors:

Himanshu Bhatt
Delhi Technological University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Himanshu Bist
Delhi Technological University
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

All content following this page was uploaded by Himanshu Bhatt on 26 March 2021.

The user has requested enhancement of the downloaded file.

HBRP PUBLICATION

Research and Development in Machine Design
Volume 4 Issue 1

Developments in Machine Tool Spindle Technology

Himanshu Bhatt*[1], Himanshu Bist[2]
[1,2]Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur Village, Rohini, Delhi

Corresponding Author
***E-mail Id:-** dbhatt.com@gmail.com*

ABSTRACT

Machine tool spindles have a great importance in machine tools. As the machining processes have developed over the years, there has been a continuous need of development of machine spindles with them. This paper gives a review of some of the developments and researches that had taken place in the machine tool spindles such as use of air-assisted bearings, development of a new cooling strategy for reducing thermal errors and a new concept for Parallel High Speed Machining (PHSC). The paper also discusses the Hybrid Spindles which are used for high-performance cuttings and a new ultrasonic spindle which uses fluctuating electromagnetic force for inducing the ultrasonic vibrations. Along with them, this paper also reviews a spindle which uses a centrifugal component for automatically adjusting the preload changes on spindle caused by the centrifugal forces and concept of a spindle which is driven by water flow and uses a hydrostatic bearing for supporting the spindle.

***Keywords:-**Machine tool spindle, hybrid spindle, ultrasonic spindle, hydrostatic bearing, controlling preload*

INTRODUCTION

The machine tool spindle is one of the important in mechanical machining processes. The spindle is a rotating shaft that holds a cutting tool that cuts the material, which is clamped and fixed to the machining tool table in the form of chips. The machine tool spindle structural property positively affects the machining productivity and its surface finish. The structural properties majorly depend on the power source like motors, tool holder system, the design of spindle assembly, bearing used, and the dimensional specification of the shaft as shown in Figure1[1].

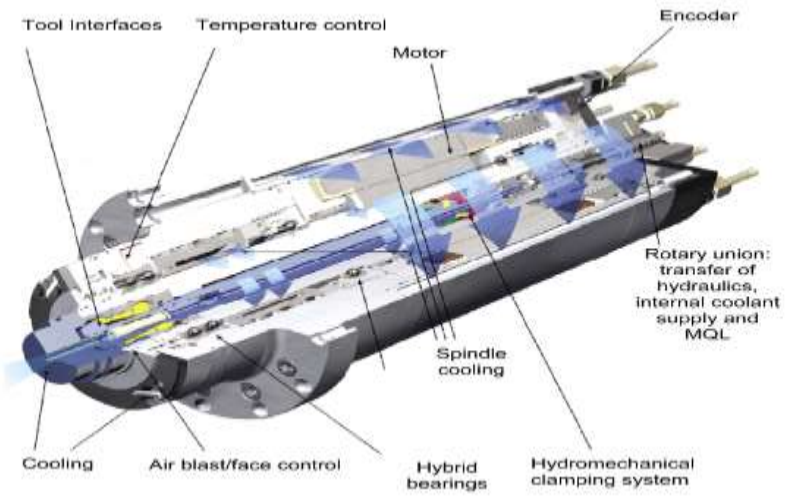


***Fig.1:-**Sectional view of a spindle [2]*

Historically, the spindles were driven by belt drives and gear drives. The rotation speed is only varied by changing either the driven poles quantity by electrified switches or the transmission ratio. But later some hydraulic and electrified controlled spindle was developed and the rotation speed changed by ward Leonard motor control system which is an adjustable rotating transformer.[3]

Nowadays, manufacturing industries have become more globalized. The manufacturers are always looking for the most efficient and economical ways to improve the machine tools' productivity by adopting convenient power source density, targeting higher speeds, more flexibility, and multitasking operations. And the only method to achieve this by continuous improvement and innovation in the machine tool spindle technology. Therefore, research for manufacturing industries is a very beneficial field of new-age researchers.

Flexible manufacturing centers capable of executing a variety of programmed tasks are modern machine tools. In all machine tools, spindles are a key component. They are used to conduct a variety of cutting operations on for example, items made of cast metal. To achieve their final shape and specification, these products involve finishing through processes such as turning, cutting, drilling, countersinking, tapping, reaming, grinding and milling. Significant fields of use include the automotive and aerospace sectors.

LITERATURE REVIEW

Greco et al. [4] proposed a spindle design for micro-machining processes. They designed a new ultrasonic-assisted air bearing spindle that can reach higher rotational speeds at low radial run-outs. They used an air-turbine in accordance with the design of a Pelton Wheel for driving the spindle's rotor. They used a thrust and a pair of journal air bearings for supporting the rotor and located one journal bearing close to the tool for minimizing the radial run-out.

A new cooling strategy has been presented by [5] for reducing the thermal errors caused due to the distortion of the spindle from its reference position because of the rise in the temperature of various components of machine tools during high speed machining. This new strategy has been referred to as CTM, Cooler Trigger Model. This operates by dynamically controlling the frequency and the working time of the compressor of standard bath recirculation unit so that the extraction of heat is in accordance with generation.

Design optimization as per the PSO, Particle Swarm Optimization, of a spindle-bearing system was performed by [6] by optimizing the natural frequencies, static stiffness and total friction torque. The design was optimized for two sets of adjacent bearing positions and preload and through simulations and experiments; improved performance of the optimized design was verified. The optimized spindle could increase static stiffness by 26% and the natural frequencies 1 and 2 by 10% and 16% respectively, under the same friction torque.

Harpaz et al. [7] proposed a new concept of spindle for Parallel High Speed Machining (PHSM) processes. This new air-lubricant based spindle is mounted as an additional spindle on the machine tool, parallelly to the main spindle. It can allow speeds upto 90,000rpm and also improves the finishing and micromachining operations. Early prototypes when used with high pressure cooling lubricant for micromachining, higher performance, good tool life and good surface quality were achieved.

Vyroubal [8] presented a new method for analyzing thermal deformation behaviour of machines due to spindle rotation thermal

load, along the direction of the spindle axis. This uses a specially designed machine frame which measures the deformation of machine column, headstock and spindle. A compensation mechanism is then developed which is based on a regression approach, but considers different thermal dynamic of each machine part to thermal load. Only temperature measurement with new measurement places like spindle cooling and workspace is used for improving the accuracy of calculation.

The ball-bearing is the pivotal component of a spindle for supporting the shaft and its characteristics are closely related to the preload on the spindle. Ciou and Lee [9] designed a controlling mechanism for spindle preload using a single piezoelectric actuator and wedge-shaped spacer ring. They tested an experimental full-sized model and results revealed that an increased preload causes reduction in vibration amplitude, which is important for high-precision spindles used in high speed and highly precised machine tools.

HYBRID SPINDLE

One of the developing technologies in the field of manufacturing technology is the hybrid spindle. Their combination of the potential spindle motor and the spindle functionality results in a wide range of milling spindle [10]. A motor-driven machining spindle is typically used for specialised operations such as High-Performance Cutting (HPC) or High-Speed Cutting (HSC). Different machining procedures necessitate different torque and cutting speed ranges as shown in Figure 2.

Fig.2:-*Hybrid Spindle[10]*

The efficiency of machining is increased by choosing optimum and efficient processes such as HPC, and HSC. There are majorly two difficulties and issues in designing the traditional spindle for HPC, or HSC. First, the spindle design of the bearing for excessive rigid and high-speed results in a inconsistency; transmitting large cutting forces with the least deformation of the bearings have to be in excessively rigid. Because of centrifugal forces, the high thermal-mechanical loads are subjected to the bearing and bearing heat up, resulting in thermal-mechanical fluctuations. As a result, the machining spindle bearing experiences a preload increase caused by the thermal expansion, which leads to loss of power in spindle bearings. Therefore, preferably a smaller diameter of the bearing and low preload are taken into account. Second, the other limiting factor is the winding system in the machine tool spindle motor. The maximum rotational speed is controlled by the output frequency and voltage of the source. The maximum torque depends upon maximum current, and also the no. of turns in a single phase is dependent upon the ratio of voltage to current. High torque and high speed are controlled by the number of turns in a single phase. High torque and high speed are controlled by the number of turns in a single

phase. For the same voltage source, a High-Speed Cutting (HSC) allows less turns in a single step than a High-Performance Cutting (HPC). The hybrid spindle offers the most effective approach for using the machining process's capabilities to solve these two problems and challenges [11].

It can be concluded that adapting the spindle bearing motor characteristics and the preload is not economical. To get the benefits of a hybrid spindle within an industry, an expensive way to adjust and manipulate the preload of bearing. Also, the characteristic of a motor has to change electrically. This technique is beneficial to allow the switching between the working modes for HPC and HSC. There are two approaches to do this. First, to approach the bearing preload bases on a cost-effective and thermal mechatronic system. The element is fully equipped with the strain gauges as a force control loop input to manipulate the preload of bearing subjected by the preload element. The current preload forces are detected by the strain gauges. However, the signals of strain induced by force are superimposed by signs of thermal-induced deformations in the preload element. Now, the temperature change is caused by the integrated Peltier elements, spindle motor, and friction as shown in Figure 3a and 3b. These factors can be diminishing by connecting the strain gauges to a Wheatstone bridge. The second approach, the spindle motors, are designed by a reconfigurable winding system. A changeable no. of branches which is parallel allows the motor properties' change and enables a transition between HSC, and HPC by decreasing the number of turns/phase.

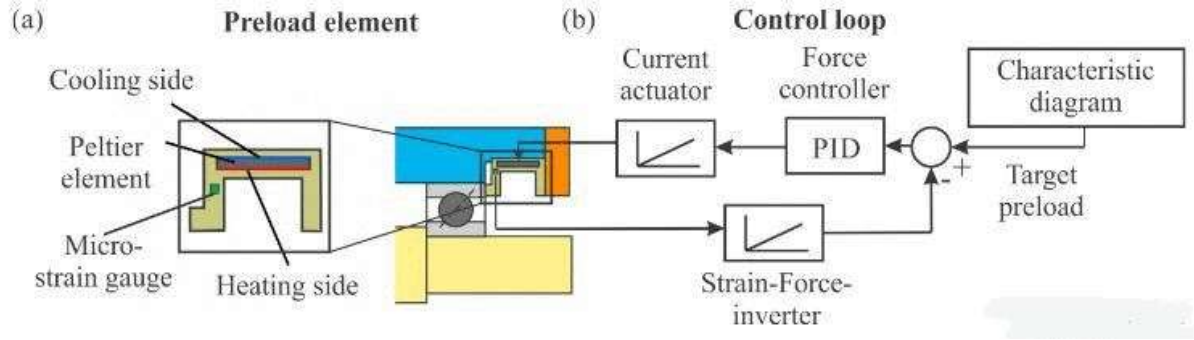


Fig.3(a):-*Thermo mechanical preload element with* ***b)*** *Control loop for preload force control[11].*

MAGNETIC FORCE INDUCED ULTRASONIC SPINDLE

Rotary ultrasonic machining is used throughout the industry for a variety of operations such as milling, drilling, threading and many more. It involves the use of an ultrasonic vibration spindle, at the end face of which the tool is attached so that the tool can be ultrasonically vibrated. The two types of transducers that are generally used for inducing ultrasonic vibrations to spindle are made up of piezoceramics and ferrite. The piezoceramics one involves use of carbon brushes and slip-rings, where the occasional poor contact between them leads to failure whereas the ferrite ones suffer high electrical losses due to high thermal generation in ferrite materials. To overcome these drawbacks, a new technique has been developed which makes use of fluctuating electromagnetic force for inducing the ultrasonic vibrations.

The new spindle [12] has a cylindrical body, with 2 driving discs and two flanges. It is supported by two housings on a base plate through the flanges and bearings. A tool can be attached to the front end of the spindle, and a motor is connected at rear end for providing rotational motion to the

spindle. The spindle's resonant vibration mode is 3rd longitudinal vibration (L3 mode), thus making 3 vibration nodes along the spindle's length. The two flanges can be placed at front and end nodes to support the spindle. The two discs are placed at the points of maximum amplitude of the L3 mode for the easy induction of ultrasonic vibrations in the spindle. Four electromagnets are placed on each of two discs with some gap for producing fluctuating magnetic forces. When all electromagnets are supplied the same AC voltage at the L3 mode of spindle, spindle begins to vibrate ultrasonically along its axis, which then vibrates the tool, because of the generation and action of electromagnetic forces on driving discs. The amplitude of vibration of spindle depends upon the frequency and amplitude of AC voltage and the gap between electromagnets as shown in Figure 4 and 5.

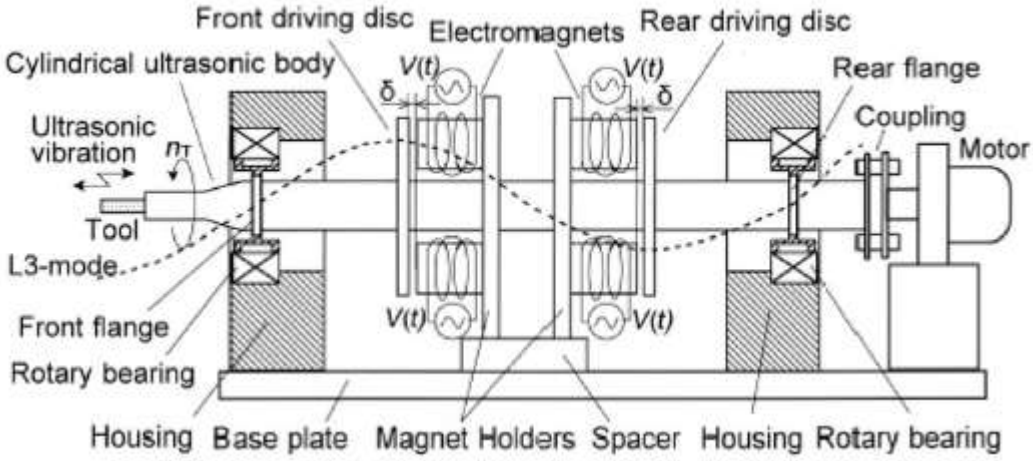


Fig.4:-*Schematic for new Ultrasonic spindle [12]*

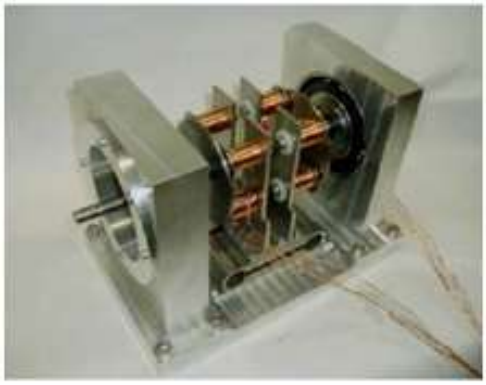


Fig.5:-*Constructed Unit [12]*

The electromagnets are responsible for the fluctuating magnetic forces which ultimately induces vibration in the spindle. Thus, it becomes essential to use electromagnets that can produce sufficient magnetic fields. Pure iron steel and enameled wire is used for core and coil materials because of their high permeability, so that they can produce enough magnetic field to produce vibrations.

CONTROLLING PRELOAD USING CENTRIFUGAL FORCE

Machine tool spindles use rolling bearings since they provide high stiffness and high load capacity which is required by spindles for high speed and high efficient machining operations. For spindles using rolling

bearings, a preload is applied initially which increases the stiffness of the spindle and limits vibrations. A greater preload helps by increasing the stiffness of the bearing, but it also destroys the oil film because of high contact stress between the bearings and the raceways, ultimately shortening the life of the bearing. When the spindle rotates at high speed, there is sufficient increase in the load because of the centrifugal force and heat generated which affects the life of bearing. In cases where lighter preload is applied, the stiffness characteristics are undermined which are important for high speed and high efficient machining.

The new preload device[13] solves this issue by automatically adjusting the preload on the spindle using the changes in the centrifugal force as shown in Figure 6. It uses a centrifugal element to adjust the preload automatically as opposed to the most commonly used hydraulic pressure method and piezoactuators. Its structure is made up of centrifugal element, main and auxiliary bearings, nuts and a constant pressure preload spring. The preload to be applied initially is found by the use of springs and fastening level of nuts.

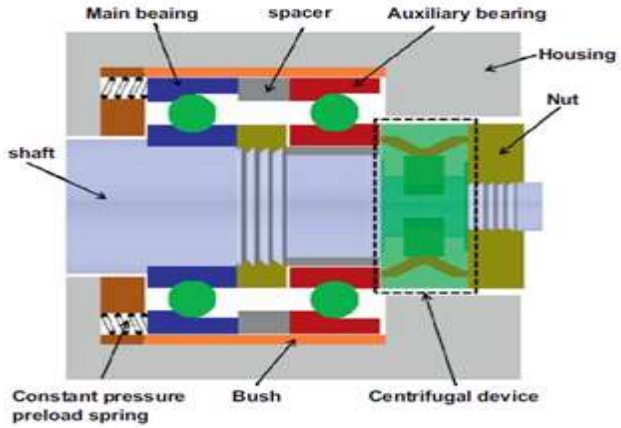


Fig.6:-*Automatic Preload Device utilizing Centrifugal force [13]*

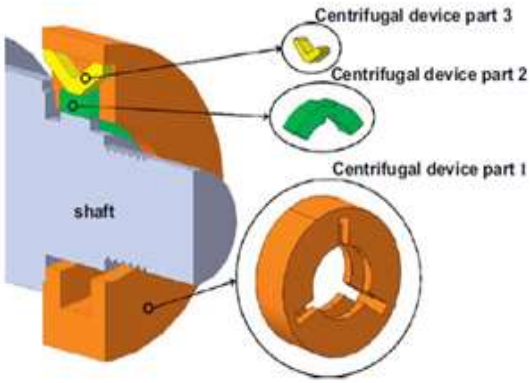


Fig.7:-*Arrangement of centrifugal component [13]*

The centrifugal component has 3 parts and is assembled about the shaft as shown in Figure 7. Part 1 is in the shape of a hollow cylinder, and it's amount of expansion in radial direction is smaller in comparison with part 2. Part 2 is a group of various sections of a cylinder. It doesn't have any constraint in radial direction due to which it has a greater expansion amount as compared to part 1. Part 3 is compressed between the 1st and 2nd parts and extends along the axial direction. When the spindle rotates, centrifugal force acting on the rotational components and difference in the conditions of constraint causes displacement of part 2 along the axial direction. Due to this displacement, the outer ring of main bearing gets pushed through the auxiliary bearing and causes decrement of preload on main bearing.

This displacement of main bearing depends upon the contact force between part 2 and auxiliary bearing, auxiliary spring's stiffness and characteristics of spring used. When the contact force increases because of centrifugal force, preload on the main bearing decreases. The decrease in the preload will be high in case of small spring force.

SPINDLE SUPPORTED BY HIGH STIFFNESS WATER HYDROSTATIC BEARING

In ultra-precise machining tools, the precision spindle plays an essential role in producing precise parts like lenses. The accuracy needed for the details reaches nanometers. Therefore, the high stiffness bearings are highly required for ultra-precise machining tools. Commonly, the accuracy of machining processes depends on the accuracy of the machining spindle's rotary motion. The machining spindle of the ultra-precise machine tools is beared by radial bearings and hydrostatic thrust, avoiding mechanical touching of the shaft and other moving parts. The hydrostatic bearing notably affects the precision of the mechanical rotational motion of the machining tool spindle. And the dependent performance index of the water hydrostatic bearing is the bearing rigidity.[14]

The stiffness of bearing stated that the outward load is divided by the spindle rotor's associated displacement along the force direction. Also, the considerable stiffness of the bearing is subjected to the spindle's small resultant displacement, and it's desirable to attain more significant precise machining. The required preciseness of machining the generated parts by the diamond turning processes is tens of nanometers of value as shown in Figure 8.

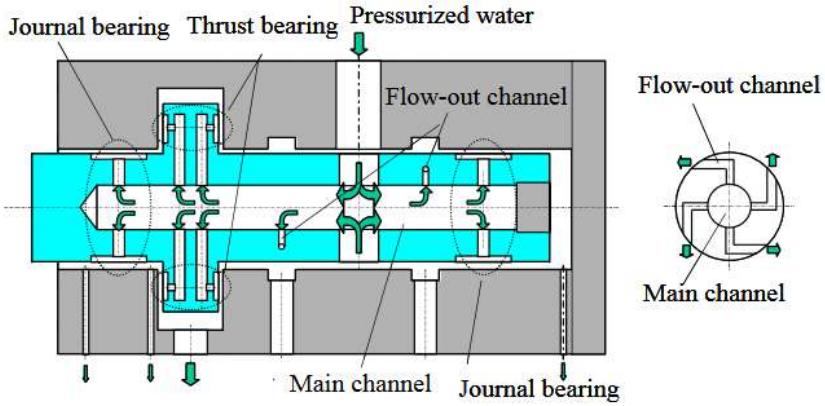


***Fig.8**:-Schematic diagram for water driven spindle with hydrostatic bearings[14]*

The spindle worked for the actual diamond rotation must be handled by radial bearings and the hydrostatic thrust. The sliding bearing supports the spindle shaft for easiness in the structural point. As we know, bearing characteristic restrictions is an essential and influential factor of stiffness. Therefore, the restrictors of the spindle are designed to maximize the bearing rigidity. Besides, the proper selection of material is necessary. For the bearing parts of the water, hydrostatic bearing because of the great lubricative property of water. However, the moving parts beared by the hydrostatic bearing don't directly touching with moving mating parts. In this case, the pieces were made up of steel majorly stainless steel. Because of the deficiency of water's lubricative property, the damages take place if the contact happens. Somebody must also consider the selections of material for the hydrostatic bearings for the bearings pads: forbearing pads, ceramics, gunmetal, carbon-based resin, and several engineering plastics.

CONCLUSION

Spindles are an important part for the machine tools and as the machining processes have developed, a lot of different developments have been made to the spindles also. From the review, we see that concepts such as using an air-turbine based on pelton wheel design for driving the spindle’s rotor, optimization of multiple objectives as per the PSO (Particle Swarm Optimization) and analyzing the thermal

deformation by using a specially designed frame and then developing a compensation mechanism based on regression approach have been developed.

Some of the developments has been the hybrid spindle, which uses a combination of the potential spindle motor & the spindle functionality, an ultrasonic spindle which instead of using piezoceramics and ferrite uses fluctuating electromagnetic force for inducing the ultrasonic vibrations and a spindle which uses a new centrifugal device for the automatic adjustment of preload as the higher or lower preload can affect proper functionality.

There is also a spindle which uses water flow for driving the spindle, supporting the spindle through hydrostatic bearing and for cooling the spindle. From the above research we can conclude that spindle technology has continuously evolved, as]the machine tools and machining processes have developed and continuous research in spindles is necessary to keep up with the other related developments.

REFERENCES

1. Maeda, O., Cao, Y., Altintas, Y. (2005). Expert spindle design system. *International Journal of Machine Tools and Manufacture, 45(*4–5), 537–548.
2. De Ciurana, J., Quintana, G., Campa, F. J. (2009). Machine Tool Spindles. *Machine Tools for High Performance Machining* (pp. 75–127).
3. Abele, E., Altintas, Y., Brecher, C. (2010). Machine tool spindle units. *CIRP Annals*, *59*(2), 781–802.
4. Greco, S., Lange, A., Kirsch, B., Aurich, J. C. (2019). Simulation-based design of an ultrasonic-assisted air bearing spindle for micro machining. P*rocedia CIRP, 82, 160*–165.
5. Grama, S. N., Mathur, A., & Badhe, A. N. (2018). A model-based cooling strategy for motorized spindle to reduce thermal errors. *International Journal of Machine Tools and Manufacture, 132*, 3–16.
6. Tong, V.-C., Hwang, J., Shim, J., Oh, J.-S., & Hong, S.-W. (2020). Multi-objective Optimization of Machine Tool Spindle-Bearing System. *International Journal of Precision Engineering and Manufacturing, 21*(10), 1885–1902.
7. Harpaz, O., Books, B., Schwaar, M., Schubert, A., & Eckert, U. (2012). Parallel High Speed Machining with a New Additional HSC Spindle for Machine Tools. *Procedia CIRP*, *1*, 673–674.
8. Vyroubal, J. (2012). Compensation of machine tool thermal deformation in spindle axis direction based on decomposition method. *Precision Engineering*, *36*(1), 121–127.
9. Ciou, Y.-S., & Lee, C.-Y. (2019). Controllable preload spindle with a piezoelectric actuator for machine tools. *International Journal of Machine Tools and Manufacture, 139*, 60–63.
10. WEISS SpindleTechnologies gmbh, A Siemens Company, Hybrid Spindle. https://www.weissgmbh.com/en/products/milling/hybrid-spindle/
11. Denkena, B., Bergmann, B., Mücke, M., Königsberg, J., Ponick, B. (2018). Hybrid Spindle – An approach for a milling machine tool spindle with extended working range for HSC and HPC. *Procedia Manufacturing, 24*, 159–165.
12. Wu, Y., Yokoyama, S., Sato, T., Lin, W., Tachibana, T. (2009). Development of a new rotary ultrasonic spindle for precision ultrasonically assisted grinding. *International Journal of Machine Tools and Manufacture, 49*(12–13), 933–938.
13. Hwang, Y. K., Lee, C. M. (2009). Development of automatic variable preload device for spindle bearing by

using centrifugal force. *International Journal of Machine Tools and Manufacture, 49*(10), 781–787.
14. Nakao Y., Masato M., Kobayashi F. (2003) Water energy drive spindle supported by water hydrostatic bearing for ultra-precision machine tool. *Proc. of ASPE 2003* Annual Meeting.

View publication stats

Eco. Env. & Cons. 27 (1) : 2021; pp. (359-364)
Copyright@ EM International
ISSN 0971–765X

Drinking water quality evaluation for groundwater of Delhi, India using GIS techniques

Ayushi Raghav* and S.K. Singh

Department of Environmental Engineering, Delhi Technological University, Delhi, India

(Received 26 June, 2020; accepted 24 July, 2020)

ABSTRACT

Groundwater is considered to be an important water resources because of it's widey use in drinking and irrigation purpose. Due to industrialization and urbanization, groundwater quality is deteriorating at a rapid rate. The present study evaluates the groundwater quality of NCT Delhi using drinking water quality parameters with the computation of water quality index (WQI) and coupled with QGIS technique. Raster interpolation technique, Inverse Distance Weighted (IDW) in QGIS software has been used to generate the spatial distribution map of water quality parameters. Six water quality parameters are used to evaluate the suitability of groundwater for the drinking purpose. WQI has been developed using eleven drinking water quality parameters. It was found that the majority of WQI of the study area lies in "poor" to "unsuitable for drinking" category.

Key words *: Groundwater quality, GIS, Water quality index (WQI), NCT Delhi and Spatial distribution.*

Introduction

Water is an essential commodity for the existence of life on planet Earth and plays a significant role in balancing the ecosystem. India annually receives about 4000 cubic km of precipitation. The total annual flow in rivers is estimated to be 1869 km^3. 50% of the urban requirements and 85% of the rural requirements are satisfied by the groundwater (Suhag, 2016). Due to rapid urbanization, industrialization and increase in population, a pressure has been developed on the groundwater to satisfy the needs. It has led to over exploitation of groundwater resources and in some cities the water table has gone several meters down below the ground level.

The water quality of any specified source like groundwater can be assessed using three parameters namely physical, chemical and biological (Ketata Mouna *et al*., 2011). Water quality index (WQI) describes the quality of water using these parameters and comparing them with their standard values. The objective of the study is to assess and map the groundwater quality of NCT Delhi for drinking purpose and to evaluate the WQI for drinking purpose of the region. For mapping purpose, QGIS software (version 3.10.2) is used. Spatial interpolation method, Inverse Distance Weighed (IDW) has been used to estimate the spatial distribution of the water quality parameters and WQI.

Study Area

Delhi, officially known as National Capital Territory of Delhi (NCT), is capital of India covering an area of 1483 sq.km. It lies from 28°24′15" to 28°53′00" North latitude and from 76°50′24" to 77°20′30" East longitudes. Delhi lies in the Northern Plains of Indian Subcontinent. The climate of Delhi varies between humid subtropical to semi-arid with long and hot summers and cold winters. The normal annual rainfall in the city is approximately 611.80 mm. The

rainfall in the city increases from southwest to the northeast.

Materials and Methodology

Input Datasets

The data used in this study is taken from Central Ground Water Board (CGWB) Water Resources Information System (WRIS) of India for the year 2015. Data from a total of 28 sampling stations were used for determining the quality of groundwater for drinking purpose. The map of NCT Delhi is geo-referenced using the geo-referencing tool available in the QGIS software. A shape file of sampling locations is prepared and joined with the respective attribute information i.e., water quality parameters.

Water Quality Index for Drinking purpose

Water quality index provides unique rating to describe the overall quality of water in a single value that helps in the selection of appropriate treatment method (Tyagi *et al.,* 2013). Hortan (1965) developed the arithmetic aggregation function for the WQI using 10 commonly measured water quality variables including pH, Dissolved oxygen (DO) and alkalinity. Similarly Brown *et al.* (1970) also employed basic arithmetic weightage for the development of WQI without using the multiplicative variables. In the present study the drinking water quality index (WQI) has been evaluated by using the drinking water quality standards recommended by Bureau of Indian Standards (BIS 2012). The arithmetic weighted index method is used here. 11 parameters were considered for computing WQI by using the following formulas.

$$WQI = \Sigma_{i=1}^{n} W_i \times Q_i \quad .. (1)$$

Where,

W_i = Relative weight of each parameter computed as:

$$Wi = \frac{Wi}{\Sigma_{i=1}^{n} wi} \quad .. (2)$$

Q_i = Quality rating and computed as:

$$Qi = \frac{(v_a - v_i)}{(s_i - v_1} \times 100 \quad .. (3)$$

Where,

w_i = unit weight assigned to each parameter

n = total number of parameters i.e n= 11

v_a and V_i = The actual and ideal values for all the selected parameters of water quality respectively. The value of is zero for all the parameters except for pH. For pH value of is 7.

S_i = Standard value of selected parameter taken from BIS **IS : 10500-2012.**

The standards and weights of the parameters are shown in Table 1.

Interpolation Technique

Geographic Information System (GIS) proved to be a robust tool for solving the water related problems which include assessment of water quality using different water quality indicators, estimation of water availability, developing better understanding

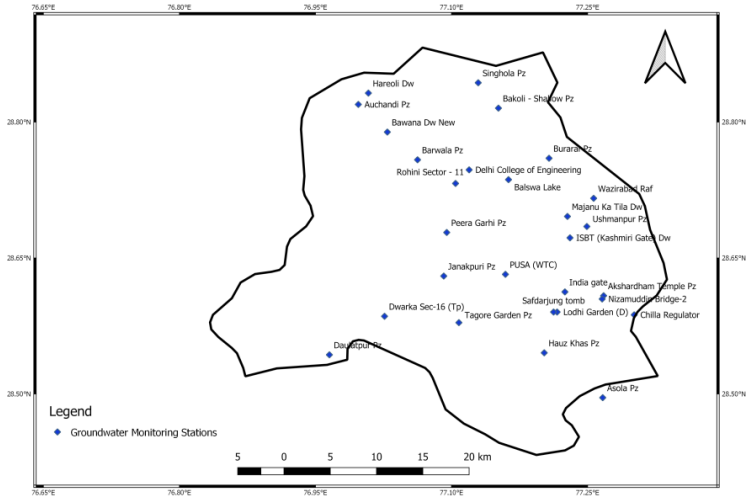


Fig. 1. Location map of the study area

of natural environment and management of water resources on a regional scale (Collet, 1996). Inverse Distance Weighted (IDW) is a raster interpolation technique available in QGIS software (version 3.10.2) and this spatial interpolation technique is used in the present study to delineate the local distribution of drinking water quality parameters and water quality index (WQI). The scale used to prepare the maps is 1:250000.

Results and Discussion

Groundwater quality for drinking purpose

For evaluation of groundwater quality for drinking purpose six parameters were considered. The statistical summary of these drinking water parameters are shown in Table 2.

pH

In study area the pH ranges from 7.55 in Barwala to 8.77 in Delhi College of Engineering. It defines that the water in the study area varied from brackish to saline water. Figure 2 shows the spatial distribution of pH in the study area.

Total Hardness

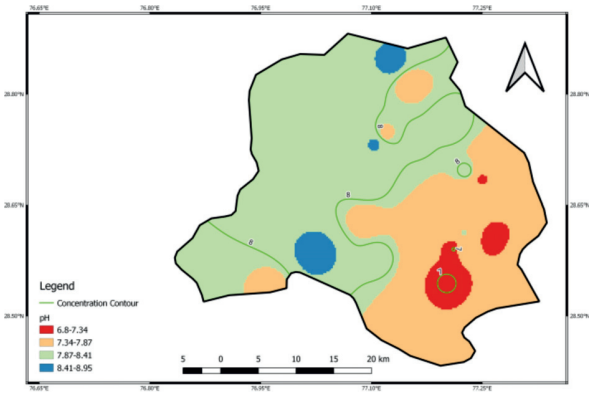


Fig. 2. Spatial Distribution of pH

In the study area the total hardness varied from a minimum value of 185 mg/L in Nizamuddin Bridge-2 to 2536 mg/L in Barwala. Figure 3 shows the spatial distribution of total hardness in the study area.

Total Alkalinity

In the study area the total alkalinity in the ground-

Table 1. The standards and weights assigned to each parameter

S. No.	Parameters	Standard limits(S_i)	Weights (w_i)
1	pH	6.5-8.5	4
2	Total Hardness(as $CaCO_3$)	200 mg/L	3
3	Total Alkalinity (as $CaCO_3$)	200mg/L	3
4	Calcium	75 mg/L	4
5	Magnesium	30 mg/L	4
6	Chloride	250 mg/L	5
7	Nitrate	45 mg/L	4
8	Fluoride	1.0 mg/L	2
9	Sulphate	200mg/L	2
10	Sodium*	200 mg/L	1
11	Potassium*	10 mg/L	1

*(In case of sodium and potassium, BIS has not prescribed any standard limit. Here for evaluation of WQI, a limit of 200 mg/L and 10 mg/L are taken for Sodium and Potassium respectively)

Table 2. Statistical Summary of Parameters of Groundwater Quality

Parameters	Units	Minimum	Maximum	Mean	Std. Deviation
Calcium	mg/L	8.0	513	155.41	154.16
Chloride	mg/L	85	2204	700.25	904.05
Nitrate	mg/L	1.61	710	100.15	129.63
pH	-	7.55	8.77	7.86	0.54
Total Hardness	mg/L	185	2536	887.75	780.57
Total Alkalinity	mg/L	2.3	441.45	229.27	132.56

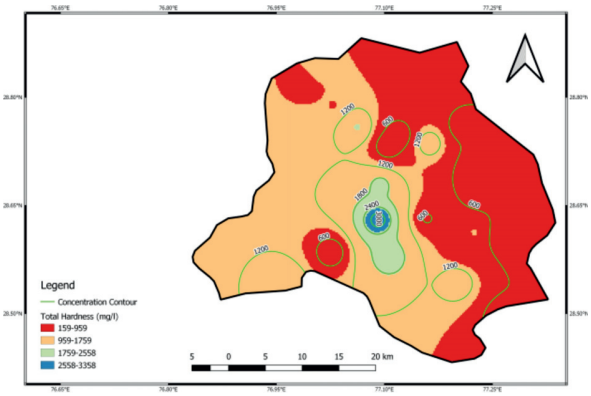


Fig 3: Spatial Distribution of TH

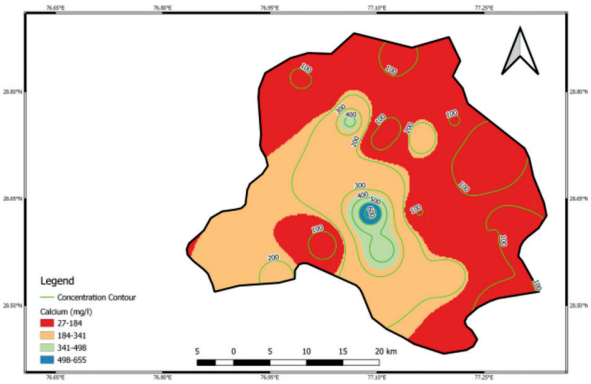


Fig. 5. Spatial Distribution of Calcium

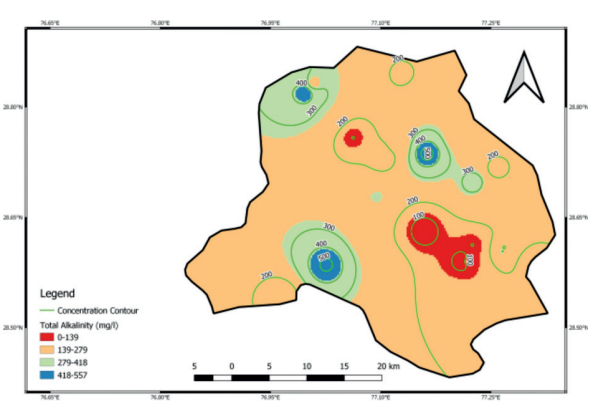


Fig. 4. Spatial Distribution of TA

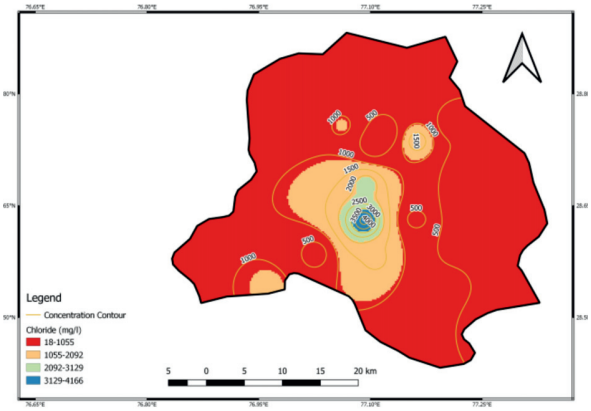


Fig. 6. Spatial Distribution of Chloride

water varied from a minimum value of 2.3 mg/L in Chilla Regulator to a maximum value of 441.45 mg/L in Dwarka Sector-16. It can be concluded that the groundwater in the study area was having alkalinity within permissible range. Figure 4 shows the spatial distribution of total alkalinity in the study area.

Calcium

The distribution of calcium in study area varied

Table 3. WQI of the stations

S.No	Stations	WQI	S.No	Stations	WQI
1	Akshardham Temple	144.79	15	Barwala	432.07
2	Chilla Regulator	46.71	16	Bawana	123.48
3	Nizamuddin Bridge-2	68.53	17	Delhi College of Engineering	124.19
4	India gate	399.73	18	Hareoli	183.14
5	Lodhi Garden	145.79	19	Rohini Sector – 11	147.73
6	Safdarjung tomb	133.23	20	Singhola	247.08
7	Burarai	45.71	21	Asola	61.90
8	ISBT (Kashmiri Gate)	202.7	22	Hauz Khas	368.98
9	Majanu Ka Tila	150.01	23	Daulatpur	302.45
10	Ushmanpur	99.85	24	Dwarka Sec-16	105.17
11	Wazirabad	76.56	25	PUSA	132.86
12	Auchandi	158.99	26	Tagore Garden	578.39
13	Bakoli	238.18	27	Janakpuri	557.03
14	Balswa Lake	303.87	28	Peera Garhi Pz	116.08

from 8 mg/L in Dwarka Sector-16 to 513 mg/L in Janakpuri. The spatial distribution of calcium in the study area is shown in Fig 5.

Chloride

The concentration of chloride in groundwater in the study area varied from 85 mg/L in Burari to 2204 mg/L in Tagore Garden. The spatial distribution of chloride in the study area is shown in Fig 6. It shows that North West part of the study area had high concentrations of chloride in groundwater. High concentrations of chloride in groundwater may be result of sewage pollution, leaching of saline residues in the soil.

Nitrate

The concentration of nitrate in study area was varying from 1.61 mg/L in Burarai to 710 mg/L in Hauz Khas. The spatial distribution of nitrate in the study area is shown in Fig 7.

Water Quality Index

In this study water quality index (WQI) is computed for drinking water using eleven water quality parameters and their BIS standards. The WQI of all the stations are given in Table 3. The spatial distribution of the WQI for the study area is shown in Fig. 8. The

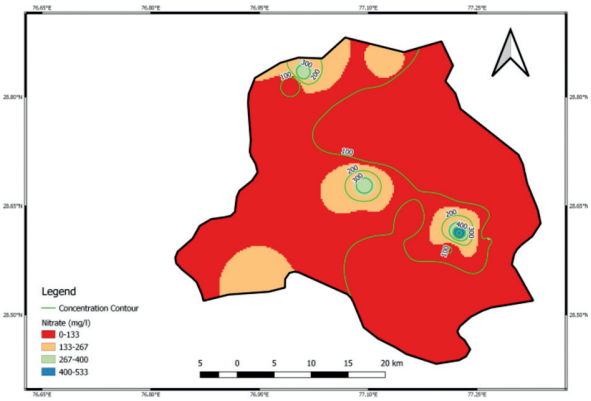


Fig 7: Spatial Distribution of Nitrate

Table 4. Water quality classification based on WQI

WQI	Water Quality	No. of Samples	Percent (%)
<50	Excellent	02	7.1
50-100	Good	04	14.3
100-200	Poor	12	42.9
200-300	Very Poor	03	10.7
>300	Unsuitable	07	25

computed water quality index values in the study area varied from 45.71 in Burarai to 578.39 in Tagore Garden. The groundwater quality can be categorized from "excellent" to "water suitable for drinking" on the basis of these computed WQI values. Majority of the sampling stations, i.e 42.9% were having poor groundwater quality. They include areas from East Delhi, New Delhi, North Delhi and North East Delhi. In these parts, the quality of groundwater might be improved by the infiltration of fresh water during the monsoon. Table 4 shows the number of samples and their percentages falling under each category.

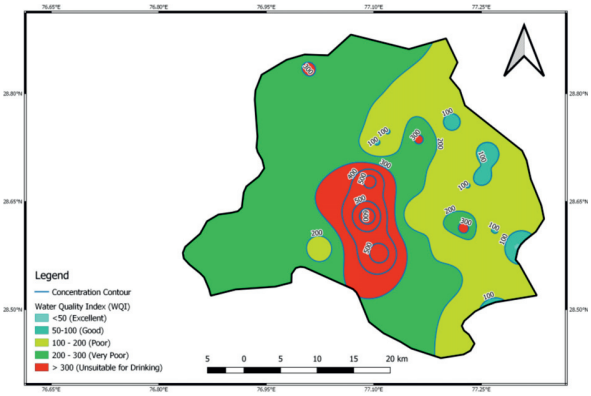


Fig. 8. Spatial distribution of WQI

Conclusion

In this study the GIS technology has been combined with the groundwater quality data of NCT Delhi obtained from CGWB for the year 2015. The generated map shows the spatial distribution of drinking water quality parameters. The spatial distribution map showed that majority of study area had pH, total alkalinity and calcium lying in the permissible limit in the study area. The spatial distribution map of total hardness reveals that the groundwater in the study area was in the category of hard to very hard water. About 78.6% of the area under the study comes under poor to unsuitable for drinking category as revealed by the WQI studies. These maps could be proved useful to the decision makers and planners for rehabilitation of groundwater quality in the area as well as for groundwater quality development.

Compliance with ethical standards

Conflict of interest statement- The authors declare that they have no conflict of interest.

References

BIS, 2012. Indian standards specifications for drinking water IS: 10500, Bureau of Indian Standards, New Delhi.

Brown, R.M., McClelland, N.I., Deininger, R.A. and Tozer, R.G. 1970. Water quality index-do we dare?. *Water Sewage Works.* 117 (10) : 339-343.

Collet, C. 1996. *Geographic Information System Needs and Software in : Geographical Information Systems in Hydrology*, 1st Edition, Kluwer Academic Publishers, Boston, USA.

Horton, R. K. 1965. An Index number system for rating water Quality. *J. Water Pollu. Contd. Fed.* 37 (3) : 300-305.

India-WRIS "Water Resources Information System of India". (n.d.). Retrieved from Central Ground Water Board.

Ketata Mouna, R., Gueddari, M. and Bouhlila, R. 2011. Use of Geographical Information System and Water Quality Index to Assess Groundwater Quality in El Khairat Deep Aquifer (Enfidha, Tunisian Sahel). *Iranica Journal of Energy & Environment.* 2 (2) : 133-144.

Suhag, R. 2016. Overview of Ground Water in India; PRS Legislative Research: http://www.prsindia.org/.

Tyagi, S., Sharma, B., Singh, P. AND Dobhal, R. 2013. Water quality assessment in terms of water quality index. *Am J Water Resour.* 1 (3) : 34–38.

Research Article

ECG Denoising Using Artificial Neural Networks and Complete Ensemble Empirical Mode Decomposition

Rajesh Birok[a], Rajiv Kapoor[b], and Mahipal Singh Choudhry[c]
[a] Deptt of Electronics & Communication Engineering, Delhi Technological University, Delhi, India
[b]Deptt of Electronics & Communication Engineering, Delhi Technological University, Delhi, India
[c]Deptt of Electronics & Communication Engineering, Delhi Technological University, Delhi, India

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Electrocardiogram (ECG) is a documentation of the electrical activities of the heart. It is used to identify a number of cardiac faults such as arrhythmias, AF etc. Quite often the ECG gets corrupted by various kinds of artifacts, thus in order to gain correct information from them, they must first be denoised. This paper presents a novel approach for the filtering of low frequency artifacts of ECG signals by using Complete Ensemble Empirical Mode Decomposition (CEED) and Neural Networks, which removes most of the constituent noise while assuring no loss of information in terms of the morphology of the ECG signal. The contribution of the method lies in the fact that it combines the advantages of both EEMD and ANN. The use of CEEMD ensures that the Neural Network does not get over fitted. It also significantly helps in building better predictors at individual frequency levels. The proposed method is compared with other state-of-the-art methods in terms of Mean Square Error (MSE), Signal to Noise Ratio (SNR) and Correlation Coefficient. The results show that the proposed method has better performance as compared to other state-of-the-art methods for low frequency artifacts removal from EEG.

Keywords: ECG, Neural Networks, CEEMD

1. Introduction

The electrocardiogram (ECG) is a quasi-periodic signal that measures the electrical activity of the heart. It is used to assess whether a patient has any type of cardiac disorder and provides information content mainly in the 0.5-120 Hz range. The P wave, QRS complex, and ST segment make up the majority of its morphology; any extra structures in the ECG indicate noise or cardiac defects. The diagnosis of these disorders is generally made by experts who examine the ECG signal for specific morphologies. As a consequence, an ECG signal's morphology is important, and it must be as noise-free as possible. In the other side, obtaining an ECG signal captures a vast range of different sounds, rendering accurate diagnosis impossible. As a consequence, correct diagnosis necessitates a filtration procedure that removes the bulk of constituent noise while preserving details about the ECG signal's morphology. Any of the noises that can be heard are power line interference (PLI), baseline wander, electrode touch noise, motion artefacts, and instrumentation noise [1]. Baseline wander artefact occurs mainly as a result of the patient's breathing, making it omnipresent in almost all ECG signals. As one of the most common kinds of noise in an ECG, it is recommended for denoising in this paper. This artefact is frequently seen during stress ECGs, especially as the patient's breathing rate rises. All of these morphological changes are normal and necessary for diagnosis, but most current denoising procedures tend to filter them out, resulting in information loss [37]. Another reason we chose baseline wander for denoising in our paper is that it hasn't gotten as much attention in the literature, so while there are fairly efficient methods for denoising PLI and other types of data, the same cannot be said for baseline wander.

2. Literature Survey

In the literature, several methods for ECG signal denoising have been proposed. For the removal of low frequency noise, digital window-based filters were first proposed [2][3]. The trouble with these is that they cause ripples in the pass band when a strong cut-off is used. As a consequence of these ripples, the morphology of the ECG signal varies, which is inappropriate from a diagnostic perspective. Furthermore, the inability to successfully use Fourier Transforms on an ECG signal due to its non-stationary and non-linear properties illustrates the uncertainty of the exact frequency cut-offs needed for a window-based filter. In [4][5] [6], adaptive filters (LMS, NLMS, BLMS, and so on) were proposed as an upgrade over window-based filters. Active signal monitoring and noise prediction based on the discrepancy between predicted and observable signals are examples of these techniques. The issues that occur when using these for ECG filtering start with the need for pure input signals in order to approximate how the signal differs. The adaptive filter's minimum Mean Square Error [MSE] criterion, as well as the fact that the reference signals are not well matched with the primary data, result in additional problems. As a consequence, although this is better than mere window-based filtering, there is still space for development. As a consequence of recent developments in nonlinear signal processing utilizing wavelets, wavelets have been more commonly utilized in ECG denoising. The thesis suggested by Donoho et al. [7] was advanced into the area of ECG denoising, and several algorithms [8] [9] [10] have been built on top of it. The standard wavelet technique, as verified by Luong et al. [11], is unsuccessful for removing baseline wander noise. A fuzzy rule-based multi wavelet methodology was proposed to solve this problem [12], which aided in the

optimization of pre and post filtering decisions for a given data collection. In his paper [13], Huang proposed Empirical Mode Decomposition as a modern method for studying non-linear non-stationary signals (EMD). This has been shown to be very effective in ECG filtration and is used in a variety of algorithms [14]. However, the majority of EMD work has been focused on the reduction of high-frequency sounds. When compared to direct denoising by EMD technique, denoising of ECG signals with the help of linear filtering (Butterworth Low pass filter) [15] yields better results. When EMD is combined with adaptive filters, the results are superior to when EMD is combined with linear filters. ECG denoising can also be accomplished using a combination of EMD and wavelets [16]. Zhang et al. [17] made even more changes to this. The energy of each IMF was analyzed in this study, and the noisy IMFs were identified as a result. The Wavelet Soft Thresholding Technique was used to denoise only the noisy IMFs. This reduces computation time and has proven to be the most efficient method to date. The problem with all of these techniques is that they are designed to deal with signals with lower IMFs or high frequency components, so they cannot be used to deal with Baseline Wander. EMD was first used for baseline wander denoising in a paper by Zhi-Dong Zhao et al. [18]. The residue signal was believed to be an approximation of baseline wander, and this residue was omitted to compensate for noise in the ECG signal. Na Pan et al. [19] made further progress in this field by decomposing the signal into 15 IMFs, with the assumption that the last three IMFs contained mostly noise components. The removal of the last three IMFs thus corrected the baseline wander. The limited nature of the signals they were tested on, as well as the lack of a suitable system to work on a varying number of IMFs, were issues with these works. Furthermore, it has been discovered that EMD's denoising property is insufficient on its own [34] [38]. As a result, it should be combined with another approach to achieve the best results. The EMD's main flaws are its noise sensitivity and the mode mixing property, which introduces error into the observed signal [35]. To address these issues, Huang suggested an Ensemble Empirical Mode Decomposition (EEMD) system [20], which was later improved [21], resulting in a numerically negligible error. Although EEMD has been used to denoise ECG [22], it has had some of the same issues with inadequate denoising strength on its own. With the help of averaging, this solves the problem of mode mixing and oscillations. However, this has a few drawbacks: it is computationally expensive, and the final amplitudes associated with the IMFs are greatly reduced due to the uneven number of modes across the ensemble [36].

Neural Networks are a type of data processing based on how the human brain works. Their use for filtering has been limited to PLI and electrode motion noise [24] and is based on signal reconstruction using existing biological signals, reducing their effectiveness. Further research, such as [25], used only the direct noisy values of the ECG signals as input, and found that networks trained on the direct values were overfit and had a diminished ability to filter novel ECG signals. For the prediction of non-linear non-stationary signals, EMD combined with Neural Networks has been widely used [26] [27]. For low frequency artefact removal from ECG signals, Mahipal Singh Choudhry et al. proposed using a cascade combination of EMD and Morphological function. Mahipal Singh Choudhry and Rajiv Kapoor suggested using CEEMD and Adaptive Morphological Function [33] for ECG denoising in a longer paper. The combination of CEEMD and Neural Network in our proposed algorithm is used to remove low frequency artefacts from ECG signals. The use of CEEMD eliminates the intrinsic flaws in EMD while also ensuring that the Neural Network is not overfit.

3. Background

A. Complete Ensemble Empirical Mode Decomposition

Any data set can be decomposed using CEEMD into a set of true Intrinsic Mode Functions (IMFs). The various IMFs obtained thus far are representations of the signal's various frequency components. It's a combination of EMD and EEMD [13]. The following is how CEEMD breaks down a signal into its constituent IMFs.

- Using EMD, K realizations of the base signal with added White Gaussian noise ($(n) + \varepsilon\ wk\ (n)$) are achieved to obtain their first modes and thereby to create an ensemble of first IMFs by taking average to obtain first true IMF, using the following formula

$$\widetilde{IMF_1[n]} = \frac{1}{K}\sum_{k=1}^{K} IMF_1^k(n) \qquad (1)$$

Where, (n) is the base signal (raw ECG signal for the first stage), $w^k(n)$ is White Gaussian noise for k^{Th} realization, $IMF_1^k(n)$is first IMF of k^{Th} realization, ε is weight coefficient of added noise and k = 1,2,3,4 … … … …K.

- At the first stage, residue is calculated by

$$r_1[n] = x[n] - \widetilde{IMF_1[n]} \qquad (2)$$

- • In the second stage of CEEMD, this residue signal is used as the base signal, and an ensemble of K residue realizations with White Gaussian noise is produced. The IMFs for each variable are first obtained, and then they are averaged to produce the second true IMF, which is provided as

$$\widetilde{IMF_2[n]} = \frac{1}{K}\sum_{k=1}^{K} E_1(r_1(n) + \varepsilon_1 E_1(w^k[n])) \quad (3)$$

- Similar to (2), the second residue is calculated and used for third true IMF.
- The above-mentioned steps are continued until further decomposition is no longer possible and all the obtained true IMFs are subtracted from base signal to obtain the final residue, given as

$$r[n] = x[n] - (IMF_1[n] + \widetilde{IMF}_2[n] + \ldots) \qquad (4)$$

Artifical Neural Network

Artificial Neural Networks (ANN) are a type of informative paradigm [28] that is based on how the human brain processes data. A neuron is the most basic component of such a system, and the processing power is determined by the number of these components and their interconnection. They necessitate initial training data, which must be carefully pre-processed to prevent overfitting. As a result, training must be done in such a way that the neural network's ability to respond to novel patterns is not harmed. The steps involved in setting up a network are

- Collection of data.
- Pre-processing of data.
- Initialization of weights of neural network and learning rate.
- Division of data set into training, testing and validation.
- Training of data: The network trains itself in such a way as to minimize a particular function such as MSE, the function to be minimized depends on the training rule to be used.
- Testing: A portion of the data set is reserved for testing of the network that has been developed. Here the output of the network is compared against a pre-determined output and its performance is checked.
- Cross Validation: This involves testing to ensure that training of the network is not leading to a loss in generalization.

4. Proposed Methodology

Flowchart of proposed method is given as Fig.1

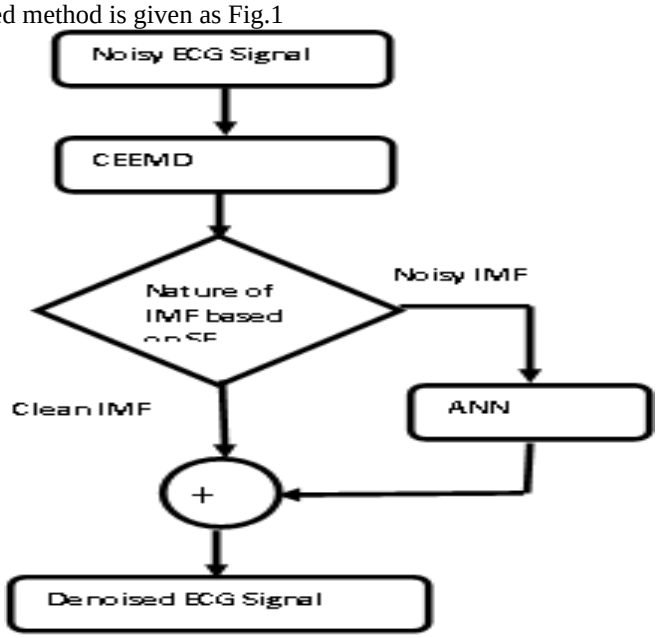


Fig.1 Proposed Methodology

The various steps of proposed method are

- Apply CEEMD to noisy ECG signal.
- IMFs are break down into two halves, those affected by baseline wander and those relatively free of it.
- A neural network is created and trained to filter each baseline affected IMF.
- The entire network is tested and cross validated.
- Results are obtained based on three output parameters on different types of noises and test subjects

The CEEMD operation is used to decompose a noisy signal into N true IMFs. These IMFs have now been grouped into two categories: (i) There are two types of IMFs: I noisy IMFs and (ii) clean IMFs. The Spectral Flatness (SF) of each individual IMF is used to calculate the number of noisy IMFs. The relationship between the Geometric mean (GM) and the Arithmetic mean (AM) is known as spectral flatness (AM). It is defined as the ratio of the power spectrum's geometric mean to its arithmetic mean.

Spectral Flatness (SF) =GM / AM (5)

Spectral Flatness measure is fuzzied to detect whether the IMFs are noisy or clean. Linguistic variable used are low, and high membership functions for Spectral Flatness.

The rules for spectral flatness are:

Rule 1: IF SF is in high range THEN IMF is clean

Rule 2: IF SF is in low range THEN IMF is noisy.

Individually, the noisy IMFs are applied to a separate Neural Network. Because each of these IMFs has a higher periodic nature than the combined full ECG signal, this performs filtering at the IMF level, making the Neural Network more efficient at filtration. This also eliminates the possibility of overfitting.

In the proposed method for Network with 10 hidden layers and 2 delays, the Levenberg–Marquardt algorithm [29] is used for the training process. A comparative analysis is used to determine the number of layers, and the number with the lowest error is chosen.

The steepest descent procedure and the Gauss–Newton algorithm are combined in the Levenberg–Marquardt algorithm. The algorithm's general concept is as follows: In areas where the curve is complex, it executes a combined training process, using the steepest descent algorithm. It switches to the Gauss-Newton algorithm when the curve can be safely estimated as quadratic, which saves a lot of time. The following is how the Levenberg–Marquardt algorithm is used to train:

1) With the randomly generated initial weights and evaluate the total error (SSE).
2) Do an update as directed by the following equation to adjust weights

$$w_{k+1} = w_k - (J_k^T J_k + \mu I)^{-1} J_k e_k \quad (6)$$

3) With the updated weights, evaluate the total error.
4) If the current total error is increased as a result of the update, then retract the step (reset the weight vector to the previous value), increase combination coefficient μ by a factor of 10 or by some other factor and then go to step 2 and try another update.
5) If the current total error is decreased as a result of the update, then accept the step (keep the weight vector as the current one), decrease the combination coefficient μ by a factor of 10 or by the same factor as step 4.
6) Go to step 2 with the new weights until the current total error is smaller than the required value.

5. Results

The MIT-BIH ST change database [30] is used to receive ECG signals for training and testing. This database contains 28 ECG recordings of varying lengths, sampled at 360 Hz, the majority of which were captured during exercise stress tests and show transient ST depression. The last five records (323 through 327) are ST elevation extracts from long-term ECG recordings. The pure ECG signal is augmented with a baseline wander noise estimation derived from the MIT BIH noise stress test database [31].

The above-mentioned data sets are then fed into a Levenberg–Marquardt algorithm with 10 hidden layers and 2 delays, which is used to train a neural network. For training, 70% of the total available data set is used. Input data from varying noisy ECG records is used to train the networks. When network generalization stops improving, cross validation is used to measure it and to stop training. Cross validation is carried out on 15% of the data set.
In order to test proposed method, Mean Square Error (MSE), Signal to Noise Ratio (SNR) and Correlation Coefficient ($R_{Xoriginal\ Xdenoised}$) are used as testing parameters. These parameters are defined as

$$\mathbf{MSE} = \boldsymbol{\Sigma}\left(\boldsymbol{X_{original}} - \boldsymbol{X_{denoised}}\right)^2 \quad (7)$$

$$\mathbf{SNR(dB)} = \mathbf{10} * \mathbf{log}\frac{\boldsymbol{\Sigma}(X_{original})^2}{\boldsymbol{\Sigma}(X_{original} - X_{denoised})^2} \quad (8)$$

$$\boldsymbol{R_{Xoriginal\ Xdenoised}} = \frac{cov\,(X_{original}, X_{denoised})}{\sigma_{Xoriginal}\ \sigma_{Xdenoised}} \quad (9)$$

Where, $\boldsymbol{X_{original}}$ is pure ECG signal, $\boldsymbol{X_{denoised}}$ is denoised ECG signal, $\boldsymbol{cov}$ is convolution function and $\boldsymbol{\sigma}$ is variance.
Fig. 2 - 4 show the plots of denoised signals using combinations of Neural Network with EMD, EEMD and CEEMD for three different ECG signals recordings.

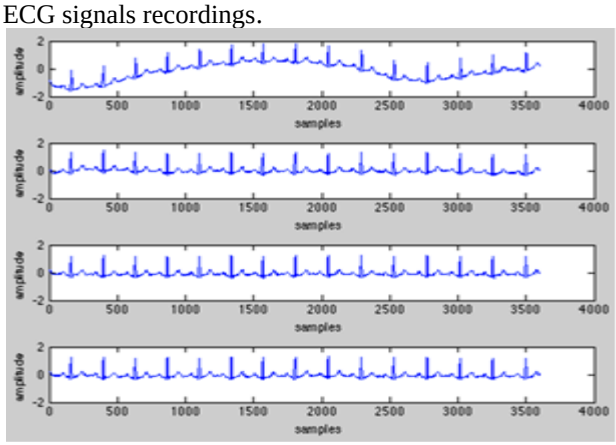


Fig.2: (a) Noisy signal (Recording-1) (b) Denoised signal using EMD and Neural Network (c) Denoised signal using EEMD and Neural network (d) Denoised signal using CEEMD and Neural network.

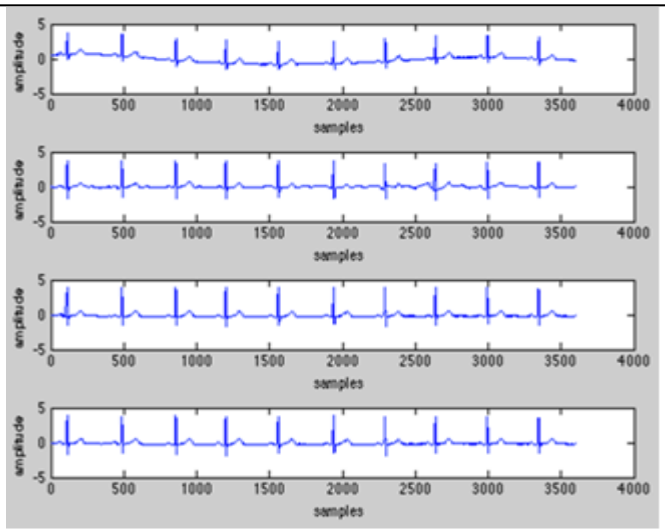


Fig.3: (a) Noisy signal (Recording-2) (b) Denoised signal using EMD and Neural Network (c) Denoised signal using EEMD and Neural Network (d) Denoised signal using CEEMD and Neural Network.

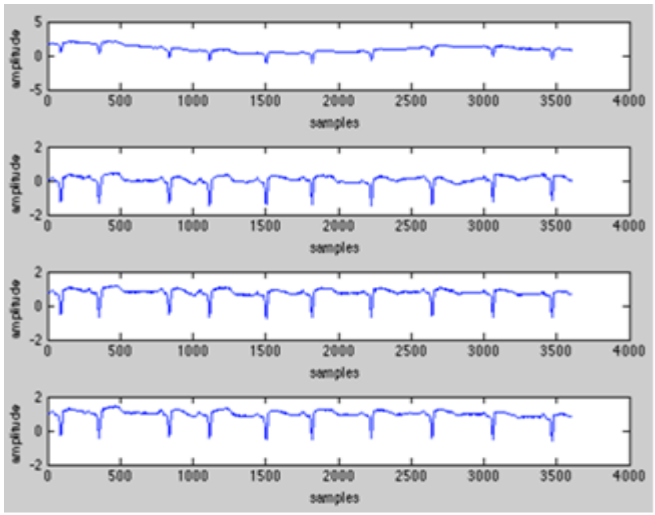


Fig.4: (a) Noisy signal (Recording-3) (b) Denoised signal using EMD and Neural network (c) Denoised signal using EEMD and Neural Network (d) Denoised signal using CEEMD and Neural Network.

The tables I-III show the results of denoising methods using combinations of Neural Network with EMD, EEMD and CEEMD for same three different ECG signal recordings.

TABLE I. RECORDING-1

Parameter	**(Input SNR = -19.0489, Input MSE = 0.3699)**		
	EMD and ANN	***EEMD and ANN***	***CEEMD and ANN***
MSE	*0.0122*	*0.0030*	*0.0016*
SNR	*14.8793*	*29.0949*	*35.3016*
Correlation Coefficient	*0.9655*	*0.9671*	*0.9898*

TABLE II. RECORDING-2

Parameter	**(Input SNR = -2.0869, Input MSE = 0.2115)**		
	EMD and ANN	***EEMD and ANN***	***CEEMD and ANN***
MSE	*0.0453*	*0.0076*	*0.0073*
SNR	*13.3992*	*31.2956*	*31.5658*
Correlation Coefficient	*0.9509*	*0.9881*	*0.9887*

TABLE III. RECORDING-3

Parameter	**(Input SNR = 15.0319, Input MSE = 0.2114)**		
	EMD and ANN	***EEMD and ANN***	***CEEMD and ANN***
MSE	*0.8182*	*0.0250*	*0.0033*

Parameter	(Input SNR = 15.0319, Input MSE = 0.2114)		
	EMD and ANN	***EEMD and ANN***	***CEEMD and ANN***
SNR	*1.5102*	*12.9336*	*56.6881*
Correlation Coefficient	*0.9066*	*0.9352*	*0.9966*

The CEEMD-based approach outperforms both EMD and EEMD in terms of results. For ECG signal recording-1, Fig.5 presents plots of denoising techniques using CEEMD and Morphological Operator, Neural Network, and a proposed combination of CEEMD and Neural Network. Table IV shows the results of the denoising technique for ECG signal recording-1 using CEEMD and Morphological Operator, Neural Network, and a proposed combination of CEEMD and Neural Network.

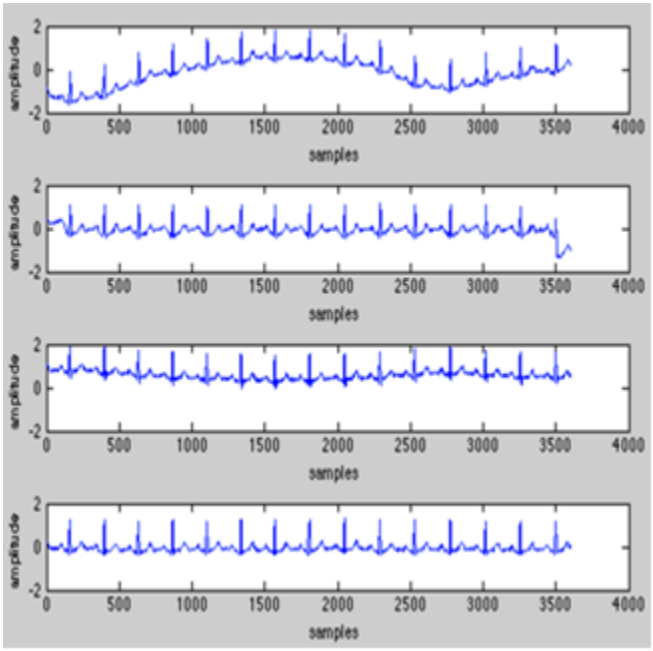


TABLE IV. RECORDING-1

Parameter	(Input SNR = -19.0489, Input MSE = 0.3699)		
	CEEMD and Morphological Operator	***ANN***	***CEEMD and ANN***
MSE	*0.0118*	*0.0095*	*0.0016*
SNR	*21.7005*	*23.7690*	*35.3016*
Correlation Coefficient	*0.9363*	*0.9457*	*0.9898*

In terms of all three parameters, the proposed technique outperforms the combination of CEEMD with Morphological Operator and Neural Network. Other ECG signal recordings yielded the same results.

6. Conclusion

When examining the results, it is clear that simple Neural Networks can produce excellent results at times, but their performance is inconsistent. The addition of CEEMD ensures that the system runs smoothly. The results show that the proposed method is clearly superior to other methods in the literature. The use of CEEMD has increased the filtration properties of Neural Networks while also preventing overfitting. Furthermore, since the filtration will be conducted mainly at lower frequencies, any high frequency cardiac defects will not be filtered, which is an important consideration.

References

1. Characterization of ECG Noise Source by Muris Mujagic.
2. Mahesh S. Chavan, * R A. Agarwala, ** M.D.Uplane ,“Application of the Chebyshev Type II Digital Filter For Noise Reduction In ECG Signal “ ,Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Malta, September 15-17, 2005 (pp1-8).
3. Mahesh S. Chavan, * R A. Agarwala, ** M.D.Uplane, “Digital Elliptic Filter Application For Noise Reduction In ECG Signal “ ,4th WSEAS International Conference on ELECTRONICS, CONTROL and SIGNAL PROCESSING, Miami, Florida, USA, 17-19 November, 2005 (pp.58-63).

4. Dr. K. L. Yadav and Sachin Singh, "Performance evaluation of different adaptive filters for ECG signal processing", International Journal On Computer Science and Engineering, vol. 40, no. 5, pp. 1880-1883, 2010.
5. Muhammad ZiaUrRahman, Rafi Ahamed Shaik and D.V.RamaKoti Roddy, "Efficient sign based normalized adaptive filtering techniques for cancelation of artifacts in ECG signals: Application to wireless biotelemetry", Journal of signal processing, vol. 91, no. 2, pp. 225-239, February 2011.
6. Ching-Haur Chang ,Kang-Ming Chang ,and Hsien-Ju Ko, "Cancellation of high frequency noise in ECG signals using Adaptive filter without external reference", Proceedings of International Conference on Biomedical Engineering and Informatics, pp. 787—790, Yantai, October 2010.
7. DL Donoho."Denoising by soft thresholding",IEEE Trans Inform Theoy, 14(3),1995, pp 612-627.
8. M Alfaouri and K Daqrouq, " Electrocardiogram(ECG) Signal denoising by Wavelet Transform thresholding", American Journal of Applied Sciences, Vol. 5 Issue 3pp.no 276-281,2008.
9. M Kania, M Fereniec, R Maniewski, " Wavelet Denoising for Multilead High resolution Electrocardiogram(ECG) Signals", Measurement Science Reiew, Volume 7, Section 2, No.4, pp no. 30-33, 2007.
10. L Chmelka and J Kozumpik, " Wavelet Based Wiener Filter for Electrocardiogram signal denoising", IEEE computers in cardiology, 2005, pp no. 771-774, 2005.
11. D.T luong, " Study on the limitations of removal of Baseline noise from electro cardiography signal in measurement using wavelet analysis", IEEE, ICUFN, 2013.
12. C.Y-F Ho ,B..W.-K, Ling, T.R-L. Wong, A.Y.-P chan. .P.-K.-S.Tam, "Fuzzy rule based multi-wavelet Electrocardiogm (ECG) denoising", Fuzzy Systems, IEEE international Conference,pp.no 1064-1068, 2008
13. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The Empirical Mode Decomposition and Hilbert Spectrum for Nonlinear and Nonstation- ary Time Series Analysis. Proceedings of the Royal Society London A., 454:903–995, 1998.
14. A.O Boudraa,J-C cexus,"EMD Based signal filtering", IEEE Transactions on instrumentation and measurement, vol 56,no. 6 ,pp 2196-2202, 2007.
15. Anil Chacko and Samit Ari, "Denoising of ECG signals using Empirical Mode Decomposition based technique" IEEE- International conference on Advances in Engineering, Science and Management (ICAESM -2012) March 30,31, 2012 pg no 6-9.
16. B. Pradeep Kumar, S Balambigai and R Asokan, "ECG denoising based on Hybrid Technique", IEEE-International conference on Advances in Engineering, Science and Management (ICAESM -2012), pp no. 285-290.
17. Changnian Zhang, XiaLi and Mengmeng Zhang, "A novel ECG signal denoising method based on Hilbert Huang Transform", 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering, CCTAE-2010, pg no 284-287.
18. Zhi-Dong Zhao, Yu-Quen Chen, "A new method for Removal of Baseline Wander and Power Line Interference in ECG signals", Proceedings if the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006,pg no. 4342-4347.
19. Na Pan, vai MingI, Mai Peng Un and Pun Sio hang,"Accurate removal of baseline wander in ECG using EMD", proceedings of NFSI & ICFBI 2007 China,pg no 177-180
20. Z. Wu and N. E. Huang, "Ensemble empirical mode de- composition: A noise-assisted data analysis method," Ad- vances in Adaptive Data Analysis, vol. 1, no. 1, pp. 1–41, 2009.
21. Mar´ıa E. Torres, Marcelo A. Colominas , Gasto´n Schlotthauer , Patrick Flandrin "A Complete Ensemble Empirical Mode Decomposition With Adaptive Noise " ICASSP 2011.
22. Gurpreet Singh, Student Member, IEEE, 2Gagandeep Kaur,3Vineet Kumar, "EC G Denoising Using Adaptive selection of IMFs through EMD and EEMD ", ICDSE 2014.
23. ECG pattern recognition and classification using non-linear transformations and neural networks: A review, Nicos Maglaverasa, , Telemachos Stamkopoulosa, Konstantinos Diamantarasb, Costas Pappasa, Michael Strintzis , International Journal of Medical Informatics,Volume 52, Issues 1–3, 1 October 1998, Pages 191–208
24. A Neural Network Approach to ECG Denoising, Rui Rodrigues (rapr@fct.unl.pt) and Paula Couto (mpcc@fct.unl.pt)
25. ECG signal denoising by Functional Link Artificial Neural Network (FLANN), International Journal of Biomedical Engineering and Technology,Volume 7, Issue 4
26. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm, Lean Yu , , Shouyang Wang, Kin Keung Lai, Energy Economics

27. Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model,Zhenhai Guoa, Weigang Zhaob, , , Haiyan Luc, Jianzhou Wangb , Renewable Energy, Volume 37, Issue1.
28. NEURAL NETWORKS by Christos Stergiou and Dimitrios Siganos
29. "Levenberg–Marquardt Training", Hao Yu, Bogdan M. Wilamowski Auburn University
30. MIT-BIH ST change Database , physiobank.org
31. MIT BIH NOISE STRESS TEST DATABASE, physiobank.org
32. Removal of baseline wander from ECG signal using cascaded Empirical Mode Decomposition and morphological functions 3rd International Conference on Signal Processing and Integrated Networks (SPIN) Feb 2016 DOI: 10.1109/SPIN.2016.7566803
33. Removal of Baseline Wander from ECG using CEEMD and Adaptive Morphological Function Journal of Chemical and Pharmaceutical Sciences Issue 4: October 2016.
34. Kumar, Shailesh & Panigrahy, Damodar & Sahu, Prasanna. (2018). Denoising of Electrocardiogram (ECG) signal by using empirical mode decomposition (EMD) with non-local mean (NLM) technique. Biocybernetics and Biomedical Engineering. 38. 10.1016/j.bbe.2018.01.005.
35. Dengyong Zhang, Shanshan Wang, Feng Li, Shang Tian, Jin Wang, Xiangling Ding, Rongrong Gong, "An Efficient ECG Denoising Method Based on Empirical Mode Decomposition, Sample Entropy, and Improved Threshold Function", Wireless Communications and Mobile Computing, vol. 2020, Article ID 8811962, 11 pages, 2020. https://doi.org/10.1155/2020/8811962
36. "Adaptive denoising of ECG using EMD, EEMD and CEEMDAN signal processing techniques", Krishna Teja1, Rahul Tiwari2 and Satish Mohanty3, Published under licence by IOP Publishing Ltd. Journal of Physics: Conference Series, Volume 1706, First International Conference on Advances in Physical Sciences and Materials 13-14 August 2020, Coimbatore, India Citation Krishna Teja et al 2020 J. Phys.: Conf. Ser. 1706 012077
37. Shi, Hao & Liu, Ruixia & Chen, Changfang & Shu, Minglei & Wang, Yinglong. (2021). ECG Baseline Estimation and Denoising With Group Sparse Regularization. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3056459.
38. Wang, Xiao & Zhou, You & Shu, Minglei & Wang, Yinglong & Dong, Anming. (2019). ECG Baseline Wander Correction and Denoising Based on Sparsity. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2902616.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/350726021

Effect of Fiber Length and Content on the Mechanical Properties of Pineapple Leaf Fiber Reinforced-Epoxy Composites

Article *in* Advances in Science and Technology · April 2021

DOI: 10.4028/www.scientific.net/AST.106.68

CITATIONS
0

READS
59

2 authors:

Mohit Mittal
Delhi Technological University
21 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Rajiv Chaudhary
Delhi Technological University
94 PUBLICATIONS **530** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Dates Extended for Full manuscript Submission at ICAPIE 2016, Delhi, India View project

Corrugation process parameters optimization View project

All content following this page was uploaded by Mohit Mittal on 08 April 2021.

The user has requested enhancement of the downloaded file.

Advances in Science and Technology
ISSN: 1662-0356, Vol. 106, pp 68-77
© 2021 Trans Tech Publications Ltd, Switzerland
Submitted: 2020-03-01
Revised: 2020-12-26
Accepted: 2020-12-30
Online: 2021-05-27

Effect of Fiber Length and Content on the Mechanical Properties of Pineapple Leaf Fiber Reinforced-Epoxy Composites

Mohit Mittal[1,a*] and Rajiv Chaudhary[2,b]

[1,2]Department of Mechanical, Production, and Industrial Engineering, Delhi Technological University, Delhi – 110042 (India)

[a]mohit.30mittal@gmail.com, [b]rch_dce@rediffmail.com

*Corresponding Author

Keywords: Pineapple leaf fiber (PALF); fiber length; fiber content; mechanical properties; polymer composites

Abstract. The behavior of a composite material under mechanical loading condition is significantly influenced by the geometrical dimensions (length and diameter) and the total content of reinforcing fiber. Therefore, this research work focused the effect of fiber length and content on the mechanical behavior of pineapple leaf fiber (PALF) reinforced-epoxy composites. In this regard, the total of four composite samples for each fiber length (10, 15, 20, and 25 mm) and content (17, 23, 34, and 43 vol.%) were developed using a hand lay-up molding technique and characterized for mechanical properties according to ASTM standards. The tensile and flexural strength of a composite was increased with the increase of PALF length and content up to 15 mm and 34 vol.% respectively. However, the composite of 25 mm fiber length with 43% fiber volume content exhibits the maximum impact strength.

Introduction

The trend of last few years confirmed the significant utilization of polymers and polymer-based composites in automobile, aerospace, and marine industries. The polymer-based composites possess wide range of advantageous properties such as low cost, low density, ease of shaped to the desired product, corrosion and chemical resistance, and low abrasiveness which leads to their utilization as an alternative to metals. However, the biodegradability and recyclability of plastic composites is still a major concern. Therefore, the scientists and technologists are putting their efforts for the bio-based materials which can be reusable at the end of life cycle and possess high mechanical strength [1-2]. Natural fibers (Jute, hemp, flax, bamboo, coir, and pineapple leaf) and some agro-industrial waste revealed the excellent shatter resistance, sonic insulation, flexibility, high specific strength and modulus. Despite their large extent of worthful properties, they show poor resistance to moisture absorption. The properties of a bio-composite are not only dictated by the composition and structure of reinforcing fibers but also a function of fiber geometry and volume fraction [3]. E. Osman *et al*.[4] reported that the water absorption of kenaf/Polyester composite was increased from 4.2% to 8.1% with the increase of kenaf fiber content from 10% to 40%. Among all types of natural fiber, the PALF is an outstanding material because it contains high cellulose (70-82%) and low lignin content (5-12%) which results in high tensile and low impact toughness [5]. S. Mishra *et al*. [6] reported that the PALF has great potential to be used in rubber industries and this was because of its high modulus (34.5-82.5 GPa). U. Wisittanawat *et al*. [7] stated that the tensile and tear strength of natural butadiene rubber (NBR) was increased significantly even after the loading of a small amount of PALF. N. Kengkhekit *et al*. [8] proved that the PALF can be effectively used as a reinforcing agent in the polypropylene matrix. The mechanical behavior of a composite material is the major function of fiber length and content. Therefore, it becomes necessary to firstly determine the optimum value of fiber aspect ratio and volume fraction for a specific pair of fiber-matrix system. S. Ozturk *et al*. [9] analyzed the effect of fiber loading on the mechanical properties of kenaf/PF and fiberfrax/PF composites. M.C.N Yemele *et al*. [10] investigated the effect of bark fiber content on the mechanical properties of Bark-HDPE composites. X. Zhao *et al*. [11] evaluated

All rights reserved. No part of contents of this paper may be reproduced or transmitted in any form or by any means without the written permission of Trans Tech Publications Ltd, www.scientific.net. (#557066604, Delhi Technological University, Bhiwani, India-08/04/21,10:26:13)

the effect of sisal fiber loading on the mechanical properties of HDPE. Y.A. El-Shekeil *et al*. [12] reported that the composite of 30% kenaf fiber content in the polyurethane matrix yields the maximum value of tensile strength than the other composition. B. Bakri *et al*. [13] stated that the Coir-Epoxy composite of 15:15 composition exhibits higher tensile and flexural properties than the other compositions (10:20; 20:10). W. Zahari *et al*. [14] studied the effect of fiber loading on tensile properties of Ijuk/PP composites. K. Zhang *et al*. [15] studied the effect of fiber length and content on flexural modulus and fracture toughness of the bamboo/epoxy composite.

From the above literature, it is proved that the fiber length and content play a major role to affect the mechanical properties of a composite material. Despite a lot of research has been done on the effect of fiber loading, there is no work available on the mechanical behavior of PALF/Epoxy composites. Therefore, an experimental investigation was done to determine the effect of fiber length and content on the mechanical behavior of PALF-Epoxy composites.

Experimental

Materials Used: Pineapple leaf fibers of different length (10, 15, 20, and 25 mm) were used as a reinforcement in polymer based composites. Initially, the fiber was of long form and cut into specified geometry by scissor. The inherent properties and surface morphology of PALF are shown in Table 1 and Fig. 1 respectively. The medium viscosity epoxy resin [DGEBA (diglycidyl bisphenol A); LY 556; density: 1.16 g/cm^3; average molecular weight (M_n): 2300 g/mol] and hardener [TETA (triethylene tetraamine); HY 951; density: 0.95 g/cm^3; average molecular weight (M_n): 146.23 g/mol] were procured from M/s Sakshi Dies and Chemicals, Delhi (India).The typical properties of cured epoxy thermoset are reported in Table 2.

Table 1 Physical properties of PALF

Density (g/cm^3)	0.98
Cellulose (%)	70-82
Hemicellulose (%)	18.8
Lignin (%)	5-12.7
Pectin (%)	1.1
Moisture content (%)	11.8
Microfibrillar angle (deg)	14
Diameter (μm)	20-80
Tensile strength (MPa)	413-1627
Young's modulus (GPa)	34.5-82.5
Elongation at break (%)	1.6

Table 2 Typical properties of epoxy polymer

The appearance of liquid resin	A clear pale yellow
Density (Kg/m^3)	1200-1250
Viscosity at 25^{0}C (cps)	550
Tensile strength (MPa)	11.14
Tensile modulus (GPa)	0.48
Flexural strength (MPa)	25.29
Flexural modulus (GPa)	1.53
Max. Elongation (%)	4.2-5.6
Impact strength (kJ/m^2)	1.2
Flashpoint	> 200^{0}C

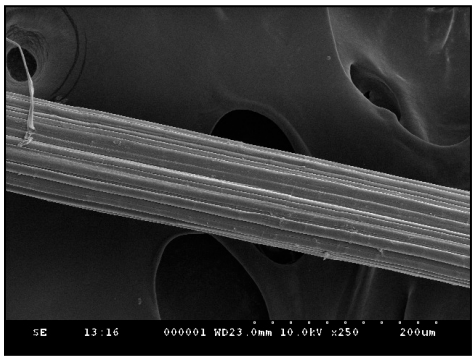


Fig.1 SEM micrograph of pineapple leaf fiber

Fabrication of Pineapple leaf fiber-Epoxy Composites: The hand lay-up molding technique was used for the fabrication of PALF-Epoxy laminate. The open glass mold [Fig. 2(a)] having dimensions of 200 X 100 X 4 mm^3 was used to fill by the fiber-matrix solution [Fig. 2(b)]. It should be taken care that the distribution of fibers and polymeric resin are uniform throughout the mold space and no air bubbles are present. During the process, the air bubbles are abstracted cautiously using a steel roller. The total four samples of each fiber length (10, 15, 20, and 25 mm) and volume content (17, 23, 34, and 43%) were fabricated. The specimens were cut from the molded sheet [Fig. 2(c)] as per ASTM D638 for tensile [Fig. 2(d)], ASTM D790 for flexural, and ASTM D256 for impact testing [Fig. 2(e)].

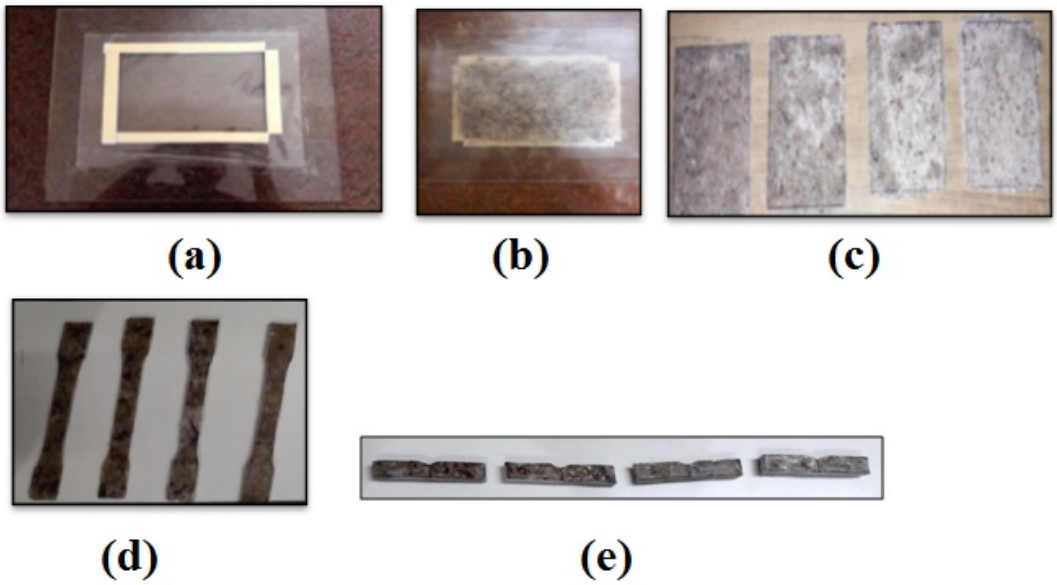


Fig. 2 (a) Glass mold covered with Teflon sheet (b) Mold with compounded material (c) PALF-Epoxy biocomposite sheets (d) Tensile specimens (e) Impact test specimens

Mechanical Tests: The tensile and flexural tests on composite samples were carried out at Table Top Tinnus Olsen H50KS, Universal Testing Machine as per ASTM D638 and ASTM D790 standard with a rate of 5 mm/min and 2 mm/min respectively. The impact strength of samples was determined using an Izod impact tester as per ASTM D256 standard.

Water Absorption Test: The water absorption capacity of composite samples was evaluated as per ASTM D570 standard at room temperature. The composite samples were immersed in distilled water up to a time period until the saturation state was reached. Initially, dry all the samples in the oven at 60 ^{0}C until the constant weight was obtained. After every 1 hr, the samples were taken out from the water bath and then weigh using a highly calibrated and reliable weighing machine. After each measurement, the percentage of water absorption was calculated using the formula:

$$Weightgain(\%) = \frac{(W_t - W_O)}{W_O} \times 100 \quad (1)$$

where W_t and W_O are the weight of the specimen after immersion time "t" and the oven dried weight respectively. The mechanism of water diffusion into the PMCs was determined by analyzing the diffusion exponent (n) according to the Eq. (2) and (3).

$$\frac{M_t}{M_m} = kT^n \tag{2}$$

$$LOG\,\frac{M_t}{M_m} = LOG(k) + nLOG(T) \tag{3}$$

where M_t and M_m are the water absorption at time t and at saturation level respectively. k is a constant parameter and n is the diffusion coefficient which defines the water diffusion mechanism.

Results and Discussion

Tensile Test: Fig. 3 shows the effect of fiber length and content on the tensile properties of PALF/Epoxy composites. It observed that the tensile strength and stiffness of composites were increased with the increase of fiber length and content up to 15 mm and 34 vol.% respectively. Further addition of fibers leads to the reduction in tensile properties and it was due to the improper wetting and agglomeration of fibers. Table 3 clearly depicts that the composites reinforced with 34% of fibers have carried higher tensile load as compared to the other composites. This result was attributed to the good interfacial adhesion between the fibers and matrix. From Fig. 3, we can also report that the composite of 25 mm fiber length with 43% fiber volume content exhibits minimum tensile strength (5.1 MPa) and tensile modulus (0.62 GPa) than the other composite samples. The ultimate tensile strength is calculated by using the formula:

$$\sigma_t = \frac{F}{wt} \tag{4}$$

Young's modulus = Tensile Stress/Tensile Strain (5)

Where,

F = Maximum load on the specimen

w = Initial width of the specimen

t = Initial thickness of the specimen

As from the results, it is clear that at an intermediate level of fiber loading, the stress distribution is uniform and fibers effectively participate to carry the load. The results obtained during the tensile test assured that the PALF is a good reinforcing agent for an epoxy thermoset matrix.

Table 3 Tensile properties of PALF/Epoxy composites

Sr. No	Fibre Length (mm)	Fibre Vol.%	Load (N)	Tensile Stress (MPa)	Change in length (mm)	Tensile Strain	Young's modulus (GPa)
1	10	17	608.4	11.7	1.1	0.011	1.02
2		23	1151.8	22.15	1.08	0.0108	2.05
3		34	1280.76	24.63	0.94	0.0094	2.61
4		43	551.2	10.6	0.43	0.0043	2.41
5	15	17	656.76	12.63	1.3	0.013	0.96
6		23	1096.16	21.08	1.17	0.0117	1.79
7		34	1954.68	37.59	1.24	0.0124	3.03
8		43	743.65	14.3	0.7	0.007	2.02
9	20	17	533.52	10.26	0.82	0.0082	1.24
10		23	680.68	13.09	0.76	0.0076	1.72
11		34	915.72	17.61	1.09	0.0109	1.61
12		43	424.32	8.16	0.77	0.0077	1.05
13	25	17	524.16	10.08	1.2	0.012	0.84
14		23	671.84	12.92	1.34	0.0134	0.96
15		34	674.96	12.98	1.15	0.0115	1.12
16		43	265.2	5.1	0.82	0.0082	0.62

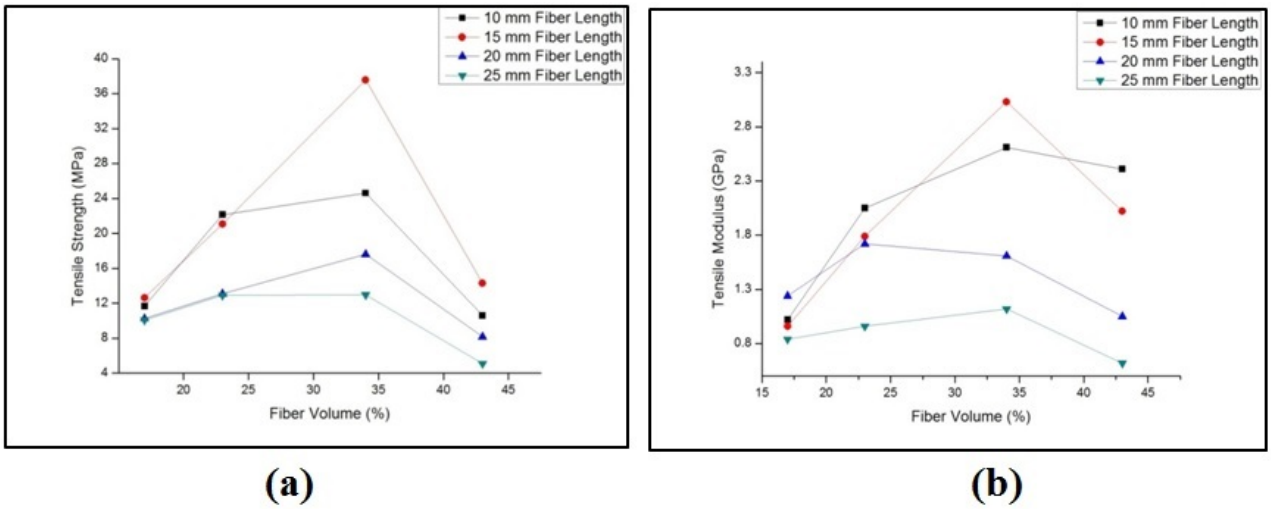


Fig.3 Effect of fiber length and content on (a) tensile strength (b) tensile modulus of PALF/Epoxy composites

Flexural Test: The flexural behavior of developed composites is shown in Table 4 and Fig. 4. It was observed that the composite of 34% PALF with 15 mm fiber length exhibited the higher flexural properties (strength and modulus) than the other composites. During flexural testing, the upper end of a specimen is subjected to a compressive force, lower end faced tensile load, and an axisymmetric plane is subjected to a shear load. Amongst all the developed specimens, the 43% fiber reinforced composite shows maximum deflection (5.05 mm) and it was due to the dominant action of fiber pull out mechanism. The flexural properties are calculated by using the formulas:

$$\sigma_f = \frac{3FL}{2wt^2} \tag{6}$$

$$Flexuralstrain = \frac{6st}{L^2} \tag{7}$$

$$Flexuralmodulus = \frac{L^3F}{4swt^2} \tag{8}$$

Where,
S = Measured deflection, L = Length of specimen

Irrespective of fiber length, the composite of 17 vol.% possess minimum flexural strength and it was due to the insufficient loading of reinforcing fibers in an epoxy matrix. However, the flexural strength of composites was reduced with the increase of fiber content from 34% to 43% and this was because of the improper wetting of fibers with viscoelastic polymeric resin.

Table 4 Flexural properties of PALF/Epoxy composites

Sr. No	Fibre Length (mm)	Fibre Vol.%	Load (N)	Flexural Strength (MPa)	Deflection (mm)	Flexural Strain	Flexural modulus (GPa)
1	10	17	108.52	43.41	3.85	0.0226	1.92
2		23	126.32	50.53	3.25	0.0191	2.64
3		34	142.77	57.11	2.03	0.0119	4.79
4		43	140.12	56.05	2.38	0.014	3.99
5	15	17	92.5	37	3.61	0.0212	1.74
6		23	130.65	52.26	2.33	0.0137	3.801
7		34	227.6	91.04	2.06	0.0121	7.5
8		43	221.25	88.5	2.38	0.014	6.305
9	20	17	92.7	37.08	2.96	0.0174	2.13
10		23	110.42	44.17	2.38	0.0141	3.13
11		34	194.25	77.7	4.42	0.0259	3
12		43	198.62	79.45	5.05	0.0296	2.68
13	25	17	104.3	41.72	3.24	0.019	2.19
14		23	107.5	43	2.83	0.0166	2.58
15		34	136.75	54.701	1.84	0.0108	5.06
16		43	133.25	53.33	1.87	0.011	4.79

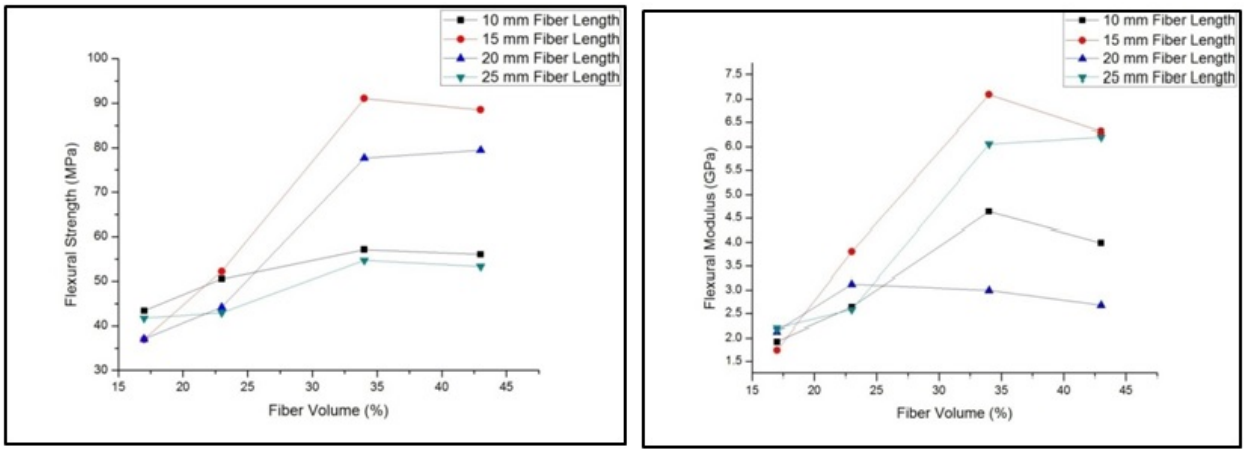


Fig.4 Effect of fiber length and content on (a) Flexural strength (b) Flexural modulus of PALF/Epoxy composites

Impact Test: For accurate estimation of impact toughness, the total 4 test run was performed. The performance of composites under impact loading condition is shown in Table 5 and Fig. 5. It observed that the impact strength was increased with the increase of fiber content. The maximum impact strength (3.15 J/m) was obtained at 25 mm fiber length with 43% fiber volume content. The average value of impact strength at 17% of PALF content is much lower (2.09 J/m) than the 43% fiber content (3.15 J/m). This was because of the predominant action of fiber fracture in spite of fiber pull out phenomena at a low level of fiber loading. When fiber content is low, fiber pull out mechanism do not play a lead role in energy absorption and the fiber deformation is small.

Table 5 Impact test results for PALF/Epoxy composites

Sr. No	Fibre Length (mm)	Fibre Vol.%	Test Run 1	Test Run 2	Test Run 3	Test Run 4	Average Value of impact strength (J/m)
1	10	17	1.61	1.61	1.76	1.82	1.7
2		23	1.72	1.62	1.89	1.93	1.79
3		34	2.23	2.05	2.89	2.89	2.51
4		43	2.34	2.01	2.73	2.67	2.43
5	15	17	1.93	1.84	2.08	2.13	1.99
6		23	2	1.9	1.97	1.86	1.93
7		34	2.71	2.68	2.75	2.83	2.74
8		43	2.91	3.08	3.07	3.11	3.04
9	20	17	2.04	2.31	2.22	2.07	2.16
10		23	1.97	2.04	2.01	2.04	2.01
11		34	2.05	2.18	2.42	2.34	2.24
12		43	2.73	2.84	2.77	3.03	2.84
13	25	17	2.13	2.34	1.78	2.11	2.09
14		23	2.02	2.07	2.02	1.97	2.02
15		34	2.88	2.67	2.91	2.91	2.84
16		43	3.01	2.91	3.47	3.24	3.15

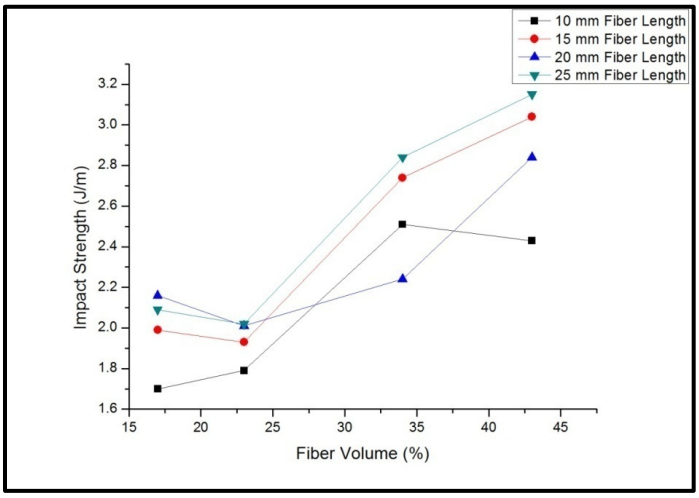


Fig.5 Effect of fiber length and content on the impact strength of PALF/Epoxy composites

Water Absorption Behavior: The effect of fiber length and content on water absorption behavior of PALF/Epoxy composites was determined as per ASTM D570 standard. Fig. 6 clearly revealed that the moisture sorption affinity of composites was increased with the increase of fiber volume fraction. Amongst all the developed specimens, the composite of 25 mm fiber length with 43% PALF content yields the maximum value of water absorption (approx. 9%) and this was due to the presence of more hydroxyl and polar functional groups. The water absorption was increased monotonically with immersion time and becomes stable after 12 hrs. Fig. 7 shows the effect of fiber length on water ingress capacity of composites. It observed that the water absorption was increased with the increase of fiber length. In all cases, the composites of 25 mm fiber length exhibit the highest affinity of water absorption and it was due to the presence of a large number of cavities and voids. Composites usually obey Fickian's diffusion mechanism and it depends on the relative mobility of penetrant (water molecules) and polymer segments. The water sorption constants (n and k) of different samples were obtained from fitting of experimental data into equation 3. The values of n and k, and water diffusion mechanism for the produced composites are presented in Table 6.

Table 6 Water sorption constants and diffusion mechanism of the PALF/Epoxy composites

Fibre Vol.%	Fibre Length (mm)	Diffusion exponent (n)	Intercept (k)	Type of diffusion mechanism
17	10	0.73	0.194	Non-Fickian Diffusion (0.5<n<1.0)
	15	0.8	0.187	
	20	0.71	0.216	
	25	0.81	0.167	
23	10	0.85	0.11	
	15	0.95	0.1	
	20	0.93	0.11	
	25	0.84	0.16	
34	10	0.46	0.214	Less Fickian Diffusion (n<0.5)
	15	0.51	0.179	Fickian Diffusion (n=0.5)
	20	0.5	0.18	
	25	0.8	0.134	Non-Fickian Diffusion (0.5<n<1.0)
43	10	0.84	0.135	
	15	0.77	0.14	
	20	0.76	0.149	
	25	0.79	0.147	

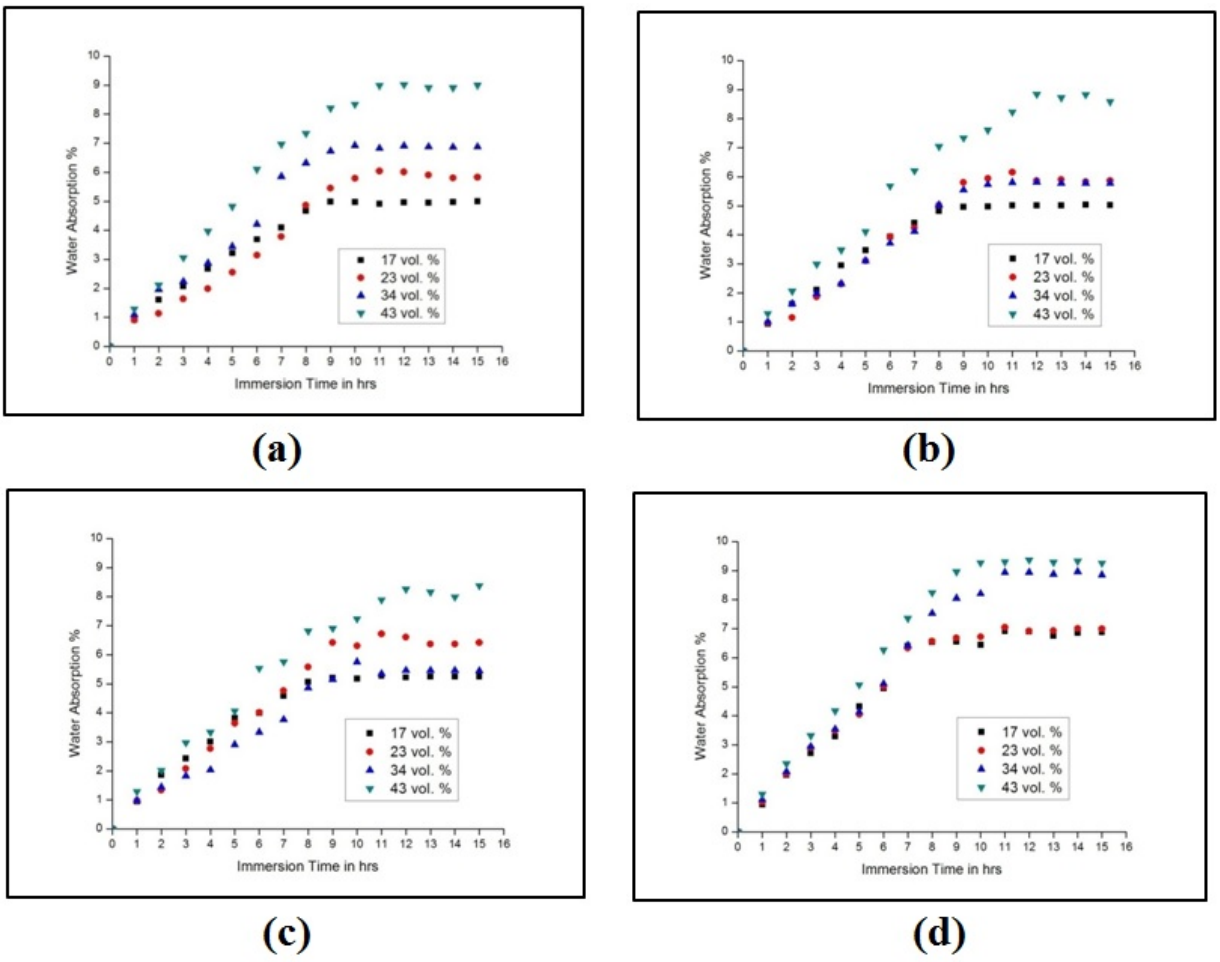


Fig.8 Effect of fiber content on water absorption of PALF/Epoxy composites: (a) 10 mm (b) 15 mm (c) 20 mm (d) 25 mm fiber length

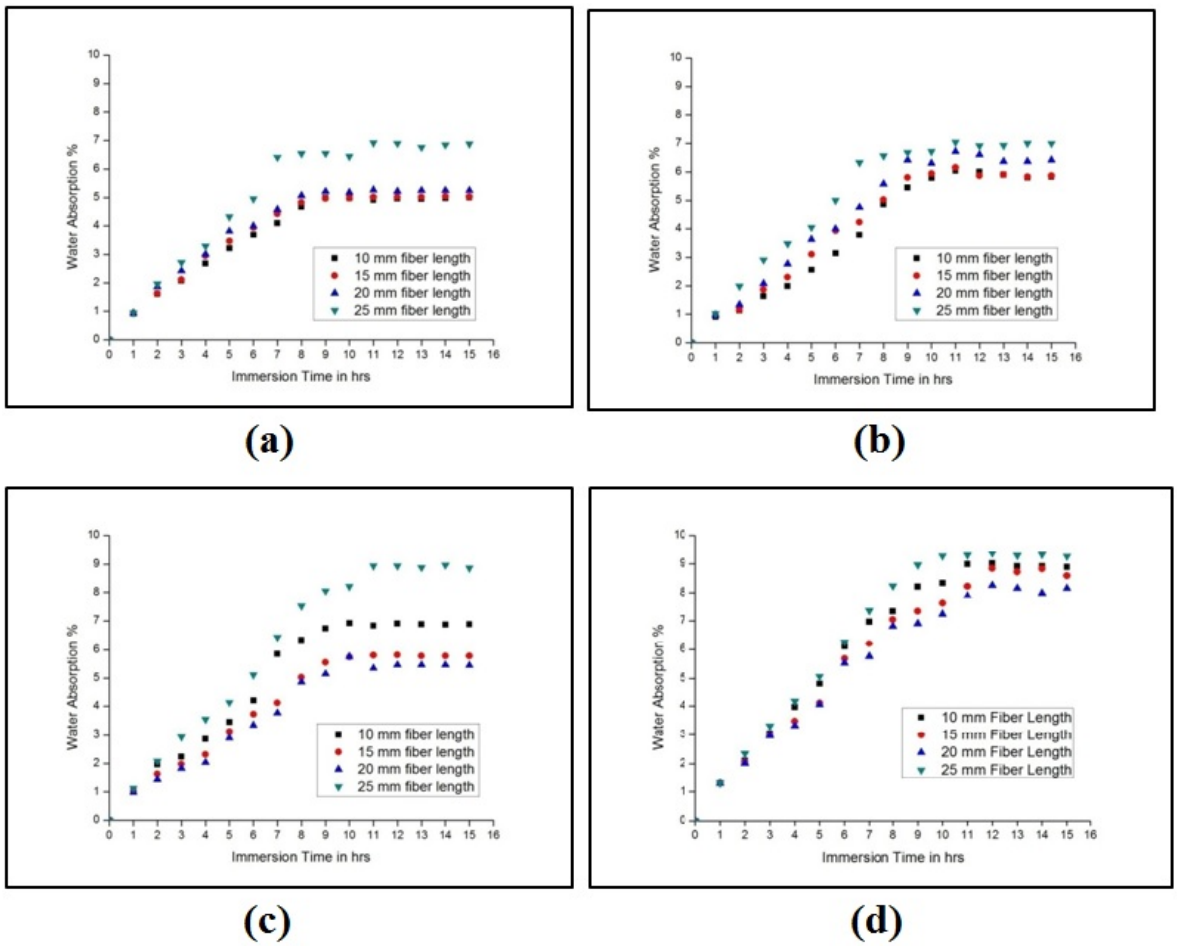


Fig.9 Effect of fiber length on water absorption of PALF/Epoxy composites (a) 17 vol.% (b) 23 vol.% (c) 34 vol.% (d) 43 vol.%

Conclusions

Based on the observations of an experimental study, the following conclusions can be made:

- The tensile and flexural strength of PALF/Epoxy composites were increased significantly with the increase of PALF length and content up to 15 mm and 34 vol. %. However, the maximum impact strength was obtained for 25 mm length and 43% PALF content.
- The degree of moisture absorption (%) was increased with the increase of fiber length and total fiber content. A large number of composites exhibited the non-fickian (anomalous) diffusion mechanism.
- The pineapple leaf fibers have significant potential to increase the mechanical strength of epoxy polymer if an optimal fiber length and content are used.

References

[1] T. Nishino, K. Hirao, M. Kotera, K. Nakamae, H. Inagaki, Kenaf reinforced biodegradable composite, Compos. Sci. Technol. 63 (2003) 1281-86.

[2] A.K. Mohanty, L.T. Drzal, M. Mishra, Novel hybrid coupling agent as an adhesion promoter in natural fiber reinforced polypropylene composites, J. Mater. Sci. Lett. 21 (2002) 1885-88.

[3] M. Ramesh, R. Logesh, M. Manikandan, N.S. Kumar, D. Pratap, Mechanical and water intake properties of banana-carbon hybrid fiber reinforced polymer composites, Mater. Res. 10 (2016) 1-12.

[4] E. Osman, A. Vakhguelt, I. Sbarski, S. Mutasher, Water absorption behaviour and its effect on the mechanical properties of kenaf natural fiber unsaturated polyester composites, 18th International Conference on Composite Materials.

[5] N. Kengkhekit, T. Amornsakchi, A new approach to "Greening" plastic composites using pineapple leaf waste for performance and cost effectiveness, Mater. Des. 55 (2014) 292-99.

[6] S. Mishra, A.K. Mohanty, L.T. Drzal, M. Mishra, G. Hinrichsen A review on pineapple leaf fibers, sisal fibers and their biocomposites, Macromol. Mater. Eng. 289 (2004) 955-74.

[7] U. Wisittanawat, S. Thanawan, T. Amornsakchai, Mechanical properties of highly aligned short pineapple leaf fiber reinforced–Nitrile rubber composite: Effect of fiber content and bonding agent, Polym. Test. 35 (2014) 20-27.

[8] N. Kengkhekit, T. Amornsakchi, Utilisation of pineapple leaf waste for plastic reinforcement: A novel extraction method for short pineapple leaf fiber, Ind. Crops. Prod. 40 (2012) 55-61.

[9] S. Ozturk, Effect of fiber loading on the mechanical properties of kenaf and fiberfrax fiber reinforced phenol formaldehyde composites, J. Compos. Mater. 44 (2010) 2265-88.

[10] M.C.N. Yemele, A. Koubaa, A. Cloutier, P. Soulounganga, M. Wolcott, Effect of bark fiber content and size on the mechanical properties of Bark/HDPE composites, Compos. Part A, 41 (2010) 131-137.

[11] X. Zhao, R. Li, S. Bai, Mechanical properties of sisal fiber reinforced high density polyethylene composites: Effect of fiber content, interfacial compatibilization, and manufacturing process, Compos. Part A, 65 (2014) 169-174.

[12] Y.A. El-Shekeil, S.M. Sapuan, K. Abdan, E.S. Zainudin, Influence of fiber content on the mechanical and thermal properties of kenaf fiber reinforced thermoplastic polyurethane composites, Mater. Des. 40 (2012) 299-303.

[13] B. Bakri, S. Chandrabakty, K. Putra, Evaluation of mechanical properties of coir-angustifolia haw agave fiber reinforced hybrid epoxy composite, J. Mech. Eng. 8 (2017) 679-685.

[14] W.Z.W. Zahari, R.N. Badri, H. Ardyananta, D. Kurniawan, F.M. Nor, Mechanical properties and water absorption behaviour of Polypropylene/Ijuk Fiber composite by using silane treatment, Procedia. Manuf. 2 (2015) 573-578.

[15] K. Zhang, F. Wang, W. Liang, Z. Wang, Z. Duan, Z. Duan, B. Yang, Thermal and mechanical properties of bamboo fiber reinforced epoxy composites, Polym. 10 (2018) 1-18.

View publication stats

International Journal of Environmental Science and Technology
https://doi.org/10.1007/s13762-021-03277-3

ORIGINAL PAPER

Energy efficient removal of COD from landfill leachate wastewater using electrocoagulation: parametric optimization using RSM

K. Faheem[1] · S. U. Khan[1] · M. Washeem[1] · S. U. Khan[2]

Received: 21 November 2020 / Revised: 17 February 2021 / Accepted: 16 March 2021
© Islamic Azad University (IAU) 2021

Abstract
In the present work, electrocoagulation method was employed using iron/aluminium electrodes for COD removal in the treatment of landfill leachate. The process was optimized with variables such as initial pH, applied current, initial COD and electrolysis time experimentally varied as per CCD design to know their influence on chosen response variables viz. COD removal efficiency and electrical energy consumption. The optimum conditions found were pH 5, applied current 0.5 A, initial COD concentration 5450 ppm and application time 9 min to achieve 64.25% efficiency with 0.59 W hour per gram removal of COD. The results prove that electrocoagulation process could be suitably used as an effective method for treatment of leachate. Operating cost of optimized condition was figured out as Indian rupee 0.211 (equivalent to 0.0028 US $) per gram of COD removed.

Keywords Landfill leachate · Electrocoagulation · Optimization · Energy consumption · Central composite design

Introduction

Increasing urbanization and industrialization have stressed the available water reserves (Khan et al. 2015). Exponentially rising population and changes in the utilization propensities, excessive generation of solid waste has grown into a serious environmental issue (Minelgaitė and Liobikienė 2019). In managing this waste, dumping at landfill sites has been the oldest and most common technique especially in the developing nations like India (Sharma and Jain 2019; Moraes et al. 2005). It is estimated that around annually 450–500 million tons of solid waste is generated globally and almost 70% of this is disposed via landfill sites (Ferronato and Torretta 2019). If such an enormous amount of solid waste is not properly dumped and safely disposed, leachate from it can migrate to soil and subsoil causing severe damage to the ecosystem of land and receiving water.

Landfill leachate is very complex in nature as it contain both oxidizable organic and inorganic compounds which directly and indirectly consume the available oxygen present in the ecosystem, thus contributing significantly to the COD value (Moreno-Casillas et al. 2007). During and after the landfilling, leachate gets generated due to the surface runoff, precipitation, infiltration, intrusion of ground water percolating through the landfill. This results in physiochemical reactions with the release of toxic contaminants into ground water. The nature and composition of leachate should therefore be investigated to minimize any risk, followed by adequate treatment (Veli et al. 2008). Leachate contains huge amount of organic and inorganic matter including chlorinated natural and synthetic matter, ammonia nitrogen which are very dangerous to human health (Kamaruddin et al. 2017). Heavy metals like Fe, Cr, Mn, Zn, Cr, Pb, Cu and Cd are also found in high concentration in leachate (Jotin et al. 2012; Elwakeel et al. 2020). The composition and concentration of leachate depends on the type of waste and landfill's age. Generally, landfills are divided into three categories based on their ages as young landfill (age less than 5 years), moderate landfill (age between

Editorial responsibility: J. Aravind.

Kashif Faheem and Saif Ullah Khan have contributed equally to this work.

✉ S. U. Khan
saif07amu@gmail.com

[1] Department of Civil Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University, Aligarh, India

[2] Department of Civil Engineering, Delhi Technological University, Delhi, India

Published online: 03 April 2021

5 and 10 years) and older landfill (age greater than 10 years) (Kang et al. 2002). Physico-chemical properties and various compounds qualities of leachate relying upon the period of landfill are shown in Table 1 (Labanowski et al. 2010).

Chemical oxygen demand (COD) is an analysis to measure the oxygen-depletion capacity of a wastewater sample polluted with organic waste material (Ren et al. 2017). It is considered to be an indicator of quality of water and an integral part of all management programs of water quality. It can assess the effect which the discharged wastewater will have on the receiving environment. COD specifically measures the equivalent amount of the oxygen necessary to oxidize soluble and particulate organic matter in wastewater; however, it oxidizes both organic and inorganic components. Additionally, COD is often used as a surrogate measure for carbon bioavailability (Pisarevsky et al. 2005). Thus, regulatory bodies and agencies have set some specific standards for greatest extent of "oxygen demand" for all discharged wastewaters (Moreno-Casillas et al.2007).

Leachates with complex chemical composition need adequate and appropriate treatment. Various treatment procedures thus employed include physiochemical methods such as advanced oxidation process, coagulation–flocculation, neutralization, ultra filtration, reverse osmosis, membrane filtration, dissolve air flotation (Kamaruddin et al. 2015). In most of these adopted methods, reagents are highly used while the solvent COD removal remains generally low. Furthermore, the chemical treatment could actuate optional contaminants because of the way that the compound reagent added may debase the treated water (Mirji et al. 2015). Aerobic and anaerobic biological treatment techniques mostly utilized for efficient treatment of leachate from streams including tannery wastewater, slaughter house wastewater, municipal wastewater, pharmaceutical wastewater, food beverage wastewater (Gotvajn et al. 2009; Ibrahim et al. 2020). However, proficiency of this treatment is influenced by the flow and composition of leachate in most of the biological treatment methods, there is general deficiency regarding expulsion of persistent organics (Ozturk et al. 2003). The main constraints with biological processes are requirement of large space, longer time durations, excessive production of sludge (Aziz et al. 2019). Therefore most of the above mentioned methods have one or the other drawbacks such as cost of treatment, low efficiency, and complex treatment operation.

With such complexities and technological advancements in conventional procedures, electrochemical assisted methods were explored. Among these, electrocoagulation (EC) is widely considered to be a simple, economical and an effective method of treating various types of waste water (Mahtab et al. 2020; Mateen et al. 2020; Naje et al. 2016; Öztürk and Özcan 2020; Jing et al. 2020; Khan et al. 2021). As compared to the traditional methods for treatment of wastewater electrocoagulation found that it offers a comparatively compact and robust treatment option in which electrochemical reactions that provide active metal cations for coagulation and flocculation are triggered by sacrificial metal anode. Electrocoagulation mechanism is quite complex with multiple reactions occurring inside a reactor simultaneously. Electrocoagulation works on the principle of coagulation chemistry resulting in production of destabilization agents which assist the removal of pollutants by neutralizing the charge in the solution (Khan et al. 2020a). Inside EC, with mostly iron or aluminium electrodes used, when current is passed, the cathode undergoes reduction while the anodic material undergoes oxidation to generate elemental metal ions Al^{3+} or Fe^{2+} as coagulant agents (Drogui et al. 2008). They form metal hydroxide like $Al(OH)_3$, $Fe(OH)_2$ and $Fe(OH)_3$ after combining with hydroxyl [OH^-] ions present in water. The floc formation of colloidal particles get started in the water sample after the coagulation is completed.

Table 1 Organic loading of landfill leachate pertaining to age factor

Leachate Age	Biodegradation (BOD_5/COD)	COD(g O_2 L^{-1})	pH	Organic load	Concentration in organic acids
Young Leachate	> 3 (medium)	> 20	< 7	Mainly unsaturated fats	> 80% of DOC* (High)
Moderate Leachate	0.1–0.3 (low)	3–15	= 7	Decrease in unsaturated and natural mixes with clear sub-atomic weight < 0.5 Da	Between 20 and 30% of DOC (Medium)
Steady Leachate	< 0.1(very low)	< 2	> 7	Basically natural mixes with obvious sub-atomic weight > 0.5 kDa	None
Crezin Leachate	0.18	0.38	7.2	70% of DOC were natural mixes with clear sub-atomic weight > 0.5 kDa	None

*DOC stands for Dissolved Organic Carbon

Some of these flocs rise to the top of the surface along with the H_2 gas bubbles produced and therefore can be easily separated (Feng et al. 2003). The generation of metal hydroxide using Fe or Al as anode and process mechanism inside the EC can be described simply from the equations [Eqs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 below] (Chen 2004; Mollah et al. 2001).

Mechanism: 1

At Anode

$$4Fe(s) \rightarrow 4Fe^{2+}(aq) + 8e^{-} \quad (1)$$

$$Al \rightarrow Al^{3+}(aq) + 3e^{-} \quad (2)$$

At Cathode

$$8H^{+}(aq) + 8e^{-} \rightarrow 4H_2 \quad (3)$$

Chemical Reactions

$$4Fe^{2+}(aq) + 10H_2O(I) + O_2 \rightarrow 4Fe(OH)_3 + 8H^{+}(aq) \quad (4)$$

$$Al^{3+}(aq) + 3H_2O \rightarrow Al(OH)_3 + 3H^{+}(aq) \quad (5)$$

Mechanism: 2

At Anode

$$Fe(s) \rightarrow Fe^{2+}(aq) + 2e^{-} \quad (6)$$

Chemical Reactions

$$Fe^{2+}(aq) + 2OH^{-}(aq) \rightarrow Fe(OH)_2 \quad (7)$$

$$Al^{3+}(aq) + 3OH^{-} \rightarrow Al(OH)_3 \quad (8)$$

At Cathode

$$2H_2O(I) + 2e^{-} \rightarrow H_2 + 2OH^{-}(aq) \quad (9)$$

Overall Reactions

$$Fe(s) + 2H_2O(I) \rightarrow Fe(OH)_2 + H_2 \quad (10)$$

$$Al^{3+} + 3H_2O \rightarrow Al(OH)_3 + 3H^{+} \quad (11)$$

The aim of this study includes multivariate optimization of EC process in achieving higher COD removal with minimum consumption of energy. Studies conducted so far to optimize the COD removal from landfill leachate generally lack consideration of energy consumption as response. Moreover, using combined Fe and Al electrodes was also tried for better efficiency and cost minimization. Process affecting factors such as initial pH, applied current, initial COD and electrolysis time were experimentally varied as per design to know their influence on chosen response variables viz. COD removal and electrical energy consumed.

Materials and methods

Study area

Ghazipur landfill site is situated in New Delhi, Indian capital which is one of the populous states of country. This landfill started functioning in 1948, has one of the tallest garbage mountains in India, spread-out in around 29 acres of land. Approximately 2000 tonnes of waste at Ghazipur landfill is dumped every day and it ascends by about 10 m per year forever to its putrid development. The site grabbed attention as tallest rubbish mountain with height ascending higher than Taj Mahal in 2020, turning into a fetid symbol for what the UN thinks about most polluted capital (Press 2020). Near to this landfill site are large markets of eatable items along with large population of slums. The water extracted by hand pumps and wells in these slums is used for all the domestic uses including drinkable purpose. Thus, landfill poses huge risk to population residing nearby with complete dependency on groundwater (Babbar et al. 2017).

The leachate was collected from Ghazipur landfill site, New Delhi. Samples were collected and stored in 5 dark water canes of capacity around 40 L each and transported to laboratory of Civil Engineering department of Aligarh Muslim University, Aligarh. With arrival, required amount of samples were taken and immediately processed for analysis. The analysis of the sample was done by determining few important parameters such as BOD_5, COD, pH, conductivity, chlorides, DO, TDS, phosphorus, nitrate and ammonia nitrogen according to Standard Methods for the Examination of Water and Wastewater. The characteristics values of leachate are shown in Table 2 below.

Table 2 Characteristics of leachate collected from Ghazipur landfill

S. No	Parameters	Observed values
1	BOD_5 (mg/L)	2310
2	PH	8.17
3	COD (mg/L)	5450
4	TDS	20,600
6	Chloride (mg/L)	7100
7	DO (mg/L)	2.41
8	Conductivity (µS/cm)	6208
9	Phosphorus (mg/L)	4.86
10	Nitrate (mg/L)	89.2
11	Ammonia nitrogen (mg/L)	504.6

Experimental set-up and procedures

The electrocoagulation reactor cylindrical in shape was made up of Plexiglas having 1.1 L capacity with an outlet port in the middle for collection of aliquots to be analyzed. In this study, a hybrid electrode made of iron and aluminum was used as anode at center of the coagulation reactor while the outer mesh made of iron was used as cathode and it is placed symmetrically around the anode [see Fig. 1a, b]. Dimension of the hybrid anode were 1.4 cm in diameter and 18 cm in length. The concentric gap between the anode and cathode was kept 2.4 cm. Reactor was placed over a magnetic stirrer for sufficient mixing and connected to a direct current source (SCIENTECH 4079) for supplying the current. pH alterations were done accordingly using HCl and NaOH and conductivity of experimental runs were adjusted using NaCl. Three samples were collected from the reactor in each run to report the mean values. The present experimental set-up and procedure adopted was almost similar to the one used in one of our previous study (Khan et al. 2020a). COD was measured as per standards (APHA 2005) by HACH reagents using DR 5000™ UV–VIS spectrophotometer.

Design of experiments

Design expert 10.0.1 (trial version) was employed to obtain a set of experimental runs according to central composite design. Factors varied were initial pH, applied current, and electrolysis time. PH was varied from 4 to 10, current from 0.2 A to 1 A, and time from 3 to 15 min. All the runs were performed at room temperature of 25 °C and a stirring speed of 350 rpm. The pH value was adjusted as per design using HCL and NaOH. Prior to each run, the electrolytic cell was washed with acetone to clear the surface of electrode. After each experimental run, sample was taken from the electrolytic cell into the test tube for measuring the COD.

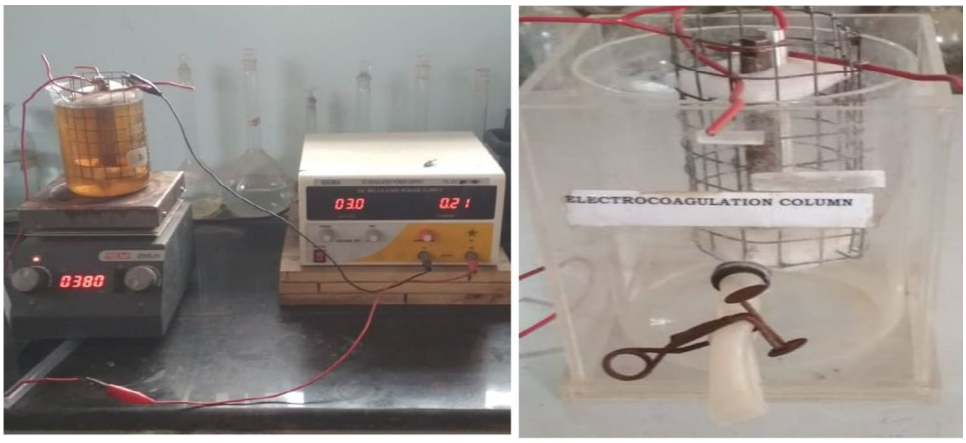
Statistical analysis and multivariate optimization

Multivariate optimization techniques are nowadays preferred for designing an experiment. This multivariate optimization is much better as compared to one factor analysis, as it is insufficient to determine the interaction among the choose variables in single factor analysis (Mahtab and Farooqi 2020). Response surface methodology (RSM) is basically a combination of statistical and mathematical means to formulate a relation between responses and variables. It functions by analyzing the process using differential calculus to obtain global minima or maxima for a given set of constraints.

Design expert analyses the experimental data for the given process variable and the model was subjected to sequential sum of square test and lack of fit to check its adequacy and obtain the ANOVA table. The fitness of the model was determined using the obtained R-square values with significant terms acknowledged at 95% confidence level while the non-significant term omitted on the basis of F values.

With the help of RSM for pre-decided conditions, quite significant models are generally obtained. A general second order model equation commonly adopted is given by the following:

$$r = \beta_o + \sum_{i=1}^{k} \beta_i p_i + \sum \sum_{i<j} \beta_{ij} p_i p_j + \sum_{i=1}^{k} \beta_{ii} p_i^2 + \in .$$



(a) (b)

Fig. 1 **a** Actual lab electrocoagulation set-up for COD removal, **b** Close up view of the electrode arrangement

where r stands for response, (ρ_i and ρ_j) are independent variables, (β_o, β_i, β_{ii} and β_{ij}) for offset term, linear, quadratic and interactive coefficients, respectively.

Experimental runs were carried out according to the central composite design for which set of data having real and encrypted values are given below in Table 3.

Results and discussion

Analysis of variance

Analysis of variance (ANOVA) was used to find the model acceptance. Significant terms were identified for the model based on F values. The value of R square obtained suggested a decent fit for the obtained models viz COD removal efficiency along with minimum energy consumption. The factors calculated are fit statistics, suggested quadratic and model applicability of 3D plots, and model satisfactoriness (see Tables 4, 5).

ANOVA tables for Quadratic models obtained for both the responses

For both the models, the predicted versus actual values were used for drawing the plots to know the models satisfactoriness. From the plots obtained as in Fig. 2, one can clearly notice that actual and predicted values are in close agreement signifying that model can be used to navigate the chosen design space.

The 3D surface plots were obtained for both the responses based on interaction between the process variables. These 3D plots have been shown below in the following Figs. 3, 4.

Process optimization

To determine the global maxima using design expert software, process optimization was carried out based on differential calculus. Input values for optimization

Table 3 Set of real and encrypted values

Parameter	Unit	Code	Real values				
			α	1	0	+1	$+\alpha$
pH	–	A	4	5.5	7	8.5	10
Current intensity	Ampere	B	0.2	0.4	0.6	0.8	1
Time	Minutes	C	3	6	9	12	15

Table 4 COD removal efficiency model

Source	Sum of squares	Df	Mean square	F value	p value	
Model	2224.29	9	247.14	31.49	< 0.0001	Significant
A-pH	1172.90	1	1172.90	149.45	< 0.0001	
B-Current Intensity	111.27	1	111.27	14.18	0.0037	
C-Time	589.39	1	589.39	75.10	< 0.0001	
AB	0.2113	1	0.2113	0.0269	0.8729	
AC	27.01	1	27.01	3.44	0.0932	
BC	2.76	1	2.76	0.3518	0.5662	
A^2	280.06	1	280.06	35.69	0.0001	
B^2	35.97	1	35.97	4.58	0.0579	
C^2	44.47	1	44.47	5.67	0.0386	
Residual	78.48	10	7.85			
Lack of Fit	67.71	5	13.54	6.29	0.0325	Significant
Pure Error	10.77	5	2.15			
Cor Total	2302.77	19				
Std. Dev	2.80		R^2			0.9659
Mean	40.96		Adjusted R^2			0.9352
C.V.%	6.84		Predicted R^2			0.7701
			Adeq Precision			20.4015

Table 5 Energy consumption model

Source	Sum of Squares	df	Mean Square	*F* value	*p* value	
Model	0.4363	9	0.0485	58.99	< 0.0001	Significant
A-pH	0.0021	1	0.0021	2.52	0.1435	
B-Current Intensity	0.2884	1	0.2884	350.95	< 0.0001	
C-Time	0.0443	1	0.0443	53.95	< 0.0001	
AB	0.0358	1	0.0358	43.54	< 0.0001	
AC	0.0049	1	0.0049	6.00	0.0343	
BC	0.0113	1	0.0113	13.72	0.0041	
A^2	0.0066	1	0.0066	8.06	0.0176	
B^2	0.0188	1	0.0188	22.87	0.0007	
C^2	0.0323	1	0.0323	39.32	< 0.0001	
Residual	0.0082	10	0.0008			
Lack of Fit	0.0073	5	0.0015	8.39	0.0179	Significant
Pure Error	0.0009	5	0.0002			
Cor Total	0.4445	19				

Std. Dev	0.0287	R^2	0.9815
Mean	0.3392	Adjusted R^2	0.9649
C.V. %	8.45	Predicted R^2	0.8652
		Adeq Precision	29.0084

Model Satisfactoriness:

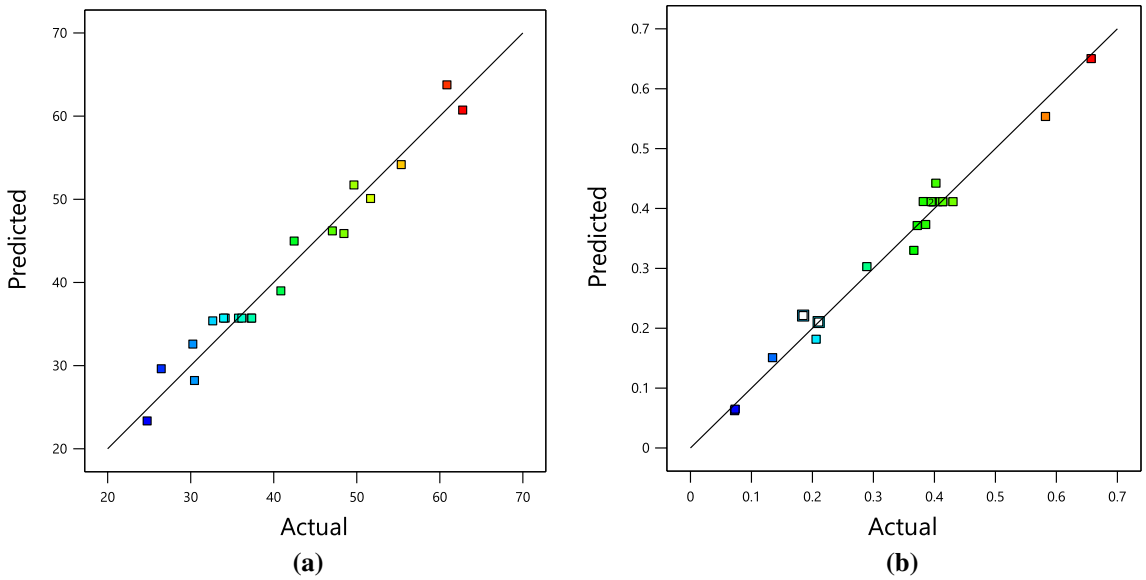


Fig. 2 Predicted versus Actual plots to know the satisfactoriness of both the models: **a** COD removal efficiency model, **b** Energy consumption model

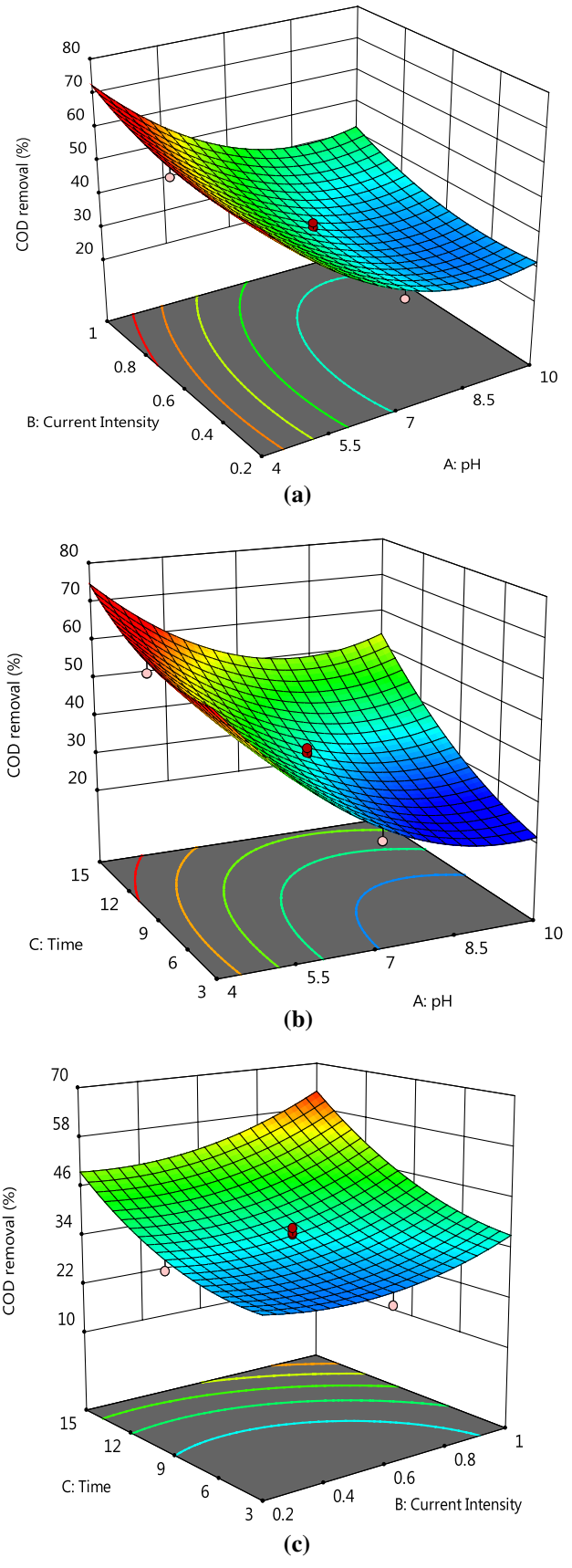


Fig. 3 3D surface plots showing the effect of interactive variables on COD removal efficiency signifying their role

as minimum, maximum, within range were applied for all the chosen process variables in the adopted range of design. The optimum conditions for desirability function value of 0.94 are shown in Table 6. Theoretical values of the optimum conditions were checked and confirmed by carrying out three set of experiments to know the mean values so as to compare them.

Effect of process parameters

Initial pH

Several studies have been done by researchers signifying the dominant effect of pH on electrocoagulation process (Akbal et al. 2011; Cañizares et al. 2009). Most of the studies carried out in past have showed that medium with acidic initial pH is favorable to achieve higher COD removal by electrocoagulation process (Tak et al. 2015). This may be due to the fact that the acid medium significantly affects coagulation mechanism, contributing to the bulk accumulation of multiple chemical species in wastewater. It may be inferred from Fig. 3a that using an acidic medium of pH 4 and applying the lowest current intensity of only 0.2A, resulted in COD removal up to 52%, while at higher pH value of 10 with maximum applied current intensity of 1 A, the COD removal attained was around 33% only, thus signifying the importance of pH. However, increasing the current intensity from minimum 0.2A to maximum 1A for the same alkaline medium at pH 10 could only enhance the COD removal from 33 to 44%. Several previous studies carried out in past validate these results with significant role of pH (Tak et al. 2015). Also, pH does not remain same, it often tends to change with time according to most of the previous investigations using EC (Parga et al. 2005; Khan et al. 2019). Based on the pH of medium, the metal hydroxides produced in the electrocoagulation process vary due to various simultaneous reactions occuring at cathode and anode (Eqs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11). Positively charged colloid particles of ferric hydroxide gets generated excessively at low pH values, with substantial decrease as pH advances toward being highly alkaline. This reduction in low generation of ferric hydroxides at higher pH values can be justified from Fe speciation diagram, resulting in lesser availability of metal hydroxide for the reaction to occur. Aluminum ions (Al^{3+}) generated by anode dissolution are quickly subjected to random hydrolysis reactions which yield different monomeric compounds across a wide range of pH (Tak et al.

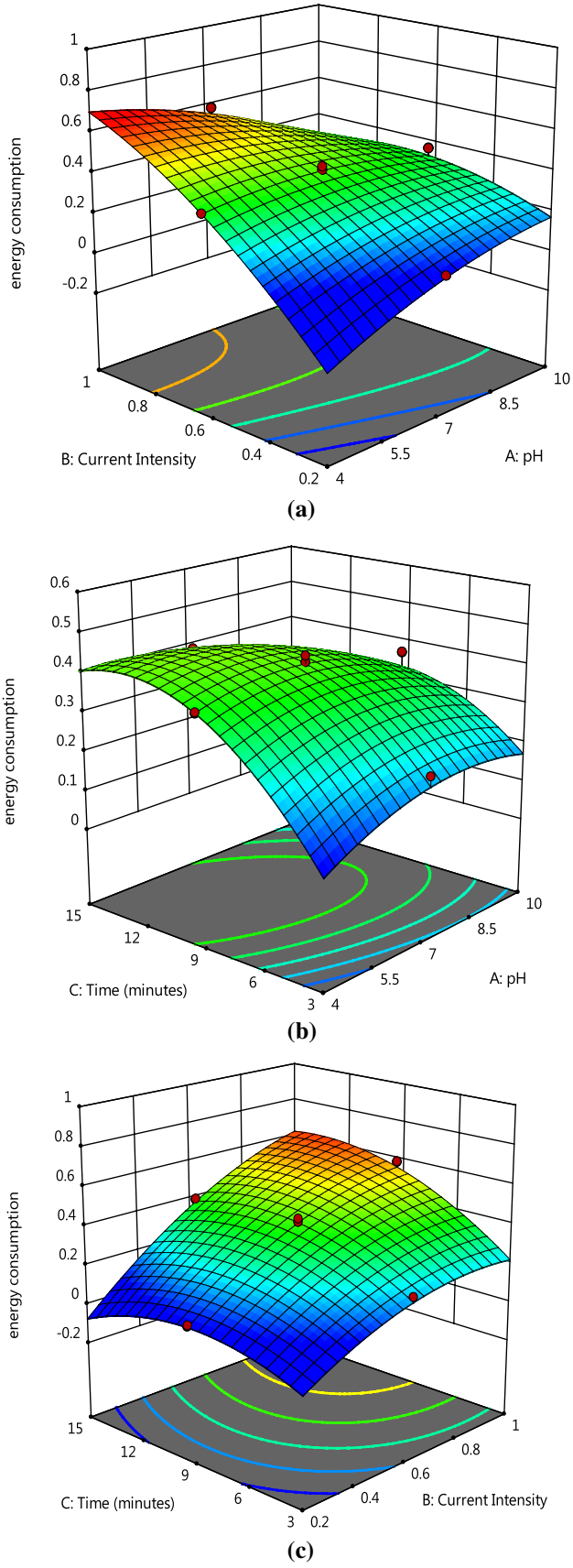


Fig. 4 3D surface plots showing the effect of process variables on energy consumption for removal of COD signifying their role

2015). At low pH, the breakdown of the aluminum anode generates Al^{3+} and $Al(OH)_2^+$ that are converted directly into $Al(OH)_3$ at acceptable pH values. The large surface area of these $Al(OH)_3$ help to absorb metal oxides and organic compounds very quickly (Daneshvar et al. 2006).

Electrolysis time

Electrocoagulation time duration or application time is also a crucial parameter which influences the entire process. The dissipated electrode mass naturally increases with the passage of time. Thus, generation of Al^{3+} ions from aluminum electrodes and Fe^{2+} from iron electrodes depends considerably on the electrolysis period, thus affecting the performance of the electrocoagulation process (Hashim et al. 2019). The COD removal efficiency get naturally enhanced by increasing the electrolysis time due to sufficient mixing and available time (Golder et al. 2010). Increasing the electrolysis time causes the increase in formation of hydroxyl radicals and ions concentration and their flocs which contribute to coagulation and flocculation resulting in higher COD removal (Asaithambi et al. 2016). It may be inferred from the obtained RSM plots in Fig. 3b that keeping the medium as acidic with pH at 4 and operating the system for just 3 min resulted in COD removal up to 60%, however, increasing this application time from 3 to 15 min at the same pH value of 4 brought a reasonable increment in removal efficiency attained as 72%. However, the effect of increasing this application time from 3 to 15 min was quite significant with initial pH value of 10, resulting in increase of COD removal from 21 to 55%.

Current intensity

Current intensity is considered as one of the most influential parameter of EC process as increasing current intensity causes increased release of ions such as Al^{3+} and Fe^{3+} from anode along with excessive forming of hydrogen bubbles, resulting in higher upward flux which ultimately enhance COD removal (Feng et al. 2007). It was noticed that at very high current intensity of 1A even for short time duration of 3 min, the COD removal achieved was significantly up to 38%, however the maximum removal efficiency up to 61% was accomplished at longer time duration of 15 min. It can be evidently seen from Fig. 3a that with initial pH 4 and applying

Table 6 Theoretical and experimental values of the optimized condition

S. No.	Process variables	Optimum conditions	Experimental mean values
1	pH	5	5
2	Current Intensity (A)	0.5	0.5
3	Electrolysis time (min)	9	9
4	COD removal efficiency (%)	64.25	63.2
5	Energy consumption (watt-hour per gram)	0.59	0.578

the lowest current intensity of 0.2A, COD removal up to 52%, while increasing this current intensity up to 1A for the same initial pH level resulted in enhanced COD removal of 72%. It may be inferred that the current intensity inside the EC system has a steadily consistent effect on the rising concentrations of electro-dissolved charges of iron and aluminum, thus resulting in higher coagulation performance and consequently increased COD removal. This was found in accordance with several past studies (Thirugnanasambandham et al. 2015), in which it was reported that an improvement in current intensity up to 0.3A resulted an increased reduction of COD from the pulp and paper industry wastewater. However, keeping higher values of current along with longer time durations result in excessive sludge production (Khan et al. 2019).

Effect on energy consumption

It is clearly evident from the obtained 3D plot in Fig. 4c that the consumption of energy was higher at higher values of current intensity and longer durations; however, at low current supply, the energy consumption was although low but time required for COD removal was quite long. As depicted from the plot 4 (a) and (c), at low current intensity variation in pH did not affect the energy consumption significantly, however increasing the current intensity at acidic pH resulted in maximum energy consumption. Moreover, the energy consumption was noticed quite high at extended time duration for low values of pH, as evident from Fig. 4b. Also, it was observed that energy consumption depends upon the strength of wastewater with high initial COD requiring more energy and the optimized current supply so as to minimize the operational cost of EC process. Optimized conditions suggest that system needs to be operated at low current values for longer durations at slightly acidic pH.

Estimation of energy consumption

The consumption of electrical energy consumed per gram removal of COD which then can be expressed as watt-hour per gram. For calculation of energy consumed, the formulae used were as follows:

Energy consumed per gram COD removal

$$= \frac{EIt \times 60}{(\mathrm{Co} - C) \times V} (\text{in joules/gram}) \tag{12}$$

Energy consumed per gram COD removal:

$$= \frac{EIt \times 60}{(\mathrm{Co} - C) \times V} \times 2.778 \times 10^{-4} (\text{in watt} - \text{hour/gram}) \tag{13}$$

where E is voltage in volt, t is time in minutes, I is current in ampere and V is the volume in liters. Co and C are initial and final concentrations in mg/l, respectively.

Dissociation of electrode can be calculated by Faraday's law (Khan et al. 2019)

$$m = \frac{I.M.T}{Z.F}, \tag{14}$$

where m (gram), I (coulomb), M (grams/mol) and T (minutes) are material dispersed at an electrode, charge, molar mass and time, respectively. Z is valency and F is Faraday's constant equal to 96,500 C/mol. The total cost of the system for electrocoagulation was calculated combining the electrode dissipation cost and cost energy consumption. Thus,

$$C_{\text{total}} = C_{\text{electrode}} + C_{\text{energy}}, \tag{15}$$

where $C_{\text{electrode}}$ is the amount for the dissipation of electrode and C_{energy} is the amount for the consumption of electrical energy per gram removal of COD (Khan et al., 2020b).

Cost estimation

In electrocoagulation process, the operational cost is estimated mainly undertaking cost of electrode and electricity charges incurred ignoring the minor factors like sludge management & maintenance. So operational cost (C_t) can be computed using the above Eq. 15.

The dissipated mass from electrodes can be computed using Faraday's law (according to Eq. 16);

$$\mathrm{m} = \frac{\mathrm{I} \times \mathrm{M} \times \mathrm{t}}{\mathrm{Z} \times \mathrm{F}} \tag{16}$$

The assumed dissoluted mass of electrode was found to be 0.0324 g per run at optimized condition. Considering the residential electricity rate norms and local market price of iron and aluminium in India, the costs were Rs

0.019 and Rs 0.194 Indian rupees, respectively, thus total cost of optimized operation was computed as Rs 0.211 (equivalent to 0.0028 US $) per gram of COD removed.

Conclusion

In this study, electrocoagulation process was applied to treat landfill leachate for the removal of COD. The EC system used was quite simple in design and operation. Few operational parameters such as applied current, pH of solution and electrolysis time were varied to find out their interactive effect on COD removal. Central composite design (CCD) was used to generate the quadratic models along with experimental runs and the models aiming minimum energy consumption per unit removal of COD aiming maximum efficiency. Acidic medium with reasonable application time duration was found most suitable and optimized condition for the EC process. The maximum COD removal (71.7%) was achieved with current supply of 1A for time duration of 14 min at pH around 5. Energy consumption for maximum removal was estimated to be 0.78 (W-hour/gram) with cost of operation for the optimized run found quite low, suggesting the need of optimization process. However, it was observed that removal efficiency increased by increasing applied current, as a result of which energy consumption also increased, thus making the process somewhat uneconomical. Thus, supplying with low current intensity in slight acidic medium with increased electrolysis time proves to be the most suitable, economical and energy efficient way for the EC process to achieve the objective.

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.

References

Akbal F, Camci S (2011) Copper, Chromium and Nickel removal from metal plating wastewater by electrocoagulation. Desalination 269(1–3):214–222

Asaithambi P, Aziz A, Raman A, Wan AZ, Daud MABW (2016) Integrated ozone—electrocoagulation process for the removal of pollutant from industrial effluent: Optimization through response surface methodology. Chem Eng Process Process Intensif 105:92–102

Aziz A, Basheer F, Sengar A, Khan SU, Farooqi IH (2019) Biological wastewater treatment (anaerobic-aerobic) technologies for safe discharge of treated slaughterhouse and meat processing wastewater. Sci Total Environ 686:681–708

Babbar P, Verma S, Mehmood G (2017) Groundwater contamination from non-sanitary landfill sites—a case study on the Ghazipur Landfill Site, Delhi (India). Int J Appl Environ Sci 12(11):1969–1991

Cañizares P, Jiménez C, Martínez F, Rodrigo MA, Saez C (2009) The pH as a key parameter in the choice between coagulation and electrocoagulation for the treatment of wastewaters. J Hazardous Mater 163(1):158–164

Chen G (2004) Electrochemical technologies in wastewater treatment. Sep Purif Technol 38(1):11–41

Drogui P, BlaisJF MG (2008) Review of electrochemical technologies for environmental applications. Recent Patents Eng 1(3):257–272

Elwakeel KZ, Elgarahy AM, Khan ZA, Almughamisi MS, Al-Bogami AS (2020) Perspectives regarding metal/mineral-incorporating materials for water purification: with special focus on Cr (vi) removal. Mater Adv 1(6):1546–1574

Feng C, Sugiura N, Shimada S, Maekawa T (2003) Development of a high performance electrochemical wastewater treatment system. J Hazard Mater 103(1–2):65–78

Feng J, Sun Y, Zheng Z, Zhang J, Li S, Tian Y (2007) Treatment of tannery wastewater by electrocoagulation. J Environ Sci 19(12):1409–1415

Ferronato N, Torretta V (2019) Waste mismanagement in developing countries: a review of global issues. Int J Environ Res Public Health 16(6):1060

Golder A, Kumar H, Samanta A, Ray S (2010) Colour diminution and COD reduction in treatment of coloured effluent by Electrocoagulation. Int J Environ Eng 2(1–3):228–238

Gotvajn AŽ, Tišler T, Zagorc-Končan J (2009) Comparison of different treatment strategies for industrial landfill leachate. J Hazard Mater 162(2–3):1446–1456

Hashim KS, Al Khaddar R, Jasim N, Shaw A, Phipps D, Kot P, Pedrola MO, Alattabi AW, Abdulredha M, Alawsh R (2019) Electrocoagulation as a greentechnology for phosphate removal from river water. Sep Purif Technol 210:135–144

Ibrahim S, El-Liethy MA, Elwakeel KZ, Hasan MA, Al Zanaty AM, Kamel MM (2020) Role of identified bacterial consortium in

treatment of Quhafa wastewater treatment plant influent in Fayuom. Egypt Environ Monit Assess 192(3):1–10
Jing G, Ren S, Gao Y, Sun W, Gao Z (2020) Electrocoagulation: a promising method to treat and reuse mineral processing wastewater with high COD. Water 12(2):595
Jotin R, Ibrahim S, Halimoon N (2012) Electro coagulation for removal of chemical oxygen demand in sanitary landfill leachate. Int J Environ Sci 3(2):921–930
Kamaruddin MA, Yusoff MS, Aziz HA, Hung YT (2015) Sustainable treatment of landfill leachate. Appl Water Sci 5(2):113–126
Kamaruddin MA, Yusoff MS, Rui LM, Isa AM, Zawawi MH, Alrozi R (2017) An overview of municipal solid waste management and landfill leachate treatment: Malaysia and Asian perspectives. Environ Sci Pollut Res 24(35):26988–27020
Kang KH, Shin HS, Park H (2002) Characterization of humic substances present in landfill leachates with different landfill ages and its implications. Water Res 36(16):4023–4032
Khan SU, Noor A, Farooqi IH (2015) GIS application for groundwater management and quality mapping in rural areas of District Agra, India. Int J Water Resour Arid Environ 4(1):89–96
Khan SU, Islam DT, Farooqi IH, Ayub S, Basheer F (2019) Hexavalent chromium removal in an electrocoagulation column reactor: process optimization using CCD, adsorption kinetics and pH modulated sludge formation. Process Saf Environ Prot 122:118–130
Khan SU, Farooqi IH, Usman M, Basheer F (2020a) Energy efficient rapid removal of arsenic in an electrocoagulation reactor with hybrid Fe/Al electrodes: process optimization using CCD and kinetic modeling. Water 12(10):2876
Khan SU, Asif M, Alam F, Khan NA, Farooqi IH (2020b) Optimizing fluoride removal and energy consumption in a batch reactor using electrocoagulation: a smart treatment technology. In: Ahmed S, Abbas S, Zia H (eds) Smart cities—opportunities and challenges. Lecture notes in civil engineering, vol 58. Springer, Singapore, pp 767–778
Khan SU, Mahtab MS, Farooqi IH (2021) Enhanced lead (II) removal with low energy consumption in an electrocoagulation column employing concentric electrodes: process optimisation by RSM using CCD. Int J Environ Anal Chem. https://doi.org/10.1080/03067319.2021.1873304
Labanowski J, Pallier V, Feuillade-Cathalifaud G (2010) Study of organic matter during coagulation and electrocoagulation processes: application to a stabilized landfill leachate. J Hazardous Mater 179(1–3):166–172
Mahtab MS, Farooqi IH (2020) UV-TiO2 process for landfill leachate treatment: optimization by response surface methodology. Int J Res Eng Appl Manag 5(12):14–18
Mahtab MS, Islam DT, Farooqi IH (2020) Optimization of the process variables for landfill leachate treatment using Fenton based advanced oxidation technique. Eng Sci Technol Int J. https://doi.org/10.1016/j.jestch.2020.08.013
Mateen QS, Khan S, Islam DT, Khan NA (2020) Copper (II) removal in a column reactor using electrocoagulation: parametric optimization by response surface methodology using central composite design. Water Environ Res 92(9):1350–1362
Minelgaitė A, Liobikienė G (2019) Waste problem in European Union and its influence on waste management behaviours. Sci Total Environ 667:86–93
Mirji G, Kalburgi PB (2015) Application of electrocoagulation mechanism for COD removal of dairy wastewater. Int J Mech Prod Eng 6(11):86–88
Mollah MYA, Parga JR, Schennach R, Cocke DL (2001) Electrocoagulation (EC)—science and applications. J Hazard Mater 29–41(84):29–41
Moraes PB, Bertazzoli R (2005) Electrodegradation of landfill leachate in a flow electrochemical reactor. Chemosphere 58(1):41–46
Moreno-Casillas HA, Cocke DL, Gomes JAG, Morkovasky P, Parga JR, Peterson E (2007) Electrocoagulation mechanism for COD removal. Sep Purif Technol 56(2):204–211
Naje AS, Chelliapan S, Zakaria Z, Abbas SA (2016) Electrocoagulation using a rotated anode: a novel reactor design for textile wastewater treatment. J Environ Manage 176:34–44
Ozturk I, Altinbas M, Koyuncu I, Arikan O, Gomec-Yangin C (2003) Advanced physico-chemical treatment experiences on young municipal landfill leachates. Waste Manage 23(5):441–446
Öztürk T, Özcan ÖF (2020) Effectiveness of electrocoagulation and chemical coagulation methods on paper industry wastewaters and optimum operating parameters. Sep Sci Technol. https://doi.org/10.1080/01496395.2020.1805465
Pisarevsky AM, Polozova IP, Hockridge PM (2005) Chemical oxygen demand. Russ J Appl Chem 78(1):101–107
Press AF (2020) Garbage mountain at Delhi's Ghazipur landfill to rise higher than Taj Mahal by 2020—India news—Hindustan Times, Hindustan Times. Available at: https://www.hindustantimes.com/india-news/garbage-mountain-at-delhi-s-ghazipur-landfill-to-rise-higher-than-taj-mahal-by-2020/story-RC0kwZdUmdHHfDs3rJGngI.html (Accessed: 14 October 2020)
Ren Q, Yu Y, Zhu S, Bian D, Huo M, Zhou D, Huo H (2017) Characterization of a novel micro-pressure swirl reactor for removal of chemical oxygen demand and total nitrogen from domestic wastewater at low temperature. Biodegradation 28(2):145–157
Sharma KD, Jain S (2019) Overview of municipal solid waste generation, composition, and management in India. J Environ Eng 145(3):04018143
Tak BY, Tak BS, Kim YJ, Park YJ, Yoon YH, Min GH (2015) Optimization of color and COD removal from livestock wastewater by electrocoagulation process: application of Box-Behnken design (BBD). J Ind Eng Chem 28:307–315

Thirugnanasambandham K, Sivakumar V, Maran JP (2015) Evaluation of anelectrocoagulation process for the treatment of bagasse-based pulp and paper industry wastewater. Environ Prog Sustainable Energy 34:411–419

Veli S, Ozturk T, Dimoglo A (2008) Treatment of municipal solid wastes leachate by means of chemical- and electro-coagulation. Sep Purif Technol 61(1):82–88

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-05559-x

RESEARCH ARTICLE-ELECTRICAL ENGINEERING

Enhanced Third-Order Generalized Integrator-Based Grid Synchronization Technique for DC-Offset Rejection and Precise Frequency Estimation

Hemant Saxena[1] · **Alka Singh**[1] · **Jitendra Nath Rai**[1]

Received: 14 September 2020 / Accepted: 13 March 2021
© King Fahd University of Petroleum & Minerals 2021

Abstract
An enhanced third-order generalized integrator (TOGI) method has been proposed for the estimation of grid frequency and the synchronizing signals. The proposed method precisely estimates grid frequency and synchronizing signals in the presence of high harmonic content, phase change, frequency change, noise, and DC-offset in the grid voltage. A robust frequency estimator based on three consecutive sampling method has been integrated with TOGI algorithm for accurate frequency estimation. The proposed technique gives zero steady-state error in case of frequency estimation under different grid voltage conditions. This paper demonstrates the developed enhanced TOGI technique for synchronization in a single-phase system. The efficacy of the proposed algorithm has been tested and analyzed in MATLAB/Simulink and verified with experimental results.

Keywords PLL · SOGI · Synchronization · Frequency estimation

1 Introduction

Distribution networks comprise different types of loads, and a high predominance of nonlinear loads may adversely affect the quality of power [1]. Voltage source converter (VSC) with different control techniques has been widely used for the power quality improvement in the distribution system [2]. All the designed control algorithm employed for shunt compensation require the estimation of synchronizing signals from grid voltage and the fundamental load component estimation from the load current signal. The simplest method to generate the synchronizing signal is the unit template method, but it works only under the ideal grid voltage condition [3,4]. However, if the grid voltage has several power quality issue such as distortion, noise and DC-offset, then new and improved methods are required to estimate synchronizing signals. Further, different phase-locked loop (PLL) techniques have been proposed by researchers and some of these do not work well under non-ideal grid conditions.

Synchronous reference frame (SRF)-PLL has been used as a conventional PLL and used by researchers for the phase and frequency estimation and realized for generation of synchronizing signals [5]. Enhanced-PLL (EPLL) has the capability of removing double-frequency oscillations [6]. Second-order generalized integrator (SOGI) PLL algorithm is widely used for its simplicity and effectiveness [7,8] in grid synchronization. It has also been realized for the estimation of the fundamental load component of the nonlinear load current also, and several variations of SOGI are fairly common [9]. The performance of the SOGI-PLL is observed to be non-satisfactory in the case of high grid distortions and DC-offset conditions. The SOGI-frequency locked loop (FLL) has been proposed for the effective working of SOGI in case of varying frequency condition [7,10,11]. The FLL block is realized to estimate the frequency of the input signal. The SOGI-FLL is a simple and effective tool for the estimation of phase, frequency and amplitude of the voltage or current signal. SOGI-FLL has been effectively used for estimation of synchronization signals as well as part of control algorithm developed to achieve shunt compensation for mitigation of power quality problems [12,13]. However, SOGI-FLL has limited filtering capacity and does not give

✉ Alka Singh
alkasingh.eed@gmail.com

Hemant Saxena
hsaxena_phd2k16@dtu.ac.in

Jitendra Nath Rai
jnrai@dce.ac.in

[1] Delhi Technological University, Delhi, India

Published online: 31 March 2021

satisfactory results in case of DC-offset in the grid voltage. Modified SOGI-FLL has been presented for the rejection of the DC-offset in grid voltage [14,15]. It is also observed that the frequency estimated by the FLL shows varying degree of oscillations under steady-state conditions under highly distorted grid. Reduced-order generalized integrator (ROGI) has been cascaded with SOGI-FLL, and it improved the filtering capability, but the frequency estimated by SOGI-FLL-ROGI shows oscillations under DC-offset conditions [16]. Cascaded delayed signal cancellation operator (CDSC)-based PLL has been investigated in three-phase system. This requires transform equations and also additional proportional integral (PI) controller [17]. Robust frequency estimation technique has been proposed under distorted grid conditions and has not been used for estimation of synchronizing signals [18,19]. Further, a method based on three consecutive sampling (3CS) has been proposed to estimate synchronizing signal in three-phase grid [20]. These techniques have been employed under frequency change and distorted conditions, and DC-offset condition has not been discussed. These techniques give zero steady-state error in case of frequency estimation except under DC-offset conditions.

TOGI-based PLL requires precise resonating frequency as input signal. This precise real-time grid frequency information is required for effective filtering operation even under grid frequency variations. TOGI block needs robust frequency estimation for satisfactory operation, especially under highly distorted grid voltage conditions and also in case of frequency change from the standard grid frequency (50 Hz). A new technique based on TOGI integrated with 3CS technique has been used effectively in this paper for frequency tracking and estimation of the synchronizing signal from the grid voltage signal. The main highlights of the paper are

1. The proposed TOGI-3CS method is a simple technique and estimates frequency with almost oscillations.
2. Conventional 3CS method could not estimate frequency precisely under DC-offset condition, while proposed method estimates without oscillations.
3. The proposed method is used to estimates phase, frequency and synchronizing signal under different grid voltage conditions.
4. Proposed method has been tested both in simulation and in experimental setup.

The paper has been organized in the following manner: Section 1 describes the introduction and the merits of the proposed algorithm. The 3CS-based frequency estimation technique is explained in Sect. 2. The proposed synchronization technique is demonstrated in Sect. 3. Sections 4 and 5 show the simulation and experimental results, respectively. After the detailed analysis of the simulation and experimental results, conclusion is described in Sect. 6.

2 3CS-Based Frequency Estimation

Under all practical conditions, the grid voltage in the power distribution network contains harmonics. The grid voltage in single-phase distribution network can be expressed as:

$$v_s(t) = V_1 \sin(\omega t + \phi_1) + h_n(t) \tag{1}$$

In (1), V_1, ω and ϕ_1 are the magnitudes of the fundamental voltage component, standard grid frequency (50 Hz) and phase angle of the fundamental component of the grid voltage, respectively. $h_n(t)$ is specified as the harmonic content in the polluted grid voltage. Equation (1) is discretized and shown as follows:

$$v_s(k) = V_1 \sin(\omega k T_s + \phi_1) + h_n(kTs) \tag{2}$$

In (2), $k = 0, 1, 2, 3....$ and T_s denotes the sampling time. For the generation of the three consecutive sampling (3CS), consecutive transfer delay of τ is applied on the polluted grid voltage signal yielding two more signals:

$$\begin{aligned} v_1(k) &= v_s(k-\tau) \\ &= V_1 \sin(\omega k T_s - \omega\tau + \phi_1) + h_n(kT_s - \tau) \end{aligned} \tag{3}$$

$$\begin{aligned} v_2(k) &= v_s(k-2\tau) \\ &= V_1 \sin(\omega k T_s - 2\omega\tau + \phi_1) + h_n(kT_s - 2\tau) \end{aligned} \tag{4}$$

Expanding Eqs. (3) and (4), we get

$$\begin{aligned} v_1(k) &= \gamma_1 V_1 \sin(\omega k T_s + \phi_1) - \gamma_2 V_1 \cos(\omega k T_s + \phi_1) \\ &\quad + h_n(kT_s - \tau) \end{aligned} \tag{5}$$

$$\begin{aligned} v_2(k) &= \gamma_3 V_1 \sin(\omega k T_s + \phi_1) - \gamma_4 V_1 \cos(\omega k T_s + \phi_1) \\ &\quad + h_n(kT_s - 2\tau) \end{aligned} \tag{6}$$

where

$$\gamma_1 = \cos\omega\tau \tag{7}$$

$$\gamma_2 = \sin\omega\tau \tag{8}$$

$$\gamma_3 = \cos 2\omega\tau \tag{9}$$

$$\gamma_4 = \sin 2\omega\tau \tag{10}$$

Applying the algebraic operations of trigonometric function on (8)–(10), relationship is obtained as:

$$\gamma_3 = 2\gamma_1^2 - 1 \tag{11}$$

$$\gamma_4 = 2\gamma_1\gamma_3 \tag{12}$$

Also, the value of $\omega\tau$ is specified as $0 < \omega\tau < \pi$ for γ_3 for $\gamma_2 \neq 0$. Now, Eqs. (5) and (6) can be written as:

$$\gamma_2 V_1 \cos(\omega k T_s + \phi_1) = \gamma_1 V_1 \sin(\omega k T_s + \phi_1) + h_n(kT_s - \tau) - v_1(k) \quad (13)$$

$$\gamma_4 V_1 \cos(\omega k T_s + \phi_1) = \gamma_3 V_1 \sin(\omega k T_s + \phi_1) + h_n(kT_s - 2\tau) - v_2(k) \quad (14)$$

Multiplying $2\gamma_1$ on both sides of Eq. (13) and substituting the values form (11), (12) and (14) in it, it gives

$$v_s(k) + v_2(k) = 2\gamma_1 v_1(k) + h_n(kT_s) - 2\gamma_1 h_n(kT_s - \tau) + h_n(kT_s - \tau) \quad (15)$$

$$v_s(k) + v_2(k) = 2\gamma_1 v_1(k) - \Delta k \quad (16)$$

In (16), Δk is the contribution denoting sum of time-delayed harmonic components of the grid voltage. In (16), γ_1 is the only parameter which is unknown and the other parameter is either known or can be obtained easily. Under ideal grid voltage condition, the harmonic component in (16) is zero, i.e., $\Delta k = 0$. Under this condition, (16) can be written as:

$$v_s(k) + v_2(k) = 2\gamma_1 v_1(k) \quad (17)$$

The frequency (ω) can be estimated as:

$$\omega = \frac{\arccos\{(v_1(k) + v_2(k))/2v_1(k)\}}{\tau} \quad (18)$$

Equation (18) estimates the frequency under standard test conditions and is valid for the ideal case when the voltage is not corrupted with harmonics. The effect of the harmonics is alleviated using a robust technique which estimates the fundamental frequency with zero steady-state error. The error is defined as:

$$e(k) = 2\hat{\gamma}_1(k)v_1(k) - v_s(k) - v_2(k) \quad (19)$$

In (19), $\hat{\gamma}_1$is the new estimated value of γ_1. Once the value of $\hat{\gamma}_1$ is estimated, the precise frequency of the fundamental component of the voltage signal can be estimated. A robust update law is developed for the estimation of $\hat{\gamma}_1$ and is expressed as (20):

$$\hat{\gamma}_1(k+1) = \hat{\gamma}_1(k) - \frac{2\alpha e(k) v_1(k)}{1 + 4\alpha V_1^2(k)} \quad (20)$$

$$\hat{\gamma}_1(k+1) = \hat{\gamma}_1(k) - \frac{2\alpha v_1(k)}{1 + 4\alpha V_1^2(k)}(2\hat{\gamma}_1(k)v_1(k) - v_s(k) - v_2(k)) \quad (21)$$

In (21), α denotes the convergence factor. Once $\hat{\gamma}_1(k)$ is estimated, the estimated frequency of the grid voltage is expressed as:

$$\omega_{est}(k) = \frac{\arccos\{\hat{\gamma}_1(k)\}}{\tau} \quad (22)$$

Considering the estimation error ($\tilde{\gamma}_1(k)$), it is expressed as:

$$\tilde{\gamma}_1(k) = \hat{\gamma}_1(k) - \gamma_1 \quad (23)$$

Also,

$$\begin{aligned}\tilde{\gamma}_1(k+1) &= \hat{\gamma}_1(k+1) - \gamma_1 \\ &= \hat{\gamma}_1(k) - \frac{2\alpha v_1(k)}{1 + 4\alpha V_1^2(k)}(2\hat{\gamma}_1(k)v_1(k) - v_s(k) - v_2(k)) - \gamma_1 \\ &= \tilde{\gamma}_1(k) - \frac{2\alpha v_1(k)}{1 + 4\alpha V_1^2(k)}(2\hat{\gamma}_1(k)v_1(k) - v_s(k) - v_2(k)) \\ &= \tilde{\gamma}_1(k) - \frac{2\alpha v_1(k)}{1 + 4\alpha V_1^2(k)}\tilde{\gamma}_1(k) \\ &= \frac{1}{1 + 4\alpha V_1^2(k)}\tilde{\gamma}_1(k)\end{aligned} \quad (24)$$

Iterating Eq. (24), it yields

$$\begin{aligned}\tilde{\gamma}_1(k) &= \frac{1}{1 + 4\alpha V_1^2(k-1)}\tilde{\gamma}_1(k-1) \\ &= \frac{1}{1 + 4\alpha V_1^2(k-1)}\frac{1}{1 + 4\alpha V_1^2(k-2)}\tilde{\gamma}_1(k-2) \\ &= \frac{1}{1 + 4\alpha V_1^2(k-1)}\frac{1}{1 + 4\alpha V_1^2(k-2)}..... \\ &\ldots\frac{1}{1 + 4\alpha V_1^2(0)}\tilde{\gamma}_1(0)\end{aligned} \quad (25)$$

In (25), $v_1(0)$ is the initial grid voltage with τ delay and $\tilde{\gamma}_1$ is the initial estimation error. In (25), there are m nonzero values of $v_1(0), v_1(1), v_1(2), ..., v_1(k-1)$ and these are assumed as $v_1(p_1), v_1(p_2), v_1(p_3), ..., v_1(p_m)$. Among all the nonzero values, the minimum is considered as $v_{1\,\min}(p_m)$.

Now, (25) can be rewritten with the above assumption as

$$\begin{aligned}\tilde{\gamma}_1(k) &= \frac{1}{1 + 4\alpha V_1^2(p_1)}\frac{1}{1 + 4\alpha V_1^2(p_2)}..... \\ &\frac{1}{1 + 4\alpha V_1^2(p_m)}\tilde{\gamma}_1(0)\end{aligned} \quad (26)$$

Applying the absolute value operation on (26) gives

$$\begin{aligned}|\tilde{\gamma}_1(k)| &= |\frac{1}{1 + 4\alpha V_1^2(p_1)}\frac{1}{1 + 4\alpha V_1^2(p_2)}..... \\ &\frac{1}{1 + 4\alpha V_1^2(p_m)}\tilde{\gamma}_1(0)| \\ &\leq \left|\left\{\frac{1}{1 + 4\alpha V_{1\,\min}^2(p_m)}\right\}^m \tilde{\gamma}(0)\right|\end{aligned} \quad (27)$$

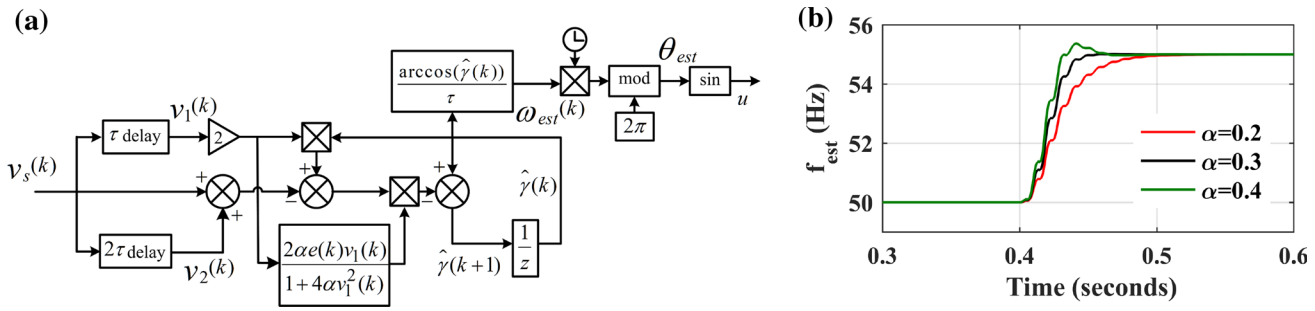


Fig. 1 **a** Structure of 3CS method for frequency estimation, **b** response of 3CS-based frequency estimator with different convergence factor for + 5 Hz frequency change

The number of the terms expressing $v_1(k)$ is infinite; hence, the number of m nonzero value is also infinite. Applying the limit in (27), it gives

$$\lim_{x\to\infty}|\tilde{\gamma}_1(k)| = \lim_{x\to\infty}|\frac{1}{1+4\alpha V_1^2(p_1)}\frac{1}{1+4\alpha V_1^2(p_2)}\ldots\ldots \frac{1}{1+4\alpha V_1^2(p_m)}\tilde{\gamma}_1(0)| \leq \lim_{x\to\infty}\left\{\frac{1}{1+4\alpha V_{1\min}^2(p_m)}\right\}^m |\tilde{\gamma}(0)| \tag{28}$$

Here, $0 < \frac{1}{1+4\alpha V_{1\min}^2(p_m)} < 1$, and (28) gives

$$\lim_{x\to\infty}|\tilde{\gamma}_1(k)| \leq \lim_{x\to\infty}\left\{\frac{1}{1+4\alpha V_{1\min}^2(p_m)}\right\}^m |\tilde{\gamma}(0)| = 0 \tag{29}$$

Hence, it can be observed that the estimation of frequency by this technique gives zero steady-state error.

The structure for the estimation of frequency by 3CS method is shown in Fig. 1a. The dynamic response of the 3CS-based frequency estimation with different convergence factors is shown in Fig. 1b for + 5 Hz step change in frequency. It is observed that a higher value of convergence factor results in a faster response and high overshoot, while a lower convergence factor shows slower response. Hence, an appropriate value of convergence factor of $\alpha = 0.3$ has been selected for a fast response and almost no-overshoot.

Once ω_{est} is estimated, then the phase information of the grid voltage is easily obtained. The estimation of in-phase synchronizing signal is obtained using the sine function. One of the advantages of the method is that the frequency is effectively tracked with zero steady-state error; hence, the estimated synchronizing signals will be distortion-free. The estimated frequency under noisy and DC-offset conditions is shown in Fig. 2. It estimates frequency effectively for the

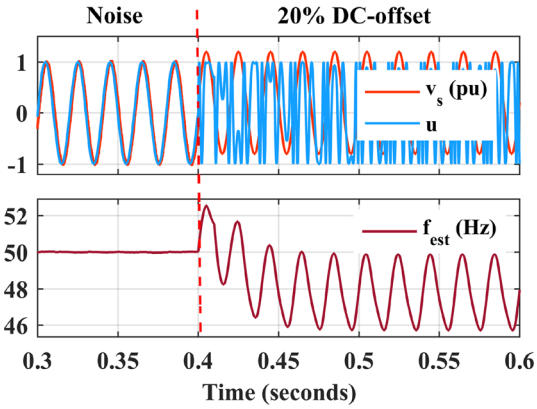


Fig. 2 Frequency and synchronizing signal estimation by 3CS method under noisy and grid voltage with DC-offset

case of phase change, frequency change, and the harmonic signal, but it gives the oscillations in estimated frequency in case of input grid voltage having DC-offset. In the case of noisy input, the frequency estimated by the 3CS method contains noise. The 3CS method is not a phase-locked loop and can generate synchronization signals only when there is no phase or frequency change. A major drawback of the 3CS based technique is it does not work well under DC-offset conditions. It does not give synchronized output signal with input in case of phase or frequency change; hence, there is a need of a hybrid algorithm which can provide precise estimation of frequency and phase angle under all conditions.

3 Proposed Enhanced TOGI Technique

SOGI is widely used as a filter and as PLL for the estimation of phase and frequency of the input signal. SOGI has some inherent advantages such as good filtering capability, but its estimation of frequency is poor; moreover, it is not capable of DC-offset rejection. One possible solution is to integrate SOGI with a DC rejection loop which makes it a filter of third order. Hence, the benefits of both TOGI and

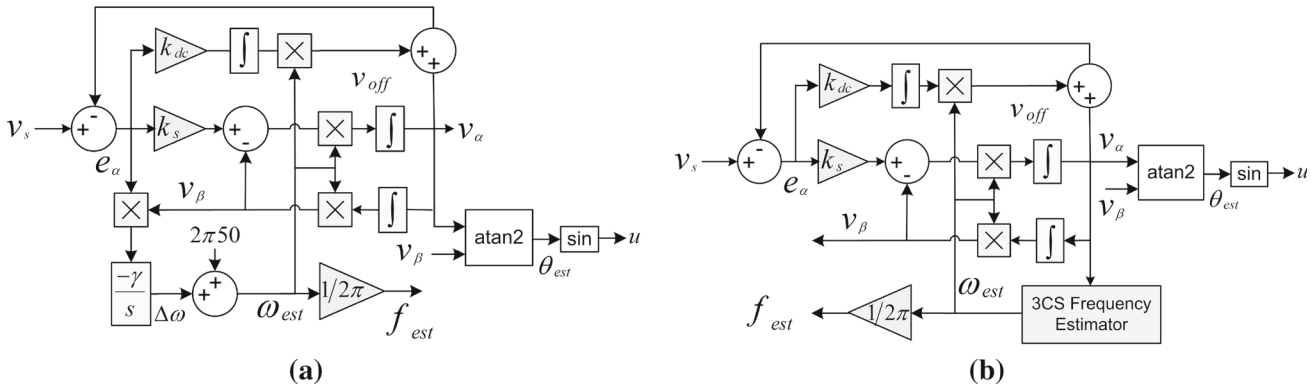


Fig. 3 Structure of **a** TOGI-FLL, **b** proposed TOGI-3CS

good frequency estimation by 3CS can be realized together in a hybrid approach to estimate the phase and the frequency of the grid voltage even under DC-offset condition. The structure of TOGI-FLL and TOGI-3CS-based synchronization is represented in Fig. 3a, b, respectively. The transfer functions of TOGI showing expressions for in-phase ($H_{\mathrm{d}}(s)$), quadrature phase ($H_{\mathrm{q}}(s)$) and the DC-offset ($H_{\mathrm{off}}(s)$) components are

$$H_{\mathrm{d}}(s) = \frac{v_{\alpha}(s)}{v_{\mathrm{s}}(s)} = \frac{k_{\mathrm{dc}}\omega_r s^2}{s^3 + (k_{\mathrm{s}} + k_{\mathrm{dc}})\omega_r s^2 + \omega_r^2 s + k_{\mathrm{dc}}\omega_r^3} \tag{30}$$

$$H_{\mathrm{q}}(s) = \frac{v_{\beta}(s)}{v_{\mathrm{s}}(s)} = \frac{k_{\mathrm{dc}}\omega_r^2 s}{s^3 + (k_{\mathrm{s}} + k_{\mathrm{dc}})\omega_r s^2 + \omega_r^2 s + k_{\mathrm{dc}}\omega_r^3} \tag{31}$$

$$H_{\mathrm{off}}(s) = \frac{v_{\mathrm{off}}(s)}{v_{\mathrm{s}}(s)} = \frac{k_{\mathrm{dc}}\omega_r (s^2 + \omega_r^2)}{s^3 + (k_{\mathrm{s}} + k_{\mathrm{dc}})\omega_r s^2 + \omega_r^2 s + k_{\mathrm{dc}}\omega_r^3} \tag{32}$$

Here, ω_r is the resonating frequency of the SOGI, and it is estimated by the 3CS method ($\omega_r = \omega_{\mathrm{est}}$). Here, k_{s} is the constant and defines the frequency bandwidth of the algorithm. Higher value of k_{s} gives a faster response, high overshoot, while a lower value of k_{s} gives a slower response and low overshoot. Hence, the value of ($k_{\mathrm{s}} = 1.414$) is appropriately considered as faster dynamic with almost no overshoot. k_{dc} is the DC-gain for the DC-rejection loop. Figure 4 shows the bode plot of $H_{\mathrm{d}}(s)$, $H_{\mathrm{q}}(s)$ and $H_{\mathrm{off}}(s)$.

The magnitude plot shows $H_{\mathrm{d}}(s)$ and $H_{\mathrm{q}}(s)$ shows the characteristics of band pass filter (BPF), which rejects the DC-offset and gives filtered in-phase and quadrature phase component. The phase plot of $H_{\mathrm{d}}(s)$ and $H_{\mathrm{q}}(s)$ shows that they are 90° apart. The characteristics of $H_{\mathrm{off}}(s)$ are of low-pass filter (LPF) which passes the DC-offset in the grid voltage signal. Once the in-phase and quadrature components have been computed, the phase angle is estimated by

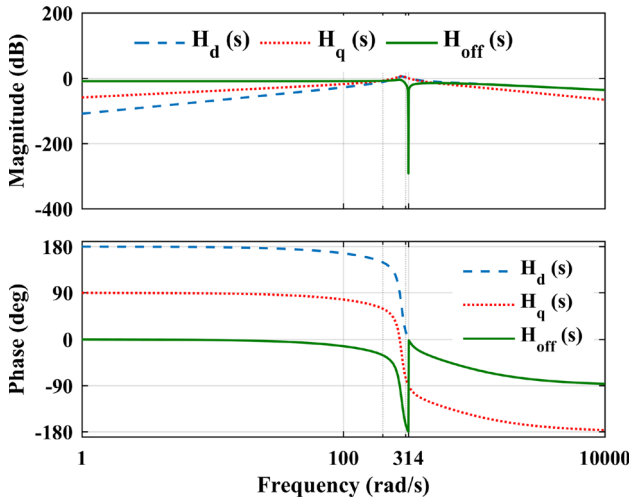


Fig. 4 Bode plot of TOGI for $H_{\mathrm{d}}(s)$, $H_{\mathrm{q}}(s)$ and $H_{\mathrm{off}}(s)$

$$\theta_{\mathrm{est}} = \tan^{-1}\left(\frac{v_{\beta}}{v_{\alpha}}\right) \tag{33}$$

The synchronization signal (u) is obtained by applying the sine operator. The estimated in-phase voltage (v_{α}) is the filtered signal and given to the 3CS-based frequency estimator. Now, the 3CS-based frequency estimator is also able to estimate frequency precisely even in case of DC-offset and in case of white noise in the voltage signal. The structure for the proposed algorithm is shown in Fig. 3b.

4 Simulation Results

The proposed algorithm has been tested under different grid voltage conditions for the estimation of synchronizing signal and frequency. Simulation results for the estimation of frequency under different grid conditions are shown in Fig. 5.

Figure 5a shows the estimation of frequency and synchronization signal under frequency change of + 5Hz with 20%

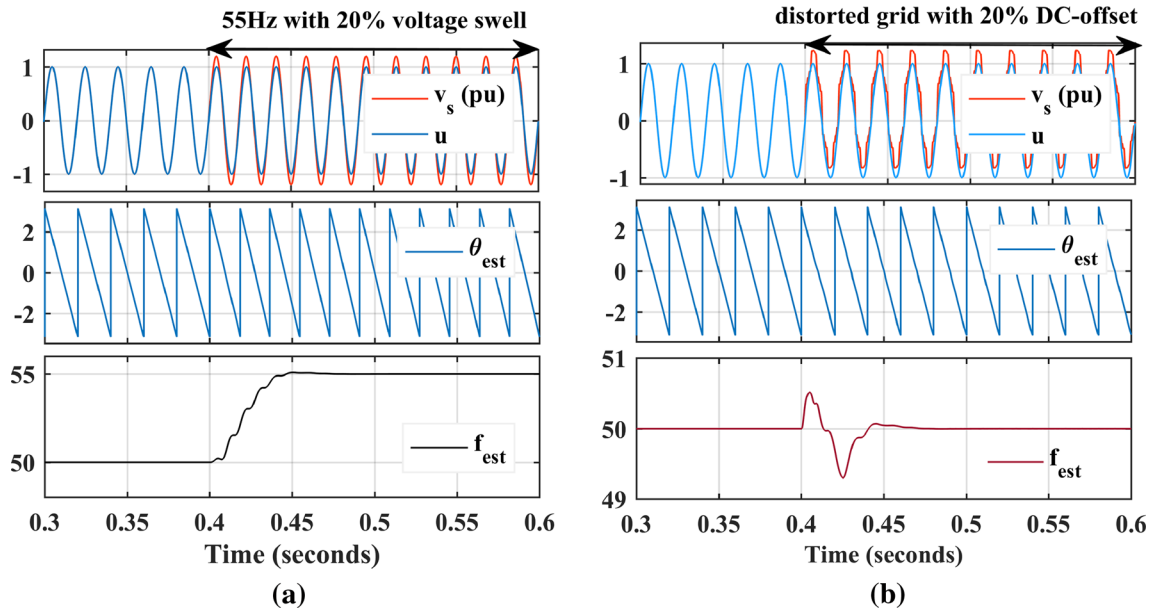


Fig. 5 Simulation performance of TOGI-3CS under **a** frequency change of + 5 Hz with 20% voltage swell, **b** distorted grid voltage with 20% DC-offset

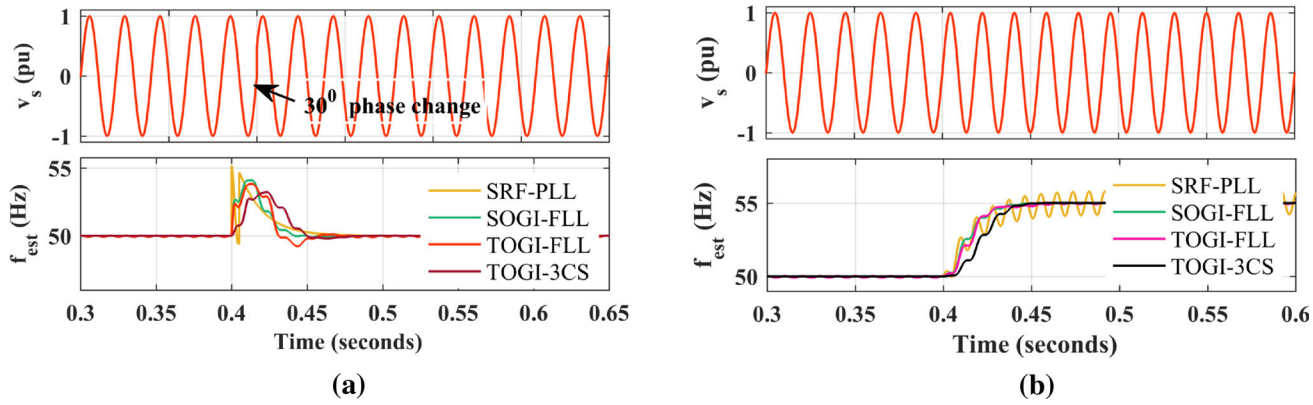


Fig. 6 Performance comparison of SRF-PLL, SOGI-FLL, TOGI-FLL and TOGI-3CS under **a** phase change of 30°, **b** frequency change of + 5 Hz

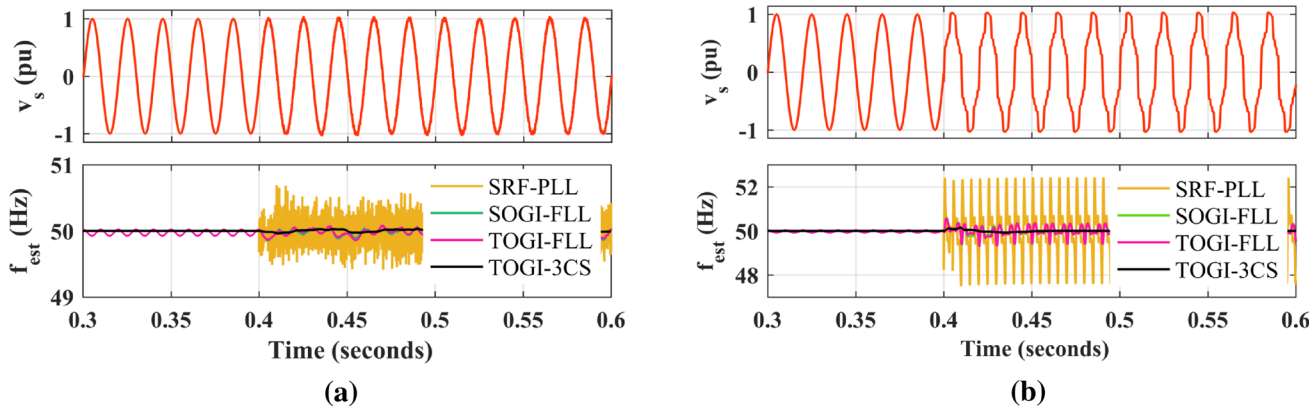


Fig. 7 Performance comparison of SRF-PLL, SOGI-FLL, TOGI-FLL and TOGI-3CS under **a** noisy grid voltage, **b** distorted grid voltage

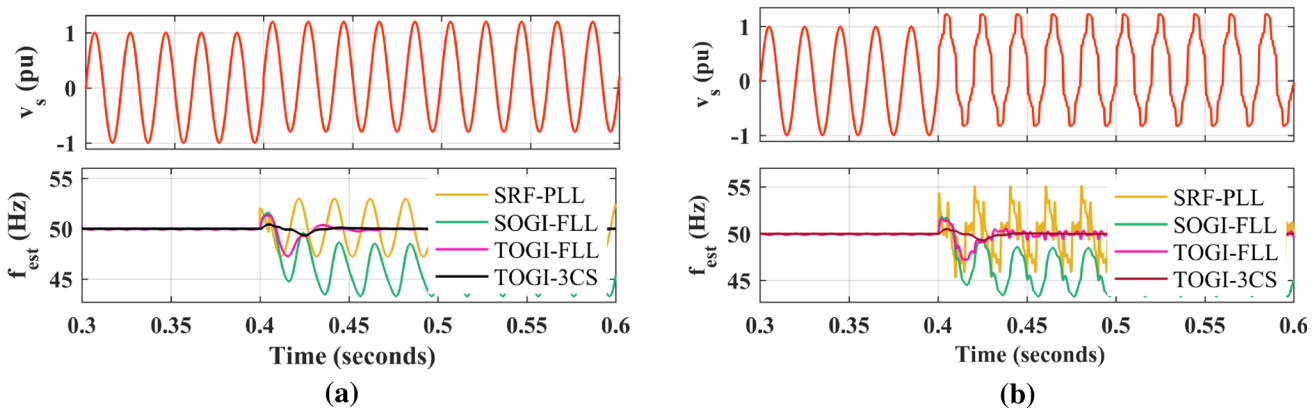


Fig. 8 Performance comparison of SRF-PLL, SOGI-FLL, TOGI-FLL and TOGI-3CS under **a** DC-offset in grid voltage, **b** DC-offset with distortion in grid voltage

voltage swell. The estimated frequency shows almost zero error in steady state, and the estimated synchronization signal is in phase with grid voltage and free of DC-offset. Figure 5b shows the estimated phase angle as well as frequency under DC-offset with distortion in grid voltage signal, and it shows the desired performance. The effectiveness of the proposed algorithm has been tested and compared with conventional and modern algorithms. The results include response with SRF-PLL, SOGI-FLL, TOGI-FLL and the proposed TOGI-3CS algorithm. Optimum gains for these algorithms have been tuned for a fair comparison. The sampling time for the all the comparing algorithm is equal ($T_s = 50\,\mu s$). Comparison has been done under different power quality issues in grid voltage considered below and compiled in the form of five cases listed below.

4.1 Case I: Phase Change of 30°

In this case (Fig. 6a), estimated frequency has lesser undershoot and zero steady-state error. At $t = 0.4$s when phase change is introduced, frequency estimated by all the methods shows deviation initially but settles down soon showing zero steady-state error. An overshoot of 5 Hz is observed for SRF-PLL, 4.09 Hz for SOGI-FLL, and 3.81 Hz for TOGI-FLL and for proposed algorithm its value is 3.2 Hz. Around 39.51 ms is required as settling time by proposed algorithm.

4.2 Case II: Large Frequency Deviation

In this case (Fig. 6b), a large frequency variation from 50 to 55 Hz is suddenly introduced at $t = 0.4$s; the estimated frequency by the proposed method is exact and shows almost zero oscillations. The frequency estimated by the SOGI-FLL and TOGI-FLL shows some oscillations, and SRF-PLL shows higher oscillations in frequency estimation, while the proposed technique shows least oscillations. Settling time of frequency estimation under large frequency deviation is not obtained as the frequency oscillates with more than 2% error band. The settling time for SOGI-FLL, TOGI-FLL and proposed algorithm is 31.66 ms, 43.381 ms and 25.766 ms, respectively, for same frequency change of + 5 Hz.

4.3 Case III: Noisy Grid Voltage

Under white noise grid conditions (Fig. 7a), the frequency obtained by SRF-PLL shows high noise, SOGI-FLL and TOGI-FLL show comparatively less noise, while the estimation by the proposed method shows negligible noise. The proposed TOGI-3CS method effectively tracks the frequency irrespective of the presence of noise in grid voltage.

4.4 Case IV: Harmonics in Grid Voltage

In this case (Fig. 7b), the THD of the input voltage signal is 14.2%. No oscillations have been observed in frequency estimated by the TOGI-3CS, while the SOGI-FLL and TOGI-FLL show much high oscillations in frequency estimation. SRF-PLL shows the worst performance.

4.5 Case V: DC-Offset in Grid Voltage

In this case (Fig. 8a), 20% DC-offset is introduced in the input voltage signal. Frequency tracking is observed using all the techniques mentioned above. It is observed that the oscillations in frequency tracking are too high with SRF-PLL and SOGI-FLL. TOGI-FLL and TOGI-3CS show less oscillation under this case. Thus, these techniques effectively fail under DC offset conditions, while the proposed algorithm shows almost zero oscillations and very good frequency tracking.

4.6 Case VI: DC-Offset with Harmonics

Here, worst-case scenario has been simulated with a high DC-offset of 20% along with harmonics (Fig. 8b). Precise frequency estimation is achieved by TOGI-3CS under these conditions also. It is also observed that the frequency estimated by the SOGI-FLL and SRF-PLL shows oscillations of 8.5 Hz and 9.8 Hz, respectively, because of the absence of DC-rejection block. SOGI-FLL fails to estimate frequency correctly, while the response of SRF-PLL is also very poor. The proposed technique achieves the best result even when grid faces multiple problems of harmonics and DC-offset together.

Based on the simulation performance, Table 1 shows the comparison of SRF-PLL, SOGI-FLL, TOGI-FLL and TOGI-3CS. It has been observed that there are negligible oscillations in the estimation of frequency by the proposed algorithm under all the six cases considered. TOGI-FLL shows good performance under frequency variation, but its performance worsens under DC-offset and DC-offset with the harmonic condition. SRF-PLL shows good performance only under ideal grid condition. Hence, the proposed algorithm performance is a much better choice than SRF-PLL, SOGI-FLL and TOGI-FLL under various voltage conditions. The response of the developed algorithm is fast under different cases considered, and the algorithm is simple and easy to implement.

5 Experimental Results

The proposed algorithm has also been executed and tested experimentally. A programmable AC/DC single-phase supply has been used for generating different grid voltages for testing. LEM make voltage sensor has been used to sense the grid voltage, and the signal is processed by dSpace 1104. The output of the proposed algorithm has been analyzed and recorded by Agilent DSO-X 2014A. The proposed algorithm has also been compared with SRF-PLL and SOGI-FLL experimentally. Sampling time (T_s) for experimental validation is 50 μs. The performance of the TOGI-3CS algorithm is displayed in Fig. 9 for two cases, viz. frequency change with voltage swell incorporated and DC-offset with distortion in grid voltage.

Figure 9a shows the estimation of synchronizing signal from the input voltage signal, phase-angle, and frequency from the grid voltage. Two disturbances have been considered together—+ 2 Hz frequency change along with 20% voltage swell. The estimated synchronizing signal is in phase with grid voltage under worst case of both disturbances occurring simultaneously. The estimated frequency shows no oscillations and quick response time. Figure 9b shows the performance of proposed algorithm under 20% of DC-offset in grid voltage. The DC-offset in the grid voltage is estimated within 20 ms; also the change in frequency (Δf) from standard frequency is almost zero. Figure 9c shows the estimation of frequency under 20% DC-offset with high distortion in grid voltage, and the TOGI-3CS algorithm performs its esti-

Table 1 Comparison between SRF-PLL, SOGI-FLL, TOGI-FLL and proposed TOGI-3CS

PI required Type		SRF-PLL Yes Non-adaptive	SOGI-FLL No Adaptive	TOGI-FLL No Adaptive	TOGI-3CS No Adaptive
Oscillations (peak to peak, Hz)	Ideal case/Case I	2.445×10^{-6}	9.357×10^{-6}	9.327×10^{-6}	00
	Case II	1.621	9.339×10^{-2}	9.314×10^{-2}	1.727×10^{-12}
	Case III	1.339	2.2×10^{-1}	2.0×10^{-1}	4.775×10^{-2}
	Case IV	4.849	9.754×10^{-1}	9.655×10^{-1}	9.298×10^{-2}
	Case V	5.741	5.229	0.1	1.026×10^{-3}
	Case VI	9.887	8.54	9.627×10^{-1}	1.495×10^{-3}
Mean frequency (Hz)	Case V	49.97	47.31	49.90	49.99
	Case VI	50.02	47.63	49.93	49.99
Dynamic response (case II)	Pre-shoot (%)	1.531	0.505	0.505	0.505
	Overshoot (%)	0.503	0.505	0.505	0.505
	Undershoot (%)	29.179	1.973	2.119	1.983
	Settling time (ms)	Not settle under 2% error band	31.660	43.381	25.766
Execution time (ms)		40	35.5	37	39
Sampling time		50 μs			

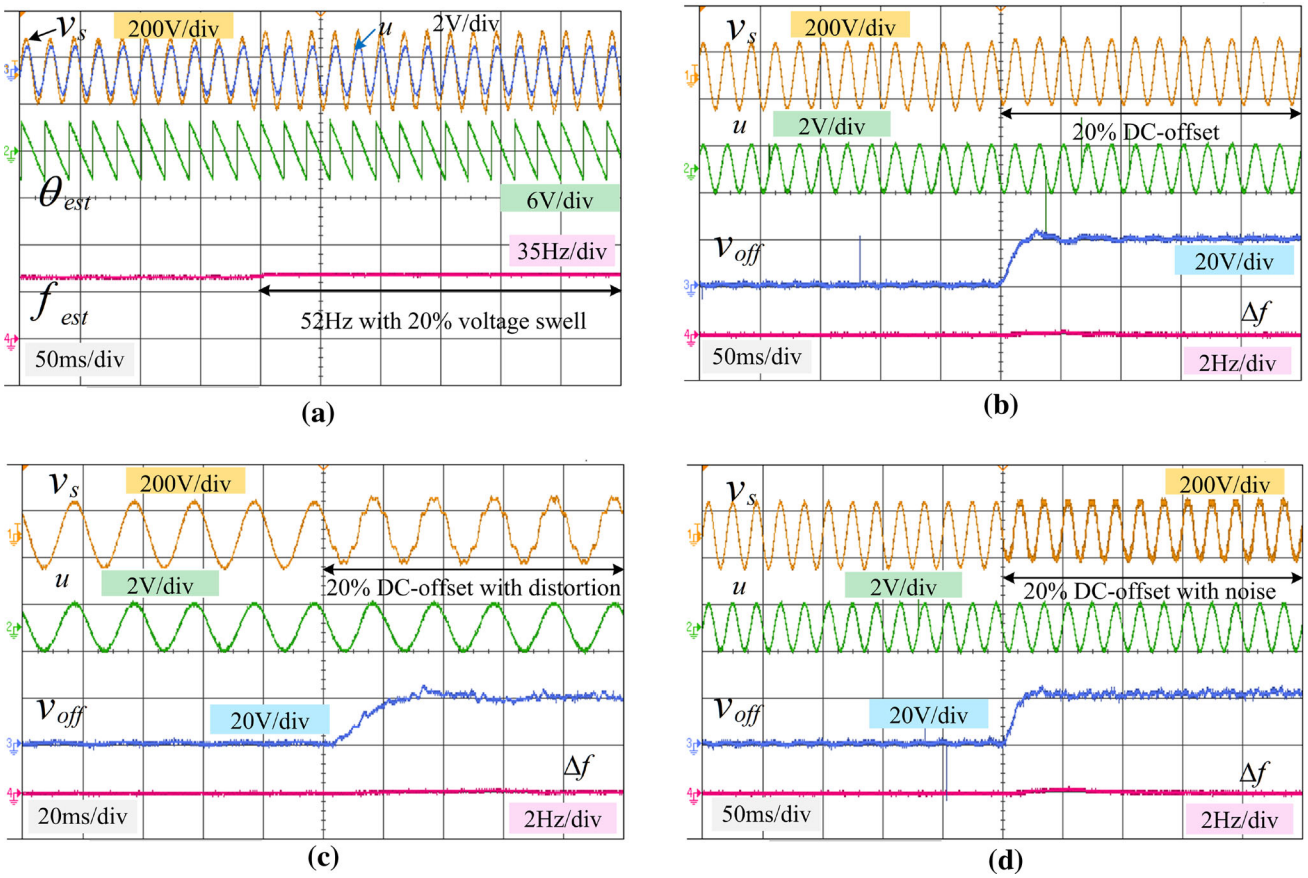


Fig. 9 Experimental performance of proposed algorithm under **a** frequency change of + 2 Hz with 20% voltage swell, (**b**) 20% DC-offset in grid voltage, **c** distorted grid voltage with 20% DC-offset, **d** noisy grid voltage with 20% DC-offset

mation task very well under this condition. Figure 9d shows the performance under 20% DC-offset with noise in grid voltage signal. Under this DC-offset condition, synchronizing signal and change in frequency are estimated accurately.

6 Conclusion

A new approach has been proposed by the enhancing the performance of TOGI-based algorithm with a robust frequency estimator. The proposed algorithm is a hybrid technique and incorporates the advantages of TOGI as a filter and 3CS method for robust frequency estimation. The performance of the proposed algorithm has been tested under several cases and analyzed using both simulation and experimental results. The proposed algorithm gives better results in frequency estimation and estimation of synchronizing signal as compared with SRF-PLL, SOGI-FLL and TOGI-FLL techniques under various grid voltage cases. The main advantage of this technique is that it can detect frequency precisely and the estimated frequency is also free from oscillations in steady-state conditions. It is very effective for use in the generation of synchronizing signal and can be easily extended to the development of a controller for power quality improvement in distribution network.

Acknowledgements The authors are thankful to the Department of Science and Technology, Science and Engineering Research Board, Government of India for the sponsored project.

References

1. Singh, B.; Chandra, A.; Al-Haddad, K.: Power Quality: Problems and Mitigation Techniques. Wiley, New York (2014)
2. Wang, J.; Xu, F.; Pan, G.; Ouyang, K.; Jin, Y.; Jin, L.; Qiu, J.: Robust control method for LCL-type shunt active power filter under weak grid condition. IET Gener. Trans. Distrib. **14**(11), 2120 (2020)
3. Chittora, P.; Singh, A.; Singh, M.: Simple and efficient control of DSTATCOM in three-phase four-wire polluted grid system using MCCF-SOGI based controller. IET Gener. Trans. Distrib. **12**(5), 1213 (2017)
4. Bangarraju, J.; Rajagopal, V.; Jayalaxmi, A.: Unit template synchronous reference frame theory based control algorithm for DSTATCOM. J. Inst. Eng. (India) Ser. B **95**(2), 135 (2014)
5. Golestan, S.; Guerrero, J.M.: Conventional synchronous reference frame phase-locked loop is an adaptive complex filter. IEEE Trans. Ind. Electron. **62**(3), 1679 (2014)

6. Singh, Y., Hussain, I., Singh, B., Mishra, S.: Real time implementation of EPLL with generalized filtering in single phase grid interfaced SPV system. In: 2016 IEEE International Conference on Electrical, Computer and Electronics Engineering (UPCON), pp. 50–54. IEEE (2016)
7. Golestan, S.; Guerrero, J.M.; Vasquez, J.C.; Abusorrah, A.M.; Al-Turki, Y.: Modeling, tuning, and performance comparison of second-order-generalized-integrator-based FLLs. IEEE Trans. Power Electron. **33**(12), 10229 (2018)
8. Xiao, F.; Dong, L.; Li, L.; Liao, X.: A frequency-fixed SOGI-based PLL for single-phase grid-connected converters. IEEE Trans. Power Electron. **32**(3), 1713 (2016)
9. Singh, Y.; Hussain, I.; Singh, B.; Mishra, S.: Implementation of grid interfaced photovoltaic system with active power filter capabilities. Int. Trans. Electr. Energy Syst. **28**(11), e2616 (2018)
10. Golestan, S.; Guerrero, J.M.; Vasquez, J.C.; Abusorrah, A.M.; Al-Turki, Y.: Single-phase FLLs based on linear Kalman filter, limit-cycle oscillator, and complex bandpass filter: analysis and comparison with a standard FLL in grid applications. IEEE Trans. Power Electron. **34**(12), 11774 (2019)
11. Saxena, H.; Singh, A.; Rai, J.: Design and analysis of different PLLs as load compensation techniques in 1-phase grid-tied PV system. Int. J. Electron. **106**(11), 1632 (2019)
12. Verma, A.; Singh, B.; Chandra, A.; Al-Haddad, K.: An implementation of solar PV array based multifunctional EV charger. IEEE Trans. Ind. Appl. **56**(4), 4166 (2020)
13. Yi, H.; Wang, X.; Blaabjerg, F.; Zhuo, F.: Impedance analysis of SOGI-FLL-based grid synchronization. IEEE Trans. Power Electron. **32**(10), 7409 (2017)
14. Zhang, C.; Zhao, X.; Wang, X.; Chai, X.; Zhang, Z.; Guo, X.: A grid synchronization PLL method based on mixed second-and third-order generalized integrator for DC offset elimination and frequency adaptability. IEEE J. Emerg. Select. Top. Power Electron. **6**(3), 1517 (2018)
15. Rodríguez, P.; Luna, A.; Candela, I.; Mujal, R.; Teodorescu, R.; Blaabjerg, F.: Multiresonant frequency-locked loop for grid synchronization of power converters under distorted grid conditions. IEEE Trans. Ind. Electron. **58**(1), 127 (2010)
16. Saxena, H.; Singh, A.; Rai, J.N.: Analysis of SOGI-ROGI for synchronization and shunt active filtering under distorted grid condition. ISA Trans. **109**, 380 (2021). https://doi.org/10.1016/j.isatra.2020.10.025
17. Wang, Y.F.; Li, Y.W.: Grid synchronization PLL based on cascaded delayed signal cancellation. IEEE Trans. Power Electron. **26**(7), 1987 (2010)
18. Dai, Z.; Fan, M.; Nie, H.; Zhang, J.; Li, J.: A Robust Frequency estimation method for aircraft grids under distorted conditions. IEEE Trans. Ind. Electron. **67**(5), 4254 (2019)
19. Vedyakov, A.A.; Vediakova, A.O.; Bobtsov, A.A.; Pyrkin, A.A.; Aranovskiy, S.V.: Frequency estimation of a sinusoidal signal with time-varying amplitude. IFAC **50**(1), 12880 (2017)
20. Dai, Z.; Zhao, P.; Chen, X.; Fan, M.; Zhang, J.: An enhanced transfer delay-based frequency locked loop for three-phase systems with DC offsets. IEEE Access **7**, 40380 (2019)

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-05670-z

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

EnPSO: An AutoML Technique for Generating Ensemble Recommender System

Garima Gupta[1] · **Rahul Katarya**[1]

Received: 27 December 2020 / Accepted: 11 April 2021
© King Fahd University of Petroleum & Minerals 2021

Abstract
With the explosive increase in data on the web, recommending items to users is becoming more complex. In recent times, the best recommender systems have come from ensemble learning, which combines many models and techniques to generate recommendations that can draw the best characteristics of the constituent models. These ensemble models can improve accuracy, and they are also able to reduce the biases that come with each model. However, the massive number of permutations in which recommendation models are combined to create an ensemble model adds another layer of complexity to an already complex problem. Thus, in line with the recent trend in Automated Machine Learning (AutoML) aimed at reducing the complexity associated with model selection, there is a need for a machine learning framework that can learn the best ensemble model for a given problem given the base models. We proposed a system Ensemble with Particle Swarm Optimization, which intelligently optimizes the recommendations by identifying the best ensemble architecture for the data at hand. Our proposed AutoML system can improve recommendations for the MovieLens dataset by combining the results from base techniques without any user effort. The major challenge in creating such a system is to develop a framework for generating ensemble models and finding an efficient way to reach the best performing model since the search space can be vast. To overcome this challenge, we have used hierarchical models to generate ensembles and Particle Swarm Optimization to find the optimal ensemble models.

Keywords AutoML · Recommender systems · EnPSO · Machine learning · Ensemble learning · Particle swarm optimization

1 Introduction

Due to the nature of digital marketplaces, where the platforms are not bound by physical space, companies are adding more and more content and products and thus increasing the options available to customers manifold. To illustrate, Amazon today has around 12 million products available, and Netflix has approximately 6000 movies and shows on its platform. It is practically impossible for the user to scroll through millions of such options. This problem makes E-commerce shopping a daunting task. To solve the problem, recommender systems have emerged as an essential tool. Using recommender systems, we can offload the problem of item selection by implementing machine learning algorithms. These algorithms learn and predict what a user is likely to buy, hence reducing the set of products that the user has to go through to find an item that is relevant to his needs. Users often visit a platform just for window shopping, where they have no particular product or need in mind. In such a scenario, recommender systems become even more critical as they bear the entire burden of customer conversion. The system has to pick a comparatively much smaller subset of items to show the user from a much bigger list of products or services. If the system does not select relevant products, it loses potential sales despite having the right products. Thus, Recommender systems are essential for working with a large pool of data. It considers the user's liking, previous purchase/like history, the social network of the user, and much more. After considering such parameters,

✉ Rahul Katarya
rahuldtu@gmail.com

Garima Gupta
garimacsdtu@gmail.com

[1] Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

Published online: 23 April 2021

the recommender system makes a systematic and sound list of products tailored to the user's likes.

The recommendations depend upon several parameters and are divided the recommender systems into three types, i.e., content based, collaborative filtering based, and hybrid based. Apart from these broad categories, several other techniques are used to build a recommender engine, depending upon the primary parameters taken into account. For instance, in this paper, we are using five recommendation techniques as a comparative benchmark analysis. One such technique is recommendations based on popular items [1]. In this technique, the recommendations are made based on the popularity of an item. So, if an item or a product has a very high rating and is positively reviewed by many users, it would be recommended. This technique is used widely across all e-commerce websites. However, the main disadvantage of this technique is that the user's recommendations may or may not be best suited for him. Although the item is popular, it may not be tailored to the user's preferences. Similarly, another technique that we have used in the paper is generating recommendations of random items to the user [2]. In this technique, the system generates random recommendations for the user. The reason behind using this technique is to get rid of the cold-start problem [3] since it is free from any bias. Whenever a new user joins an e-commerce site, this technique is used. Nevertheless, the main disadvantage of this technique is that the recommendations might not be tailored to the user's preferences, and the recommendation technique will eventually lose its purpose.

Another example is the user-based collaborative filtering technique [4]. In this technique, the items are recommended to the user based on the preferences of the neighbors of the user. If the user's neighbor likes a product, the recommender system is likely to recommend that product. This technique plays on the theory that if my friend likes an item, there is a high chance that I also might like that item. However, this technique comes with a significant disadvantage. The theory upon which this technique is based is not always correct. Suppose the user's neighbor likes the product range $\{x_1, x_2,\ldots..,x_n\}$ and the user likes the product range $\{y_1, y_2,\ldots..,y_n\}$, then the user will be recommended the product range $\{x_1, x_2,\ldots..,x_n\}$. The recommended product list will be completely different from the user's liking, and the application of the recommender system will lose its meaning. These were a few examples of the recommendation techniques commonly used. There are many such techniques, and each such technique suffers from one or the other limitation.

If we use any one of these techniques, our recommender system will have some short-comings, and to get rid of this problem, ensembles are a great tool. However, creating ensembles that effectively combine individual techniques is a complex problem in its rights. The number of permutations in which techniques can be combined is too large for any Machine Learning practitioner to explore all on his own. Thus, he has to rely on the knowledge of ensemble learning to make informed decisions and come up with an ensemble model that works. This becomes a big chore in itself in the development process of a recommender system. Our work aims to eliminate this process by introducing an Automated Machine Learning (AutoML) framework that provides an optimized ensemble structure to the user without any oversight or the requirement of advanced knowledge about creating ensembles.

AutoML was first introduced [5] to automate the training of a Recurrent Neural Network by Reinforcement Learning to identify the most optimum architecture. AutoML was defined as a combination of automation and Machine Learning [6]. The authors in [7] explained how, without much intervention from experts, the ML applications could be implemented by the domain experts using AutoML. As explained in [8], the implementation of AutoML encompasses all the machine learning processes, i.e., *data preparation, feature engineering, model generation,* and *model evaluation.* In this paper, we have emphasized the *Model Generation* segment of the process. In our technique, Ensemble with Particle Swarm Optimization (EnPSO), we use Particle Swarm Optimization to search the model parameter space to find the best performing model. The model's generation is done by identifying and combining intelligent operations like intersection, concatenation, and many more. The system then generates the structure of the model. After the generation of the model, the system performs the optimization of parameters. In this stage, the system identifies the most optimum model by continually tweaking the parameters in a large search space.

1.1 Major Difficulties and Challenges

The science involved in generating recommendations comprises complex computational problems like machine learning, deep learning, and others. As discussed before, a recommender system using just one Machine Learning technique suffers from one or more limitations. Hence, it is crucial to use an array of techniques.

Generating efficient ensembles is crucial to mitigate the shortcomings of the individual recommender techniques. However, creating an ensemble is a challenging task on its own. There are endless ways to combine techniques and their outputs. For an ill-defined model, combining the base techniques in the search space for the parameters can be very large and require extensive training data and a lot of time to optimize. To solve this problem, we use a hierarchical model which, without blowing up the search space, offers us efficient ways to combine the base recommenders. The second challenge is optimization. Even with the right model, the search will still have a large number of dimensions across

which we are required to search for the optimal ensemble. To solve this shortcoming, we have used PSO, which performs better for larger search spaces in terms of time taken to converge.

Employing a single recommendation technique is already a complex problem that requires optimizing the model parameters and the hypermeters for the given data. On top of that, combining multiple such techniques using either ensemble learning or some other method adds another layer of complexity and requires the practitioner to have advanced knowledge of ensemble learning. To solve this issue, we have applied AutoML, which automates this entire process and reduces such unnecessary layers of complex computations to generate accurate recommendations. Another alternative way to reduce the complexity of the model selection is to use standard frameworks for creating ensembles. However, such a system would not be optimized and hence would not provide the best results. Attempt to optimize the ensemble will require expert knowledge and will hence introduce complexity. Our system provides an easy way to find a well-performing ensemble without the associated complexities.

1.2 Major Contributions of Our Manuscript

1. We proposed an AutoML framework for Ensemble Learning in recommender systems to automate the process of generating the ensembles.
2. We employed the evolutionary algorithm Particle Swarm Optimization (PSO) on the generated ensembles for finding the best model in the search space.
3. We performed the analysis of the results obtained on three publicly available benchmark MovieLens datasets (100 K, 1 M, 10 M).
4. We implemented five benchmark machine learning techniques as base recommenders to evaluate our system.

The paper has been organized as follows. Section 2 states the related work done in the field. Section 3 elucidates the system architecture, and Sect. 4 gives an introduction to the base recommendation classifiers to compare our proposed algorithm. Section 5 explains the concept of ensemble learning. In Sect. 6, the preliminary behind the implementation and optimization of generated ensembles using AutoML is provided. Results and experimental analyses have been provided in Sect. 7 of the paper. The paper concludes with the Conclusion and Future Work in Sect. 8, followed by References.

2 Related Work

Much work has already been done in the domain of recommender systems and in implementing ensemble learning on recommender systems, but no work has been done in employing AutoML to automate the entire process. This section covers the research studies done in this domain of work.

For improving movie recommendations, [9] implemented the boosting technique of ensemble learning. In another research, [10] combined the content-based and collaborative recommender system by creating an ensemble of both techniques using a hierarchical Bayesian approach. Authors in [11] created an ensemble regressor for the prediction of missing ratings in recommender systems. Similarly, in another study [12], the authors created an ensemble of unimodal generated rankings, hence creating multimodal interactions resulting in improved recommendations. [13] used a multi-view ensemble system called MV-DEM for detecting shilling attacks using base classifiers. Authors in [14] created an ensemble of matrix decomposition methods combined with SVD++ to study its effects on collaborative filtering. To study and remediate the problem of cold start in recommender systems, authors in [15] applied Gradient Boosting Decision Tree ensemble learning on access logs. In another study concerning online product recommendations, authors [16] implemented three ensemble approaches based on user feedback's multimodal interactions. To improve music recommendations, authors in [17] applied a bio-inspired cluster ensemble of swarm intelligence and fuzzy clustering on UBCF (user-based collaborative filtering). Wu [18] created an ensemble of regularized, maximum margin, and nonnegative matrix factorization on collaborative filtering. Table 1 enlists other research work done in this area in a concise format. Authors in [19–21] used TLBO (Teaching Learning-Based Optimization) and Jaya algorithms to optimize the results, and authors in [22] used grey wolf optimization algorithm to optimize their results. A few different applications of ensemble learning implemented in several other domains are [23–28].

In the research studies cited in Table 1, the main drawback that we found common was that after performing ensemble learning, any optimization technique was not applied to the ensemble structures to find the best results. In addition to that, major missing work was using the concept of AutoML in any recommender system. Hence, in this paper, we have worked towards all these shortcomings to present a novel work.

Table 1 Research studies featuring related work

Paper	Objective	Method
[7]	Without much intervention from experts, the ML applications can be implemented by the domain experts	Applied AutoML on the largest number of frameworks as compared to other research studies
[6]	Defined AutoML as a combination of automation and machine learning	Presented an exhaustive survey on AutoML
[5]	Introduced Neural Architecture Search and implemented it on CIFAR-10 dataset	The authors introduced the concept of AutoML with automating the training of a Recurrent Neural Network by reinforcement learning to identify the most optimum architecture
[17]	To enhance the recommender technique of collaborative filtering approaches	Applied an ensemble of fuzzy-based clustering and swarm intelligence on user-based collaborative filtering
[29]	Applying ensemble learning to enhance user's recommendations	Applied Logistic Regression, Decision Trees, and Artificial neural Networks and implemented Bagging and Bumping to improve performance
[30]	Improving the dynamic nature of news recommendation to keep the user constantly updated	Implemented and evaluated three different ensemble approaches
[31]	Implementing the multi-criteria feedback to improve collaborative filtering recommendations	Application of Hypergraph Partitioning Algorithm (HGPA) on clusters
[32]	Proving that cluster ensembles provide better recommendations	Applied three ensemble methods on two cluster techniques (k-means and SOM)
[33]	To predict the side effects of drugs	Created an ensemble of Neighborhood-based method and Restricted Boltzmann machine-based method
[9]	Improving movie recommendations	Implemented boosting ensemble learning for improved recommendations
[10]	Combining content-based and collaborative filtering to improve recommendations	Created an ensemble of content-based and collaborative filtering such as a hierarchical Bayesian approach
[12]	Improving the recommendations using multimodal interactions	Ensemble of unimodal generated rankings based on interaction types
[13]	Detection of shilling attacks in recommender systems	Implemented a multi-view ensemble detection framework by merging base classifiers
[14]	To study the effects of matrix decomposition methods for collaborative filtering	Created an ensemble of matrix decomposition methods combined with SVD++
[15]	To study and work on the cold start problem in recommender systems	Apply Gradient Boosting Decision Tree ensemble learning on access logs
[16]	To improve online products recommendations	Implementation of three ensemble approaches based on user feedback's multi-modal interactions
[18]	Appling matrix factorization on collaborative filtering	Created an ensemble of regularized, maximum margin and nonnegative Matrix factorization
[34]	To improve collaborative filtering-based recommender system by implementing clustering techniques	Created bio-inspired clustering techniques of swarm intelligence and fuzzy-based clustering for user-based recommendations
[35]	Improving the accuracy of text recommender systems	Application of text mining-based ensemble learning in a post-processing technique
[36]	Solving the problem of dynamic recommender systems for constantly updating the feed	Implementation of online bagging ensemble learning
[37]	Deliver compound recommendations for online social network	Included social trust ensemble, topic modeling and probabilistic matrix factorization
[38]	Recommending user-specific, latest and relevant news items	Creating a hierarchical ensemble for user groups based on common interests

3 System Architecture

The employment of the system starts from uploading the MovieLens dataset and then performing data cleaning to filter out the columns which are essential to us. This can also be referred to as feature extraction, where the system extracts all the features or columns from the dataset which are relevant to our system. In this study, one of the most crucial columns is *movies*. In this column, all the movies present in the dataset are listed. After extracting this feature, the EnPSO checks for NULL rows. All such NULL rows are discarded. Afterward, the system extracts the feature *rating*. Similarly, it checks for NULL rows and discards such data points. Similarly, the system extracts the feature *tags* and performs feature extraction on these column values. After all the data cleaning and feature extraction, EnPSO

applies all the recommendation classifiers, i.e., POPULAR, RANDOM, IBCF, UBCF, and SVD. After this, the ensembles are created, and the generation of the ensembles is performed using the hierarchical ensemble technique about which we discussed earlier. Then the EnPSO optimizes the ensemble using Particle Swarm Optimization (PSO) and checks whether the ensemble is optimized or not. After generating the optimized ensemble, the content-based recommender engine is applied by the EnPSO to generate recommendations.

Five different techniques are used in the paper as base recommenders, IBCF, UBCF, POPULAR, RANDOM, and SVD. An ensemble structure ENSEMBLE is created, which takes the set of techniques and the set of ensemble methods as the parameters. In the later step, Particle Swarm Optimization (PSO) is applied to function F for identifying the most optimum recommendation results. Finally, we get the results and the ensemble structure, which provided the most optimum results. The architecture of the system is shown in Fig. 1.

4 Base Recommenders

To prevail over the limited efficiency of the recommender systems, we have incorporated ensemble learning in our paper. Ensemble Learning is a technique wherein multiple algorithms are trained, and their outputs are combined, forming a "committee." The formation leads to the calculation of results in such a way that the most optimal result is obtained by the end of implementing the ensemble [39]. The ensemble method primarily increases the effectiveness of the model. Weights are assigned to the outputs of multiple classifier systems, and the opinions and results of other individual decisions by the classifier are combined. Hence, the final result of the ensemble learning of these classifiers provides the most optimum result. The ensemble model designs itself in such a way that the risk of making a weak solution is avoided. For example, for undergoing surgery, the patient consults a few doctors and then decides which is best for him. This decision that he makes depends upon the prescriptions provided by each doctor. Similarly, in ensemble learning, different techniques provide different results, and individual weights are assigned to the results. According to these results, the ensemble optimizes the output of the recommender by combining the results of base classifiers in an optimal way. It can also be said that in ensemble learning, different training parameters for each classifier generate different decision boundaries. The following are different benchmark techniques for making recommendations that we have used in our paper to compare with our proposed algorithm EnPSO.

We have used the following five base recommendation techniques in the AutoML. The results obtained in this paper will lead to the implementation of other techniques in our future research. The first technique that we used is based on *popular items*. In this technique, a vast pool of popular items, i.e., the items with maximum positive reviews, are maintained. The users are recommended items from this pool so that they are recommended the most highly rated

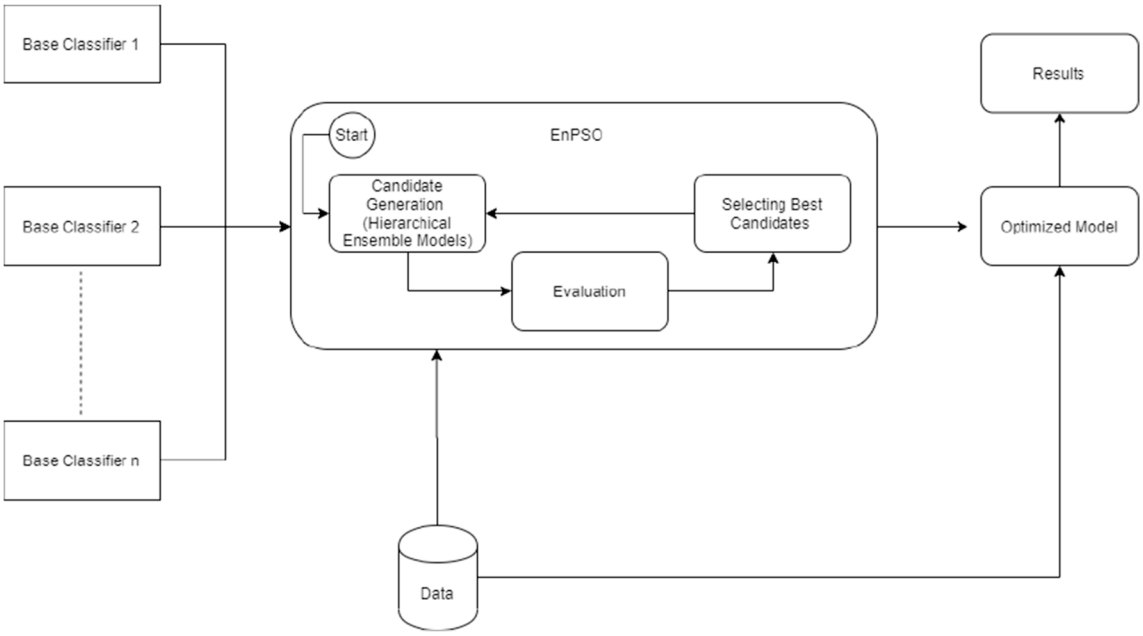


Fig. 1 Architecture of the proposed AutoML ensemble recommender system

items, and the probability of purchase increases. This technique is best in the case of cold-start users [40]. The cold-start problem in users is one of the most common problems witnessed in recommender systems. Whenever a new user is introduced, the recommender system does not have any idea about the choice and liking of the user [1]. So, it is very cumbersome for the recommender system to recommend appropriate items. Thus, given that case, this technique of recommending the most popular items is used, and the chances of the user liking the items increase.

The second classifier is *random items*. In this technique, depending on the size of the dataset, random data points are selected from the entire pool of data. This dataset consists of random items. When recommendations take place from this dataset, items are recommended randomly to the users. This recommendation technique is primarily used whenever the problem of cold-start occurs for items. Whenever a new item is launched in the market, it gets challenging to label that data item as there are no previous reviews for that data [2]. When this technique is applied, random items are recommended to the users, and there emerges a high probability that this new item is randomly chosen to be a part of the to-be-recommended dataset. Hence, this technique effectively solves the problem of the cold-start problem for items.

The third classifier is *User-Based Collaborative Filtering (UBCF).* User-based collaborative technique is a neighborhood-based approach in the world of recommender systems [41]. There exist numerous factors which are taken into account to construct an efficient recommender system. In the previous subsections, we read about the recommendations taking place based on the popularity of the items and based on the randomness of the recommender engine. However, to create an algorithmic approach, the recommendations in collaborative filtering depend on the neighbor of the user. In user-based collaborative filtering, if the user's neighbor likes an item X, the recommender engine recommends Y item to the user, given that X and Y are related to each other in some way [4]. This type of generation of recommendations is specific to the users. However, the next subsection describes such recommendations generating based on the items.

Major notations used in the paper are summarized in Table 2:

The similarity between two users a and b can be given by Eq. 1 [42]:

$$\mathrm{sim}(a,b) = \frac{\sum_{i\in I}\left(r_{a,i}-\bar{r}_a\right)\left(r_{b,i}-\bar{r}_b\right)}{\sqrt{\sum_{i\in I}\left(r_{a,i}-\bar{r}_a\right)^2}\sqrt{\sum_{i\in I}\left(r_{b,i}-\bar{r}_b\right)^2}} \tag{1}$$

here, $r_{a,i}$ is the rating of user a for item i and similarly $r_{b,i}$ is the rating of user b for item i. I is the list of all the items rated by the users a and b.

Table 2 Major notations used in the paper

Notation	Remarks
$r_{a,i}$	Rating of user a for item i
$r_{b,i}$	Rating of user b for item i
$\mathrm{sim}(a,b)$	Similarity between two users a and b
$\hat{r}_{ui}$	Expected rating
I	List of items rated by the users a and b
λ	Regularization vector
E	Bootstrapped estimator
C_{soc}	Social attractiveness
C_{cog}	Cognitive attractiveness

The fourth classifier is the *Item-Based Collaborative Filtering (IBCF).* As given in the previous subsection, item-based collaborative filtering is also a type of neighborhood-based approach as it is a collaborative filtering-based recommendation technique. However, unlike in the user-based approach, the item-based approach provides recommendations to the user based on the similarity between the items. To give an example, if the user likes product P and if product P is similar to product Q in one or the other way, there is a high probability of the user liking product Q. Hence, once the recommender engine knows that the user likes product P, it generates such product recommendations such that certain feature vectors of those items are common to product P [43]. Equation 2 determines which items are to be recommended to the user from a list of items [44].

$$\mathrm{score}(u,i) = \sum_{j}^{I} \frac{\mathrm{similarity}(i,j)\left(r_{(u,j)}-\bar{r}_j\right)}{\mathrm{similarity}(i,j)} + \bar{r}_i \tag{2}$$

here, u is the user for whom the recommendations are generated, i represents the item concerned, and j represents the list of items similar to item i.

The fifth and final classifier used is *Singular Value Decomposition (SVD).* This technique aims at reducing the dimensionality of the dataset. The features in the datasets that are used for computation are enormous in number, and they are often not restricted to just one dimension. This technique helps in that scenario and hence extracts such features which are used in comparing the users [45]. To effectively perform the recommendation, in the first step, the dataset is spread across a user-item matrix, and the entries in the cells are the individual ratings. In the second step, the empty cells are filled with average rating values. In the last step, the dimensionality of the user-item matrix is decreased by using the SVD algorithm, and the rating values can be predicted from the previous matrix.

Let each item in the user-item matrix be represented by vector i and each user be represented as vector u. The expected rating can be Eq. 3:

$$\hat{r}_{ui} = i^T u \tag{3}$$

here i^T is the transpose of the item vector I, to find the vectors i and u, the minimum of the user-item matrix can be found using Eq. 4:

$$\min(u,i) \sum_{(u,i)\in K} \left(r_{ui} - i^T.u\right)^2 \tag{4}$$

To avoid our model to over-fit the training set, we have introduced a regularization vector, λ to Eq. 5 [46]:

$$\min(u,i) \sum_{(u,i)\in K} \left(r_{u,i} - i^T.u\right)^2 + \lambda\left(||i||^2 + ||u||^2\right) \tag{5}$$

These classifiers are used as hyperparameters of the ensemble learning engine. We will study the details of the ensemble learning in Sect. 4.

There are other kinds of recommender techniques that can be used as base recommenders in our system [47], e.g., trust-aware-based recommender, location-aware recommender, content-based filtering, and others. However, the aim of this paper is to provide an insight into the effect of ensemble learning on generating recommendations and how the implementation of an evolutionary algorithm efficiently chooses the best performing technique by generating ensembles. We could have used other techniques, but we used the afore-mentioned base recommenders as they are standard across the domain in recommender systems and serves as the right choice for a proof of concept for our system, which is designed to be generic and easily extensible to other domains.

5 Ensemble Technique

Recommender systems are primarily of three types, i.e., *collaborative filtering based*, *content-based,* and *hybrid*. Collaborative filtering technique involves the neighbors of the user, and depending upon the past purchase history of the neighbor and his likes or dislikes, the user is recommended items [9]. In content-based filtering, the recommendations occur based on the content of items the user previously purchased or liked [10]. However, the hybrid recommender system is an amalgamation of both content-based and collaborative filtering-based recommender systems. It works on the properties of both content-based and collaborative filtering recommender systems.

Ensemble Learning is a technique where many base models are combined to form one optimized model. It is beneficial for either an extensive dataset or a very small dataset. In the former case, the dataset is divided into smaller subsets of data, and different classifiers work on these different individual datasets. In the latter case, the concept of *bootstrapping* is used [48]. In this technique, one subset of data is selected at random, and then another subset of a dataset is selected with replacement from the entire pool of data points. Ensemble learning is the most efficient way of creating a hybrid recommender system.

5.1 Hierarchical Learning

Hierarchical learning is a two-step process. In the first step, each base learner learns or processes the information either separately or after communicating. In the second step, the predictions summarized by the trained classifiers are combined, considering the functional hierarchy. The systematic structure of parametric optimization of the Hierarchical Ensemble using PSO is provided in Fig. 2.

There are two types of ensemble systems according to their functionality, *parallel ensemble system* and *sequential ensemble system.* Figure 3 represents the architecture of the parallel ensemble systems. In parallel systems, the execution of all the recommenders involved in the system is independent of each other, and the outputs of each recommender are combined using an aggregator. *Weighted* and *switching* are two types of parallel ensemble techniques. Figure 4 represents the architecture of the sequential recommender systems. In these systems, the input of $(n-1)$ recommenders depend upon the output of the previous recommender, given that there are n numbers of recommenders. *Cascade* and *Feature Augmentation* are two types of sequential ensemble techniques.

There are several types of ways to perform ensemble learning by forming ensembles. The first type is *Bagging* or *Bootstrap Aggregating.* It is a technique of ensemble learning wherein bootstrapping and aggregation are combined to form one model. Given the data, several bootstrapped samples are formed, and a decision tree is formed on every bootstrapped subsample. Afterward, a decision tree is formed after aggregating all the decision trees to find the most optimum solution [49]. It is represented as follows:

For every bootstrap sample (X_1^*, Y_1^*),….. (X_n^*, Y_n^*),

Compute the bootstrapped estimator using Eq. 6:

$$E = h_n\left(\left(X_1, Y_1\right), \ldots..\left(X_n, Y_n\right)\right)(\cdot) \tag{6}$$

where function $h_n(\bullet)$ defines the function of the data.

Another type of ensemble learning is called *Boosting*. Unlike Bagging, the boosting ensemble methods are sequential, where the weights depend on the previous fitness functions. The base models considered for boosting have low variance and high bias because boosting aims at reducing bias. In addition to this, the boosting models are computationally expensive to fit. The task of fitting the weak learners to form a final sequential model can be of two types, *Adaboost* and *gradient boosting*. The significant difference

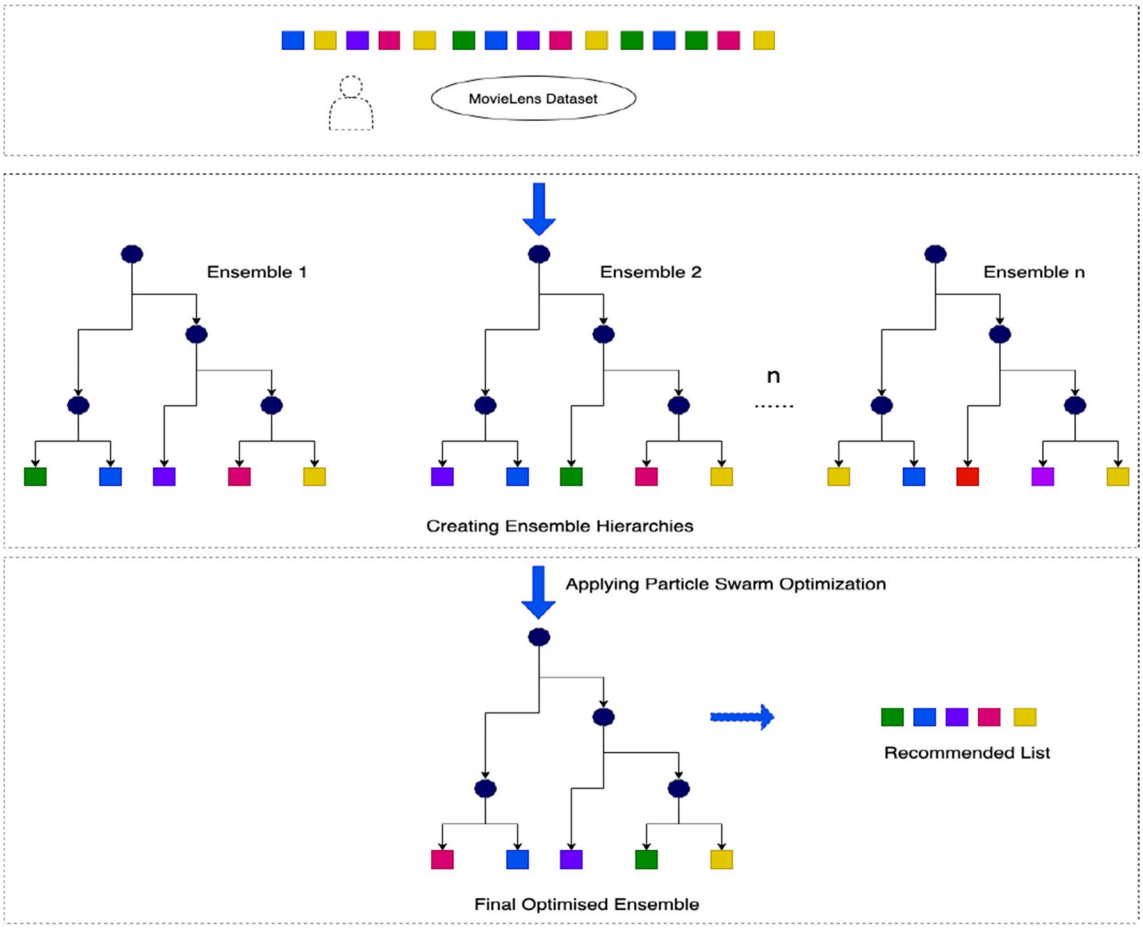


Fig. 2 Systematic structure of the parametric optimization of Hierarchical Ensemble using PSO

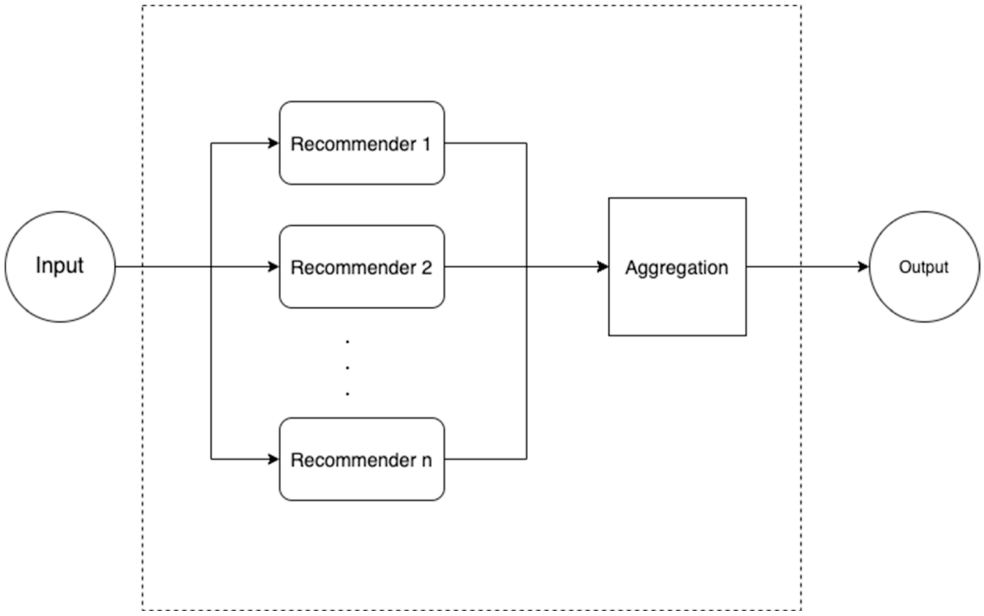


Fig. 3 Architecture of the parallel ensemble systems

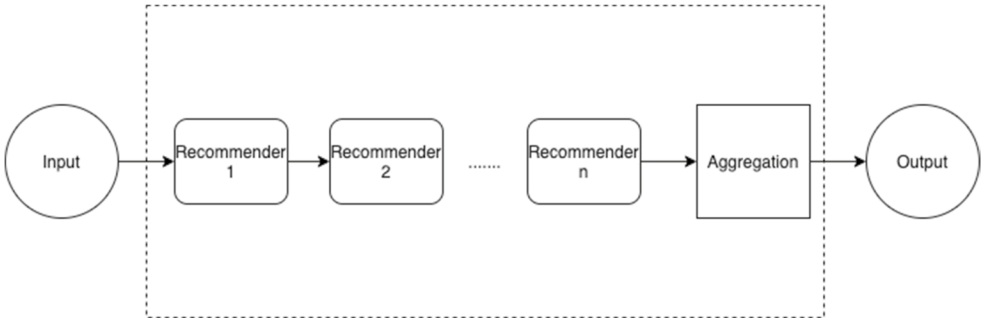


Fig. 4 Architecture of the sequential ensemble systems

between the two models is the technique of creating and aggregating the weak learners in the sequential process. In adaptive boosting, the weights attached to the observations of the training dataset are updated, and in gradient boosting, the values of the observations are updated.

5.1.1 Adaptive Boosting (Adaboost)

1. The Adaboost ensemble model can be defined as a weighted sum of L weak learners by Eq. 7 [50]:

$$s_L(.) = \sum_{l=1}^{L} c_l \times w_l(.) \tag{7}$$

Here, c_l represent the coefficients and w_l represents the weak learners.

Since this is an additive problem and requires both coefficients and weak learners, we employ iterative optimization technique. It can be given as following Eq. 8:

$$s_l(.) = s_{l-1}(.) + c_l \times w_l(.) \tag{8}$$

This can further be denoted in Eq. 9:

$$\begin{aligned}\left(c_l, w_l(.)\right) &= \arg\min_{c,w(.)} E\left(s_{l-1}(.) + c \times w(.)\right) \\ &= \arg\min_{c,w(.)} \sum_{n=1}^{N} e\left(y_n, s_{l-1}\left(x_n\right) + c \times w\left(x_n\right)\right)\end{aligned} \tag{9}$$

here, $E(.)$ represent the fitness error of the model, and $E(.,.)$ represent the loss function.

5.1.2 Gradient Boosting

In this method, the additive problem is visualized such that it is casted into gradient descent. The gradient descent process can be written as Eq. 10 [51]:

$$s_l(.) = s_{l-1}(.) - c_l \times \nabla_{s_{l-1}} E\left(s_{l-1}\right)(.) \tag{10}$$

here $E(.)$ is the fitting error, and c_l is the coefficient of the step size.

5.1.3 Bayes' Optimal Classifier

It is a classification technique that makes the computations more feasible by presuming that the data is conditionally independent. Each hypothesis is given a vote equivalent to the possibility that the training data is sampled if that hypothesis were true. It is given by Eq. 11 [52]:

$$x = \arg\max \sum_{c_j \in C, h_i \in H} P\left(c_j|h_i\right) P\left(T|h_i\right) P\left(h_i\right) \tag{11}$$

here, $x \in$ predicted class, $C \in$ set of all possible classes, H is the hypothesis space, and $T \in$ training data.

The ensembles are said to have two types of designs, i.e., *parallel* and *sequential,* as shown in Fig. 5. For example, weighted and switching designs can be considered as parallel designs because, in these designs, all the recommenders can work on problems to either provide a weighted average or conquering tasks using different recommenders, respectively. However, cascade and feature augmentation designs

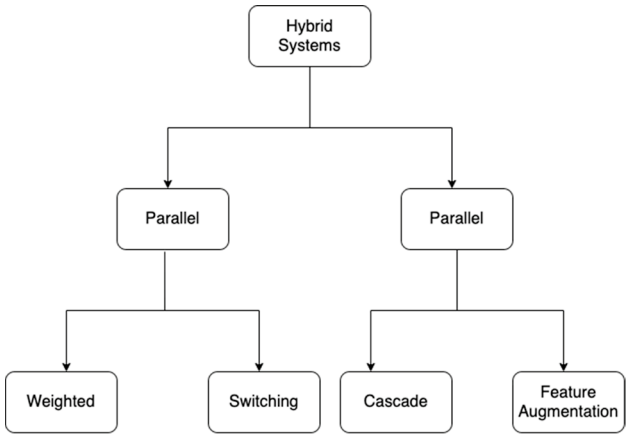


Fig. 5 Classification of ensemble hybrids

are sequential since they operate sequentially, and the execution of one recommender system and generating its output is essential to carry forward the rest of the process.

5.2 Advantages of Ensemble Learning

There are primarily three reasons for performing ensemble learning. The first reason is *Statistical*. Due to the lack of appropriate data, it gets challenging to represent the data distribution efficiently. The second reason is *Computational*. Among many available models, it gets difficult to decide which model is to be chosen. The third reason is *Representational*. This reason occurs when the decided model is not able to provide adequate results because of providing improper decision boundaries.

We have used hierarchical learning because it produces a *high modularity* system. The two-steps included in this technique increase the *efficiency* of the system. After creating ensembles, in the next step, we applied the evolutionary Particle Swarm Optimization on the generated ensembles to find the most optimum solution.

6 EnPSO: AutoML Generation of Optimal Ensemble Model

In our study, the AutoML has optimized the hierarchical ensemble recommender system using *Particle Swarm Optimization (PSO)*. This algorithm uses an evolutionary strategy for model optimization. The algorithm starts with a set of models and uses evolutionary techniques to select and evolve the best model from the set. AutoML achieved this evolutionary nature of algorithm by using a known evolutionary algorithm, Particle Swarm Optimization [53]. It revolves around the concept of moving *particles* or solutions around in the search-space dimension depending upon the particle's position and velocity. In PSO, every particle has a memory that is vital to the working of the algorithm. Every particle has an independent velocity of its own and using these velocities. Such particles keep updating themselves dynamically. The particles help in transmitting the information among each other, and that is how the information processing takes place. Not every particle holds power to transmit information. Only the best lot of particles are allowed to do so. The spread of information is a one-way transmission, and the best solution leads to the evolution.

Its main advantage is that it can search in vast spaces of particles. Evolutionary Algorithms are meta-heuristic algorithms that aid in solving hard problems. These algorithms mimic the working of nature in various forms. Such processes of nature are replicated to perform computations, and hence, they help in achieving the most optimum results [54].

6.1 Alternatives of Particle Swarm Optimization algorithm

There are several different bio-inspired evolutionary algorithms used to improve the computation results. Some of these algorithms are *Genetic Algorithms* [55], *Ant-Colony Optimization* [56], *Artificial Bee Colony Algorithm* [57], *Bees Algorithm*, *Cuckoo Search* [58], *Firefly Algorithm* [59], *Particle Swarm Optimization* [53], *Hunting Search* [60], etc.

6.2 Advantages of using PSO

Since our problem has large dimensions, we found PSO to be better suited than the other methods because PSO performs better in problems with large dimensions and reaches the result more rapidly than the alternative methods. PSO also works with the fitness functions that are not differentiable or difficult to model mathematically, which happens to be the case with the problem on hand. PSO also converges faster than the alternatives, and since our search space is very large, PSO is more convenient than the alternatives in the context of time complexity.

In comparison with the other evolutionary algorithms, the key advantages of using PSO are its uncomplicated mechanism of implementation, and the hyperparameters which need to be adjusted are not enormous. In our described model EnPSO, we have used Particle Swarm Optimization as the evolutionary algorithm because its advantages gel perfectly with the requirements of the solution to be provided. There are a large number of ensembles that are created using ensemble learning, and hence to identify the best performing ensemble or the solution, the search space also becomes enormous, and PSO performs this task brilliantly. In addition to this, the implementation of this algorithm is easier compared to other evolutionary algorithms, and the optimization of the hyperparameters strives for lesser time complexity.

6.3 Optimizing the Ensemble Model

There are the following four steps to perform evolutionary computation, (1) Initialization, this step deals with creating an initial population of solutions. The population is a collection of all the possible solutions to the problem [61]. (2) Selection, after the population is created, the members of the population, i.e., all the possible solutions to the given problem, are evaluated using a *Fitness Function*. This function decides how efficient a solution is [62]. And hence, after using this fitness function, the most efficient members are selected. (3) Genetic Operation, for carrying out the evolutionary computation, genetic operations are required, which mimic the biological processes. This operation takes place in two steps. The first one is **crossover** [63]. In this part, all the selected top members are used to create the next generation.

This is done so that the best features of both the parents are inherited by the children so that the results keep improving in every iteration. The second step is **mutation** [64]. Here, some changes or mutations are introduced in the children as produced in the previous step. This is done so that the children do not turn out to be exact replicas of their parents. If this happens, there will be no evolved generation.

PSO works on a population (**swarm**) of candidate solutions (**particles**). These particles move and interact in the search-space dimension adhering to a set of formulae. Another interesting concept in PSO is *topology*. A topology of the swarm can be defined as the subset of particles wherein each particle can trade information. This adds to the advantage of the algorithm as it promotes particle–particle communication [53]. Thus, the entire swarm agrees upon the same best position from a single particle. However, there is a downside to this concept that the swarm might get trapped in local minima. The working of the model is as follows.

In the first step, the AutoML generates a random position for the particles within an initialization region. In the second step, the velocities of the particles are initialized, and this can further be done in three ways. (1) within an initialized region. (2) zero. (3) assign random values to disallow particles from exiting the search-space in the first iteration [65]. The particles can also move to regions outside the feasible search space [66]. The activity of the swarm is characterized by two parameters, C_{soc} which determine the *social attractiveness* and C_{cog} which determines the *cognitive attractiveness* [67]. The velocity update of each particle i is given by Eq. 12 [66]:

$$\boldsymbol{v}_i = w.\boldsymbol{v}_i + C_{\text{soc}}.\boldsymbol{r}_1 \circ \left(x_i - g\right) + C_{\text{cog}}.\boldsymbol{r}_2 \circ \left(x_i - b_i\right) \tag{12}$$

where, $\exists\, i \in (1, \ldots .., n)$, n is the swarm size, $\mathbf{r}_1$ and $\mathbf{r}_2$ are the randomly generated vectors to stop the convergence to local minima, b_i defines the best position established by the particle i, g_i is the global best position established by the swarm for that particle i.

After every iteration, the position of the particle gets updated depending upon its velocity and this is denoted by Eq. 13:

$$x_i = x_i + v_i \tag{13}$$

Algorithms

Algorithm 1 (PSO)

Steps:

1. ***for*** *each particle p in a swarm population P:*
2. ***initialize*** *random* x_p
3. ***initialize*** *random* v_p
4. ***Evaluate*** *fitness function* $f(x_p)$
5. $Local_p = x_p$
6. ***initialize*** $global_p$ *as* x_p *having best fitness*
7. ***repeat until*** *stopping criteria is satisfied*
8. ***for*** *each particle p:*
9. ***update*** v_i *and* x_i *according to Equations 12 and 13*
10. ***Evaluate*** *fitness function* $f(x_i^t)$
11. ***If*** $f(local_i) < f(x_i^t)$
12. $Local_i \leftarrow x_i^t$
13. ***If*** $f(global_i) < f(x_i^t)$
14. $global_i \leftarrow x_i^t$

Algorithm 1 states the steps for implementing the PSO algorithm.

Algorithm 2 (EnPSO)

Input: *Users (u_1, u_2, u_3,…,u_n) who rated movies (m_1, m_2, m_3,…,m_4)*
Output: *Optimized ensemble generating recommendation results*

Steps:

```
15. for R in rows(dataset) :
16.     if R.movies == NULL
17.         R.delete
18.     end
19.     if R.rating == NULL
20.         R.delete
21.     end
22.     if R.tags = NULL
23.         R.delete
24.     end
25. end for
26. ibcf = ibcf.train (dataset)
27. ubcf = ubcf. train (dataset)
28. random  = random  . train (dataset)
29. popular = popular. train (dataset)
30. svd = svd. train (dataset)
31. ensemble = ensemble.create(ibcf, ubcf, random, popular, svd, initial_params)
32. objective_function = ensemble.results.MSE
33. pso(ensemble, objective_function, parameter_range)
34. optimized _params = pso.optimize()
35. return ensemble(ibcf, ubcf, random, popular, svd, optimized_params)
```

In Algorithm 2, our AutoML system creates a hierarchical ensemble of recommendation techniques. Any two or more recommendation models can be combined using voting or mixing, which leads to many different types of hierarchical ensembles that can be formed. Each structure will have its own set of hyper-parameters, which will need to be optimized. Different structures with their hyper-parameters will form a search space where the system finds the best performing model. To perform the search, we have used Particle Swarm Optimization. We performed the implementation of the EnPSO algorithm in R using the recommenderlab R package [68].

6.3.1 Time Complexity

The time complexity of Algorithm 1 is dependent on the stopping criteria. *Stopping criteria* is defined by the user as the number of iterations or as a threshold with respect to the objective function. We can define the computational complexity of PSO per iteration as follows [69]:

$$C_m = dNn + 3dn + 3 \tag{14}$$

$$C_a = dNn + 5dn + 3 + (N - 1) \tag{15}$$

here, C_m represents the number of complex multiplications per iteration (Eq. 14), C_a represents the number of complex additions per iterations (Eq. 15), d represents the

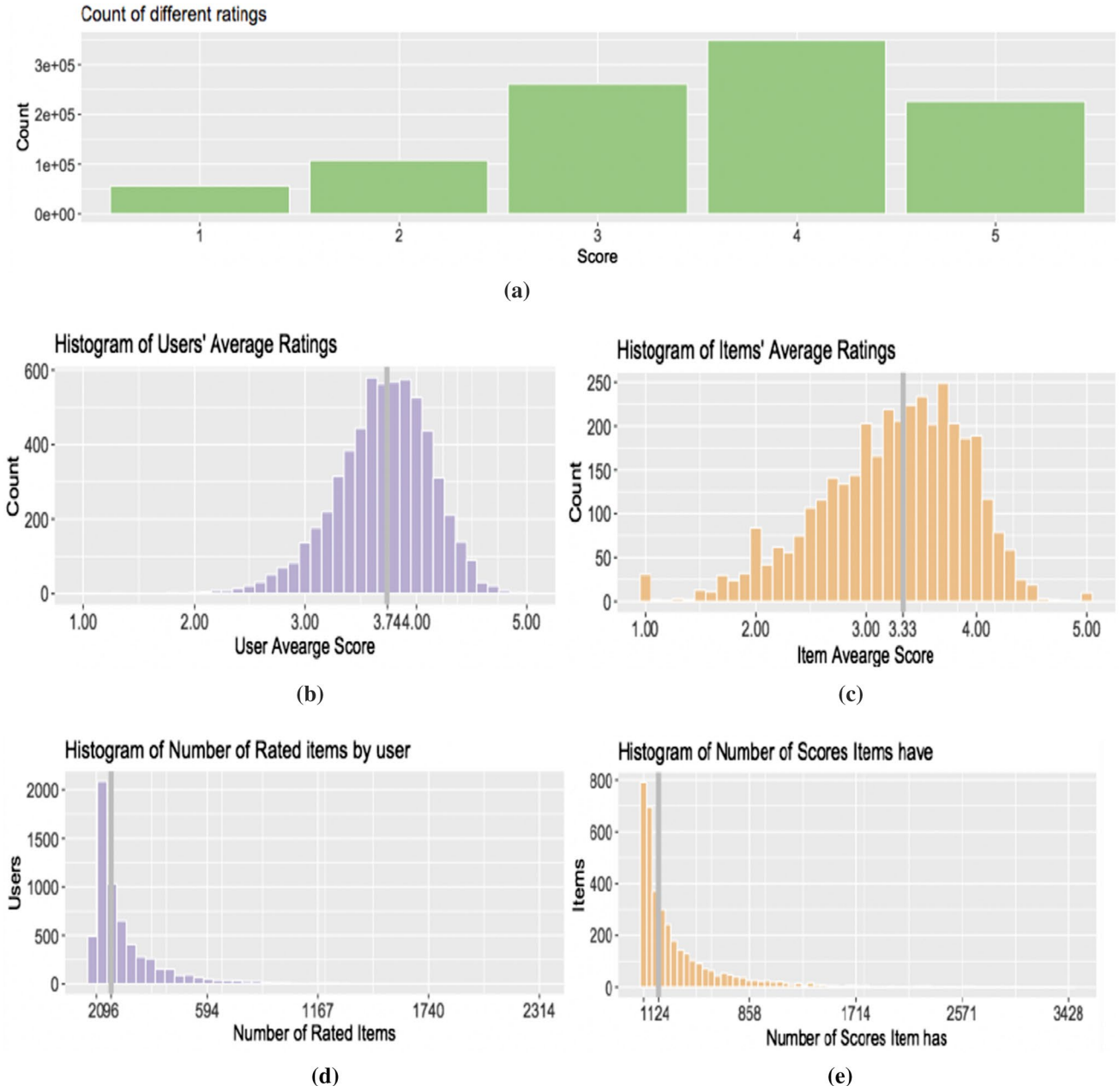


Fig. 6 **a** Study of the dataset based on number of different ratings, **b** study of the dataset based on average ratings of users, **c** study of the dataset based on the average ratings of the item, **d** study of the dataset based on the number of related items grouped by users, **e** study of the dataset based on the number of scores items have

dimensions, n represents the number of particles, and N represents the size of underlying data.

Algorithm 2 involves preparing data for training which is a one-time operation, and its complexity depends on the size of the data. In the second step, we train various base recommender systems, the time complexity of which is determined by the size of the training data. In the third step, we create the ensembles and try to find the optimal solution using Particle Swarm Optimization. The time complexity of this algorithm is determined as per the *stopping criteria* and the computation per iteration, as mentioned in Eqs. 14 and 15. Comparing our algorithm with the previous algorithm, creating ensembles through trial and error relies on expert knowledge. EnPSO eliminates the need for expert knowledge and practitioner and much quickly generates an optimized ensemble. There is no other AutoML technique we could find in the literature that would enable the practitioner to create ensemble recommenders.

7 Experiments and Results

In this section, a detailed explanation of the results obtained is provided. The information about the datasets used has been given, and the analysis of the working of the techniques is provided.

7.1 Dataset

We have used the benchmark MovieLens dataset for the implementation. This dataset consists of 3 data files encoded in UTF-8 format < ratings, tags, movies > . Figure 6a provides an analysis of the dataset based on the count of different ratings. Figure 6b provides the average rating of the users. Figure 6c shows the average rating of the items. In Fig. 6d, the number of related items grouped by users is shown. Figure 6e provides the analysis of the dataset based on the number of scores items have. We have done the implementation on 3 versions of MovieLens dataset:

1. 100 K: It consists of 100,000 ratings from 943 users on 1682 movies.
2. 1 M: consists of 1,000,209 anonymous ratings of approximately 3900 movies made by 6040 MovieLens users who joined MovieLens in 2000.
3. 10 M: consists of 10,000,054 ratings and 95,580 tags applied to 10,681 movies by 71,567 users

Other datasets could also have been used for the experimentation like Jester, Netflix, and others, but we used this dataset due to the availability of its vast range of data points and its benchmark usage in major recommender system papers, which is ideal for a proof of concept. Using this dataset made our work easily accessible and reproducible. Datasets other than MovieLens can also be used for the implementation as future scope of this paper, and we can expect similar results those datasets.

7.2 Metrics

To analyze the results obtained by the system, we have used the three *error metrics*, i.e., RMSE, MSE, and MAE. These are the standard error metrics used for evaluating a typical user-item recommender system like the one in this paper, i.e., movie recommender system. Metrics that can be used to evaluate the recommender system are precision, recall, and f1-score. However, these metrics vary greatly based on the domain and underlying data. Hence, they are more suited towards evaluating individual recommender systems within the same domain. However, for our case, we are presenting a proof of concept applicable across all domains. Hence, *RMSE*, *MSE,* and *MAE* are better-suited metrics.

7.2.1 Root-Mean-Square Error (RMSE)

It is defined as the standard deviation of the prediction errors. It is calculated by Eq. 16:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(\left(\widehat{x}_i \right) - \left(x_i \right) \right)^2}{n}} \tag{18}$$

For Eqs. 16, 17 and 18, Here, $\widehat{x}_i$ is the predicted value and, x_i is the observed value for the ith observation.

Table 3 Error metrics for 100 K MovieLens dataset

	UBCF	IBCF	SVD	RANDOM	POPULAR	EnPSO
RMSE	1.0906673	1.266997	1.1077915	1.591971	1.0347649	0.9183021
MSE	1.1895551	1.555479	1.2272020	2.425979	1.0707384	0.9837294
MAE	0.8905862	1.130137	0.9066842	1.174222	0.8278716	0.7529134

Table 4 Error metrics for 1 M MovieLens dataset

	UBCF	IBCF	SVD	RANDOM	POPULAR	EnPSO
RMSE	1.0906673	1.366997	1.0752522	1.510546	0.9732996	0.8837047
MSE	1.1895551	1.98438	1.1561673	3.689639	0.9473122	0.9183024
MAE	0.8905862	1.196394	0.8598624	1.413785	0.7677965	0.7192942

Table 5 Error metrics for 10 M MovieLens dataset

	UBCF	IBCF	SVD	RANDOM	POPULAR	EnPSO
RMSE	1.0543233	1.329419	1.0636744	1.488290	0.9638857	0.8528932
MSE	1.1115976	1.943237	1.1314032	2.527348	0.9290757	0.8920123
MAE	0.8388887	1.075388	0.8483209	1.0922772	0.7560278	0.7421234

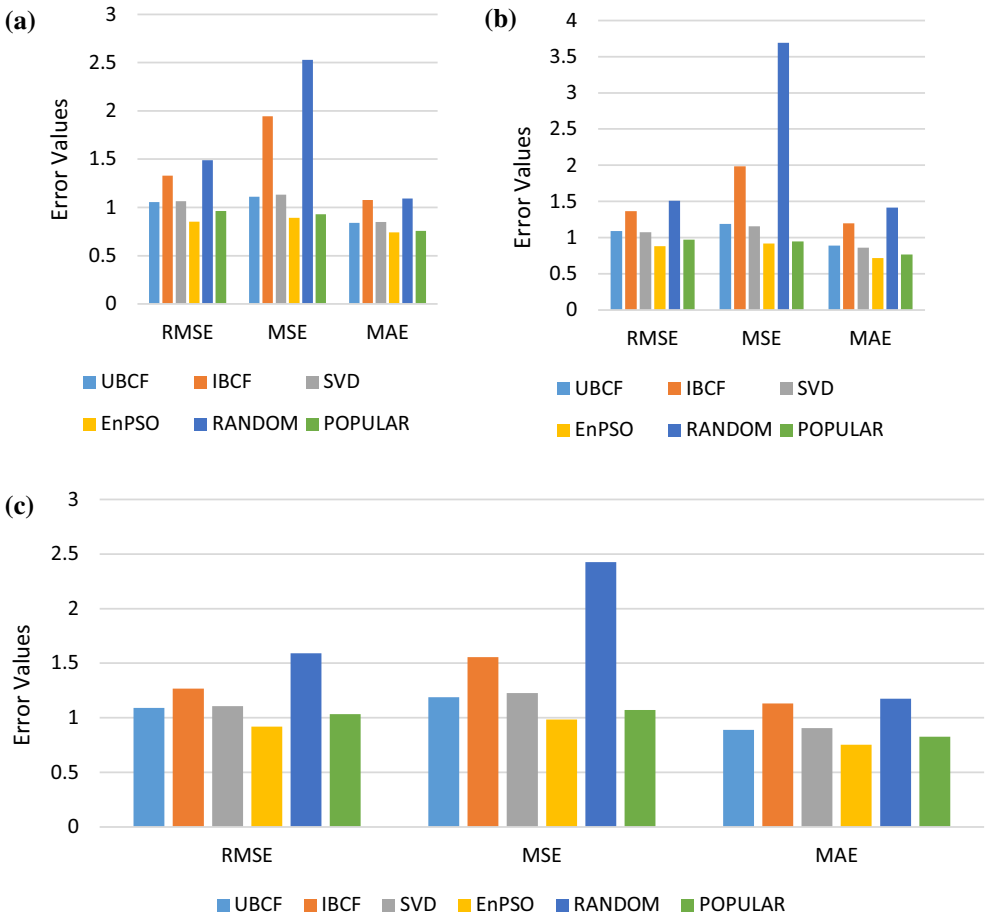


Fig. 7 **a** Error analysis of 10 M dataset, **b** error analysis of 1 M dataset, **c** error analysis of 100 K dataset

7.2.2 Mean Square Error (MSE)

MSE is used to detect the quality of an estimator by calculating the mean of its squared error. It is calculated by Eq. 17:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i)^2 \tag{17}$$

7.2.3 Mean Absolute Error (MAE)

MAE is the average of all the absolute errors. It is given by Eq. 18:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{18}$$

7.3 Results

The final results after execution have been shown in this section. Table 3 depicts the comparative analysis of error metrics between the other base recommendation techniques and EnPSO on the 100 K MovieLens dataset. Table 4 depicts error metrics on the 1 M MovieLens dataset, and similarly, Table 5 depicts error metrics on the 10 M MovieLens dataset. The results of the 1 M MovieLens dataset have been shown graphically using bar charts in Fig. 7a. Similarly, the results of the 10 M MovieLens dataset are shown in Fig. 7b, and the results of the 100 K MovieLens dataset are shown in Fig. 7c.

It can be seen in Table 3, 4, and 5 that the MSE is greater than RMSE and MAE. Furthermore, for all the recommendation techniques, our proposed algorithm EnPSO has the least error metrics. The highest error values are of Random and IBCF recommendation techniques.

It is evident from the results that our proposed algorithm makes the minimum error compared to the rest of the machine learning techniques. In this paper, we have calculated the error metrics, i.e., RMSE, MSE, and MAE, for 100 K, 1 M, and 100 M datasets. However, for other metrics, we have analyzed the results on 1 M and 10 M datasets.

It can be seen from the results that our model proved to be better than all the other techniques. The error metrics of

Table 6 Error metrics for level 1 ensemble for 100 K dataset

	IBCF + SVD	IBCF + RANDOM	IBCF + UBCF + POPULAR	IBCF + POPULAR	EnPSO
RMSE	0.9619	0.9384	0.9327	0.9214	0.9183
MSE	0.9611	0.9806	0.9840	0.9856	0.9837
MAE	0.9649	0.9328	0.8664	0.7635	0.7529

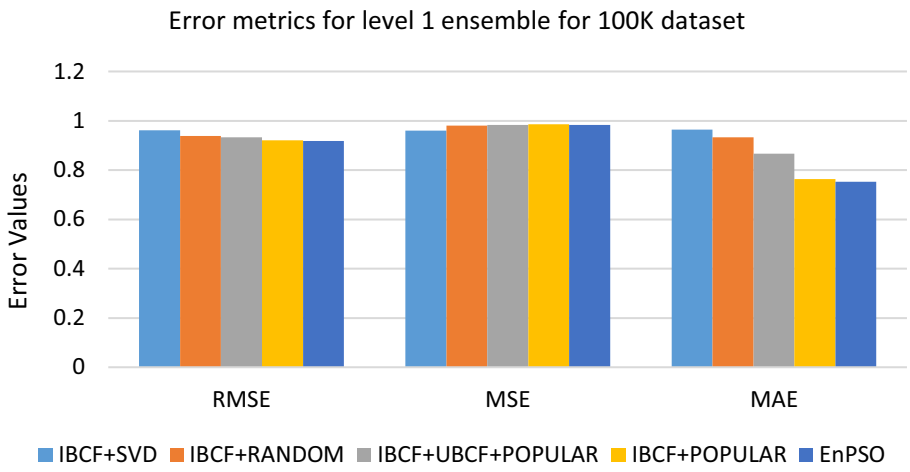


Fig. 8 Graphical representation of error metrics for level 1 ensemble for 100 K dataset

our technique were lower than the error metrics of every other technique. We used three error metrics, namely, Root-Mean-Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). For the 100 K MovieLens dataset, the RMSE value of the proposed algorithm was 0.91, the value of MSE was 0.98, and the value of MAE was 0.75. These error values were the least among all the other techniques.

In the 1 M MovieLens dataset, the value of RMSE was 0.88, the value of MSE was 0.91, and the value of MAE was 0.71. These error values were also lesser than the error values of other techniques. For the 10 M MovieLens dataset, the RMSE value of our proposed algorithm was 0.85, the error value of MSE was 0.89, and the error value of MAE was 0.74. These error values of our proposed model are also lesser than the error values of all other techniques.

To verify the results, we created 1 and 2—level ensembles by combining any two or more recommenders and calculating their error metrics. <RMSE, MSE, MAE>. Among these the 4 best performing ensembles were: IBCF + SVD, IBCF + RANDOM, IBCF + UBCF + POPULAR, and IBCF + POPULAR for 100 K MovieLens dataset.

The values are given in Table 6, and the graphical representation is presented in Fig. 8. Whenever ensembles are created, an approach of combining all possible algorithms is followed, and afterward, the ensembles generating the best results are shortlisted. In this paper, we have automated this process of ensemble selection using Particle Swarm Optimization. To confirm whether our automated system EnPSO performs better than the conventional method or not, we manually computed the error metrics of all the possible ensembles and shortlisted four best-performing ensembles. As seen from Table 6, the error metrics of all the 1-level ensembles are more than the error metrics of the multi-level ensemble, i.e., EnPSO. Since the other techniques yielded higher errors while computing the results, we can safely say that our proposed model provides us with the most optimum results.

There is a limitation to our work. Being a proof of concept, it does not quantitatively inform us about the possible performance gains in different domains and using different base recommenders. Also, since we are using Particle Swarm Optimization for optimizing the ensemble, our algorithm EnPSO may not always be able to find a well-performing ensemble providing significant performance gains within the time constraints or stopping conditions for some cases.

8 Conclusion and Future Work

Our results have shown that EnPSO can effectively combine the base techniques automatically to create an ensemble model that can outperform the base models. To carry out this task, our technique uses a hierarchical model to create ensembles and Particle Swarm Optimization to efficiently find the optimal model in the search space. The optimized model not only performs better than the base models but is also free from the individual biases and shortcomings of

the techniques. Through our approach, someone with little or no knowledge of ensemble models can generate well-performing optimized ensembles from the base methods. Our work is a proof of concept which can be extended to different domains, datasets, and base techniques. We used five base recommendation techniques for our experiments. The experiments were performed on the MovieLens data, a standard benchmark dataset in the field of recommender systems [70]. As for the scope of future work, our system functions as a proof of concept where we have to show the use of AutoML techniques to create recommender systems. Results can be improved by using more advanced models for the ensemble with hyper-parameter optimization. Also, more complex evolutionary strategies can be considered for finding the best models. Apart from working on model generation, research on AutoML can also be done in *data preparation, feature engineering,* and *model evaluation.* Thus, creating a complete platform to make recommender systems very trivial to implement for the user by encapsulating all the complexities within an AutoML framework.

References

1. Liang, H.; Tao, X.; Xu, Y.; Nayak, R.; Li, Y.: Connecting users and items with weighted tags for personalized item recommendations. 51 (2010). https://doi.org/10.1145/1810617.1810628
2. Jamali, M.; Ester, M.: *TrustWalker*: a random walk model for combining trust-based and item-based recommendation. In: KDD '09 Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 397–406 (2009).
3. Yuan, J.; Shalaby, W.; Korayem, M.; Lin, D.; AlJadda, K.; Luo, J.: Solving cold-start problem in large-scale recommendation engines: {A} deep learning approach. CoRR. abs/1611.0, 1901–1910 (2016)
4. Nilashi, M.; Ibrahim, O.; Bagherifard, K.: A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. Expert Syst. Appl. **92**, 507–520 (2018). https://doi.org/10.1016/j.eswa.2017.09.058
5. Zoph, B.; Le, Q.V.: Neural architecture search with reinforcement learning. Cvpr **2019**, 1–11 (2019)
6. Yao, Q.; Wang, M.; Chen, Y.; Dai, W.; Li, Y.F.; Tu, W.W.; Yang, Q.; Yu, Y.: Taking human out of learning applications: a survey on automated machine learning (2018). arXiv:1810.13306
7. Zöller, M.A.; Huber, M.F.: Benchmark and survey of automated machine learning frameworks. J. Artif. Intell. Res. **70**, 409–472 (2021)
8. He, X.; Zhao, K.; Chu, X.: AutoML: a survey of the state-of-the-art. Knowl.-Based Syst. **212**, 106622 (2021)
9. Lili, C.: Recommender algorithms based on boosting ensemble learning. Int. J. Smart Sens. Intell. Syst. **8**, 368–386 (2015). https://doi.org/10.21307/ijssis-2017-763
10. Yu, K.; Schwaighofer, A.; Tresp, V.; Ma, W.; Zhang, H.: Collaborative ensemble learning: combining collaborative and content-based information filtering via hierarchical bayes. In: UAI'03 Proceedings of the Ninth Conference of Uncertainty in Artificial Intelligence, pp. 353–360 (2003).
11. Schclar, A.; Tsikinovsky, A.; Rokach, L.; Meisels, A.; Antwarg, L.: Ensemble methods for improving the performance of neighborhood-based collaborative filtering. In: RecSys'09—Proceedings of the 3rd ACM Conference on Recommender Systems, pp. 261–264 (2009).
12. Fortes, A.; Manzato, M.: Ensemble learning in recommender systems: combining multiple user interactions for ranking personalization. In: WebMedia 2014—Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, pp. 47–54 (2014). https://doi.org/10.1145/2664551.2664556
13. Hao, Y.; Zhang, P.; Zhang, F.: Multiview ensemble method for detecting shilling attacks in collaborative recommender systems. Commun. Netw. Secur. (2018) https://doi.org/10.1155/2018/8174603
14. Srikanth, T.; Shashi, M.: A scalable ensemble architecture for collaborative filtering in recommender systems. Int. J. Appl. Eng. Res. **11**, 5103–5109 (2016)
15. Ayaki, T.; Yanagimoto, H.; Yoshioka, M.: Recommendation from access logs with ensemble learning. Artif. Life Robot. **22**, 163–167 (2017). https://doi.org/10.1007/s10015-016-0346-x
16. Da Costa, A.F.; Manzato, M.G.: Exploiting multimodal interactions in recommender systems with ensemble algorithms. Inf. Syst. **56**, 120–132 (2016). https://doi.org/10.1016/j.is.2015.09.007
17. Tiemann, M.; Pauws, S.; Vignoli, F.: Ensemble learning for hybrid music recommendation. In: Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, pp. 179–180 (2007)
18. Wu, M.: Collaborative filtering via ensembles of matrix factorizations. In: KDD Cup and Workshop 2007, pp. 43–47 (2007)
19. TahsinÖztürk, H.; Dede, T.; Türker, E.: Optimum design of reinforced concrete counterfort retaining walls using TLBO. Jaya Algor. Struct. **25**, 285–296 (2020). https://doi.org/10.1016/j.istruc.2020.03.020
20. Atmaca, B.; Dede, T.; Grzywinski, M.: Optimization of cables size and prestressing force for a single pylon cable-stayed bridge with Jaya algorithm. Steel Compos. Struct. **34**, 853–862 (2020). https://doi.org/10.12989/scs.2020.34.6.853
21. Dede, T.; Grzywinski, M.; Selejdak, J.: Continuous size optimization of large-scale dome structures with dynamic constraints. Struct. Eng. Mech. **73**, 397–405 (2020). https://doi.org/10.12989/sem.2020.73.4.397
22. Kalemci, E.N.; İkizler, S.B.; Dede, T.; Angın, Z.: Design of reinforced concrete cantilever retaining wall using Grey wolf optimization algorithm. Structures **23**, 245–253 (2020). https://doi.org/10.1016/j.istruc.2019.09.013
23. Sharafati, A.; Asadollah, S.B.H.S.; Hosseinzadeh, M.: The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. Process. Saf. Environ. Prot. **140**, 68–78 (2020). https://doi.org/10.1016/j.psep.2020.04.045
24. Tahir, M.F.; Haoyong, C.; Mehmood, K.; Larik, N.A.; Khan, A.; Javed, M.S.: Short term load forecasting using bootstrap aggregating based ensemble artificial neural network. Recent Adv. Electr. Electron. Eng. (Formerly Recent Patents Electr Electron. Eng. **13**, 980–992 (2019). https://doi.org/10.2174/2213111607666191111095329
25. Shamshirband, S.; JafariNodoushan, E.; Adolf, J.E.; Abdul Manaf, A.; Mosavi, A.; Chau, K.: Wing: ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. Eng. Appl. Comput. Fluid Mech. **13**, 91–101 (2019). https://doi.org/10.1080/19942060.2018.1553742
26. Alizadeh, M.J.; JafariNodoushan, E.; Kalarestaghi, N.; Chau, K.W.: Toward multi-day-ahead forecasting of suspended sediment concentration using ensemble models. Environ. Sci. Pollut. Res. **24**, 28017–28025 (2017). https://doi.org/10.1007/s11356-017-0405-4
27. Homsi, R.; Shiru, M.S.; Shahid, S.; Ismail, T.; Harun, S.B.; Al-Ansari, N.; Chau, K.W.; Yaseen, Z.M.: Precipitation projection

using a CMIP5 GCM ensemble model: a regional investigation of Syria. Eng. Appl. Comput. Fluid Mech. **14**, 90–106 (2020). https://doi.org/10.1080/19942060.2019.1683076
28. Wu, C.L.; Chau, K.W.: Prediction of rainfall time series using modular soft computingmethods. Eng. Appl. Artif. Intell. **26**, 997–1007 (2013). https://doi.org/10.1016/j.engappai.2012.05.023
29. Lee, Y.; Kim, K.-J.: Product recommender systems using multi-model ensemble techniques. J. Intell. Inf. Syst. **19**, 39–54 (2013). https://doi.org/10.13088/jiis.2013.19.2.039
30. Lommatzsch, A.: Real-time news recommendation using context-aware ensembles. Lecture Notes Computer Science (Including Subseries Lecture Notes Artificial Intelligence. Lecture Notes Bioinformatics). 8416 LNCS, pp. 51–62 (2014). https://doi.org/https://doi.org/10.1007/978-3-319-06028-6_5
31. Nilashi, M.; Bagherifard, K.; Rahmani, M.; Rafe, V.: A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. Comput. Ind. Eng. **109**, 357–368 (2017). https://doi.org/10.1016/j.cie.2017.05.016
32. Tsai, C.F.; Hung, C.: Cluster ensembles in collaborative filtering recommendation. Appl. Soft Comput. J. **12**, 1417–1425 (2012). https://doi.org/10.1016/j.asoc.2011.11.016
33. Zhang, W.; Zou, H.; Luo, L.; Liu, Q.; Wu, W.; Xiao, W.: Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing **173**, 979–987 (2016). https://doi.org/10.1016/j.neucom.2015.08.054
34. Logesh, R.; Subramaniyaswamy, V.; Malathi, D.; Sivaramakrishnan, N.; Vijayakumar, V.: Enhancing recommendation stability of collaborative filtering recommender system through bio-inspired clustering ensemble method. Neural Comput. Appl. (2018). https://doi.org/10.1007/s00521-018-3891-5
35. Manzato, M.G.; Domingues, M.A.; Fortes, A.C.; Sundermann, C.V.; D'Addio, R.M.; Conrado, M.S.; Rezende, S.O.; Pimentel, M.G.C.: Mining unstructured content for recommender systems: an ensemble approach. Inf. Retr. J. **19**, 378–415 (2016). https://doi.org/10.1007/s10791-016-9280-8
36. Vinagre, J.; Jorge, A.M.; Gama, J.: Online bagging for recommender systems. Expert Syst. **35**, 1–13 (2018). https://doi.org/10.1111/exsy.12303
37. Wu, H.; Yue, K.; Pei, Y.; Li, B.; Zhao, Y.; Dong, F.: Collaborative topic regression with social trust ensemble for recommendation in social media systems. Knowl. Based Syst. **97**, 111–122 (2016). https://doi.org/10.1016/j.knosys.2016.01.011
38. Zheng, L.; Li, L.; Hong, W.; Li, T.: PENETRATE: personalized news recommendation using ensemble hierarchical clustering. Expert Syst. Appl. **40**, 2127–2136 (2013). https://doi.org/10.1016/j.eswa.2012.10.029
39. Gomes, H.M.; Barddal, J.P.; Enembreck, F.; Bifet, A.: A survey on ensemble learning for data stream classification. ACM Comput. Surv. **50**, 1–36 (2017). https://doi.org/10.1145/3054925
40. Bauman, K.; Liu, B.; Tuzhilin, A.: Aspect based recommendations. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discover Data Mining—KDD '17, pp. 717–725 (2017). https://doi.org/10.1145/3097983.3098170
41. Kant, S.; Mahara, T.: Merging user and item based collaborative filtering to alleviate data sparsity. Int. J. Syst. Assur. Eng. Manag. **9**, 173–179 (2018). https://doi.org/10.1007/s13198-016-0500-9
42. Paper, C.: Cosine similarity metric learning for face for face verication. Accv (2015). https://doi.org/10.1007/978-3-642-19309-5
43. Shmueli, E.; Tassa, T.: Secure multi-party protocols for item-based collaborative filtering. In: Proceedings of the Eleventh ACM Conference on Recommender Systems—RecSys '17, pp. 89–97 (2017). https://doi.org/10.1145/3109859.3109881
44. Wang, J.; De Vries, A.P.; Reinders, M.J.T.: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and development in information 2006, pp. 501–508 (2006). https://doi.org/10.1145/1148170.1148257
45. Vozalis, M.G.; Margaritis, K.G.: Applying SVD on generalized item-based filtering. Int. J. Comput. Sci. Appl. Technomath. Res. Found. **3**, 27–51 (2006)
46. Marlin, B.M.: Modeling user rating profiles for collaborative filtering. In: Advances in Neural Information Processing Systems, vol. 16, pp. 627–634 (2003)
47. Praveen Kumar, D.; Amgoth, T.; Annavarapu, C.S.R.: Machine learning algorithms for wireless sensor networks: a survey. Inf. Fusion. **49**, 1–25 (2019). https://doi.org/10.1016/j.inffus.2018.09.013
48. Lopes, M.E.: Estimating the algorithmic variance of randomized ensembles via the bootstrap. Ann. Stat. **47**, 1088–1112 (2019). https://doi.org/10.1214/18-AOS1707
49. Bühlmann, P.: Bagging, boosting and ensemble methods. In: Handbook of Computational Statistics, pp. 985–1022. Springer, Berlin, Heidelberg (2012)
50. Al-Stouhi, S.; Reddy, C.K.: Adaptive boosting for transfer learning using dynamic updates. Lecture Notes Computer Science (including Subseries Lecture Notes Artificial Intelligence Lecture Notes Bioinformatics). 6911 LNAI, pp. 60–75 (2011). https://doi.org/10.1007/978-3-642-23780-5_14
51. Beygelzimer, A.; Hazan, E.; Kale, S.; Luo, H.: Online gradient boosting. Adv. Neural Inf. Process. Syst. **2015**, 2458–2466 (2015)
52. Dembczyñski, K.; Cheng, W.; Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML 2010—Proceedings, 27th International Conference Machine Learning, pp. 279–286 (2010)
53. Zhang, Y.; Wang, S.; Ji, G.: A comprehensive survey on particle swarm optimization algorithm and its applications. Math. Probl. Eng. (2015). https://doi.org/10.1155/2015/931256
54. Vikhar, P.A.: Evolutionary algorithms : a critical review and its future prospects, pp. 261–265 (2016). https://doi.org/https://doi.org/10.1109/ICGTSPICC.2016.7955308
55. Maulik, U.; Bandyopadhyay, S.: Genetic algorithm-based clustering technique. Pattern Recognit. **33**, 1455–1465 (2000). https://doi.org/10.1016/S0031-3203(99)00137-5
56. Dorigo, M.; Stützle, T.: The Ant colony optimization metaheuristic: algorithms, applications, and advances. Handb. Metaheur. (2006). https://doi.org/10.1007/0-306-48056-5_9
57. Karaboga, D.; Akay, B.: A comparative study of artificial bee colony algorithm. Appl. Math. Comput. **214**, 108–132 (2009). https://doi.org/10.1016/j.amc.2009.03.090
58. Shehab, M.; Khader, A.T.; Al-betar, M.A.: A survey on applications and variants of the cuckoo search algorithm. Appl. Soft Comput. J. (2017). https://doi.org/10.1016/j.asoc.2017.02.034
59. Wang, H.; Wang, W.; Zhou, X.; Sun, H.; Zhao, J.; Yu, X.; Cui, Z.: PT US CR. Inf. Sci. (Ny). (2016). https://doi.org/10.1016/j.ins.2016.12.024
60. Oftadeh, R.; Mahjoob, M.J.; Shariatpanahi, M.: A novel meta-heuristic optimization algorithm inspired by group hunting of animals: hunting search. Comput. Math. Appl. **60**, 2087–2098 (2010). https://doi.org/10.1016/j.camwa.2010.07.049
61. Jin, Y.; Zhang, Q.; Zhou, A.; Sendhoff, B.; Tsang, E.: Prediction-based population Re-initialization for evolutionary dynamic multi-objective optimization. Evol. Multi Criterion Optim. (2007). https://doi.org/10.1007/978-3-540-70928-2_62
62. Büche, D.; Schraudolph, N.N.; Koumoutsakos, P.: Accelerating evolutionary algorithms with Gaussian process fitness function models. IEEE Trans Syst. Man Cybern. Part C Appl. Rev. **35**, 183–194 (2005). https://doi.org/10.1109/TSMCC.2004.841917
63. Tsutsui, S.; Yamamura, M.; Higuchi, T.: Multi-parent recombination with simplex crossover in real coded genetic algorithms. Proc. Genet. Evol. Comput. Conf. **1**, 657–664 (1999)

64. Corus, D.; Oliveto, P.S.: Standard steady state genetic algorithms can Hillclimb faster than mutation-only evolutionary algorithms. IEEE Trans. Evol. Comput. **22**, 720–732 (2018). https://doi.org/10.1109/TEVC.2017.2745715
65. Wang, D.; Tan, D.; Liu, L.: Particle swarm optimization algorithm: an overview. Soft Comput. **22**, 387–408 (2018). https://doi.org/10.1007/s00500-016-2474-6
66. Marini, F.; Walczak, B.: Particle swarm optimization (PSO). A tutorial. Chemom. Intell. Lab. Syst. **149**, 153–165 (2015). https://doi.org/10.1016/j.chemolab.2015.08.020
67. Cazzaniga, P.; Nobile, M.S.; Besozzi, D.: The impact of particles initialization in PSO: parameter estimation as a case in point. In: 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology CIBCB (2015). https://doi.org/10.1109/CIBCB.2015.7300288
68. Hahsler, M.: recommenderlab: a framework for developing and testing recommendation algorithms (2015). http://elib.ict.nsc.ru/jspui/bitstream/ICT/1861/1/recommenderlab.pdf
69. Iqbal, N.; Zerguine, A.; Al-Dhahir, N.: Decision feedback equalization using particle swarm optimization. Signal Process. **108**, 1–12 (2015). https://doi.org/10.1016/j.sigpro.2014.07.030
70. Gupta, G.; Katarya, R.: Research on understanding the effect of deep learning on user preferences. Arab. J. Sci. Eng. (2020). https://doi.org/10.1007/s13369-020-05112-2

Environment, Development and Sustainability
https://doi.org/10.1007/s10668-021-01353-9

Evaluation of urban transport-environment sustainable indicators during Odd–Even scheme in India

Ravindra Kumar[1] · Rajeev Kumar Mishra[2] · Satish Chandra[1] · Asif Hussain[2]

Received: 19 March 2020 / Accepted: 22 March 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Identification of the most relevant parameters of a sustainable transportation system is regarded as a significant choice that influences the success of policy-related intervention to improve efficiency. A hybrid-based approach has been used to analyze the data in multiple steps to improve accuracy and objectivity. In the first step, evidence on social, economic, and environmental indicators was formed for each opinion and then combined. The second step proposed a mathematical framework to treat indecision on the indicators collected from multiple sources, viz. expert and public opinion survey, models, and actual measurement during pre- and post-implementation stages of the Odd–Even scheme. By setting up, a reliability function and consistency index are used to evaluate and rank the Odd–Even transportation scheme introduced by the Government of Delhi. A case study in South Delhi's surrounding areas has been examined to demonstrate the proposed hybrid approach and evaluate the parameters in terms of the index. The result shows that the indicators like trip cost, trip time, safety and security, user's satisfaction, congestion level, parking demand, and fuel consumption affected positively by 1766.6, 22.9, 283.3, 19.1, 187.5, 200, and 17.2 percent, respectively, after the implementation of the scheme. The overall increase in the index was found as 28.2%, which favors the "Odd–Even" scheme as a sustainable transportation measure in Delhi.

Keywords Analytical hierarchy process · D–S theory · Odd–Even scheme · Sustainability index

✉ Rajeev Kumar Mishra
rajeevctrans@gmail.com

Ravindra Kumar
ravindra261274@gmail.com

Satish Chandra
satisfce@gmail.com

Asif Hussain
asifhussain009@gmail.com

[1] Central Road Research Institute, New Delhi 110025, India

[2] Department of Environmental Engineering, Delhi Technological University, Delhi 110042, India

1 Introduction

Presently, Delhi has the dubious distinction of being one of the most polluted cities in the world. More than 10 million vehicles are plying on the roads of Delhi (Economic Survey of Delhi, 2018–2019). Vehicular pollution contributes 22% of total emissions to the environment. Air pollution has now become a big environmental evil in mega-cities and Delhi is one of them, where more than 16 million people are exposed to severe high pollutant concentrations (WHO, 2016). Environmental Pollution-Prevention and Control Agency (EPCA), New Delhi, found that many vehicles are plying on Delhi road without mandatory 'Pollution Under Control (PUC)' certificates. This results in potentially imperiling the health of millions of people and causes various problems like air and noise pollution, traffic congestion, security and safety measures, higher living costs, longer duration of journey time. The reasons behind such environmental issues are poor urban transport management strategies, lack of eco-friendly policies, and lack of implementation of sustainable indicators that can control vehicular emission with the economy and social context of transportation measures. Sustainable transport strategies and infrastructure play a vital role in climate change, immediate health benefits to the majority of the world's population, large equity benefits for vulnerable groups, and far-reaching reductions in traffic-related health risks due to air and noise pollution as well as injuries (Hosking et al., 2011). To overcome the severe problem of air pollution and traffic congestion, the Government of Delhi introduced an Odd–Even vehicle rationing scheme (Jiang et al., 2015; Michael et al., 2017; Singhania et al., 2016) in the year 2016. During the implementation of this scheme, private cars with odd/even number plates of registration were allowed to ply in the city on an alternate day. However, the principal objective of this scheme was to reduce air pollution and traffic congestion.

Nevertheless, the fundamental right of access of persons and communities should meet safety, steady flow, and environmental friendly for both in a sustainable way (CST, 2005). Those parameters are addressed by developing Transport Sustainability Index (TSI) and evaluated during and after implementation of the Odd–Even scheme. During this study, three major arterial roads, namely Captain Gaur Marg (From Lajpat Nagar Xing to Modi Mill), Outer Ring Road (Modi Mill to Nehru Place Flyover), and Lala Lajpat Rai Marg (Nehru Place Flyover to Moolchand Flyover), were selected to collect the data related to evaluation in terms of sustainability which also encircle the busiest South Delhi central district Nehru Place. The objective of this paper is to evaluate the transportation sustainability for the transportation system during the implementation period of the "Odd–Even scheme" along with its impact on the sustainability of the transportation system. Many researchers have developed the sustainability indicator and index using an independent approach that requires only one or a few data sets. Sustainability index development of transport cannot be limited to one data source (Buzási & Csete, 2015; Haghshenas & Vaziri, 2012) neither cover the holistic view of the index. To assess the sustainability of the implemented Odd–Even scheme in Delhi, a hybrid approach, based on multi-criteria decision-making approach as analytical hierarchy process (AHP) (Thanki et al., 2016; Yang et al., 2016) and Dempster–Shafer (DS) theory of evidence was adopted (Dempster, 1968; Deng et al., 2015; Jiang et al., 2015; Kang et al., 2019; Shafer, 1976; Waheed et al., 2011). It comprises several steps, but in the foremost stage, the criteria for evaluating transportation sustainability measures (Odd–Even scheme) were identified, whereas AHP was used to organize and rank the parameters by their weightage.

Further, the basic probability function to manage uncertainty was used to assess the change due to the implementation of the scheme (Miller et al., 2017). The data were collected from various domains like experts, public opinion surveys, sensors/actual measurements, models, etc. The acquired data were pooled and fused using the D–S theory, and the selected criteria for the evaluation were tested. After that, the sustainability of the Odd–Even scheme was estimated for indicative purposes. As it is necessary to build a composite index to cover the environmental, economic, and social dimensions (Haghshenas and Vaziri, 2012) at city level application, these indicators were used to build a composite sustainable transportation index (CSTI) to determine few implementable effective sustainability factors to maintain an ecological balance of the city (Zito & Salvo, 2011).

2 Literature review

All policy development and transport-related demand need to address several indicators related to very concise information systems quantitatively and qualitatively. The fundamental aim is to understand their trends occur in specified contexts (Bell and Morse, 2008).

2.1 Sustainability indicators

Hak et al., (2007) reported the sustainability indicators for a wide range of economic, social, or environmental sources. They also set up policy analysis procedure in five stages, i.e., (i) Clarifying goals, (ii) Describing trends, (iii) Analyzing conditions, (iv) Projecting developments, and (v) Inventing, evaluating, based on this analysis, the selecting alternatives was easier and comfortable to interpret (Clark, 2002). Several set indicators are available for developing Sustainability Indices, the most prominent ones being given by Propolis (2014) which is based on 35 parameters, other by Dobranskyte-Niskota et al. (2009) and Konsult (2008). Table 1 shows the list of sustainable indicators used in various studies.

2.2 Assessment methodologies

Sustainable public transport infrastructure was the core focus area seeking equilibrium among environment care, economic development, and social justice (Yang et al., 2016). The identification, selection, assessment, and adoption of such transportation policy measures are very comprehensive and time-consuming and usually involve several stages. Verma et al. (2015) addressed the necessities of a sustainable transportation system for the Indian cities and summarized the research issues concerning planning/modeling, non-motorized vehicles, public transport, driver's behavior, road safety, and traffic management. They also proposed a model to assess the impact of various transportation policies and projects which was based on variation in three pillars of sustainability, i.e., environmental, economic, and social. Mohan and Tiwari (2000) suggested the use of buses and a non-motorized transport system for India. However, these were not identified as generic indicators that can affect the sustainability of urban public transport.

Wellar (2009) presented 42 techniques to demonstrate a city's transportation sustainability, including life cycle analysis, cost–benefit analysis, impact assessment, multi-criteria decision-making (MCDM) approach, and Delphi techniques surveys and indexing. Ramani et al. (2011), Yedla and Shrestha (2012), Reisi et al. (2014), and Attiyah et al. (2016) developed performance-based approaches to evaluate sustainable transportation

Table 1 List of sustainable indicators used in various studies

Sustainability dimensions	Indicators			
	Propolis (2014)	Dobranskyte-Niskota et al. (2009)	Konsult (2008)	Reisi et al. (2014)
Environmental	GHGs from transport	VOCs emissions	Vibration	Depletion of non-renewable resources
	Acidifying gases	PM10 and PM2.5 emissions	Level of air quality pollutants	GHG emissions
	Land coverage	Fuel consumption	Visual intrusion	Air pollutants
	Volatile organic compounds from transport	Polluting accidents	Townscape quality	Land consumption for transport
	Quality of open space	Land taken by transport	CO2 emissions	
	Consumption of mineral oil	Age of vehicle fleets	Fuel consumption	
	Fragmentation of open space		Severance	
Social	Exposure to PM	Infrastructure	Safety and security	Accessibility
	Exposure to NO2	Accessibility and Mobility	Equity	Fatalities and injuries related to traffic accidents
	Exposure to traffic noise	Risks and Safety	Accessibility	Mortality effect of air pollutants
	Traffic deaths	Health Impacts	Accessibility	Mortality effect of air pollutants
	Traffic injuries	Affordability	Economic efficiency	Car ownership cost
	Housing standard	Employment		
	Investment cost	Transport demand and intensity		
Economic	Users benefit, investment cost	Transport cost and prices, transport demand and intensity	Economic regeneration, economic efficiency	Vehicle and general cost of accidents, car ownership cost
	Government benefits, users benefits	Occupancy of transportation, transport cost and prices	Economic regeneration	Vehicle and general cost of accidents
	Accident cost, government benefits	Technology status, occupancy of transportation	Economic regeneration	Vehicle and general cost of accidents
	Emission cost, accident cost	Technology status		
	GHG cost, emission cost	Technology status		
	Noise cost, GHG cost			
	Noise cost			

systems as an index using a multi-criteria decision approach in Texas, Delhi, Melbourne, and Tabuk city, respectively. Nijenhuis (2012) used the AHP and ELASTIC framework to set the guidelines for evaluating the sustainability of the transportation system in the context of medium-sized cities in Indonesia. A mixed approach based on AHP and DST was applied by Awasthi and Chauhan (2011) that reduced uncertainty, fuzziness, and incompleteness for assessing transportation sustainability (Ercan, 2009; Rodrigues et al., 2010). Sustainability performance and epistemic uncertainties in its environmental performance evaluation were considered by Shokravi and Kurnia (2014). Epistemic uncertainty, fuzziness, and incompleteness of data cannot be well catered by using MCDM approach like AHP, Fuzzy logic, etc., alone, but it can be well treated for further analysis with the help of a hybrid combination of MCDM approach and belief theory (Omrani et al., 2009). Zhou et al. (2019) developed the SSC-VIKOR approach to prioritize vehicle brand candidates from a big data analytical (BDA) viewpoint. Zhou et al. (2018) developed a novel integrated MCDM "fuzzy DEMATEL-AEW-FVIKOR" approach to choose the best recycling partner for small and medium enterprises. A systematic literature review on applications of MCDM methods on corporate sustainability by Chowdhury and Paul (2020) provides a useful directive on the integration of MCDM methodologies. An integrated MCDM approach was used by Yasmina et al. (2020) to explore BDA capabilities' interdependence and its impact on firm performance. A strong relationship of BDA capabilities was found to the operational performance of the firm.

Salavati et al. (2016) applied AHP and clustering approaches for public transportation decision making in Isfahan city. Bouraima et al. (2020) combined the SWOT matrix and AHP to evaluate the priority factors and employ them to develop the railway transportation system's strategies. Ma et al. (2017) evaluated urban green transportation planning based on central point triangle whiten weight function and entropy-AHP. Fuzzy AHP-based performance evaluation of massive online courses provider for online learners was done by Chandna et al. (2021). Pham et al. (2021) generated the hazard, exposure, and vulnerability maps using AHP combined with a deep neural network. Luís and Cabral (2021) used AHP and GIS tools for small dams/reservoirs site location analysis in a semi-arid region of Mozambique. Arantes et al. (2021) defined weights of the environmental attributes using AHP. Drivers, enablers, and barriers are prioritized using the AHP methodology (Mojumder & Singh, 2021) to adopt green supply chain management in India. Wanke et al. (2021) models include (a) the fuzzy Rasch model, which combines item response theory (IRT) and fuzzy set theory; (b) the Fuzzy AHP model; and (c) the crisp AHP model, based on eight different judgment scales concerning the relative evolution of each criterion for low-carbon operations in an emerging market, and Caner and Aydin (2017) selected shipyard site by raster calculation method and AHP in GIS environment for İskenderun, Turkey.

The above survey of the literature indicates that various efforts have been made in the area of sustainable transportation index development, but not a single generalized and detailed model using the fusion of data from different sources or tool is developed that can assess the Odd–Even scheme on suitable transportation characteristics of an Indian city. Therefore, the present work is focused on developing the TSI of driver behavior restraint policy measures. The hybrid approach of AHP (Analytic Hierarchy Process) and D–S (Dempster–Shafer) theory is used to determine the sustainability index of a transportation measure, i.e., Odd–Even scheme in Delhi, which can be further used as an effective tool to quantify the sustainability of the urban transportation system. The results of this sub-index have been compared internationally to review the policy for further improvement in sustainability and adopt a hybrid approach.

3 Methodology

In the present study, AHP and D–S theory are used to evaluate the transportation sustainability of the Odd–Even scheme in Delhi. This hybrid combination of AHP and D–S theory was adopted because of their ability to deal with multiple decision-makers, heterogeneous data types and ignorance and missing information which is very likely the case in realistic transportation situations. The location of the study area is shown in Fig. 1.

3.1 Analytic Hierarchy Process (AHP)

AHP is a multi-criteria decision-making technique proposed by Saaty (1980). The various steps of AHP are as follows.

1. Defining the problem and determining its goal.
2. Structuring the hierarchy from the top (the objective) through the intermediate level (criteria) to the lowest level (alternatives).
3. Constructing a set of pair-wise comparison (size n x n), for each of the lower levels for one matrix for each element in the level immediately above by using the relative scale measurement as shown in Table 2. The pair-wise comparison is made in terms of preferences of one over the other.
4. There are $n\ (n-1)/2$ judgments per matrix to develop a set of matrices in step 3. The reciprocals are automatically assigned in each pair-wise comparison.
5. Having made all pair-wise comparisons, the consistency is determined by using the eigenvalue λ_{max} to calculate the consistency index (CI) where $CI = (\lambda_{max} - n)/(n-1)$, where n is the matrix size. Judgment consistency can be checked by seeing the value of consistency ratio CR for the appropriate matrix value in the Table 3. If the $CR \leq 0.1$, the judgment matrix is acceptable; otherwise, it is considered inconsistent. To obtain a consistent matrix, the options should be reviewed and improved.

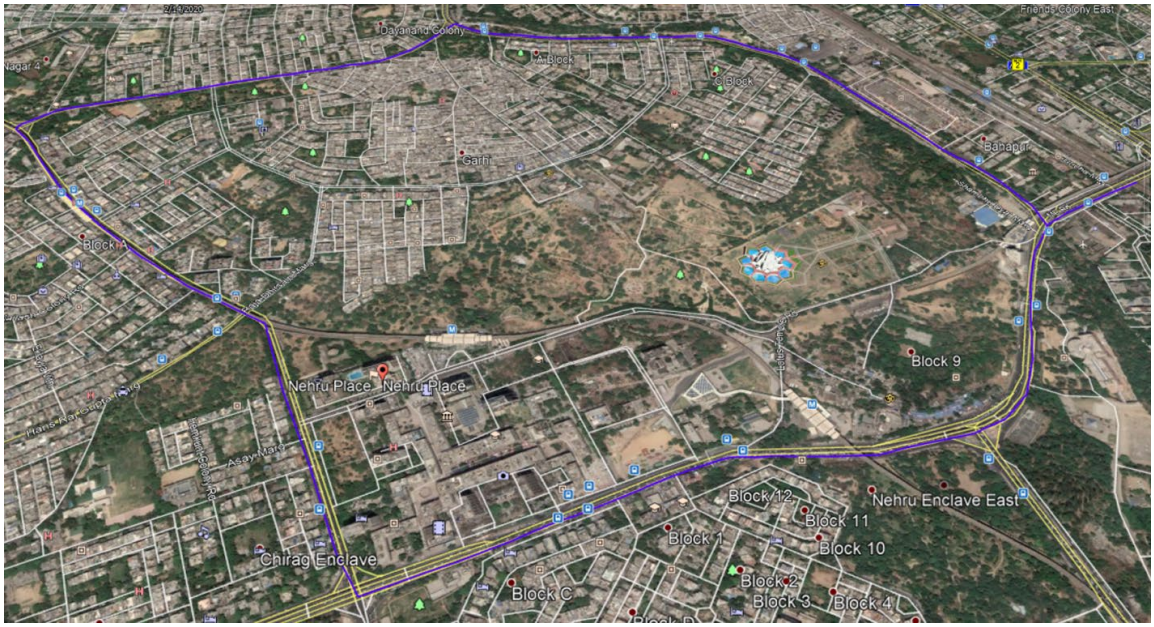


Fig. 1 Location of the study area (Source: Image@ 2020 Maxar technologies)

Table 2 Pair-wise comparison scale for AHP (Saaty 1990)

Numerical rating	Scale of importance
1	Equal importance
3	Moderate importance
5	Strong importance
7	Very strong importance
9	Extreme importance
2, 4, 6, 8	Intermediate values between the adjacent judgments
Reciprocals	When activity *i* compared to *j* is assigned one of the above numbers, activity *j* compared to *i* is assigned its reciprocal

Table 3 Average random consistency ratio

Size of matrix	1	2	3	4	5	6	7	8	9	10
Random consistency	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

6. Hierarchical synthesis is now used to weight the normalized eigenvector by the weights of criteria. Then, the sum is taken over all weighted eigenvector entries corresponding to those in the next lower hierarchy level. A nonzero vector "C" is called an eigenvector of any matrix say "A" if and only if there exists a number (real or complex) λ such that AC = λC. if such a number λ exists it is called the eigenvalue of "A" and the vector "C" is called eigenvector associated with eigenvalue λ.

The strength of AHP is that it allows the verification of the transitivity property of criteria weight that is if criterion "a" has a higher criteria weight than criterion "b" which has a higher weight than criteria "c," then criterion a will always have higher weight than criteria "c." This is the reason why it is chosen over other simple methods of weight allocation.

The AHP is used to assign weights to the chosen criteria for evaluating the Odd–Even scheme, whereas the D–S theory is used for data fusion or combining information from multiple sources (Calderwood et al., 2017; Huadong et al., 2002; Wang & Xiao, 2019).

3.2 Dempster–Shafer (D–S) theory

The Dempster–Shafer theory is related to Bayesian probability theory in the sense that both deal with subjective belief. It includes the Bayesian probability theory as a special case with later not being able to deal with ignorance. D–S theory has been widely applied to artificial intelligence, expert systems, pattern recognition, information fusion, risk assessment, and multiple attribute decision analysis, etc. The basic strengths of this approach lie in its ability to treat heterogeneous, uncertain, and incomplete data coming from multiple information sources. This approach is very useful for transport decision-making situations with limited, heterogeneous data.

Let (H) = {H1, H2,…HN) be a collectively exhaustive and mutually exclusive set of hypotheses or propositions, called the frame discernment. A basic probability assignment (BPA) is a function m: 2(H) → [0, 1], which is called mass function, satisfying

$$m(\phi) = 0$$

and

$$\sum_{A(H)} m(A) = 1 \tag{1}$$

where ϕ is an empty set A is any subset of (H) and 2(H) is the power set of (H) which consists of all the subsets of (H), i.e.,

$$2(H) = \{\phi, \{\mathrm{H1}\}, \ldots \{\mathrm{HN}\}, \{\mathrm{H1}, \mathrm{H2}\}, \{\mathrm{H1}, \mathrm{HN}\}, \ldots, (\mathrm{H})\} \tag{2}$$

The assigned probability (also called probability mass) m(A) measures the belief exactly assigned to A and represents how strong evidence supports A. All the assigned probabilities sum to unity, and there is no belief in empty set ϕ. The assigned probability to (H), i.e., m((H)), is called the degree of ignorance. Each subset A (H) such that m(A) > 0 is called a focal of m. All the related focal elements are collectively called the body of evidence.

Evidence associated with each BPA symbolizes the belief function (Bel), and the plausibility measure, PI which is both functions: 2(H) → [0,1] and given by Bel(A) = $\sum_{BA} m(B)$ and PI(A) = $\sum_{AB\neq\phi} m(B)$ where A and B are a subset of (H), Bel(A) represents the exact support to A, i.e., the belief of hypothesis A being true; PI(A) represents the possible support to A, i.e., total amount of belief that could be potentially placed in A, [Bel(A), PI(A)] constitute the interval of support to A and can be seen as the lower and upper bounds of the probability to which A is supported. The two functions are related to each other by PI(A) = 1 − Bel($\overline{A}$) where $\overline{A}$ denotes the complement of A. The difference between the belief and plausibility of set A describes the ignorance of the assessment for set A.

Since m(A), Bel(A), and PI(A) are in one-to-one correspondence, they can be seen as three facets of the same information. There are other several such as the commonality function and doubt function, which can be used to represent evidence. They all represent the same information and provide flexibility to match a variety of reasoning applications.

The evidence from different sources is combined using Dempster's rule of combination. The rule assumes that the information sources are independent and uses the orthogonal sum to combine multiple belief structures:

$$m = m1 \oplus m2 \oplus m3 \oplus \cdots \oplus mk \tag{3}$$

where ⊕ represents the operator of the combination, for two belief structures m1 and m2, and the Dempster's rule of combination is given by:

$$[m1 \oplus m2]C = \begin{cases} 0 & c = \emptyset \\ \frac{\sum_{A\cap B=C} m1(A)m2(B)}{1-\sum_{A\cap B=\emptyset} m1(A)m2(B)} & c \neq \emptyset \end{cases} \tag{4}$$

where A and B are focal elements and [$m1$⊕ $m2$] (C) is a BPA. The denominator $1-\sum_{AB\neq\phi} m1(A)m2(B)$ is called the normalization factor and $\sum_{AB\neq\phi} m1(A)m2(B)$ is denoted by k is called the degree of conflict, which measures the conflict between pieces of evidence. The larger the k, the more the sources are conflicting and lesser in a sense in their combination. If $k=0$, this shows complete compatibility, and if $0<k<1$, it shows partial

compatibility. Finally, the orthogonal sum does not exist when $k=1$. In this case, the sources are completely contradictory.

The Dempster's rule of combination is proved to be both commutative and associative, i.e., $m1 \oplus m2 = m2 \oplus m1$ commutatively and $(m1 \oplus m2) \oplus m3 = m1 \oplus (m2 \oplus m3)$ associatively. These two properties shows that evidence can be combined in any order. Therefore, in multiple belief structures, the combination of evidence can be carried out in a pair-wise way.

The reasons behind choosing the techniques mentioned above are their capability to deal with multiple decision-making and heterogeneous data types like perception, expert opinion, actual measurement/sensors, and models (Baloian et al., 2019; Beynon et al., 2001). The research methodology opted for this study is presented in Fig. 2.

3.3 Criteria selection

After going through relevant literature (Litman, 2019; Dobranskyte-Niskota et al., 2007; Rassafi & Vaziri, 2005; Gudmundsson, 2001, 2004), discussions with subject-related transportation and environmental experts, and considering the transport-related environmental problems in Delhi, 11 criteria are identified to evaluate the Odd–Even scheme (Table 4). Using AHPs pair-wise comparison methodology, weights of different selected criteria

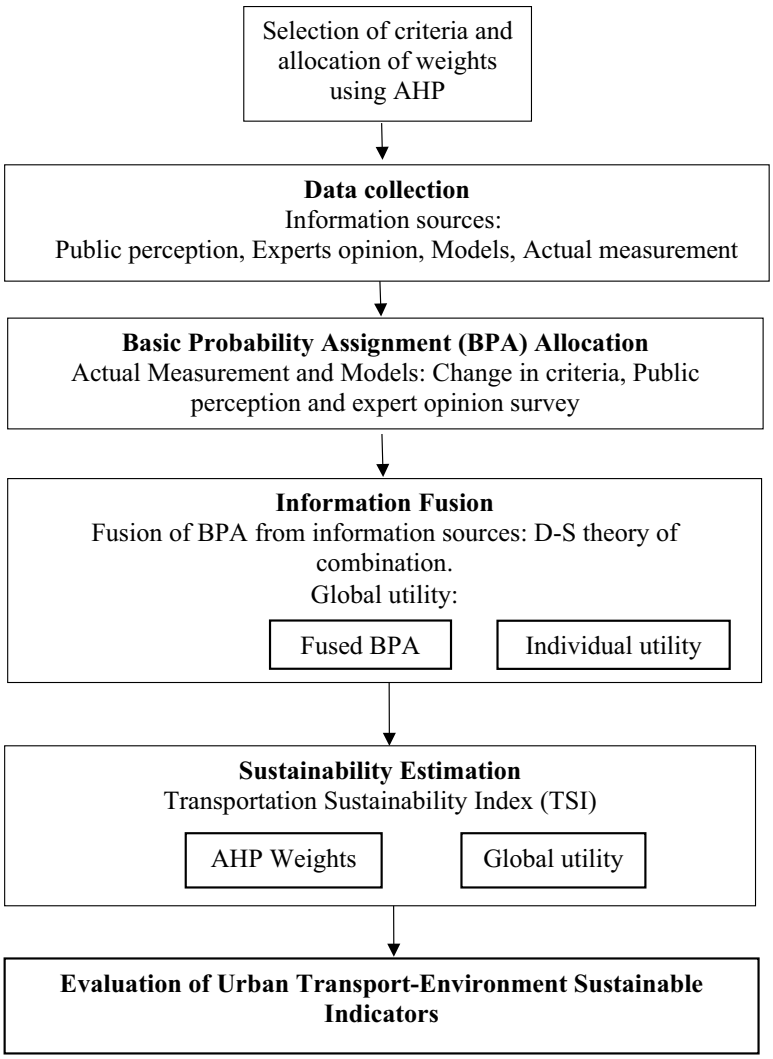


Fig. 2 Research methodology of the study

Table 4 List of selected criteria to evaluate the Odd–Even scheme

Criteria levels	Sustainability assessment criteria
C1	Trip cost
C2	Safety and security
C3	Accidents
C4	Users satisfaction
C5	Trip time
C6	Congestion level
C7	Parking demand
C8	Para-transit demand
C9	Air quality
C10	Noise level
C11	Fuel consumption

are computed using the BPMSG AHP calculator (http://bpmsg.com/academic/ahp.php). For that, thirty expert opinion surveys on the questionnaire of AHP are done; further, the weights of indicators using the above calculator have been estimated. The mean value of all the selected samples is tabulated below.

3.4 Field data collection and information fusion

Three transport corridors, namely Captain Gaur Marg, Outer Ring Road, and Lala Lajpat Rai Marg (Fig. 1), were selected to collect the study area data. The data for evaluation criteria were collected from expert opinion, public opinion, measuring devices (GPS, Sound Level Meter, etc.), primary and secondary sources, which were directly used to quantify the parameters. The sample size of different types of surveys and their techniques are presented in Table 5. In addition to this, the summary of the quantifying parameters for criteria quantification is also shown in Table 6 (Survey proforma is attached as Supplementary Material 1).

BPA value is directly taken from public rating survey for safety and security criteria, using a stated preference (0–10) scale. The details are mentioned in Supplementary Material 2. In Table 7, positive values represent an increase and a negative value represents a decrease in that selected quantifying parameters. The difference in the model's quantified values and actual measurement information source may be due to some variation in the model's input data and actual measurement.

The aggregate percentage change due to the implementation of policy in different criteria is depicted in Table 8. The data were extracted from models, sensors, people and

Table 5 Sample size of different type of survey and its technique

Survey type	Number of sample size before and after implementation of the Odd–Even scheme	Technique
Public questionnaire survey	120	Stated preference
Expert opinion survey for BPA allocation	30	Stated preference
Measurement of traffic volume	3 locations	Videography
Noise level measurement	3 locations	Sound level meter
Air quality	Secondary data sources	CPCB monitoring station

Table 6 Description of the quantifying parameters for criteria

Quantifying parameters for criteria quantification	Before Odd–Even scheme implementation	After Odd–Even scheme implementation
Distance (km) (total sum of the distance traveled by surveyed vehicles)	3161	2455
Time (min) (total sum of the time taken by surveyed vehicles in the origin–destination journey)	7460	6715
Total number of cars possessed by the surveyed vehicle owner	125	140
Total number of cars actually used daily by surveyed owner	124	83
Total sum of times spent by vehicle in traffic jams	2925	1695
Total number of cars surveyed or observed	50	50
Car occupancy (sum of the number of occupants of car surveyed or observed)	77	113
Total number of autos surveyed or observed	26	21
Total number of taxis surveyed or observed	20	10
Total number of e-vehicles surveyed or observed	5	7
Autos occupancy (sum of the number of occupants of autos surveyed or observed)	59	59
Taxis occupancy (sum of the number of occupants of taxis surveyed or observed)	45	31
E-vehicles occupancy (sum of numbers of occupants of e-vehicles surveyed or observed)	23	35
Autos trip (sum of trips of all autos surveyed or observed)	197	234
Taxis trip (sum of trips of all taxis surveyed or observed)	132	99
E-vehicles trip (sum of trips of all e-vehicles surveyed or observed)	52	99

Table 7 Change in criteria due to implementation of the Odd–Even scheme

Criteria	Quantifying parameters	Captain Gaur Marg	Outer Ring Road	Lala Lajpat Rai Marg
Trip cost				
Model	Car occupancy	47% (increase)	47% (increase)	47% (increase)
	Daily car usage	41% (decrease)	41% (decrease)	41% (decrease)
Actual measurement	Fuel wastage in traffic	42% (decrease)	42% (decrease)	42% (decrease)
Trip time				
Model	Speed	16% (increase)	16% (increase)	16% (increase)
	Fuel wastage in traffic	29% (decrease)	29% (decrease)	29% (decrease)
Actual measurement	Actual time	10% (increase)	10% (increase)	10% (increase)
Safety and security	Subjective survey data			
Users satisfaction				
Model	Change in other criteria	21% (increase)	21% (increase)	21% (increase)
Actual measurement	Change in satisfaction	17% (increase)	17% (increase)	17% (increase)
Accidents				
Model	Netherland model	3% (decrease)	10% (decrease)	6% (decrease)
Actual measurement	Netherland model	3% (decrease)	10% (decrease)	6% (decrease)
Congestion level				
Model	Speed	16% (increase)	16% (increase)	16% (increase)
	Time spent in traffic	42% (decrease)	42% (decrease)	42% (decrease)
Actual measurement	Traffic volume	12% (decrease)	19% (decrease)	17% (decrease)
Parking demand				
Model	Traffic volume	12% (decrease)	19% (decrease)	17% (decrease)
Actual measurement	Actual data	33% (decrease)	33% (decrease)	33% (decrease)
Para-transit demand				
Model	Trip and occupancy	27% (increase)	27% (increase)	27% (increase)
Actual measurement	Actual vehicle count	17% (increase)	15.5% (increase)	19.5% (increase)
Fuel consumption				
Model	Tiwari model	12% (decrease)	19% (decrease)	17% (decrease)
Actual measurement	Pump data	16% (decrease)	16% (decrease)	16% (decrease)
Noise level				
Model	Lam and Tam model	2% (decrease)	2% (decrease)	2% (decrease)
Actual measurement	Actual data	3% (decrease)	3% (decrease)	3% (decrease)
Air quality				
Model	Emission model	12% (decrease)	22% (decrease)	22% (decrease)
Actual measurement	TERI report	6% (decrease)	6% (decrease)	6% (decrease)

Table 8 Aggregated change in criteria

S.no	Criteria	Percent change (%)	
		Model	Actual
1	Trip cost	−44	−42
2	Trip time	−29	−10
3	Safety and security	Subjective Survey Data	
4	Accidents	−7	−7
5	Users satisfaction	21	17
6	Congestion level	−29	−16
7	Parking demand	−16	−33
8	Para-transit demand	27	18
9	Fuel consumption	−16	−16
10	Noise level	−2	−3
11	Air quality	19	6

specialist perception using questionnaire surveys for all the selected 11 criteria. The models are used to quantify the criteria, and after analyzing the change, BPA is estimated accordingly. The BPA function from different information origins was finally aggregated using the D–S theory (Kang et al., 2019).

3.5 BPA Allocation

The basic probability assignment (BPA) is a primitive of evidence theory. A basic probability belief function was allocated to bring into homogeneous basic probability assignment observed in a different unit in terms of increase, decrease, and no change for all the 11 criteria for both pre-implementation and post-implementation stage of Odd–Even schemes in Delhi using D–S Theory. In the first stage, BPA was allocated for observing the above changes in aggregate obtained from the other two information sources: expert and public opinion survey (Boudraa et al., 2004; Cheng & Chen, 2015). The final BPA is fused for all four sources in the case of the pre- and post-implementation stage. The example of BPA allocation is attached as a Supplementary Material 3. Based on the above-calculated changes, the BPA is allocated using the presume threshold scale (Table 9) and if the change observed is 40%, the BPA would be 0.4 for the change evaluation level and the remaining 0.6 would be equally distributed to the other two evaluation levels so that biasness is avoided in the allocation.

3.6 Information fusion

This study's novel combination of AHP and D–S theory is designed, which is hardly used in the transportation section (Wu et al., 2019). The basic probability assignments form all the information sources merged/fused by Dempster–Shafer rule of combination to determine the final fused BPA (fusion example is presented as Supplementary Material 4). The final BPA has been combined with assumed individual utilities to find out the global utilities of all the criteria to get the TSI. The summary of the fused BPA is discussed in the result section of this paper.

Table 9 Presumed threshold scale

Presumed threshold scale	
Percent change	BPA
100	1
90	0.9
80	0.8
70	0.7
60	0.6
50	0.5
40	0.4
30	0.3
20	0.2
Up to 10	0.1

3.7 Estimation of the transportation sustainability index

Reisi et al. (2014), Zito and Salvo (2011), and Attiyah et al. (2016) developed sustainability indicators and indexes based on the transportation needs of different countries. For TSI estimation in this study, the individual utilities for all the 11 criteria were assigned by local transportation experts as presented in Table 10. These are allocated in three sets of evaluation levels ranged as an increase (I), decrease (D), and no change (N). The global utility for a parameter is estimated using Eq. 5.

$$u_i = \sum_{k=1}^{p} u(Hk) \times bpa(Hk) \tag{5}$$

where Hk is the assessment level.

Hk€ {I,N,D}, u(Hk) is the individual utility of an evaluation level Hk.

BPA (Hk) is the basic probability assignment or mass function related to each assessment level Hk, and.

p represents the number of assessment levels.

$$p = 3 \text{ for Hk\{I,N,D\}}.$$

The utility value ranges from 0 (decrease) to 1(increase), with a middle value of 0.33 showing no change. The utility values are shown in Table 10 for each criterion.

Global utilities for the parameters C_1, C_2……, C_N at time t_n is given by $u_1(t_n)$, $u_2(t_n)$,……., $u_n(t_n)$, then the TSI is given in Eq. (6)

$$\text{TSI}\ (t_n) = u_1(t_n) \times W_1 + u2(t_n) \times W_2 + \ldots uN(t_n) \times WN \tag{6}$$

where *W1, W2*…., *WN* represent the weight of criteria *C1*, *C2*……, *CN* obtained from AHP.

3.8 Impact assessment and evaluation

Verma et al. (2015) and Bueno et al. (2015) presented a methodology based on the assessment of change in the composite index before and after implementing the transportation

Table 10 Utility values for selected criteria

Criteria level	Name of criteria	Utility values		
		u(I)	u(N)	u(D)
C1	Trip cost	0	0.3	1
C2	Trip time	0	0.3	1
C3	Safety and security	1	0.3	0
C4	Accidents	0	0.3	1
C5	Users satisfaction	1	0.3	0
C6	Congestion level	0	0.3	1
C7	Parking demand	0	0.3	1
C8	Para-transit demand	0	0.3	1
C9	Fuel consumption	0	0.3	1
C10	Noise level	0	0.3	1
C11	Air quality	1	0.3	0

measure. In this case, also, the transportation sustainability index of the Odd–Even scheme is evaluated by observing the change in efficiency in term of TSI for before (n-$_1$) and after (n) the implementation of the vehicle rationing scheme as presented by equation Eq. 7.

$$\Delta \mathrm{TSI}(t_n, t_{n-1}) = \mathrm{TSI}(t_n) - \mathrm{TSI}(t_{n-1}) \tag{7}$$

Based on this evaluation, if $\Delta \mathrm{TSI}\ (t_n, t_{n\text{-}1}) > 0$, the effect of the vehicle rationing scheme (Odd–Even scheme) in the city has been considered as having a positive impact on sustainability, and the measure is recommended for implementation. If $\Delta \mathrm{TSI}\ (t_n, t_{\ n\text{-}1}) \leq 0$, then the Odd–Even scheme measure is rejected in evaluation.

4 Results and discussion

The implementation of the Odd–Even scheme has complexity due to site-specific parameters, lack of homogeneity in detection, measure, store, and monitoring of standardized performance indicators to analyze the effects of adopted policy measures (Zito & Salvo, 2011) quantitatively. This study has eleven criteria: trip cost, trip time, safety and security, accidents, user's satisfaction, congestion level, parking demand, para-transit demand, fuel consumption, noise level, and air quality were selected to evaluate the "Odd–Even scheme" in Delhi. The opinion of 30 experts from different academic and research institutes was obtained, and weights of different criteria were selected (Table 11) using AHP's pairwise comparison methodology (Blachowski, 2015; Emrouznejad & Marra, 2017).

The changes in the criteria mentioned above were taken as the measures for evaluating the "Odd–Even scheme" in Delhi. Based on the change mentioned above in the criteria, BPA has been allocated. The model underestimated noise and parking demand. Trip time, user satisfaction, congestion level, para-transit demand, and air quality were overestimated by the model up to 33 to 55%. The data obtained from model, field, public perception, expert opinion have large variation within their domain parameters. It needs further fusion of all the observed data to assess the more realistic estimates (Table 12).

A basic probability belief function was allocated to bring into homogeneous basic probability assignment observed in a different unit in terms of increase, decrease, and no change for all the 11 criteria for both pre-implementation and post-implementation stage of

Table 11 Weight of selected criteria

S. no	Criteria	Mean
1	Air quality	0.144
2	Accidents	0.141
3	Safety and security	0.117
4	Congestion level	0.116
5	Trip time	0.104
6	Noise level	0.099
7	Users satisfaction	0.089
8	Fuel consumption	0.076
9	Trip cost	0.064
10	Para-transit demand	0.042
11	Parking demand	0.036

Table 12 Allocation of BPA for pre-implementation stage of scheme

Evaluation criteria	Expert opinion survey			Model			Public opinion			Actual measurement		
	I	N	D	I	N	D	I	N	D	I	N	D
Trip cost	0.62	0.2	0.18	0.44	0.28	0.28	0.63	0.18	0.19	0.42	0.29	0.29
Trip time	0.58	0.23	0.19	0.29	0.355	0.355	0.61	0.2	0.19	0.1	0.45	0.45
Safety and security	0.32	0.25	0.43	0.34	0.22	0.44	0.34	0.22	0.44	0.34	0.22	0.44
Accidents	0.55	0.26	0.19	0.1	0.45	0.45	0.6	0.2	0.2	0.1	0.45	0.45
Users satisfaction	0.28	0.27	0.45	0.395	0.395	0.21	0.29	0.22	0.49	0.415	0.42	0.17
Congestion level	0.66	0.18	0.16	0.29	0.355	0.355	0.66	0.18	0.16	0.16	0.42	0.42
Parking demand	0.62	0.21	0.17	0.16	0.42	0.42	0.7	0.14	0.16	0.33	0.36	0.36
Para-transit demand	0.36	0.25	0.39	0.365	0.365	0.27	0.32	0.23	0.45	0.365	0.37	0.18
Fuel consumption	0.63	0.21	0.16	0.16	0.42	0.42	0.64	0.2	0.16	0.16	0.42	0.42
Noise level	0.56	0.21	0.23	0.1	0.45	0.45	0.58	0.2	0.22	0.1	0.45	0.45
Air quality	0.29	0.22	0.49	0.405	0.405	0.19	0.25	0.2	0.55	0.45	0.45	0.1

the Odd–Even scheme in Delhi using D–S theory. In the first stage, BPA was allocated for observing the above changes in aggregate obtained from the other two information sources that are expert and public opinion survey (Boudraa et al., 2004; Cheng & Chen, 2015). The final BPA is fused for all four sources in the pre- and post-implementation stage (Tables 12 and 13). The result shows that parking demand decreased considerably, whereas para-transit demand and user satisfaction are increased considerably. This indicates that policy has an impact on opinion as well as measurement.

4.1 Information fusion

The basic probability assignments form all the information sources merged/fused by the Dempster–Shafer rule of combination to find out the final fused BPA (Wu et al., 2019). The final BPA has been combined with assumed individual utilities to find out the global utilities of all the criteria to find out the TSI. The summary of the fused BPA is presented in Table 14.

Table 13 Allocation of BPA for post-implementation stage

Evaluation criteria	Experts			Model			Public opinion			Actual measurement		
	I	N	D	I	N	D	I	N	D	I	N	D
Trip cost	0.24	0.22	0.54	0.28	0.28	0.44	0.27	0.19	0.54	0.29	0.29	0.42
Trip time	0.24	0.2	0.56	0.355	0.355	0.29	0.27	0.2	0.53	0.45	0.45	0.1
Safety and security	0.59	0.21	0.2	0.61	0.21	0.18	0.61	0.21	0.18	0.61	0.21	0.18
Accidents	0.26	0.23	0.51	0.45	0.45	0.1	0.25	0.23	0.52	0.45	0.45	0.1
Users satisfaction	0.63	0.19	0.18	0.21	0.395	0.395	0.63	0.21	0.16	0.17	0.42	0.42
Congestion level	0.2	0.22	0.58	0.355	0.355	0.29	0.23	0.21	0.56	0.42	0.42	0.16
Parking demand	0.28	0.24	0.48	0.42	0.42	0.16	0.24	0.22	0.54	0.35	0.36	0.33
Para-transit demand	0.54	0.2	0.26	0.27	0.365	0.365	0.59	0.2	0.21	0.18	0.37	0.37
Fuel consumption	0.26	0.24	0.5	0.42	0.42	0.16	0.25	0.21	0.54	0.42	0.42	0.16
Noise Level	0.3	0.22	0.48	0.45	0.45	0.1	0.27	0.2	0.53	0.45	0.45	0.1
Air Quality	0.54	0.23	0.23	0.19	0.405	0.405	0.59	0.21	0.2	0.1	0.45	0.45

Table 14 Value of BPA of evaluation criteria before and after the implementation of scheme

Evaluation criteria	Pre-implementation stage			Post-implementation stage		
	I	N	D	I	N	D
Trip cost(C1)	0.9268	0.0375	0.0357	0.0841	0.0543	0.8616
Trip time(C2)	0.4389	0.3144	0.2467	0.4084	0.2521	0.3395
Safety and security(C3)	0.2425	0.0513	0.7062	0.9773	0.0142	0.0085
Accidents(C4)	0.1533	0.4892	0.3575	0.4962	0.4038	0.1000
Users satisfaction(C5)	0.4305	0.3149	0.2546	0.5572	0.2572	0.1856
Congestion level(C6)	0.7813	0.1867	0.1475	0.2380	0.2390	0.5230
Parking demand(C7)	0.8295	0.1587	0.1468	0.3591	0.2821	0.4905
Para-transit demand(C8)	0.3700	0.1847	0.2056	0.3977	0.1369	0.1868
Fuel consumption(C9)	0.4640	0.3330	0.2030	0.4205	0.3260	0.2535
Noise level(C10)	0.1342	0.3515	0.4234	0.5888	0.3199	0.0913
Air quality(C11)	0.5014	0.3043	0.1943	0.2605	0.3788	0.3607

From the above table, it is clear that trip cost, congestion level, parking demand, trip time, and accident and air quality have been decreased by 2300%, 254%, 234%, 37%, 223%, and 85%, respectively, due to scheme intervention. On the other hand, safety and security and user satisfaction have also been increased by 300% and 30% correspondingly compared to the pre-implementation stage. This quite profound decrease in trip cost may be due to the use of low-cost transportation services or shared trips with available cars/transport or shared hired taxi, or auto by the people to reach the office on time. The congestion level and parking demand remained low because of a smaller number of vehicles on roads. Due to the presence of many volunteers on roads, people felt safe and secure, whereas the accident has increased due to an increase in speed. Overall, the satisfaction level increased due to the pro-activeness of implementing such a scheme by the government. Although the scheme was implemented for air pollution reduction, it was not as high as the other parameters like trip cost or safety, or security. However, the total sustainability

Table 15 Global utilities for pre- and post-test stage

Evaluation criteria	Pre-test stage	Post-test stage
Trip cost(C1)	0.047	0.878
Trip time(C2)	0.341	0.415
Safety and security(C3)	0.258	0.982
Accidents (C4)	0.504	0.221
Users satisfaction(C5)	0.525	0.634
Congestion level(C6)	0.204	0.595
Parking demand(C7)	0.194	0.575
Para-transit demand(C8)	0.261	0.228
Fuel consumption(C9)	0.303	0.351
Noise level(C10)	0.529	0.187
Air quality(C11)	0.593	0.374

is not dependent on a single parameter, but it was estimated to take all the parameters into account. Therefore, sustainability was calculated based on the global utility and individual utility-based criteria as given in Table 15.

From the above table, it is clear that except for three parameters, namely accidents, noise level, and air quality, for all individual parameters, the global utility has increased from 20 to 1700% after the implementation of the Odd–Even scheme in Delhi.

4.2 Transportation sustainability index (TSI)

The TSI criteria are calculated by multiplying the weights allocated by AHP and global utilities-based theory (Fernández et al., 2018) as given in Eq. (1). The TSI and ΔTSI of all the selected criteria are presented in Table 16.

Considering the lengthiness of information fusion as evident from the tabular calculation of information fusion as in Table 14, instead of individual road-wise sustainability index calculation, subsequently, a composite sustainability index is calculated, which collectively reflects the change in the sustainability of all the selected corridors.

In Table 16, the trip cost shows a substantial increase in TSI value. It may be due to the rise in the number of car-sharing as a result of the scheme's implementation. The 'trip time' shows a moderate increase in TSI value as an increase in speed and decrease in time spent in traffic jams is not substantial. The third criterion, i.e., safety and security, shows a very substantial increase in TSI value (Olofsson et al., 2011) as it is calculated based on a subjective survey in which people acknowledge the implemented scheme as a positive initiative based on enhancement in their safety and security (sample data are attached as Supplementary Material 2). The criterion of 'accident' indicates a decrease in sustainability because the reduction in accidents is found very low and BPA for the decline is also found deficient which in turn increase the BPA of other two evaluation level, that is, decrease and no change. The other criterion, 'user satisfaction' shows a small change in TSI value because change brought by scheme implementation is not very substantial. The sixth criterion, i.e., congestion level, presents a moderate change in TSI value as the change observed in speed and time spent in traffic jams is moderate. The parking demand criterion shows an increase in TSI value because a substantial decrease in parking demand is observed from the parking demand data tabulation. The para-transit demand criteria represent a decrease in sustainability as its

Table 16 TSI and ΔTSI of individual criteria

	TSI (pre-implementation stage)	TSI (post-implementation stage)	ΔTSI (% change)
Trip cost(C1)	0.003	0.056	1766.6%
Trip time(C2)	0.035	0.043	22.9%
Safety and security(C3)	0.030	0.115	283.3%
Accidents (C4)	0.071	0.031	− 56.3%
Users satisfaction(C5)	0.047	0.056	19.1%
Congestion level(C6)	0.024	0.069	187.5%
Parking demand(C7)	0.007	0.021	200%
Para-transit demand(C8)	0.011	0.010	− 9.1%
Fuel consumption(C9)	0.023	0.027	17.4%
Noise level(C10)	0.052	0.019	− 63.5%
Air quality(C11)	0.085	0.054	− 36.5%

demand increase in the post-implementation stage, which results in a negative impact on sustainability. The criterion of fuel consumption shows an increase in TSI values as it is evident from the fuel consumption data that fuel consumption has been increased during the scheme. The noise level, as well as air quality criteria, shows a decrease in TSI values. This may be due to the decline in these two criteria found significantly less. Its BPA is also found very less for positive evaluation, which in turn increases the BPA values for negative evaluation levels.

The change in TSI of the Odd–Even scheme transportation measures in both stages is observed to recommend for the adoption in the city (Airy & Chandiramani, 2016). The estimated pre-TSI value is found as 0.39, and the post-TSI value is found as 0.50. The change in TSI value is observed as 0.11(22.8% increase). As the Δ TSI > 0, it may be recommended for the adoption in the city, but it also depends on the adoption authority that if the authority has set some threshold of change brought by the measure, then adoption is based upon that threshold and if the change is greater or equal to that threshold, then the measure would be adopted otherwise rejected.

ESCAP (2018) and Haghshenas and Vaziri (2012) developed the Sustainable Urban Transport Index (SUTI) for different cities of the world. These indexes are found useful for comparing the performance of policy adopted by any city and improving sustainability. Figure 3 shows the relative comparison of the five considered parameters with Greater Jakarta, Kathmandu, Colombo, Hanoi, Taiwan, and Atlanta and freight-specific sustainability. After implementing the Odd–Even policy, the criteria user's satisfaction performs very similar to Hanoi and Atlanta, whereas, in terms of air quality, it is found similar to Colombo and Taiwan. Trip cost performs for the considered country very similar that shows that economy of transport is one the critical factors in sustainability for all the countries. Atlanta, Jakarta, Kathmandu, Hanoi, and Colombo perform very well in terms of accident parameters for adding value to sustainability.

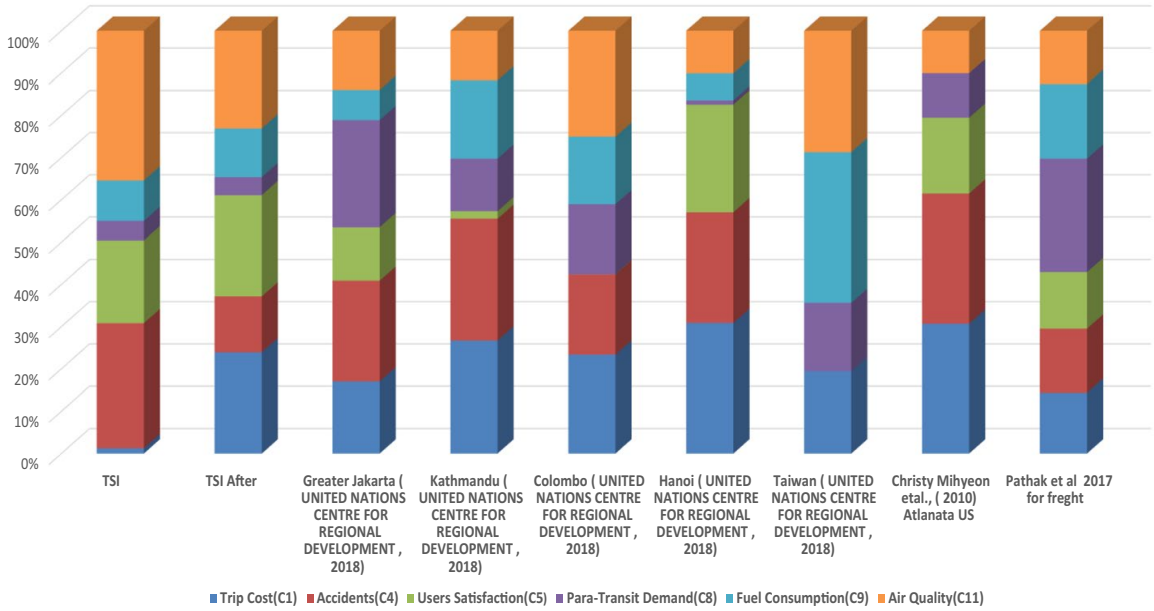


Fig. 3 Comparison of normalized transportation sustainability index of different countries

5 Conclusion

In this study, the vehicle rationing scheme (Odd–Even scheme) was evaluated to understand its impact on different transport-environment sustainable indicators. In this evaluation, the transportation sustainability index at pre- and post-implementation stages was calculated by adopting a multi-criteria hybrid approach based on AHP and D–S theory. In this evaluation, the selected indicators were quantified and the change brought about by the scheme in these indicators was calculated in both quantitative as well as in TSI terms. In this evaluation, quantitative change has been observed in various selected indicators after implementing the scheme. A reduction of 42%, 33%, 16%, 10%, and 7% was observed in the trip cost, parking demand, congestion level, trip time, and traffic accidents, respectively. Due to this scheme, the fuel consumption and noise level were decreased by 16% and 3%, respectively. The Transportation Sustainability Index, the trip cost, safety and security, congestion level, and parking demand increased by 1766.6, 283.3, 187.5, and 200%, respectively, due to this scheme. At the same time, the rest of the indicators did not show a significant improvement. After evaluating the entire data, an overall increase of 28.2% was found in the Transportation Sustainability Index after the implementation of the Odd–Even scheme in the capital city and recommended it as a sustainable transportation measure. International comparison of different sustainability parameters shows that trip cost is one of the significant parameters affecting transportation sustainability in all countries after accidents. Improving the Delhi transportation system more sustainable, the government must implement these to meet the internationally acclaimed sustainable transportation measures in combination. In the future, the present study may be used as a base case to assess the sustainability of several transportation measures in combination.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10668-021-01353-9.

References

Airy, A., & Chandiramani, J. (2016). The good and bad of 'Odd-Even Formula': Case study of Delhi and alternative measures towards sustainable transport. *Advances in Economics and Business Management, 3*(1), 31–35.

Arantes, L. T., Carvalho, A. C. P., Carvalho, A. P. P., Lorandi, R., Moschini, L. E., & Lollo, J. A. D. (2021). Surface runoff associated with climate change and land use and land cover in southeast region of Brazil. *Environmental Challenges, 3*, 100054.

Attiyah, M., Atawi, AI, Kumar, R., & Saleh, W. (2016). Transportation sustainability index for Tabuk city in Saudi Arabia: an analytic hierarchy process. *Transport, 31*(1), 47–55.

Awasthi, A., & Chauhan, S. S. (2011). Using AHP and Dempster-Shafer theory for evaluating sustainable transport solutions. *Environmental Modeling & Software, 26*(6), 787–796.

Baloian, N., Frez, J., Pino, J. A., Peñafiel, S., Zurita, G., & Abarca, A. (2019). Technology support for collaborative preparation of emergency plans. *Sensors, 19*, 5040.

Bell, S., & Morse, S. (2008). Sustainability indicators: Measuring the immeasurable? Earthscan, London, Sterling, VA, UK, ISBN-13: 978-1-84407-299-6.

Beynon, M., Cosker, D., & Marshall, D. (2001). An expert system for multi-criteria decision making using Dempster Shafer theory. *Expert Systems With Applications, 20*, 357–367.

Blachowski, J. (2015). Methodology for assessment of the accessibility of a brown coal deposit with analytical hierarchy process and weighted linear combination. *Environmental Earth Sciences, 74*, 4119–4131.

Boudraa, A. O., Bentabet, A., Salzenstein, F., & Guillon, L. (2004). Dempster-Shafer's basic probability assignment based on fuzzy membership functions. *Electronic Letters on Computer Vision and Image Analysis, 4*(1), 1–9.

Bouraima, M. B., Qiu, Y., Yusupov, B., & Ndjegwes, C. M. (2020). A study on the development strategy of the railway transportation system in the West African Economic and Monetary Union (WAEMU) based on the SWOT/AHP technique. *Scientific African, 8*, e00388.

Bueno, P. C., Vassallo, J. M., & Cheung, K. (2015). Sustainability assessment of transport infrastructure projects: A review of existing tools and methods. *Transport Reviews, 35*(5), 622–649.

Buzási, A., & Csete, M. (2015). Sustainability indicators in assessing urban transport systems. *Periodica Polytechnica Transportation Engineering, 43*(3), 138–145.

Calderwood, S., McAreavey, K., Liu, W., & Hong, J. (2017). Context-dependent combination of sensor information in Dempster-Shafer theory for BDI. *Knowledge and Information Systems, 51*, 259–285.

Caner, H. I., & Aydin, C. C. (2021). Shipyard site selection by raster calculation method and AHP in GIS environment, İskenderun. *Turkey. Marine Policy, 127*, 104439.

Chandna, R., Saini, S., & Kumar, S. (2021). Fuzzy AHP based performance evaluation of massive online courses provider for online learners. *Materials Today: Proceedings*, https://doi.org/10.1016/j.matpr.2021.02.255.

Cheng, Y. H., & Chen, S. Y. (2015). Perceived accessibility, mobility, and connectivity of public transportation systems. *Transportation Research Part A: Policy and Practice, 77*, 386–403.

Chowdhury, P., & Paul, S. K. (2020). Applications of MCDM methods in research on corporate sustainability. *Management of Environmental Quality: An International Journal, 31*(2), 385–405.

CST (2005). A report on "Defining Sustainable Transportation", Centre for Sustainable Transportation. Available online at http://cst.uwinnipeg.ca/documents/Defining_Sustainable_2005.pdf. Accessed on 13th Dec, 2019.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological), 30*(2), 205–247.

Deng, Y., Liu, Y., & Zhou, D. Y. (2015). An improved genetic algorithm with initial population strategy for symmetric TSP. *Mathematical Problems in Engineering, 2015*(3), 1–6.

Dobranskyte-Niskota, A., Perujo, A., Pregl, M. (2007). Indicators to Assess Sustainability of Transport Activities Part 1: Review of the Existing Transport Sustainability Indicators Initiatives and Development of an Indicator Set to Assess Transport Sustainability Performance, Joint Research Centre Institute for Environment and Sustainability.

Dobranskyte-Niskota, A., Perujo, A., Jesinghaus, J., Jensen, P. (2009). Indicators to assess sustainability of transport activities. Joint Research Centre (JRC) Scientific and Technical Reports, EUR 23041 EN/2.

Economic Survey of Delhi, (2018–2019). Available online at http://delhiplanning.nic.in/sites/default/files/Final%20Economy%20survey%20English.pdf. Accessed 12 Jan 2020.

Emrouznejad, A., & Marra, M. (2017). The state of the art development of AHP (1979–2017): a literature review with a social network analysis. *International Journal of Production Research, 55*(22), 6653–6675.

Ercan, T. (2009). Sustainability analysis of intelligent transportation systems. Master of Science Thesis, Department of Civil and Environmental Engineering, College of Engineering and Computer Science, University of Central Florida, Orlando, Florida.

ESCAP (2018). Available online at https://www.unescap.org/announcement/sustainable-urban-transport-index-suti. Accessed on 14th Feb, 2020.

Fernández, M. S., Serrano, G. A., Falero, M., & Pascual, E. (2018). Comparison of AHP and a utility-based theory method for selected vertical and horizontal forest structure indicators in the sustainability assessment of forest management in the sierra de Guadarrama National Park, Madrid Region. *Sustainability, 10*, 4101.

Gudmundsson H. (2001). Indicators and performance measures for Transportation, Environment and Sustainability in North America. Report from a German Marshall Fund Fellowship 2000 individual study tour October 2000. National Environmental Research Institute, Denmark. 87 p. Research Notes from NERI No. 148.

Gudmundsson, H. (2004). Sustainable transport and performance indicators. *Issues in Environmental Science and Technology, 20*, 35–63.

Haghshenas, H., & Vaziri, M. (2012). Urban sustainable transportation indicators for global comparison. *Ecological Indicators, 15*, 115–121.

Hak, T., Dahl, A. L., & Molden, B. (2007). *Sustainability Indicators. A scientific assessment*. Washigton, DC: Island Press.

Hosking, J., Mudu, P., & Dora, C. (2011). Health co-benefits of climate change mitigation—Transport sector, Health in the green economy. World Bank Report, 1–144.

Huadong, W., Siegel, M., Stiefelhagen, R., & Yang, J. (2002). Sensor fusion using Dempster-Shafer theory. In *Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.00CH37276)*, Anchorage, AK, USA, 7–12.

Jiang, W., Yang, Y., Luo, Y., & Qin, X. Y. (2015). Determining basic probability assignment based on the improved similarity measures of generalized fuzzy numbers. *International Journal of Computers Communications & Control, 10*(3), 333–347.

Kang, B. Y., Zhang, P., Gao, Z., Chhipi-Shrestha, G., Hewage, K., & Sadiq, R. (2019). Environmental assessment under uncertainty using Dempster-Shafer theory and Z-numbers. *Journal of Ambient Intelligence and Humanized Computing, 11*, 2041–2060.

Konsult. (2008). Available online at http://www.konsult.leeds.ac.uk. Accessed 12 Dec 2019.

Litman, T., 2019. Well Measured Developing Indicators for Sustainable and Livable Transport Planning, Victoria Transport Policy Institute, Todd Alexander Litman © 2006–2019.

Luís, A.d.A., & Cabral, P. (2021). Small dams/reservoirs site location analysis in a semi-arid region of Mozambique. *International Soil and Water Conservation Research*. https://doi.org/10.1016/j.iswcr.2021.02.002.

Ma, F., He, J., Ma, J., & Xia, S. (2017). Evaluation of urban green transportation planning based on central point triangle whiten weight function and entropy-AHP. *Transportation Research Procedia, 25*, 3634–3644.

Michael, G., Santosh, H., Pande, R., Sudarshan, A., 2017. Clearing the air on Delhi's odd-even program. Available online at https://epic.uchicago.in/wpcontent/uploads/2017/07/Odd-Even-EPW-draft-02072017.pdf. Accessed on 7th Jan, 2020.

Miller, P., de Barros, A. G., Kattan, L., & Wirasinghe, S. C. (2017). Managing uncertainty in the application of composite sustainability indicators to transit analysis. *Transportation Research Procedia, 25*, 4003–4018.

Mohan, D., & Tiwari, G. (2000). Mobility, environment and safety in megacities: Dealing with a complex future. *IATSS Research, 24*(1), 39–46.

Mojumder, A., & Singh, A. (2021). An exploratory study of the adaptation of green supply chain management in construction industry: The case of Indian Construction Companies. *Journal of Cleaner Production, 295*, 126400.

Nijenhuis, G., (2012). Setting guidelines for identifying sustainable transport challenges in medium-sized cities in Indonesia. Master of Science Thesis, Department of Civil Engineering and Management, University of Twente.

Olofsson, Z., Varhelyi, A., Koglin, T., & Angjelevska, B. (2011). *Measuring sustainability of transport in the city-development of an indicator-set*. . Lund University Faculty of Engineering.

Omrani, H., Awasthi, A., Ion, L., & Trigano, P. (2009). Hybrid approach for evaluating environmental impacts for urban transportation mode sharing. *Journal of Decision Systems, 18*(2), 185–201.

Pham, B. T., Luu, C., Dao, D. V., Phong, T. V., Nguyen, H. D., Le, H. P., Meding, J. V., & Indra Prakash, I. (2021). Flood risk assessment using deep learning integrated with multi-criteria decision analysis. *Knowledge-Based Systems, 219*, 106899.

Propolis (2014). Developing sustainability indicators for urban transportation system. Available online at http://www.spiekermann-wegener.de/pub/pdf/BSKSMW_CUPUM2005.pdf. Accessed on 12th Dec, 2019.

Ramani, T. L., Zietsman, J., Knowles, W. E., & Quadrifoglio, L. (2011). Sustainability enhancement tool for state departments of transportation using performance measurement. *Journal of Transportation Engineering, ASCE, 137*(6), 404–415.

Rassafi, A. A., & Vaziri, M. (2005). Sustainable transport indicators: Definition and integration. *International Journal of Environmental Science and Technology, 2*, 83–96.

Reisi, M., Aye, L., Rajabifard, A., & Ngo, T. (2014). Transport sustainability index: Melbourne case study. *Ecological Indicators, 43*, 288–296.

Rodrigues, V. S., Potter, A., & Naim, M. M. (2010). The impact of logistics uncertainty on sustainable transport operations. *International Journal of Physical Distribution & Logistics Management, 40*, 61–83.

Saaty, T. L. (1980). *The analytic hierarchy process*. New York (NY): McGraw-Hill International.

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26.

Salavati, A., Haghshenas, H., Ghadirifaraz, B., Laghaei, J., & Eftekhari, G. (2016). Applying AHP and clustering approaches for public transportation decision making: A case study of Isfahan City. *Journal of Public Transportation, 19*(4), 38–55.

Shafer, G. (1976). *A Mathematical Theory of Evidence Princeton University Press*. .

Shokravi, S., & Kurnia, S. (2014). A step towards developing a sustainability performance measure within industrial networks. *Sustainability, 6*(4), 2201–2222.

Singhania, K., Girish, G. P., & Emodi, N. V. (2016). Impact of Odd-Even rationing of vehicular movement in Delhi on air pollution levels. *Low Carbon Economy, 7*(4), 151–160.

Thanki, S., Govindan, K., & Thakkar, J. (2016). An investigation on lean-green implementation practices in Indian SMEs using analytical hierarchy process (AHP) approach. *Journal of Cleaner Production, 135*(1), 284–298.

Verma, A., Rahul, T. M., & Dixit, M. (2015). Sustainability impact assessment of transportation policies—A case study for Bangalore city. *Case Studies on Transport Policy, 3*, 321–330.

Waheed, B., Khan, F. I., Veitch, B., & Hawboldt, K. (2011). Uncertainty-based quantitative assessment of sustainability for higher education institutions. *Journal of Cleaner Production, 19*, 720–732.

Wang, Z., & Xiao, F. (2019). An improved multi-source data fusion method based on the belief entropy and divergence measure. *Entropy, 21*, 611.

Wanke, P.F., Jabbour, C.J.C., Antunes, J.J.M., Jabbour, A.B.L.de S., Roubaud, D., Sobreiro, V.A., Gonzalez□, E. (2021). An original information entropy-based quantitative evaluation model for low-carbon operations in an emerging market. *International Journal of Production Economics, 234*, 108061.

Wellar, B. (2009). Methodologies for Identifying and Ranking Sustainable Transport Practices in Urban Regions. Transport Canada Project, Research Report 3, Wellar Consulting INC., Ottawa, Ontario.

WHO (World Health Organization) (2016). Available online at http://www.who.int/phe/health_topics/outdoorair/databases/AAP_database_summary_results_2016_v02.pdf?ua=1. Accessed 13 Jan 2020.

Wu, C., Zhang, Z., & Zhong, W. (2019). A Group Decision-Making Approach Based on DST and AHP for New Product Selection under Epistemic Uncertainty. Mathematical Problems in Engineering, 1–16.

Yang, C. H., Lee, K. C., & Chen, H. C. (2016). Incorporating carbon footprint with activity-based costing constraints into sustainable public transport infrastructure project decisions. *Journal of Cleaner Production, 133*, 1154–1166.

Yasmina, M., Tatoglua, E., Kilicb, H. S., Zaimc, S., & Delend, D. (2020). Big data analytics capabilities and firm performance: An integrated MCDM approach. *Journal of Business Research, 114*, 1–15.

Yedla, S., & Shrestha, R. M. (2012). Application of analytic hierarchy process to prioritise urban transport options—comparative analysis of group aggregation methods. *World Review of Science, Technology and Sustainable Development, 9*(1), 15–33.

Zhou, F., Wang, X., Lim, M. K., He, Y., & Li, L. (2018). Sustainable recycling partner selection using fuzzy DEMATEL-AEW-FVIKOR: A case study in small-and-medium enterprises (SMEs). *Journal of Cleaner Production, 196*, 489–504.

Zhou, F., Lim, K. M., He, Y., & Pratap, S. (2019). What attracts vehicle consumers' buying: A Saaty scale-based VIKOR (SSC-VIKOR) approach from after-sales textual perspective? *Industrial Management & Data Systems, 120*(1), 57–78.

Zito, P., & Salvo, G. (2011). Toward an urban transport sustainability index: an European comparison European. *Transport Research Review, 3*(4), 179–195.

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

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Evolving Intelligent System for Trajectory Tracking of Unmanned Aerial Vehicles

Rupam Singh, *Graduate Student Member, IEEE*, and Bharat Bhushan, *Senior Member, IEEE*

***Abstract*— This article develops an evolving type-2 quantum fuzzy neural network (eT2QFNN) control scheme for achieving trajectory tracking with unmanned aerial vehicles (UAVs). The proposed approach involves quantum membership functions, automatic rule growing process, and parameter adjustment learning scenario to deal with the problems of inadequacy, uncertainties, and noise in conventional control techniques. Furthermore, the proposed approach is operated in a parallel structure with the proportional derivative (PD) controller to compensate the transients in performance and learn the dynamic characteristics of the system. Besides, a sliding theory-based adaptive law is equipped with the control approach to compensate for the nonlinearity of the UAV. To assess the performance, numerical simulations and real-time experiments are carried for pitch and yaw axes control of two degrees of freedom (2DoF) helicopter test rig with the proposed approach. The simulations and experiments are aimed at achieving an offline path tracking with an objective to minimize the deviation error and improve the time response characteristics of the UAVs. The results depict the robustness of the proposed approach in terms of integral time absolute error for a helicopter following various trajectories.**

***Note to Practitioners*—This article addresses the problem of trajectory tracking and attitude control in a two-rotor UAV system. In practical application, there are multiple end users for an efficiently controlled UAV system. Generally, the trajectory tracking and attitude control are associated with the capability of UAVs to perform vertical take-off, landing, maneuver, and cyclic rotation as per the change in path. Furthermore, the control of UAVs for trajectory tracking and attitude provides a unified framework in efficiently following the desired path and maintaining the desired attitude. This has potential applications in the field of search and rescue operations, surveillance, military applications, environmental exploration, and aerial cinematography. Although trajectory tracking and attitude control problems have been studied a lot, the drawbacks due to uncertainties, immediate response to trajectory changes, and attitude settling are still open and challenging. This article proposed an eT2QFNN for a two rotor of UAV system with PD controller using automatic rule growing process. Sufficient trajectories are developed to make the UAV follow them under the proposed approach. Both simulation and real-time experiments were conducted and the results of the developed controller are compared with conventional approaches.**

Manuscript received February 7, 2021; accepted April 7, 2021. This article was recommended for publication by Associate Editor V. Stojanovic and Editor C. Seatzu upon evaluation of the reviewers' comments. *(Corresponding author: Rupam Singh.)*

Rupam Singh was with the Department of Electrical Engineering, Delhi Technological University, New Delhi 110042, India. She is now with the Institute for Intelligent System Technologies, Alpen-Adria-Universität Klagenfurt, 9020 Klagenfurt, Austria (e-mail: singhrupam99@gmail.com).

Bharat Bhushan is with the Department of Electrical Engineering, Delhi Technological University, New Delhi 110042, India.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2021.3072339.

Digital Object Identifier 10.1109/TASE.2021.3072339

***Index Terms*— Evolving type-2 quantum fuzzy neural network (eT2QFNN), proportional–derivative (PD), quantum membership function (QMF), two degrees of freedom (2DoF), unmanned aerial vehicles (UAVs).**

I. Introduction

In THE past few years, a growing interest was exhibited in the field of automation-based robotics. In fact, various industries are choosing robots when it comes to dealing with dangerous and difficult tasks. Considering this, the unmanned robotic systems, especially the aerial vehicles are one of the most desirable aspects in the field of search and rescue operations, surveillance, military applications, environmental exploration, mapping, and others. The major advantage of the unmanned aerial vehicles (UAVs) in these applications is their capability of vertical take-off, landing, and efficient enough to turn suddenly as per the change in trajectory. Furthermore, many efforts have been made in the literature to achieve autonomous navigation, especially in the field of attitude stabilization and path tracking for UAVs.

A. Literature Review

Conventionally, attitude stabilization of quadrotor UAVs has been achieved with proportional–integral–derivative (PID) controllers [1] and linear quadratic regulators [2]. But it is identified that the PID is restricted to control around a fixed set point and unable to work efficiently when the system is highly nonlinear in nature [3]. Alternatively, LQR control tracks a reference trajectory more accurately compared to PID and responded effectively for parametric uncertainties, but showed limited robustness against unstructured uncertainty. To overcome these drawbacks, the $H2/H\infty$ control method has been discussed to reduce the oscillations and regulate the rotational moments in the presence of unmodeled dynamics and parametric uncertainties [4]. Furthermore, the complexities with the $H2/H\infty$ controller in defining and varying the weighting functions with respect to the real-time performance of UAV have limited its application in this field [5]. In [6], the immersion and invariance-based adaptive tracking control for UAVs is developed to achieve attitude tracking and position control. He *et al.* [7]–[9] developed various control approaches for trajectory tracking of rigid body flexible wing micro aerial vehicles. In [7], an adaptive control scheme is developed with a bounded position controller and a hyperbolic tangent function to extract the pitch command. Furthermore, the extracted pitch command is tracked with a radial basis function neural

1545-5955 © 2021 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 June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

network-based attitude controller to achieve the unknown aerodynamic torque. In [8], the boundary conditions-based control approaches are developed with partial and ordinary differential equations derived through the Hamiltonian principles. The control strategy is developed to restrain the vibrations in flexible wing operation and achieve desired angular position for trajectory tracking. Similarly, in [9], two sets of dual-loop iterative learning approaches are developed to achieve constant trajectory tracking by suppressing vibrations and rejecting rigid body displacement with micro aerial vehicles. In [10], the recurrent fuzzy neural network scheme is developed to achieve a stable close chain system for uncertain system parameter of dual-arm space robot. In [11], a tracking controller for pneumatic muscle actuators is developed using nonlinear model predictive controller and echo state Gaussian process. Furthermore, in [12], the noise control for real-time manipulator control is achieved with modified primal-dual neural network by taking the advantage of phase information and unknown amplitude of the harmonic signals. However, this method cannot be applied with general periodic disturbance conditions. In [13], the uncertainties in robotic systems are solved with an integral barrier Lyapunov function along with a soft saturation function. In [14], a finite time dissipative filtering approach is developed with hidden Markov models, and mode-dependent condition probability matrix to solve the conic-type nonlinearities of Markov jump systems. The developed approach achieved effective finite time boundedness for two jumping modes of the Markov jump systems by analyzing the trajectories of the output errors. In [15], the unstructured uncertainty of spatially interconnected systems is controlled by developing a PD-type iterative learning approach. The developed approach adapts a state feedback-based learning approach which cannot be implemented with the systems where the state measurements are not available. Furthermore, in [16], an adaptive fuzzy full state feedback control is developed to improve the tracking accuracy of robotic systems with dynamic uncertainties. Similarly, in [17], an impedance learning approach is used to develop the adaptive neural network control technique for identifying the uncertain plant models. The approaches exhibited poor convergence properties while operating under disturbances and uncertainties.

In [18]–[20], the fuzzy-based approaches are developed to achieve trajectory tracking with the UAVs by minimizing the system error based on the inputs and the antecedent fuzzy sets of nonsingleton fuzzy logic controllers (FLCs). But the inability of these techniques to perform while dealing with higher order uncertainties is considered as a major drawback. Furthermore, in [21], an event-triggered scheme was employed with a fuzzy model to overcome the drawback of missing data or events for a closed-loop operation of a network-controlled system. Generally, these methods increase the complexity of the system and can be simply handled by utilizing upper and lower bounded membership functions. In another kind, the type-2 FLCs (T2FLCs) were widely adapted to cope up with the higher order uncertainties. Besides, the increasing fuzzy sets have improved the degree of freedom for T2FLC to handle uncertainties conveniently [22]. Furthermore, the parameter uncertainties in nonlinear systems can be easily captured by the type-2 fuzzy systems which involve membership functions with upper and lower bounds [23]. But the type reduction scenario made the general T2FLCs computationally intensive.

Furthermore, the hybridization of artificial neural networks with T2FLC has improved the abilities due to their inherent learning ability and uniform approximation of nonlinear system [24]. This resulted in the development of interval type-2 fuzzy neural networks (IT2FNNs) [25], [26]. In [25], the deterministic and nondeterministic approaches with fuzzy and neural network controllers for navigation control of robotic system are compared to analyze the robustness, response time, precision, and computational complexity. Furthermore, in [26], the precision control of two-axis motion control system is achieved with IT2FNN. The adaptive laws of the developed approach are achieved with a Lyapunov stability theorem which significantly improve the robustness of the system with the developed approach. Furthermore, the operation of fuzzy neural networks is enhanced with the sparse deep belief approaches to achieve fast and efficient weight initialization and perform supervised modeling of nonlinear systems [27]. In [28], a density peak-based clustering approach is adapted to process nonlinearity and obtain suitable fuzzy rules for the adaptive fuzzy neural network approach. In [29], an adaptive tracking control is developed using a reinforcement learning algorithm to tolerate the faults in a class of nonlinear systems with less learning parameters. Similarly, in [30], a fault-tolerant control approach and a sensor fault detection mechanism are developed for discrete time model-free systems using echo state networks. Despite their better performance and prediction capabilities regarding the future dynamics of the system, the development of neural network control law can be cumbersome due to its drawbacks with computation of type reduction. In [31], the generalized fuzzy hyperbolic model is used to investigate the predefined performance problems for multiple-input and multiple-output systems. The developed approach constructs the adaptive laws on the basis of predistortion error to solve the approximation error of generalized fuzzy hyperbolic model. In [32], the parameter optimization of IT2FNNs is achieved by eliminating the backpropagation computation with big bang-big crunch and particle swarm optimization methods. Most studies in the literature are based on an assumption that the plant accurate model is known. Generally, this assumption is unrealistic, as many complex system dynamics are really tough to be explicitly formulated in real world [33]. Besides, the drawbacks of conventional fuzzy sets with self-organizing process, and inadequacy to adapt for changes in inputs raised the demand for innovation in the field of computation intelligence [34].

B. Motivation

In light of the above issues, this article develops an evolving type-2 quantum fuzzy neural network (eT2QFNN) targeted at overcoming the drawbacks of conventional fuzzy systems and achieving path tracking with UAVs. The proposed approach overcomes the effects of parametric uncertainties in the UAVs with the help of a quantum rule growing scenario. Besides, the drawbacks of conventional control techniques

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

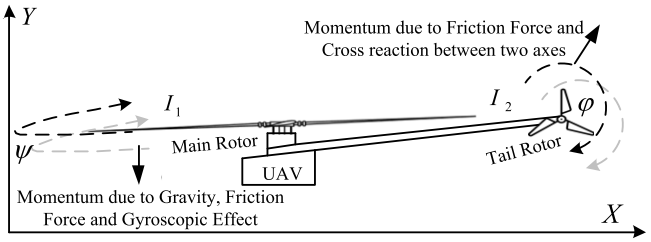


Fig. 1. Twin rotor model for UAV.

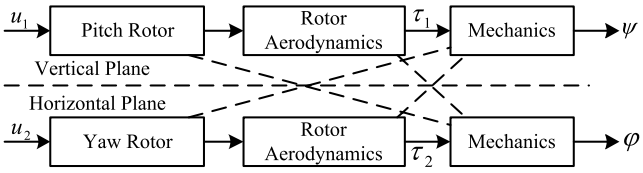


Fig. 2. Block diagram of couplings in the twin rotor UAV system.

with harmonic noises, and initializing weights for achieving trajectory tracking are overcome by the automatic generation of the rules in a single pass learning mode. Moreover, the adapted quantum membership functions (QMF) tackle the problem of inadequacy and noise in conventional membership functions of fuzzy sets. Furthermore, the two learning policies encompassed with the proposed approach, i.e., growing scenario and parameter adjustment, deal with the drawbacks of self-organizing process in conventional fuzzy techniques. Moreover, to achieve the trajectory tracking with the proposed approach, an unmanned helicopter system with two degrees of operation is operated with the proposed controller in the laboratory environment. The major contributions of this article can be summarized as follows.

1) An eT2QFNN controller is proposed to achieve attitude control and trajectory tracking for two degrees of freedom (2DoF) operation of UAVs with proportional derivative (PD) controller.
2) QMFs utilize the upper and lower bounded membership functions tackle the unprecedent events, and noise problems of conventional fuzzy sets.
3) The uncertain jump functions of QMFs identify the uncertainties in the inputs of the controller to achieve efficient control action.
4) The eT2QFNN is equipped with a rule growing mechanism which automatically generates the rules in single-pass learning mode and deals with the degree of uncertainty in type-2 fuzzy sets.
5) A sliding surface-based adaptive law compensates the nonlinearity of the UAV operating with eT2QFNN and PD controller to obtain the desired performance.

Further sections of this article are arranged as follows. Section II discusses the dynamic modeling of the laboratory setup imitating the bidirectional operation of UAVs. Section III provides a detailed overview of design of eT2QFNN controller. The learning policy of the designed controlled is addressed in Section IV. In Section V, the adaptive law and stability analysis for the developed controller are discussed, and the simulation and experimental analysis are carried out in Section VI.

II. Modeling and Control

A. Modeling

In this article, the twin rotor model of a UAV representing an unmanned helicopter is modeled as shown in Fig. 1. The pitch/main rotor is responsible for pitch motion around x-axis and z-axis by generating the vertical thrust, whereas yaw/tail rotor is responsible for yaw motion around the y-axis by generating horizontal thrust. The two dc motors are operated by voltage input (u_1, u_2) which returns torque τ_1 for main rotor and torque τ_2 for tail rotor. The coupling between the main and tail rotor of the helicopter is shown in Fig. 2. The cross coupling between both the rotors is indicated by dotted lines and the output of the two planes is taken from the mechanic's block. The vertical plane output corresponds to the pitch angle ψ and horizontal plane output corresponds to the yaw angle φ.

The dynamic modeling of the 2DoF operation of an unmanned helicopter is achieved by dividing the system dynamics into electrical and mechanical parts [35]. This approach aided in identifying the complex behavior of the twin rotor operation and eased the development of control action based on the nested feedback loops. Besides, the Euler–Lagrange method is adapted for modeling the mechanical aspects of the helicopter system. This allowed the higher gain adjustment when compared with the Newtonian approaches especially for real-time control scenarios. The modeling of the helicopter with reference to the Euler–Lagrange method to obtain the equations of motion is summarized on the basis of resolution of forward kinematics, and by evaluating of kinetic and potential energy. The driving torques are produced by the propellers, and the rotation can be described in principle as the motion of a pendulum. The mechanical and electrical aspects of the motors are characterized with the differential equations in the following manner.

1) For the Main Rotor and Vertical Movement:

$$\frac{d}{dt}\psi = \dot{\psi} \tag{1}$$

$$\frac{d}{dt}\dot{\psi} = \frac{a_1}{I_1}\tau_1^2 + \frac{b_1}{I_1}\tau_1 - \frac{M_g}{I_1}\sin\psi + \frac{0.0326}{2I_1}\sin(2\psi)(\dot{\varphi})^2 - \frac{B_{1\psi}}{I_1}\dot{\psi} - \frac{k_{gy}}{I_1}a_1\cos(\psi)\dot{\varphi}\tau_1^2 - \frac{k_{gy}}{I_1}b_1\cos(\psi)\dot{\varphi}\tau_1 \tag{2}$$

$$\frac{d}{dt}\tau_1 = -\frac{T_{m1}}{T_{m2}}\tau_1 + \frac{k_1}{T_{m2}}u_1. \tag{3}$$

2) For Tail Rotor and Horizontal Movement:

$$\frac{d}{dt}\varphi = \dot{\varphi} \tag{4}$$

$$\frac{d}{dt}\dot{\varphi} = \frac{a_2}{I_2}\tau_2^2 + \frac{b_2}{I_2}\tau_2 - \frac{B_{1\varphi}}{I_2}\dot{\varphi} - \frac{1.75}{I_2}k_c a_1\tau_1^2 - \frac{1.75}{I_2}k_c b_1\tau_1 \tag{5}$$

$$\frac{d}{dt}\tau_2 = -\frac{T_{t1}}{T_{t2}}\tau_2 + \frac{k_2}{T_{t2}}u_2. \tag{6}$$

For a twin rotor multiple-input and multiple-output helicopter system [35], the parameters $a_1, a_2, b_1,$ and b_2

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

are the static characteristics parameters with values $0.0135, 0.02, 0.0924$, and 0.09, respectively, I_1 and I_2 are the moment of inertia of main and tail rotor, M_g is the gravity momentum, $B_{1\psi}$ and $B_{1\varphi}$ are friction momentum function parameters for main and tail motors, k_{gy} is the parameter of gyroscopic momentum, T_{m1}, T_{m2}, T_{t1}, and T_{t2} are denominator parameters for main and tail motors, k_1 and k_2 are the motor gains for main and tail rotors, respectively, and k_c is the cross-reaction momentum gain.

Now, the modeling of the helicopter can be expressed as a continuous-time linear system by linearizing (1)–(6) around the point $\bar{l} = \begin{bmatrix} \bar{u}_1 & \bar{u}_2 & \bar{\psi} & \dot{\bar{\psi}} & \bar{\varphi} & \dot{\bar{\varphi}} \end{bmatrix}^{\mathrm{T}}$, gives

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \tag{7}$$

where $x \in R^6 = \begin{bmatrix} \psi & \dot{\psi} & \varphi & \dot{\varphi} & \tau_1 & \tau_2 \end{bmatrix}^{\mathrm{T}}$ corresponds to the states, $u \in R^2 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ is the control input, and $y \in R^2 = \begin{bmatrix} \psi \\ \varphi \end{bmatrix}$ is the measured output.

Furthermore, the matrices A, B, C, and D can be estimated by linearizing equations (1)–(6) as shown at the bottom of the next page.

To determine the relationship between input and output variables, a relative gain analysis is performed [36] which reveals a low interaction between the inputs and outputs of the vertical and horizontal systems. Considering this, the differential equations in (1)–(6) are approximated with the individual states of the helicopter responsible for achieving the vertical and horizontal motion.

3) For Pitch Model: The representation in (1)–(3) approximates the states of the helicopter involved into the output of vertical motion when $u_2 = 0$. Here, the pitch variables $\psi, \dot{\psi}$, and τ_1 have a representation into the state space as

$$\frac{d}{dt}\psi = \dot{\psi} \tag{8}$$

$$\frac{d}{dt}\dot{\psi} = \frac{a_1}{I_1}\tau_1^2 + \frac{b_1}{I_1}\tau_1 - \frac{M_g}{I_1}\sin\psi - \frac{B_{1\psi}}{I_1}\dot{\psi} \tag{9}$$

$$\frac{d}{dt}\tau_1 = -\frac{T_{m1}}{T_{m2}}\tau_1 + \frac{k_1}{T_{m2}}u_1$$

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\psi} \\ \dot{\boldsymbol{\psi}} \\ \boldsymbol{\tau_1} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u_1} \\ \boldsymbol{0} \end{bmatrix}, \quad \text{and } \boldsymbol{y} = \begin{bmatrix} \boldsymbol{\psi} \end{bmatrix}^{\boldsymbol{T}}. \tag{10}$$

4) For Yaw Model: Similar to the pitch subsystem, the model approaches of representations (4)–(6) for the horizontal output states are achieved by considering $u_1 = 0$. The variables of yaw subsystem $\varphi, \dot{\varphi}$, and τ_2 have a representation into the state space as

$$\frac{d}{dt}\varphi = \dot{\varphi} \tag{11}$$

$$\frac{d}{dt}\dot{\varphi} = \frac{a_2}{I_2}\tau_2^2 + \frac{b_2}{I_2}\tau_2 - \frac{B_{1\varphi}}{I_2}\dot{\varphi} \tag{12}$$

$$\frac{d}{dt}\tau_2 = -\frac{T_{t1}}{T_{t2}}\tau_2 + \frac{k_2}{T_{t2}}u_2$$

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\varphi} \\ \dot{\boldsymbol{\varphi}} \\ \boldsymbol{\tau_2} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{u_2} \end{bmatrix}, \text{ and } \boldsymbol{y} = [\varphi]^{\boldsymbol{T}}. \tag{13}$$

This representation of approximated state-space results in a linearized model for formulating the control objective. Hence, based on the principle of offline path tracking which aims at minimizing the deviation error and without the loss of generality, the control scheme for the helicopter needs to be developed.

B. Control Scheme

The PD controller is widely designed to regulate the elevation and travel angles of the helicopter operating with 2DoF. Furthermore, it can be operated as a nonlinear feed forward controller which compensates for the gravitational torque. Besides, the PD controller holds responsibility for the asymptotic stability of the helicopter and accommodates intelligent control techniques like T2FNN with enough time for online learning of the system dynamics. Considering these advantages, the generalized representation of PD law for the control of a particular rotor model is adapted as

$$y_{PD} = K_P e + K_D \dot{e} \tag{14}$$

where e and $\dot{e}$ correspond to the feedback error and its time derivative, respectively, between the desired and measured pitch or yaw angle value, and K_P and K_D correspond to the proportional and derivative gains of the PD controller, respectively. Furthermore, while operating the UAV with the PD controller, the disadvantages due to high control gain signal, and increasing accumulated error are observed. Besides, the differentiator leads to noise amplification which effects the flight of the UAVs and deviates them from following the predefined trajectory. To overcome these drawbacks, the advantages of self-learning techniques are identified to be widely applicable. As a part of the self-learning process, IT2FNN techniques are adapted in this article.

C. Type-2 Fuzzy Neural Network

The development of IT2FNN in this research adapts the antecedent type-2 and consequent crisp number (A2–C0) model [37]. The general structure of the A2–C0 fuzzy system for an ijth rule with two inputs is given as

$$\text{If } e \text{ is } \tilde{A}_{1i} \text{ and } \dot{e} \text{ is } \tilde{A}_{2j}, \text{ Then, } y_{NN} = \Omega_{ij} \quad \forall i = 1, \ldots, I \text{ and } j = 1, \ldots, J \tag{15}$$

where e and $\dot{e}$ correspond to the input variables for the type-2 fuzzy, I and J correspond to the number of membership functions, $\tilde{A}_{1i}$ and $\tilde{A}_{2j}$ are the fuzzy sets of IT2FNN, and their corresponding membership functions are $\tilde{\mu}_{1i}$ and $\tilde{\mu}_{2j}$, respectively. y_{NN} defines the output variable of the fuzzy neural network and Ω_{ij} are the adaptive crisp values. Furthermore, the upper and the lower membership functions of $\tilde{\mu}_{1i}$ are $\bar{\mu}_{1i}$ and $\underline{\mu}_{1i}$ and similarly for $\tilde{\mu}_{2j}$ are $\bar{\mu}_{2j}$ and $\underline{\mu}_{2j}$. The system output with IT2FNN can be written as [37]

$$y_{NN} = q\frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\underline{\mathcal{W}}_{ij}\Omega_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{I}\underline{\mathcal{W}}_{ij}} + (1-q)\frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\overline{\mathcal{W}}_{ij}\Omega_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{I}\overline{\mathcal{W}}_{ij}} \tag{16}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

where $\underline{\mathcal{W}}_{ij} = \underline{\mu}_{1i}\underline{\mu}_{2j}$, $\overline{\mathcal{W}}_{ij} = \bar{\mu}_{1i}\bar{\mu}_{2j}$, and q reflects the weighting parameter which shares the contribution of the upper and the lower membership functions.

III. Evolving Type-2 Quantum Fuzzy Neural Network Architecture

In general, the most commonly used multiple-input and single-output IT2FLC structure has I inputs ($x_1 \in X_1, \ldots, x_I \in X_I$) and one output ($y \in Y$) with K term nodes for each input. Furthermore, any given property of fuzzy can be expressed with multiple fuzzy characteristic functions. This provides a choice for the expert system designer, based on the understanding about the problem which may sometimes lead to certain cases having vague predicates. This can happen either due to the limited knowledge about the modeling of real-time system or due to insufficiency related to rule formulation. Considering these aspects, an arena of quantum computation is introduced in this article through the quantum fuzzy sets [38]. The quantum fuzzy sets are defined as a sense superpositions of various fuzzy sets which can be seen either as an extension of quantum properties, or quantum predicates. This provides an advantage for the development of a quantum fuzzy neural network approach which overcomes the drawbacks in conventional fuzzy sets. Besides, the QMF modifies the fuzzy controller into an improved nonstationary fuzzy set using minimum and maximum operators for the intersection and union. This helps in reducing the computational complexity of the fuzzy controller. Furthermore, the quantum fuzzy sets formulate an appropriate shape and number of membership functions which reduces the cost of training and decreases the computational complexity of the network. The details regarding the development of quantum theory in fuzzy approach and their advantages with objectivity, and preciseness are discussed in [38]. The complete architecture of eT2QFNN is shown in Fig. 3.

The layer 1 represents the input nodes or variables and layer 2 expresses the input variables as the fuzzy linguistic variables by calculating the quantum membership values. The layer 3 corresponds to the compensatory rule nodes which are equated with the number of compensatory fuzzy sets of each external linguistic input variable. The links before and after layer 3 represent the preconditions and consequences of the rules, respectively. The layer 4 is an output processing layer whose nodes are a linear function of the input variables. Generally,

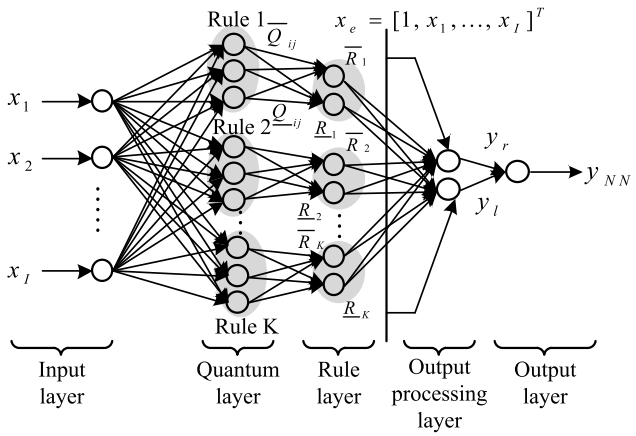


Fig. 3. Architecture of eT2QFNN.

these nodes are known as the consequent nodes. Furthermore, the last layer corresponds to an output layer and the nodes are defined as output nodes which act as a defuzzifier. The rule premise comprises of interval type-2 QMF (IT2QMF) as an antecedent to conventional IT2FNN and variable crisp numbers as the resultant part. For K term nodes, the fuzzy rules are expressed as follows:

$$R_j\text{: If } x_1 \text{ is } \tilde{Q}_{1j} \text{ and } x_i \text{ is } \tilde{Q}_{iK}, \text{ Then, } y_j = X_e\tilde{\Omega}_j \quad (17)$$

where x_i is the ith input, $\tilde{Q}_{ij} = [\bar{Q}_{ij}, \underline{Q}_{ij}]$ is the set of upper and lower linguistic terms of QMFs, y_j is the regression output in jth rule, and $\tilde{\Omega}_j = [\overline{\Omega}_j, \underline{\Omega}_j]$ is the set of upper and lower weight parameters defined as $\tilde{\Omega}_j \in \Re^{2(I+1)}$. Here, the QMF is extended into IT2FLC structure with uncertain jump functions for identifying uncertainties in the inputs and achieve efficient control. The output of IT2QMF for ith input and jth rule is given as (18), as shown at the bottom of the next page, where m_{ij} is the mean of ith input in jth rule, β is the slope factor, $\tilde{\theta}_{ij} = [\overline{\theta}_{ij}, \underline{\theta}_{ij}]$ is the uncertain jump position set defined as $\tilde{\theta}_{ij} \in \Re^{2\times I\times n_s\times K}$, and n_s defines the number of grades. For any case, if the upper jump position ($\overline{\theta}_{ij} = [\overline{\theta}^1_{1j} \ldots \overline{\theta}^{n_s}_{1j}; \ldots; \overline{\theta}^1_{Ij} \ldots \overline{\theta}^{n_s}_{Ij}]$) is greater than the lower jump position ($\underline{\theta}_{ij} = [\underline{\theta}^1_{1j} \ldots \underline{\theta}^{n_s}_{1j}; \underline{\theta}^1_{Ij} \ldots \underline{\theta}^{n_s}_{Ij}]$), which is defines as $\overline{\theta}_{ij} > \underline{\theta}_{ij}$, the output of IT2QMF (18) leads to IT2 inference scheme producing footprint of uncertainties (FOUs). Furthermore, the union function U with $U(x; a, b) = \begin{cases} 1, \text{if } a \le x_i < b \\ 0, \text{otherwise} \end{cases}$ implies that $\dot{x_i}, m_{ij}, \infty$ have

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -\frac{M_g}{I_1} & -\frac{B_{1\psi}}{I_1} & 0 & 0 & \frac{b_1}{I_1} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{B_{1\psi}}{I_2} & -1.75k_c\frac{b_1}{I_2} & \frac{b_2}{I_2} \\ 0 & 0 & 0 & 0 & -\frac{T_{m1}}{T_{m2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{T_{t1}}{T_{t2}} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{k_1}{T_{m2}} & 0 \\ 0 & \frac{k_2}{T_{t2}} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{and } D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the properties of commutativity, associativity, monotonic, and continuous [39].

A detailed operation of each layer in eT2QFNN is discussed in detail as follows.

Layer 1 *(Input Layer):* The input signals are fed into this layer where the data are directly propagated into the next layer without performing any computation. For an nth observation, the input x_i is defined as $X_n \in \Re^{1\times I}$ and the output is given defined as

$$u_i = x_i. \tag{19}$$

Layer 2 *(Quantum Layer):* In this layer, the fuzzification process is performed. The QMFs calculate the degree of membership for X_n in each rule. The total number of rules initially defined are denoted as K. The outputs of the quantum layer are mathematically obtained as

$$\overline{Q}_{ij} = \tilde{Q}_{ij}\left(x_i, \beta, m_{ij}, \overline{\theta}_{ij}\right) \tag{20}$$

$$\underline{Q}_{ij} = \tilde{Q}_{ij}\left(x_i, \beta, m_{ij}, \underline{\theta}_{ij}\right). \tag{21}$$

Layer 3 *(Rule Layer):* This layer calculates the spatial firing strength by combining the membership degrees of jth rule which is given as follows:

$$\tilde{R}_j = [\overline{R}_j, \underline{R}_j] \tag{22}$$

where $\tilde{R}_j$ is the spatial firing strength, $\overline{R}_j = [\overline{R}_1 \ldots \overline{R}_K]$ is the set upper firing strength, and $\underline{R}_j = [\underline{R}_1 \ldots \underline{R}_K]$ is the set of lower firing strength. The product T-norm is adapted with the IT2QMF as given in the following equations for performing this operation:

$$\overline{R}_j = \Pi_{i=1}^{I} \overline{Q}_{ij} \tag{23}$$

$$\underline{R}_j = \Pi_{i=1}^{I} \underline{Q}_{ij}. \tag{24}$$

Layer 4 *(Output Processing Layer):* The upper and lower crisp outputs, i.e., y_r and y_l, respectively, are calculated in this layer. The type reduction process is carried out by employing design factors $[q_r, q_l]$ for converting IT2 variables into type 1 variable. The design factors control the upper and lower membership function proportion in such a way that $q_r > q_l$. This process of type reduction requires lesser epochs when compared to conventional type reduction techniques [40]. Here, the decoupled extended Kalman filter [40] is used to adjust the design factor such that the upper and lower crisp outputs keep adapting to the changes in the inputs. The upper and lower crisp outputs are calculated as

$$y_r = \frac{(1-q_r)\sum_{j=1}^{K} \underline{R}_j \overline{\Omega}_j x_e^{\mathrm{T}} + q_r \sum_{j=1}^{K} \overline{R}_j \overline{\Omega}_j x_e^{\mathrm{T}}}{\sum_{j=1}^{K} (\overline{R}_j + \underline{R}_j)} \tag{25}$$

$$y_l = \frac{(1-q_l)\sum_{j=1}^{K} \underline{R}_j \underline{\Omega}_j x_e^{\mathrm{T}} + q_l \sum_{j=1}^{K} \overline{R}_j \underline{\Omega}_j x_e^{\mathrm{T}}}{\sum_{j=1}^{K} (\overline{R}_j + \underline{R}_j)} \tag{26}$$

where $\overline{\Omega}_j = \left[\overline{w}_{ij}, \ldots, \overline{w}_{(I+1)j}\right]$ is the weight parameter of upper layer for jth rule, $\underline{\Omega}_j = \left[\underline{w}_{ij}, \ldots, \underline{w}_{(I+1)j}\right]$ is the weight parameter of lower layer for jth rule, and $x_e = [1, x_1, \ldots, x_I]$ is the extended input vector. The 1 in extended input prevents untypical gradient and intercepts the rule consequent.

Layer 5 *(Output Layer):* This layer calculates the crisp output of the network by summing the upper and lower crisp outputs obtained at the output processing layer as follows:

$$y_{NN} = y_r + y_l. \tag{27}$$

IV. Learning Algorithm

The eT2QFNN works on an online learning policy and consists of two scenarios, the rule growing criteria and the parameter update scenario. These scenarios are executed for every iteration.

A. Growing Criterion

Initially, the learning process has an empty rule base and the network structure keeps on updating as the input data changes. This eT2QFNN automatically evolves its fuzzy rules by forming a hypothetical rule from a newly obtained data sample. The hypothetically formed rules need to be significantly evolved before adding it to the network structure. For a jth rule, the significance is evaluated using a Gaussian mixture model (GMM) [41] to approximate the complex changes in data. This is mathematically formulated as

$$\hat{E}_{\mathrm{sig}}(j) = \left\|\omega_j\right\| \left(\pi^{\frac{I}{2}} \det\left(\Sigma_j\right)^{\frac{1}{2}} N_j A^{\mathrm{T}}\right)^{1/2} \tag{28}$$

where $\Sigma_j = \mathrm{diag}\left(\sigma_{1,j}^2, \ldots, \sigma_{Ij}^2\right)$ is the positive definite weighting matrix, $A = [\alpha_1, \ldots, \alpha_H]$ is the GMM mixing coefficients vector with Gaussian coefficients $\alpha_1, \ldots, \alpha_H$, and N_j is further represented as (29), as shown at the bottom of the next page.

Here, $m_j = [m_{1,j}, \ldots, m_{I,j}]$ is the mean vector of jth rule.

Furthermore, to improve the estimating accuracy of the IT2QMF, the interval type-2 Gaussian membership functions (IT2GMFs) are adapted as follows:

$$\tilde{Q}_{i,j}\left(x_i, \beta, m_{i,j}, \tilde{\theta}_{i,j}\right) \approx \tilde{\mu}_{i,j} = \exp\left(-\frac{\left(x_i - m_{i,j}\right)^2}{\tilde{\sigma}_{i,j}}\right) \tag{30}$$

where $\tilde{\sigma}_{i,j} = \left[\underline{\sigma}_{i,j}, \overline{\sigma}_{i,j}\right]$ with $\underline{\sigma}_{i,j} = \min \underline{\theta}_{i,j}$ and $\overline{\sigma}_{i,j} = \min \overline{\theta}_{i,j}$. In this process, both the means of IT2GMF and IT2QMF are defined equal and the minimum value of $\tilde{\theta}_{i,j}$ is used to define the width of upper and lower IT2GMF. The illustration of IT2GMF in estimating the significance of IT2QMF is given in Fig. 4. By following the above-discussed criteria, it can be observed that the IT2GMF fits within the area of IT2QMF resulting in a good approximation for IT2QMF.

$$\tilde{Q}_{ij}\left(x_{ij}, \beta, m_{ij}, \tilde{\theta}_{ij}\right) = \frac{1}{n_s}\sum_{r=1}^{n_s}\left[\left(\frac{1}{1+\exp\left(-\beta x_i - m_{ij} + \left|\tilde{\theta}_{ij}^r\right|\right)}\right) U\left(\dot{x}_i, -\infty, m_{ij}\right) + \left(\frac{\exp\left(-\beta\left(x_i - m_{ij} - \left|\tilde{\theta}_{ij}^r\right|\right)\right)}{1+\exp\left(-\beta\left(x_i - m_j - \left|\tilde{\theta}_{ij}^r\right|\right)\right)}\right) U\left(\dot{x}_i, m_{ij}, \infty\right)\right] \tag{18}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

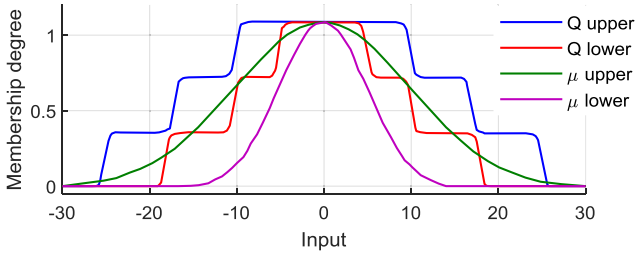


Fig. 4. Comparison of IT2QMFs and Gaussian membership functions.

By applying the modified IT2QMF estimation, the rule significance can be estimated using (31) and the design factor given by

$$\hat{E}_j = \left|\hat{E}_{j,r}\right| + \left|\hat{E}_{j,l}\right|. \tag{31}$$

This adds the hypothetical rule as a new rule R_{K+1} to the network if it contributes over the existing rules. This formulates the condition of network for rule growing criteria as $E_{K+1} \geq \rho \sum_{j=1}^{K} E_j$ with the vigilance parameter $\rho \in (0, 1]$ [42].

B. Parameter Update

The process of adding a hypothetical rule to a network structure is based on the factors of satisfying the conditions of the network structure. If the condition is satisfied, the process is called the initialization of fuzzy rule, and if it is not satisfied, the parameters of the network are adjusting according to the change in input data. This strategy is a winning rule update strategy.

1) Initializing Fuzzy Rules: The growing criterion in Section IV-A estimated the significance of a hypothetical rule according to the change in input data. For a data at nth time step X_n, a new mean is assigned to IT2QMF as follows:

$$m_{K+1} = X_n. \tag{32}$$

This resulted in a new jump position which is achieved by the distance-based formulation. Furthermore, the application of mixed mean of GMM $\hat{v}$ modified the distance as a new one and the updated jump positions are given as

$$\begin{aligned} \overline{\theta}_{i,1} &= \frac{1}{\left(\frac{n_s+1}{2}\right)} r\overline{\sigma}_{i,1} \\ \underline{\theta}_{i,1} &= \frac{1}{\left(\frac{n_s+1}{2}\right)} r\underline{\sigma}_{i,1}. \end{aligned} \tag{33}$$

As the GMM approximates the mean and variance for complex changes in inputs, the eT2QFNN uses the diagonal positive definite matrix of mixed variance as follows:

$$\begin{aligned} &\overline{\sigma}_{i,1} = \hat{\sigma}_i, \underline{\sigma}_{i,1} = \delta_1 \overline{\sigma}_{i,1} \\ &\hat{\Sigma} = \sum_{h=1}^{H} \Sigma_h \cdot v_h, \hat{\Sigma} = \operatorname{diag}\left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_I^2\right) \end{aligned} \tag{34}$$

where δ_1 creates the FOU.

Here, the consequent rule parameters of the hypothetical rule are like the winning rule as given by (35). The process of obtaining the winning rule is discussed in further sections. Finally, if the hypothetical rule satisfies the condition of the network structure, a new rule R_{K+1} is added and its covariance matrix is formulated as in (36). Furthermore, to accommodate new rules, the covariance matrix needs to be adjusted [43], which is achieved by multiplying $(K^2+1)/(K^2)$ with the covariance matrices of other rules as given in (37)

$$\tilde{\Omega}_{K+1} = \tilde{\Omega}_{j_w} \tag{35}$$

$$P_{K+1}(n) = I_{Z\times Z} \tag{36}$$

$$P_j(n) = \left(\frac{K^2+1}{K^2}\right) P_j(n-1). \tag{37}$$

2) Winning Rule Update: If the hypothetical rules do not satisfy the conditions of network, the performance of eT2QFNN needs to be maintained by adjusting the network parameters according to the changes in input data. This is achieved by formulating a winning rule which is dependent on the highest spatial firing strength. The spatial firing strength assesses the satisfaction rate of the rule antecedent part according to changes in the input data. The mathematical formulation for winning rule update is given as follows:

$$j_w = \arg \max_{j=1,\ldots,k} \hat{P}\left(R_j|X\right) \tag{38}$$

$$\tilde{R}_j = \left(\overline{R}_j + \underline{R}_j\right)/2. \tag{39}$$

Previously, the decoupled extended Kalman filter is used to adjust the design factor such that the upper and lower crisp outputs keep adapting to the changes in the inputs. This helps in grouping the local parameters which results in the formation of a block diagonal covariance matrix $P(n)$ (40). For every time step, only one bock covariance matrix is updated, i.e., $P_{j_w}(n)$. This property of localizing parameters enhances the efficiency by maintaining the robustness of the system. Remaining mathematics corresponding to decoupled extended Kalman filter is given in (41)–(43)

$$\tilde{\boldsymbol{P}}(\boldsymbol{n}) = \begin{bmatrix} \boldsymbol{P_1(n)} & \ldots & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \ddots & & & \vdots \\ \mathbf{0} & & \boldsymbol{P_j(n)} & & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{0} & \ldots & \boldsymbol{P_K(n)} \end{bmatrix} \tag{40}$$

$$\begin{aligned} G_{j_w}(n) = {} & P_{j_w}(n-1)H_{j_w}(n) \\ & \times \left[\eta I_{M\times M} + H_{k_w}^{\mathrm{T}}(n) P_{j_w}(n-1) H_{j_w}(n)\right]^{-1} \end{aligned} \tag{41}$$

$$P_{j_w}(n) = \left[I_{Z\times Z} - G_{j_w}(n) H_{j_w}^{\mathrm{T}}(n)\right] P_{j_w}(n-1) \tag{42}$$

$$\vec{\theta}_{j_w}(n) = \vec{\theta}_{j_w}(n-1) + G_{j_w}(n)[t(n) - y(n)] \tag{43}$$

where $G_{j_w}(n)$ is the Kalman gain matrix, H_{k_w} is the Jacobian matrix, and $\vec{\theta}_{j_w}(n)$ is the parameter vector at nth iteration.

$$N_j = \left[\mathcal{N}\left(m_j - v_1; 0, \frac{\Sigma_j}{2} + \Sigma_1\right), \mathcal{N}\left(m_j - v_2; 0, \frac{\Sigma_j}{2} + \Sigma_2\right), \ldots, \mathcal{N}\left(m_j - v_H; 0, \frac{\Sigma_j}{2} + \Sigma_H\right)\right]. \tag{29}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

V. Adaptive Law

To compensate the nonlinearity of the UAV operating with the PID controller, the eT2QFNN developed in Section III is trained with a nonlinear regulator. In this research, a sliding theory-based adaption law is equipped based on the Lipschitz condition [44] to develop a sliding surface $s_P(e, \dot{e})$ as

$$s_p(e, \dot{e}) = \dot{e} + \chi e, \quad \chi > 0. \tag{44}$$

To achieve zero convergence of the parameter error within a finite time, a persistence excitation condition as defined in [45] and [46] is adapted. The design process of the control structured is summarized with a theorem as follows.

Theorem: The consequent part parameters of the eT2QFNN with the adaptation laws are chosen as

$$\dot{\mathcal{F}}(t) = -K_D \mathcal{P}(t)\tilde{\mathcal{W}}(t)s_p(t), \quad \mathcal{F}(0) = \mathcal{F}_0 \in R^{nx1} \tag{45}$$

where $P(t)$ is recursively updated as

$$\dot{\mathcal{P}}(t) = -\mathcal{P}(t)\tilde{\mathcal{W}}(t)\tilde{\mathcal{W}}^{\mathrm{T}}(t)\mathcal{P}(t), \quad \mathcal{P}(0) = \mathcal{P}_0 \in R^{n\times n}. \tag{46}$$

In a compact set where the system under control is of second-order, the sliding theory-based adaption law guarantees the stability of the learning process.

Proof: The proof of the above state is initiated by identifying the control signals associated with the UAV. The output of the PD controller from (14) can be rewritten as follows:

$$y_{PD}(t) = K_D\left(\dot{e}(t) + \frac{K_P}{K_D}e(t)\right). \tag{47}$$

Furthermore, the overall control action is defined by

$$Y = -y_{NN}(t) + y_{PD}(t). \tag{48}$$

The representation of the PD controller output in (47) with sliding surface in (44) gives $\chi = (K_P/K_D)$. Considering this, the control output of PD is defined as

$$y_{PD}(t) = K_D s_P(t) \tag{49}$$

Furthermore, the control output of eT2QFNN from (27) is rewritten as

$$y_{NN}(t) = \sum_{j=1}^{K}\left(\left((1-q_r)\tilde{\bar{R}}_j\bar{\Omega}_j\right) + \left((1-q_l)\underline{\tilde{R}}_j\underline{\Omega}_j\right)\right). \tag{50}$$

Furthermore, by defining $\tilde{\mathcal{W}}(t)$ and $\mathcal{F}(t)$ as

$$\tilde{\mathcal{W}}(t) = \Big[(1-q_r)\tilde{\bar{R}}_1(t) + (1-q_1)\underline{\tilde{R}}_j(t) \ldots (1-q_r)\tilde{\bar{R}}_K(t) + (1-q_l)\underline{\tilde{R}}_K(t)\Big]^{\mathrm{T}} \tag{51}$$

$$\mathcal{F}(t) = \left[\bar{\Omega}_1 + \underline{\Omega}_1 \ldots \bar{\Omega}_K + \underline{\Omega}_K\right]^{\mathrm{T}} \tag{52}$$

(50) can be summarized as

$$y_{NN}(t) = \mathcal{F}^{\mathrm{T}}(t)\tilde{\mathcal{W}}(t). \tag{53}$$

Hence, the cost function with the adaption law can be defined as

$$\mathcal{J}(t) = \int_0^t s_P^2(\tau)d\tau = \frac{1}{K_D^2}\int_0^t (y(\tau) + y_{NN}(\tau))^2 d\tau = \int_0^t \left(y(\tau) + \mathcal{F}^{\mathrm{T}}(t)\tilde{\mathcal{W}}(\tau)\right)^2 d\tau. \tag{54}$$

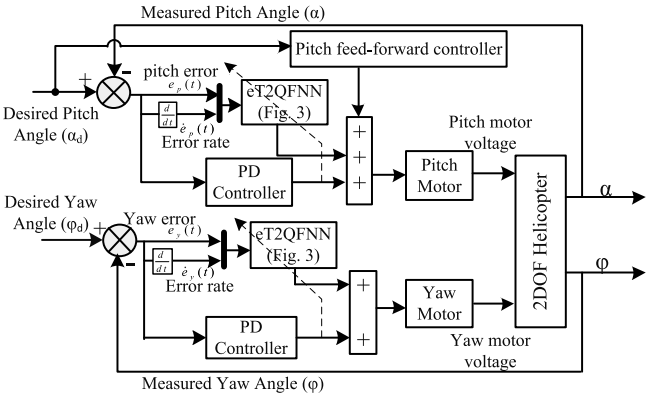


Fig. 5. Block diagram for action of developed controller on 2DoF helicopter system.

Considering the formulations in [47], the gradient of cost function $\mathcal{J}$ concerning $\mathcal{F}$ is derived as

$$\begin{aligned}&\nabla_{\mathcal{F}}\mathcal{J}(t)\\ &\Longrightarrow \int_0^t \tilde{\mathcal{W}}(\tau)y(\tau)d\tau + \mathcal{F}(t)\int_0^t \tilde{\mathcal{W}}(\tau)\tilde{\mathcal{W}}^T(\tau)d\tau = 0\\ &\Longrightarrow \mathcal{F}(t) = -\left[\int_0^t \tilde{\mathcal{W}}(\tau)\tilde{\mathcal{W}}^T(\tau)d\tau\right]^{-1}\int_0^t \tilde{\mathcal{W}}(\tau)y(\tau)d\tau.\end{aligned} \tag{55}$$

Furthermore, $\mathcal{P}$ is defined as

$$\mathcal{P}(t) = \left[\int_0^t \tilde{\mathcal{W}}(\tau)\tilde{\mathcal{W}}^T(\tau)d\tau\right]^{-1} \tag{56}$$

$$\Rightarrow \mathcal{P}^{-1}(t)\mathcal{F}(t) = -\int_0^t \tilde{\mathcal{W}}(\tau)y(\tau)d\tau \tag{57}$$

$$\Rightarrow \dot{\mathcal{P}}(t) = -\mathcal{P}(t)\tilde{\mathcal{W}}(t)\tilde{\mathcal{W}}^{\mathrm{T}}(t)\mathcal{P}(t). \tag{58}$$

From (57), it is observed that $\dot{\mathcal{P}}(t)$ is negative definite and the $\mathcal{P}(t) \in \ell_\infty$ decreases over time. By performing time derivative of (57) and by adapting the algebraic manipulations, the consequent part parameters of the controller are obtained as

$$\begin{aligned}\dot{\mathcal{F}}(t) &= \dot{\mathcal{P}}(t)\mathcal{P}^{-1}(t)\mathcal{F}(t) - \mathcal{P}(t)\tilde{\mathcal{W}}(t)y(t)\\ &= -K_D\mathcal{P}(t)\mathcal{W}(t)s_P(t).\end{aligned} \tag{59}$$

A. Stability Analysis

Since the action of the developed controller is realized by performing trajectory tracking for a 2DoF helicopter, two identical eT2FNN controllers are developed for controlling the pitch and yaw motors as shown in Fig. 5. These eT2FNN controllers operate in parallel with the PD controller which is initially defined in (14). As the eT2QFNN is an approximator of general function, there exists F^* such that

$$y(t) = \mathcal{F}^{*T}\tilde{\mathcal{W}}(t). \tag{60}$$

Defining $\tilde{\mathcal{F}}(t)$ as

$$\tilde{\mathcal{F}}(t) = \mathcal{F}(t) - \mathcal{F}^* \tag{61}$$

$$s_P(t) = \tilde{\mathcal{F}}^T(t)\tilde{\mathcal{W}}(t) \tag{62}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

it is identified that, for the inputs to the fuzzy system $e(t)$ and $\dot{e}(t)$, the approximation of general function is a second order one.

Lemma:

$$\frac{d\left(\mathcal{P}^{-1}\tilde{\mathcal{F}}(t)\right)}{dt} = \dot{\mathcal{P}}^{-1}(t)\tilde{\mathcal{F}}(t) + \mathcal{P}^{-1}(t)\dot{\tilde{\mathcal{F}}}(t) = 0. \tag{63}$$

This indicates $\mathcal{P}^{-1}(t)\tilde{\mathcal{F}}(t)$ does not change over the time and hence $\mathcal{P}^{-1}(t)\tilde{\mathcal{F}}(t) = \mathcal{P}^{-1}(0)\tilde{\mathcal{F}}(0), \forall t > 0$

$$\lim_{t\to\infty} \tilde{\mathcal{F}}(t) = \lim_{t\to\infty} \mathcal{P}(t)\mathcal{P}^{-1}(0)\tilde{\mathcal{F}}(0). \tag{64}$$

Since $\mathcal{P}(t)$ is decreasing and $\tilde{\mathcal{F}}(t) \in \ell_\infty$, $\mathcal{F}(t) \in \ell_\infty$. Considering the following Lyapunov function [48]

$$V = \frac{1}{2}\tilde{\mathcal{F}}^{\mathrm{T}}(t)\mathcal{P}^{-1}\tilde{\mathcal{F}}(t) \tag{65}$$

$$\begin{aligned} V &= \tilde{\mathcal{F}}^{T}(t)\mathcal{P}^{-1}\dot{\tilde{\mathcal{F}}}(t) + \frac{1}{2}\tilde{F}^{\mathrm{T}}(t)\dot{\mathcal{P}}^{-1}\tilde{\mathcal{F}}(t) \\ &= \tilde{\mathcal{F}}^{\mathrm{T}}(t)\tilde{\mathcal{W}}(t)s_P(t) - \frac{1}{2}\tilde{\mathcal{F}}^{\mathrm{T}}(t)\mathcal{P}^{-1}(t)\mathcal{W}(t)\mathcal{W}^{\mathrm{T}}(t)\tilde{\mathcal{F}}(t) \\ &= s_P^2(t) - \frac{1}{2}s_P^2(t) = \frac{1}{2}s_P^2(t). \end{aligned} \tag{66}$$

To perform the stability analysis of the developed controller, some assumptions need to be considered.

Assumption 1: The inputs of eT2FNN and their derivatives are bounded [49], [50]

$$|e(t)| \le B_e, |\dot{e}(t)| \le \dot{B}_e \quad \forall t \tag{67}$$

where X and $\dot{X}$ are the known positive constants.

Assumption 2: The outputs of the eT2FNN controller are bounded [49], [50]

$$|y_{NN}(t)| \le B_y, |\dot{y}_{NN}(t)| \le \dot{B}_y \quad \forall t \tag{68}$$

$$|y_{PD}(t)| \le B_y, |\dot{y}_{PD}(t)| \le \dot{B}_y \quad \forall t \tag{69}$$

$$|y(t)| \le B_y, |\dot{y}(t)| \le \dot{B}_y \quad \forall t \tag{70}$$

where Y and $\dot{Y}$ are the known positive constants.

Assumption 3: An inequality is considered for positive design parameters [51]

$$\Sigma > \dot{y}. \tag{71}$$

Considered that the system inputs and the controller actions are bounded to positive constants, the Lyapunov function candidate in (65) is solved as

$$V = \frac{1}{2K_D^2}y_{PD}^2. \tag{72}$$

By applying differentiation, the following equation is formulated as

$$\dot{V} = \frac{1}{K_D^2}y_{PD}\dot{y}_{PD} = \frac{1}{K_D^2}y_{PD}(\dot{y} + \dot{y}_{NN}). \tag{73}$$

This guarantees the conditions $V > 0$ and $\dot{V} < 0$. For an arbitrary initial condition $e(0) \neq 0$, the output $y_{PD}(t)$ and the error $e(t)$ will approach 0 within finite time t_f. Therefore, from Lyapunov stability theorem, the closed-loop system achieves global stability, when the controller closely tracks the desired trajectory. This also achieves guaranteed position convergence and minimized tracking error.

VI. Experiment

In this section, the eT2QFNN method is evaluated for the 2DoF helicopter system with coupling among output channels. The trajectory of 2DoF helicopter has been defined by pitch angle and yaw angle movements. The main rotor lift is responsible for change in pitch, and pulling forces of tail rotor are responsible for yaw angle movements. The objective is to decouple the two typical output channels such that the output responses follow a desired trajectory. From the dynamic model of the 2DoF helicopter system, it has been identified that the pitch and yaw axes can be controlled by appropriately selecting the voltages of the pitch and yaw motors, respectively. Due to the linear dependence of each rule on the input variables, the developed controller is considered to be ideal for acting as an interpolating supervisor for the PD controller. From the block diagram in Fig. 5, desired and measured pitch and yaw position caused to generate an error which further applied to fuzzy input, where pitch error $e_p(t) = \alpha_d - \alpha$, derivative of pitch error $\dot{e}_p(t)$, yaw error $e_y(t) = \varphi_d - \varphi$, and derivative of yaw error is $\dot{e}_y(t)$. Since the principal idea of developing the eT2QFNN is based on QMF with uncertain jump positions, the controller benefits from flexible rule formulation with nine initial rules. Further, these initial membership functions are designed in a closed interval with the universe of discourse lying between $[-0.2, 0.2]$. This range is further evolved during finding an optimal position. Apart from the rules and universe of discourse, the vigilance parameter for performing rule growing mechanism is set at 0.65 for simplicity, and the FOU is created at 0.7. At the initial instant, two control rules are added at $e_p(t)$ and $e_y(t)$ corresponding to the criterion discussed in algorithm 1. Furthermore, the rule growing process inherently develops the rules based on the heuristic motion of the helicopter system.

A. Simulation Results

The response of pitch and yaw motion of the 2DoF helicopter with eT2QFNN controller is analyzed using Simulink/MATLAB software. The reference trajectory of amplitude 5 is considered for pitch and yaw path tracking assessment. To replicate the behavior of disturbance conditions during the simulation process, an automated white Gaussian noise of 20 dB is considered along with the inputs driving the system. To assess the superiority of the developed controller, the results obtained are compared with conventional interval type-2 FLC (IT2FLC) applied with the same setup. The corresponding results are shown in Fig. 6.

From Fig. 6, it is observed that the pitch angle for eT2QFNN in Fig. 6(a) reach an angle of -5.1° at 0.91 s for the first negative step. The control action is so strong that it can uplift the body within 0.4 s and able to hover very fast. The helicopter body is at rest position and tilted at -40° and so the pitch motor required strong thrust at the start to uplift the body. Due to strong control action at 0.4 s the pitch velocity in Fig. 6(b) reaches to its highest value of 159°/s and will be able to generate enough force as per requirement. The pitch voltage also stops varying after reaching its stable state and gives no oscillations. The pitch voltage varies up to

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

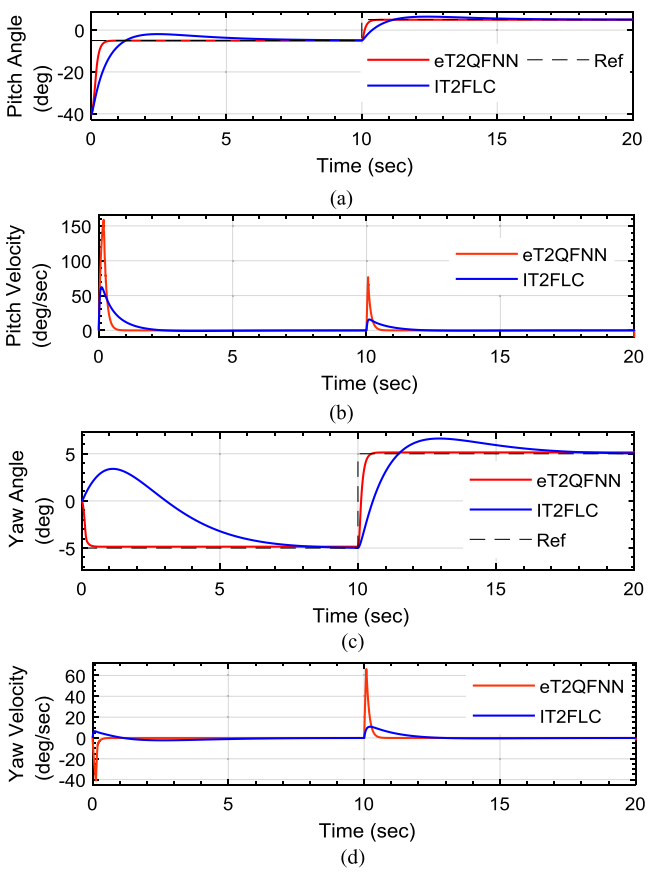


Fig. 6. Output responses of the 2DoF helicopter system for square trajectory. (a) Pitch angle, (b) pitch velocity, (c) yaw angle, and (d) yaw velocity for square trajectory in simulation.

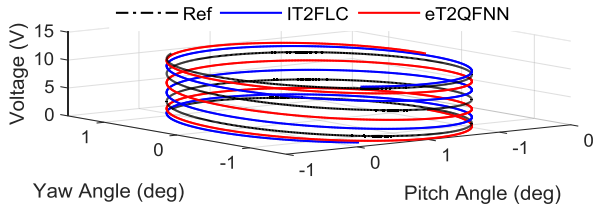


Fig. 7. Output responses of the 2DoF helicopter system for circular helix trajectory.

25 V between 0.2 and 0.4 s and reduces in decreasing order resulting in smooth operation. The yaw angle in Fig. 6(c) for eT2QFNN controller reaches to its trajectory of angle $-4.8°$ at 0.6 s. To maintain the coupling between pitch and yaw rotor, the pitch velocity is comparatively less, and the yaw propeller are rotating with yaw velocity as shown in Fig. 6(d). The variation is observed to be around $-41°$/s at 0.2 s. In this case, the yaw voltage varies up to 3.12 V between 0.1 and 0.3 s. When compared to the conventional IT2FLC, the proposed controller depicted fast response to variation in trajectory with less settling time.

Furthermore, to identify the action of the developed controller two different trajectories are developed by formulating (74)–(77) in MATLAB/Simulink. The development of circular helix trajectory is achieved by controlling the pitch

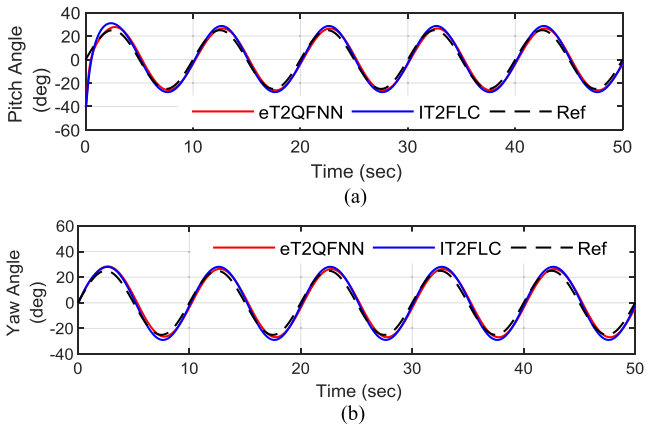


Fig. 8. (a) Variation of pitch angle in following the circular helix trajectory. (b) Variation of yaw angle in following the circular helix trajectory.

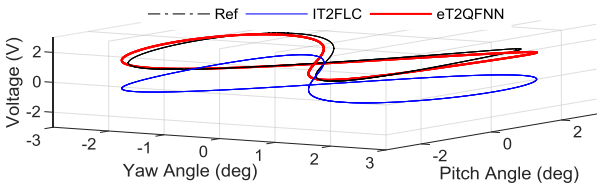


Fig. 9. Output responses of the 2DoF helicopter system for infinite trajectory.

and yaw motors of the helicopter according to (74) and (75), respectively. The corresponding circular helix trajectory and the movement of helicopter along the trajectory are depicted in Fig. 7. Furthermore, the variation of pitch angle and yaw angle for following the defined trajectory is shown in Fig. 8(a) and (b), respectively,

$$\psi_{r1} = \sin(t) \tag{74}$$

$$\varphi_{r1} = \cos(t) \tag{75}$$

where $t = 0 : \text{pi}/50 : 4\,\text{pi}$

The development of infinite trajectory is achieved by controlling the pitch and yaw motors of the helicopter according to (76) and (77), respectively,

$$\psi_{r2} = \cos(t) \tag{76}$$

$$\varphi_{r2} = \frac{\sin(2t)}{2}. \tag{77}$$

The corresponding infinite trajectory and the movement of helicopter along the trajectory are depicted in Fig. 9. Furthermore, the variation of pitch angle and yaw angle with the defined trajectory is shown in Fig. 10(a) and (b), respectively.

Furthermore, the action of the developed controller is tested in the laboratory on a 2DoF helicopter test rig.

B. Real-Time Results

The test rig for 2DoF operation of the unmanned helicopter system is shown in Fig. 11. The test rig is developed by Quanser interactive labs [52], and consists of a helicopter model with two propellers driven by dc motors. The dc motors are power with a volt-PAQ power amplifier which provides ± 24 V for pitch motor control and ± 15 V for yaw motor

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

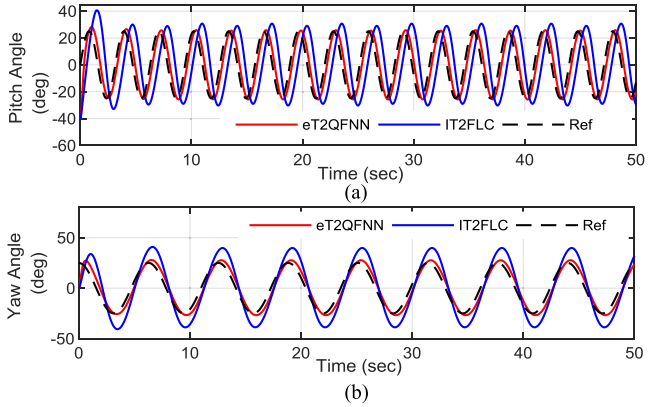


Fig. 10. (a) Variation of pitch angle in following the infinite trajectory. (b) Variation of yaw angle in following the infinite trajectory.

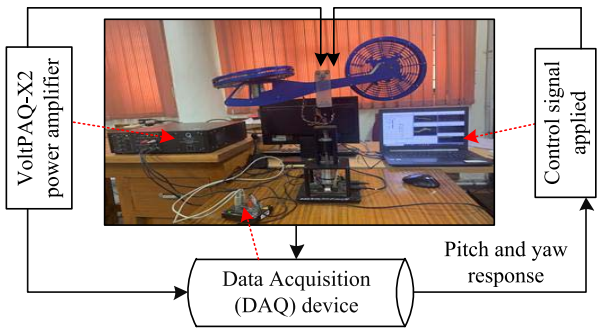


Fig. 11. Experimental setup of a helicopter system.

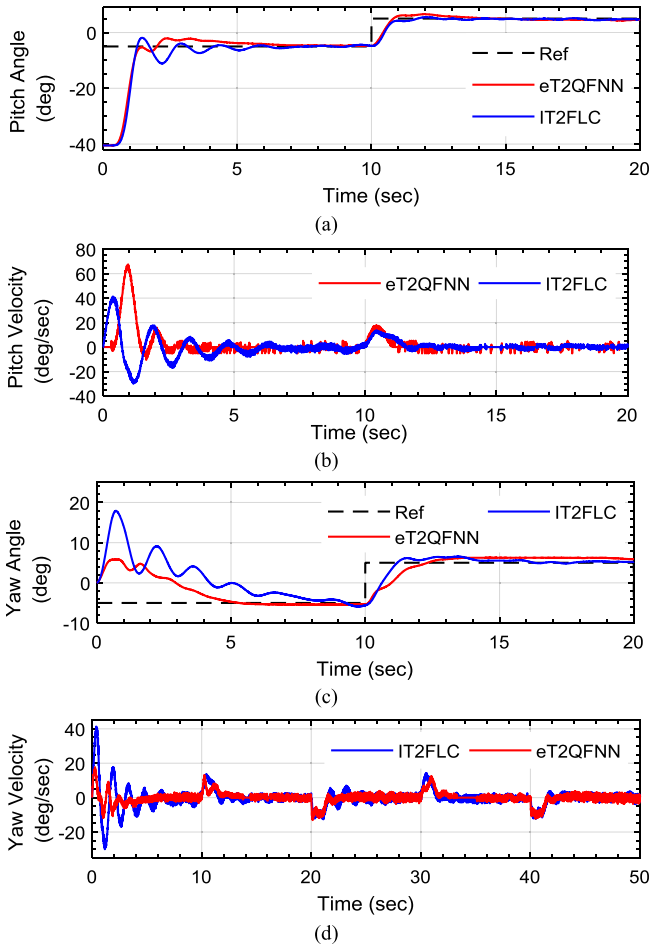


Fig. 12. Output responses of the 2DoF helicopter test rig for square trajectory. (a) Pitch angle, (b) pitch velocity, (c) yaw angle, and (d) yaw velocity for square trajectory in real-time.

control. The helicopter elevation is controlled by the main propeller about the pitch axis and the side motion is controlled by the tail propeller about the yaw axis. Furthermore, the angular variations in pitch and yaw are measured with optical encoders of resolutions 8192 and 4096, respectively. The pitch and yaw angles are measured using high-resolution encoders. The pitch encoder and motor signals are transmitted via a slip ring. The measured angles are acquired into the Quanser-MATLAB integration through a Q8USB-DAQ card. The DAQ has eight digital inputs and modulated outputs with 4−kHz sampling rate. Furthermore, the features from DAQ are merged with the Simulink models through C code generation and the control signals are transferred to the test rig through a hardware-in-the loop application processing interface.

Fig. 12 shows the output response of the pitch angle movement, yaw angle movement, pitch velocity, and yaw velocity for the developed controller on square trajectory. To assess the superiority of the developed controller, the action of IT2FLC on the same setup is carried out and the results obtained are compared. The pitch angle due to action of eT2QFNN in Fig. 12(a) reaches an angle of -5.1° at $t = 1.5$ s for the first time and touches the reference trajectory. Further, it settles to the angle of -4.9° at $t = 5.41$ s and stabilizes the system. The pitch velocity due to the action of eT2QFNN in Fig. 12(b) reaches to its highest value of 69°/s at $t = 0.8$ s. The velocity reduces to 69°/s in real time as no disturbances were considered in simulation. The pitch

TABLE I

TRACKING ERROR COMPARISON OF eT2QFNN-PD AND IT2FLC-PD CONTROLLERS FOR DIFFERENT TRAJECTORIES IN SIMULATION AND REAL TIME

Parameter (→)	**Pitch** (ψ)		**Yaw** (φ)	
Controller(→)	*eT2QFNN-PD*	*IT2FLC-PD*	*eT2QFNN-PD*	*IT2FLC-PD*
Trajectory(↓)				
Square (Sim)	6.32e+02	3.83e+03	1.56e+04	2.19e+4
Circular helix (Sim)	4.31e+04	3.19e+05	3.21e+04	2.47e+5
Infinite (Sim)	1.12e+04	3.28e+04	0.64e+04	2.8e+04
Square (RT)	7.91e+03	4.34e+04	2.12e+04	3.16e+04
Circular helix (RT)	5.81e+04	4.2e+05	4.61e+04	1.15e+05
Infinite (RT)	1.42e+04	4.87e+04	1.09e+04	3.2e+04

*Sim=Simulation, *RT=Real Time

voltage varies between 23–11.2 V between 0.6 and 3.4 s and stops varying after reaching a stable state. Similarly, the yaw angle variation due to action of eT2QFNN in Fig. 12(c) settles at the trajectory of angle -5.8° at $t = 5.6$ s and maintains the constant trajectory. As mentioned earlier, to keep the proper coupling between pitch and yaw rotor, the pitch velocity is comparatively less than the yaw velocity depicted in Fig. 12(d). The yaw velocity goes up to 19°/s at $t = 0.6$ s

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
COMPARISON OF THE DEVELOPED APPROACH WITH THE CONVENTIONAL TECHNIQUES

Control	Properties			
	Adaptive law	*MFs*	*Rule Process*	*Merits and Drawbacks*
Adaptative FNN [28]	Improved Levenberg Marquardt	Gaussian	Self-organizing	*Merits:* Updates the parameters without back propagation, Least identification error, Higher Identification performance. *Drawbacks:* Low accuracy, Less resiliency to uncertainties, and dynamic behavior of the plant.
IT2FNN [26]	Adaptive lumped uncertainty estimation	Gaussian	Self-organizing	
IT2FNN [32]	Particle swarm	Gaussian, Ellipsoidal, and Exponential	Self-organizing	
eT2QFNN (Proposed)	Sliding surface theory, and Extended Kalman filter	Quantum	Rule growing process	*Merits:* Automatic rule generation, Compensates the nonlinearity of the plant, Tackles the drawbacks of inadequacy, uncertainties, and noise

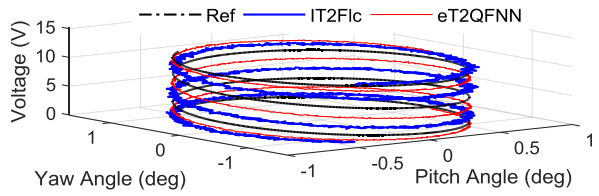


Fig. 13. Output responses of the 2DoF helicopter test rig for circular helix trajectory.

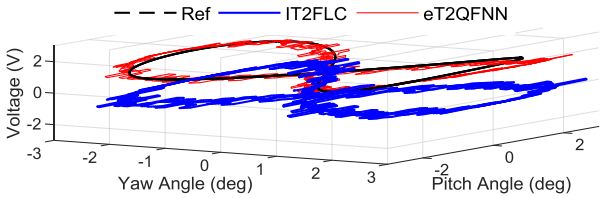


Fig. 14. Output responses of the 2DoF helicopter test rig for infinite trajectory.

and the yaw voltage varies up to 4.3 V between 0.4 and 3 s. In a similar observation, the IT2FLC underperformed while controlling both the pitch and yaw motors. The comparison between action of both the controllers on the test rig depicted that IT2FLC has large settling time, and high oscillations, whereas the proposed controller depicted fast response to variation in trajectory with less settling time.

Furthermore, the reference circular helix trajectory formed from (74) and (75) and the infinite trajectory formed from (76) and (77) are tested for assessing the action of eT2QFNN controller on the test rig and the results are depicted in Figs. 13 and 14, respectively.

Furthermore, to highlight the controller fitness and improvement in the tracking control, the tracking error of both eT2QFNN-PD and IT2FLC-PD for different trajectories is assessed with the integral time absolute error (ITAE) as

$$\text{ITAE} = \int t|e(t)|dt \tag{78}$$

where t is the time and e is the tracking error between reference and measured trajectory. The ITAE for simulation and hardware is illustrated in Table I and the advantages of the developed approach concerning the techniques developed in the literature are discussed in Table II.

VII. CONCLUSION

In this article, the eT2QFNN control scheme is developed for achieving attitude tracking control in UAVs. Initially, the modeling of 2DoF helicopter is discussed from pitch and yaw angle point of view for UAV operation. Furthermore, a parallel structure controller with the conventional PD controller and the eT2QFNN scheme are developed to achieve the attitude tracking. Both pitch motor and yaw motor are individually controlled by operating the controller with a feedback error. The controller generated the required rules through rule growing mechanism and parameter adjustment learning scenario. Furthermore, a sliding surface-based adaptive law is equipped to compensate the nonlinearity of the UAV operating with eT2QFNN and PD controller. The stability of the developed approach is assessed using the Lyapunov stability method. Besides, simulation and experimental analysis are conducted with the developed approach and modeled UAV for offline path tracking with square, circular helical, and infinite shaped trajectories. The corresponding results are compared with the conventional interval type-2 fuzzy logic control technique with Gaussian membership function which depicted the robustness of the proposed approach in achieving attitude tracking. The fitness of the controller while operating in different trajectories is calculated by the means of integral time-weighted absolute error (ITAE) which depicted minimum error while operating with the proposed approach.

The main limitations of the developed approach are the adaptive learning of QMFs comes at a cost of the increased number of parameters that needs to be learned. Setting of initial precomputed rule set for practical conditions is a difficult task due to the unpredictability of the external disturbances. Furthermore, the implementation of the developed approach is only tested for two degree freedom of operation of aerial vehicles and its testing with an increased degree of operation is a planned future work. The research can be extended to improve the tracking efficiency with steady-state performance with additional parameters. The developed winning rule approach will be updated with winners take all rules

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

approach as discussed in [53] to update the quantum fuzzy neural network control process.

REFERENCES

[1] P. E. I. Pounds and A. M. Dollár, "Stability of helicopters in compliant contact under PD-PID control," *IEEE Trans. Robot.*, vol. 30, no. 6, pp. 1472–1486, Dec. 2014, doi: 10.1109/TRO.2014.2363371.

[2] H. Liu, G. Lu, and Y. Zhong, "Robust LQR attitude control of a 3-DOF laboratory helicopter for aggressive maneuvers," *IEEE Trans. Ind. Electron.*, vol. 60, no. 10, pp. 4627–4636, Oct. 2013, doi: 10.1109/TIE.2012.2216233.

[3] V. Bagyaveereswaran and P. Arulmozhivarman, "Gain scheduling of a robust setpoint tracking disturbance rejection and aggressiveness controller for a nonlinear process," *Processes*, vol. 7, no. 7, p. 415, Jul. 2019, doi: 10.3390/pr7070415.

[4] J. P. Ortiz, L. I. Minchala, and M. J. Reinoso, "Nonlinear robust H-infinity PID controller for the multivariable system quadrotor," *IEEE Latin Amer. Trans.*, vol. 14, no. 3, pp. 1176–1183, Mar. 2016, doi: 10.1109/TLA.2016.7459596.

[5] X. C. Méndez Cubillos and L. C. G. de Souza, "Using of H-infinity control method in attitude control system of rigid-flexible satellite," *Math. Problems Eng.*, vol. 2009, May 2009, Art. no. 173145, doi: 10.1155/2009/173145.

[6] B. Zhao, B. Xian, Y. Zhang, and X. Zhang, "Nonlinear robust adaptive tracking control of a quadrotor UAV via immersion and invariance methodology," *IEEE Trans. Ind. Electron.*, vol. 62, no. 5, pp. 2891–2902, May 2015, doi: 10.1109/TIE.2014.2364982.

[7] W. He, T. Wang, X. He, L.-J. Yang, and O. Kaynak, "Dynamical modeling and boundary vibration control of a rigid-flexible wing system," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 6, pp. 2711–2721, Dec. 2020, doi: 10.1109/TMECH.2020.2987963.

[8] W. He, X. Mu, L. Zhang, and Y. Zou, "Modeling and trajectory tracking control for flapping-wing micro aerial vehicles," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 1, pp. 148–156, Jan. 2021, doi: 10.1109/JAS.2020.1003417.

[9] W. He, T. Meng, X. He, and C. Sun, "Iterative learning control for a flapping wing micro aerial vehicle under distributed disturbances," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1524–1535, Apr. 2019, doi: 10.1109/TCYB.2018.2808321.

[10] J. Cheng and L. Chen, "The fuzzy neural network control scheme with H_∞ tracking characteristic of space robot system with dual-arm after capturing a spin spacecraft," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 5, pp. 1417–1424, Sep. 2020, doi: 10.1109/JAS.2018.7511180.

[11] Y. Cao and J. Huang, "Neural-network-based nonlinear model predictive tracking control of a pneumatic muscle actuator-driven exoskeleton," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 6, pp. 1478–1488, Nov. 2020, doi: 10.1109/JAS.2020.1003351.

[12] S. Li, M. Zhou, and X. Luo, "Modified primal-dual neural networks for motion control of redundant manipulators with dynamic rejection of harmonic noises," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4791–4801, Oct. 2018, doi: 10.1109/TNNLS.2017.2770172.

[13] W. He, C. Xue, X. Yu, Z. Li, and C. Yang, "Admittance-based controller design for physical human–robot interaction in the constrained task space," *IEEE Trans. Automat. Sci. Eng.*, vol. 17, no. 4, pp. 1937–1949, Oct. 2020, doi: 10.1109/TASE.2020.2983225.

[14] X. Zhang, S. He, V. Stojanovic, X. Luan, and F. Liu, "Finite-time asynchronous dissipative filtering of conic-type nonlinear Markov jump systems," *Sci. China Inf. Sci.*, vol. 64, no. 5, May 2021, Art. no. 152206, doi: 10.1007/s11432-020-2913-x.

[15] L. Zhou, H. Tao, W. Paszke, V. Stojanovic, and H. Yang, "PD-type iterative learning control for uncertain spatially interconnected systems," *Mathematics*, vol. 8, no. 9, p. 1528, Sep. 2020, doi: 10.3390/math8091528.

[16] X. Yu, W. He, H. Li, and J. Sun, "Adaptive fuzzy full-state and output-feedback control for uncertain robots with output constraint," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Feb. 3, 2020, doi: 10.1109/TSMC.2019.2963072.

[17] W. He and Y. Dong, "Adaptive fuzzy neural network control for a constrained robot using impedance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1174–1186, Apr. 2018, doi: 10.1109/TNNLS.2017.2665581.

[18] M. Innocenti, L. Pollini, and D. Turra, "A fuzzy approach to the guidance of unmanned air vehicles tracking moving targets," *IEEE Trans. Control Syst. Technol.*, vol. 16, no. 6, pp. 1125–1137, Nov. 2008, doi: 10.1109/TCST.2008.917224.

[19] B. Celen and Y. Oniz, "Trajectory tracking of a quadcopter using fuzzy logic and neural network controllers," in *Proc. 6th Int. Conf. Control Eng. Inf. Technol. (CEIT)*, Oct. 2018, pp. 1–6, doi: 10.1109/CEIT.2018.8751810.

[20] C. Fu, A. Sarabakha, E. Kayacan, C. Wagner, R. John, and J. M. Garibaldi, "Input uncertainty sensitivity enhanced nonsingleton fuzzy logic controllers for long-term navigation of quadrotor UAVs," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 2, pp. 725–734, Apr. 2018, doi: 10.1109/TMECH.2018.2810947.

[21] Z. Chen, B. Zhang, V. Stojanovic, Y. Zhang, and Z. Zhang, "Event-based fuzzy control for T-S fuzzy networked systems with various data missing," *Neurocomputing*, vol. 417, pp. 322–332, Dec. 2020, doi: 10.1016/j.neucom.2020.08.063.

[22] H. A. Hagras, "A hierarchical type-2 fuzzy logic control architecture for autonomous mobile robots," *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 4, pp. 524–539, Aug. 2004, doi: 10.1109/TFUZZ.2004.832538.

[23] Y. Pan and G.-H. Yang, "Event-triggered fuzzy control for nonlinear networked control systems," *Fuzzy Sets Syst.*, vol. 329, pp. 91–107, Dec. 2017, doi: 10.1016/j.fss.2017.05.010.

[24] J. I. Mulero-Martinez, "Robust GRBF static neurocontroller with switch logic for control of robot manipulators," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1053–1064, Jul. 2012, doi: 10.1109/TNNLS.2012.2196053.

[25] F. Gul, W. Rahiman, and S. S. N. Alhady, "A comprehensive study for robot navigation techniques," *Cogent Eng.*, vol. 6, no. 1, pp. 1–25, 2019, doi: 10.1080/23311916.2019.1632046.

[26] F.-J. Lin and P.-H. Chou, "Adaptive control of two-axis motion control system using interval type-2 fuzzy neural network," *IEEE Trans. Ind. Electron.*, vol. 56, no. 1, pp. 178–193, Jan. 2009, doi: 10.1109/TIE.2008.927225.

[27] G. Wang, Q.-S. Jia, J. Qiao, J. Bi, and C. Liu, "A sparse deep belief network with efficient fuzzy learning framework," *Neural Netw.*, vol. 121, pp. 430–440, Jan. 2020, doi: 10.1016/j.neunet.2019.09.035.

[28] J. Qiao and H. Zhou, "Modeling of energy consumption and effluent quality using density peaks-based adaptive fuzzy neural network," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 5, pp. 968–976, Sep. 2018, doi: 10.1109/JAS.2018.7511168.

[29] L. Liu, Z. Wang, and H. Zhang, "Adaptive fault-tolerant tracking control for MIMO discrete-time systems via reinforcement learning algorithm with less learning parameters," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 299–313, Jan. 2017, doi: 10.1109/TASE.2016.2517155.

[30] L. Liu, Z. Wang, X. Yao, and H. Zhang, "Echo state networks based data-driven adaptive fault tolerant control with its application to electromechanical system," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 3, pp. 1372–1382, Jun. 2018, doi: 10.1109/TMECH.2018.2817495.

[31] L. Liu, Z. Wang, Z. Huang, and H. Zhang, "Adaptive predefined performance control for MIMO systems with unknown direction via generalized fuzzy hyperbolic model," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 3, pp. 527–542, Jun. 2017, doi: 10.1109/TFUZZ.2016.2566803.

[32] J. Wang and T. Kumbasar, "Parameter optimization of interval type-2 fuzzy neural networks based on PSO and BBBC methods," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 1, pp. 247–257, Jan. 2019, doi: 10.1109/JAS.2019.1911348.

[33] D. Liu and G.-H. Yang, "Prescribed performance model-free adaptive integral sliding mode control for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2222–2230, Jul. 2019, doi: 10.1109/TNNLS.2018.2881205.

[34] C.-H. Chen, C.-J. Lin, and C.-T. Lin, "An efficient quantum neuro-fuzzy classifier based on fuzzy entropy and compensatory operation," *Soft Comput.*, vol. 12, no. 6, pp. 567–583, Apr. 2008, doi: 10.1007/s00500-007-0229-0.

[35] L. M. Belmonte, R. Morales, A. Fernández-Caballero, and J. A. Somolinos, "Nonlinear cascade-based control for a twin rotor MIMO system," in *Nonlinear Systems—Design, Analysis, Estimation and Control*. Rijeka, Croatia: InTech, 2016.

[36] Y. Arkun and J. Downs, "A general method to calculate input-output gains and the relative gain array for integrating processes," *Comput. Chem. Eng.*, vol. 14, no. 10, pp. 1101–1110, Oct. 1990, doi: 10.1016/0098-1354(90)85006-V.

[37] M. Biglarbegian, W. W. Melek, and J. M. Mendel, "On the stability of interval type-2 TSK fuzzy logic control systems," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 40, no. 3, pp. 798–818, Jun. 2010, doi: 10.1109/TSMCB.2009.2029986.

[38] S. Ishikawa, "A quantum mechanical approach to a fuzzy theory," *Fuzzy Sets Syst.*, vol. 90, no. 3, pp. 277–306, Sep. 1997, doi: 10.1016/S0165-0114(96)00114-5.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[39] U. S. Tiwary. (2009). *Fuzzy Sets (Type-1 and Type-2) and Their Applications*. [Online]. Available: http://www.iitk.ac.in/ee/courses/archives/2013/intel-info/d1pdf3.pdf
[40] W. W. Tan and D. Wu, "Design of type-reduction strategies for type-2 fuzzy logic systems using genetic algorithms," in *Advances in Evolutionary Computing for System Design* (Studies in Computational Intelligence), vol. 66, L. C. Jain, V. Palade, and D. Srinivasan, Eds. Berlin, Germany: Springer, 2007, doi: 10.1007/978-3-540-72377-6_7.
[41] R. Chellappa *et al.*, "Gaussian mixture models," *Encyclopedia Biometrics*, no. 2, pp. 659–663, 2009, doi: 10.1007/978-0-387-73003-5_196.
[42] L. E. Brito da Silva, I. Elnabarawy, and D. C. Wunsch, "A survey of adaptive resonance theory neural network models for engineering applications," *Neural Netw.*, vol. 120, pp. 167–203, Dec. 2019, doi: 10.1016/j.neunet.2019.09.012.
[43] M. Kordestani, M. Dehghani, B. Moshiri, and M. Saif, "A new fusion estimation method for multi-rate multi-sensor systems with missing measurements," *IEEE Access*, vol. 8, pp. 47522–47532, 2020, doi: 10.1109/ACCESS.2020.2979222.
[44] Z. Kai, S. Tielong, and Y. Yu, "Filippov solutions on a Lipschitz continuous surface," in *Proc. 27th Chin. Control Conf.*, Jul. 2008, pp. 577–581, doi: 10.1109/CHICC.2008.4605726.
[45] M. N. Mahyuddin, S. G. Khan, and G. Herrmann, "A novel robust adaptive control algorithm with finite-time online parameter estimation of a humanoid robot arm," *Robot. Auto. Syst.*, vol. 62, no. 3, pp. 294–305, Mar. 2014, doi: 10.1016/j.robot.2013.09.013.
[46] C. Cao, A. M. Annaswamy, and A. Kojic, "Parameter convergence in nonlinearly parameterized systems," *IEEE Trans. Autom. Control*, vol. 48, no. 3, pp. 397–412, Mar. 2003, doi: 10.1109/TAC.2003.809146.
[47] E. Sanchez and A. Alanis, "Adaptive approximation based control: Unifying neural, fuzzy and traditional adaptive approximation approaches (Farrell, J.A. and Polycarpou, M.M. [Book review])," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 731–732, Apr. 2008, doi: 10.1109/TNN.2008.921577.
[48] J.-B. Hu, G.-P. Lu, S.-B. Zhang, and L.-D. Zhao, "Lyapunov stability theorem about fractional system without and with delay," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 20, no. 3, pp. 905–913, Mar. 2015, doi: 10.1016/j.cnsns.2014.05.013.
[49] E. Kayacan, E. Kayacan, and M. Ahmadieh Khanesar, "Identification of nonlinear dynamic systems using type-2 fuzzy neural networks—A novel learning algorithm and a comparative study," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1716–1724, Mar. 2015, doi: 10.1109/TIE.2014.2345353.
[50] E. Kayacan and O. Kaynak, "Sliding mode control theory-based algorithm for online learning in type-2 fuzzy neural networks: Application to velocity control of an electro hydraulic servo system," *Int. J. Adapt. Control Signal Process.*, vol. 26, no. 7, pp. 645–659, Jul. 2012, doi: 10.1002/acs.1292.
[51] J. S. Arora, "Optimum design problem formulation," in *Introduction to Optimum Design*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 19–70.
[52] *2-DOF Helicopter*, Quanser, Markham, ON, Canada, 2011.
[53] S. Li, M. Zhou, X. Luo, and Z.-H. You, "Distributed Winner-take-all in dynamic networks," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 577–589, Feb. 2017, doi: 10.1109/TAC.2016.2578645.

Rupam Singh (Graduate Student Member, IEEE) received the B.Tech. degree in electrical and electronics engineering from the Hindustan College of Science & Technology, Mathura, India, in 2013, the M.Tech. degree in control system from Amity University, Noida, India, in 2016, and the Ph.D. degree in intelligent control and robotics from the Department of Electrical Engineering, Delhi Technological University, New Delhi, India, in 2021.

She is currently a Pre-Doctoral Fellow with the Institute for Intelligent System Technologies, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria. She has many publications in peer-reviewed journals and presented her research articles in several international conferences. Her area of research is artificial intelligence, machine learning, control systems, condition monitoring, and their application in robotics and unmanned vehicles.

Bharat Bhushan (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Delhi College of Engineering, University of Delhi, New Delhi, India, in 1992, 1996, and 2012, respectively.

He is currently a Professor with the Department of Electrical Engineering, Delhi Technological University, New Delhi. His area of research is control systems, intelligent control, soft computing techniques, and bioinspired algorithms.

Dr. Bhushan is a Life Member of ISTE, a fellow of IETE and IE (India) and published many research articles in conferences and journals. He has organized the First IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES) in 2016 at Delhi Technological University. He also has organized two faculty development programs on Intelligent Control Techniques, and Their Applications in 2014 and Nature-Inspired Algorithms and Their Applications in 2015.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:55:21 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/350616440

Experimental investigations of performance and smoke emission of Karanja biodiesel blends & diesel in a diesel engine

Conference Paper · April 2021

CITATIONS
0

READS
9

4 authors, including:

Sanjay Mohite
National Institute of Technology, Kurukshetra
20 PUBLICATIONS **24** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Biodiesel View project

PERFORMANCE CHARACTERISTICS AND ANALYSIS OF DIESEL & BIODIESEL BLENDS USING ENERGY AUDIT TECHNIQUE View project

All content following this page was uploaded by Sanjay Mohite on 04 April 2021.

The user has requested enhancement of the downloaded file.

Experimental investigations of performance and smoke emission of Karanja biodiesel blends & diesel in a diesel engine

Sanjay Mohite
National Institute of Technology
Kurukshetra
Haryana
India

Dr. Sagar Maji
Delhi Technological University
New Delhi
India

Dr. Sudhir Kumar
National Institute of Technology
Kurukshetra
Haryana
India

Dr. Amit Pal
Delhi Technological University
New Delhi
India

ABSTRACT

An experimental investigation has been carried out to analyze the performance & smoke emission characteristics of diesel engine fuelled with Karanja biodiesel blends (10%, 20% & 30%) and diesel at constant speed. Brake thermal efficiencies of Karanja biodiesel blends have been found to be comparable with diesel. Increase in brake specific fuel consumption of Karanja biodiesel blends has been observed while comparing with diesel. Exhaust gas temperature of Karanja biodiesel blends has also increased. Slight improvement has been observed in mechanical efficiencies of Karanja biodiesel blends with respect to diesel. Considerable reduction has been found in smoke opacity of Karanja biodiesel blends while comparing with diesel. Experimental investigation has revealed that B20 biodiesel blend with diesel can be used in an unmodified diesel engine in an effective manner.

INTRODUCTION

Petroleum resources are depleted rapidly due to increase in its uses which may create severe energy crisis in near future. Researchers are finding vegetable oil as alternative fuel which is renewable also [1]. Vegetable oil can not be used as fuel in a diesel engine because of its high viscosity. Unsaturated fatty acids and high molar masses are the main causes for high viscosity of vegetable oils. If it is used in diesel engine, it may cause poor atomization of fuel resulting in incomplete combustion, carbon deposits on injector and valve seat, causing engine fouling, injector choking and dilution of lubrication oil [2].

Biodiesel is mono alkyl fatty acid ester which is produced from vegetable oils or animal fats [3]. Biodiesel is produced by the process of transesterification of vegetable oils with alcohol in the presence of catalyst [4]. Biodiesel has certain advantages over diesel as higher cetane number, absence of aromatic compounds and sulphur and also has about 10% oxygen [5].

Indian forests have large varieties of non-edible oilseed plants such as Karanja, Kusum, Ratanjyot, Neem, Pilu, Sal, Mahua, Phulware, Simarouba and many more. Karanja is also one of the non-edible oils which is available in India. Pongamia glabra is its botanical name [6].

Karanja tree is drought resistant and it can tolerate salinity. This can be grown on soils of various types like stony, sandy and clay and areas having annual rainfall ranging from 500 mm to 2500 mm are suitable for its cultivation [7]. Annual production of karanja oil is 1, 35,000 metric tons in India. Its seed contains 30-40% oil content [8].

Thermal efficiency of karanja biodiesel is found to be lower than that of diesel. This may be attributed due to low volatility and higher viscosity causing poor atomization and combustion [9]. Higher BSFC was observed for higher biodiesel blend than diesel and this may be attributed to lower heating value of biodiesel [10]. Exhaust gas temperature of an engine may be a representation of heat conversion into work. Because of high viscosity, Karanja biodiesel results in poor combustion causing higher exhaust gas temperature than that of diesel. It was also investigated that BTE and mechanical efficiencies were better than diesel for some biodiesel blends under certain loads [11]. Smoke opacity of biodiesel blends were reported to be lower than that of diesel [12].

NOMENCLATURE

BTE: Brake Thermal Efficiency
EGT: Exhaust Gas Temperature
BSFC: Brake Specific Fuel Consumption
B10: 10% Biodiesel blend by volume with diesel
B20: 20% Biodiesel blend by volume with diesel
B30: 30% Biodiesel blend by volume with diesel

METHODOLOGY

The experiment of performance & smoke emission characteristics were conducted on a single cylinder, four stroke, water cooled, direct injection diesel engine. The engine was connected to eddy current dynamometer for loading. The diesel engine has stroke of 110mm and bore of 87.5 mm. The experiment used stand- alone panel box consisting of air box, dual fuel tank for fuel test, manometer, fuel measuring unit, transmitters for air & fuel flow measurements, process indicator and engine indicator. Rotameters were provided for cooling water and calorimeter water flow measurement. Lab view based Engine Performance Analysis software package “Engine soft LV” was used for on line performance evaluation.

Blends of Karanja biodiesel with diesel fuel were prepared in the Laboratory. The blends were prepared on the volume percentage basis of biodiesel and diesel. Three blends 10% (B10), 20% (B20) and 30% (B30) biodiesel with diesel were used.

Readings were observed by running the engine with fuel at compression ratio of 18 and varying the load from 0.5 kW to 3.5 kW load in steps of 0.5 kW at constant speed of 1500 rpm. ‘Engine Soft’ software was used to record the engine performance characteristics. Then, the engine was run on these blends and the parameters were recorded as above. The calorific value and density values of the fuels were entered accordingly in the Engine Soft software. The smoke opacity was recorded by AVL Smoke meter.

RESULTS AND DISCUSSIONS

Various performance and smoke emission parameters were obtained and the results of this experiment are discussed as given below:-

Brake thermal efficiency increases due to reduction of heat loss and increase in brake power at higher loads as shown in figure 1. In case of B10 and B30 biodiesel blends, slight reduction in thermal efficiencies are observed due to lower heating value as compared to diesel, poor volatility and high viscosity of biodiesel which results in poor atomization of fuel leading to slow combustion. But Brake thermal efficiency for B20 has been observed slightly higher than diesel and this may be attributed due to proper combustion of B20 blends due to its oxygenated nature.

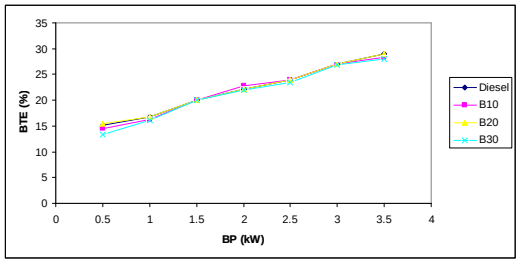


Fig.1. Variation of brake thermal efficiency with brake power

BSFC decreases with increase in load (brake power) as shown in figure 2. In all cases, the BSFC of biodiesel blends are higher than diesel due to its low calorific value and higher density. Higher mass of fuel is injected for same volume at the same injection pressure of injection pump for biodiesel as compared to diesel due to its higher density of biodiesel blends. The BSFC depends on specific gravity, viscosity and calorific value of fuel. Lower calorific value of biodiesel fuel requires more amount of fuel to produce the same amount of energy as that of diesel and therefore the BSFC of biodiesel increases as compared to diesel.

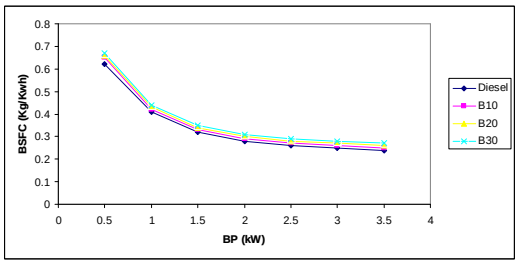


Fig.2. Variation of brake specific fuel consumption with brake power

It is observed that mechanical efficiencies go on increasing from low loads to high loads for diesel and all biodiesel blends as shown in figure 3. B30 blend has higher mechanical efficiency as compared to diesel, B10 and B20. This may be attributed due to high lubricity quality of biodiesel which reduces friction thereby increasing mechanical efficiency. Therefore, mechanical efficiency depends upon the proportions of biodiesel and it increases with the increase in proportions of biodiesel in blends.

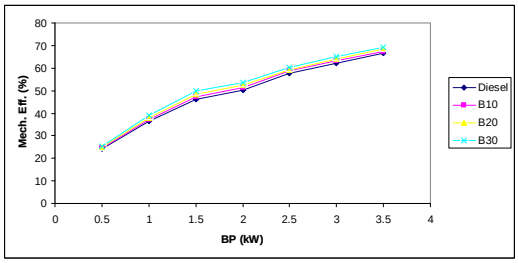


Fig.3. Variation of mechanical efficiency with brake power

It is observed that exhaust gas temperature goes on increasing with respect to increase of load as shown in figure 4. Exhaust gas temperature increases with increase in loads as more fuel is required to generate more brake power. Exhaust gas temperature is also increased with increase in compositions of biodiesel blends due to lower heat transfer rate in biodiesel which may be understood from lower brake thermal efficiency of biodiesel. But B20 has lowest EGT and this may be attributed due to its higher brake thermal efficiency than diesel.

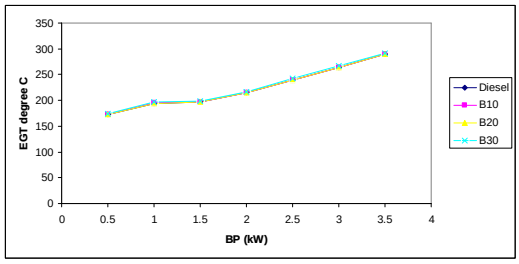


Fig.4. Variation of exhaust gas temperature with brake power.

It is observed that smoke opacity goes on increasing with respect to increase of load as shown in figure 5. This may be attributed due to more fuel injection and their combustion at higher loads. Decrease in the smoke opacity has been observed while using blends of biodiesel as compare to diesel. The reduction in smoke opacity has been observed when the proportion of biodiesel in blends increases. B30 biodiesel blends show least level of smoke opacity among all fuels used in this experiment. This may be attributed due to oxygenated nature of biodiesel which leads to proper combustion.

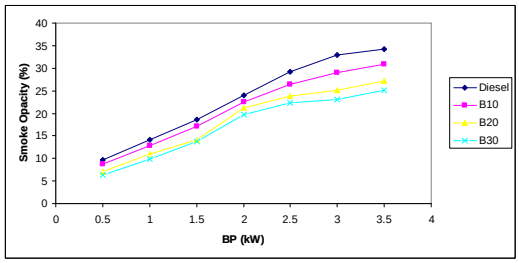


Fig.5. Variation of smoke opacity with brake power

CONCLUSIONS

Experimental investigation has revealed that B20 karanja biodiesel blend with diesel can be used in an unmodified diesel engine in an effective manner which has slightly higher brake thermal efficiency and considerable reduction in smoke opacity than diesel. B20 has also slightly higher brake specific fuel consumption and mechanical efficiency and slightly lower exhaust gas temperature than that of diesel fuel.

REFERENCES

[1] Jidon Janaun, Naoko Ellis: Perspectives on bodiesel as a sustainable fuel. Renewable & Sustainable energy reviews. 14, 1312-1320 (2010).

[2] R.D. Misra, M.S.Murthy: Straight vegetable oils usage in a compression ignition engine- A review. Renewable and Sustainable Energy Reviews. 14, 3005-3013 (2010).

[3] A.B.Chhetri and M.R.Islam: Towards producing a truly green biodiesel. Energy Sources Part-A. 30,754-764 (2008).

[4] L.C.Meher, Vidya S.S.Dharmagadda, S.N.Naik: Optimization of alkali-catalyzed transesterification of Pongamia pinnata oil for production of biodiesel. Bioresource Technology. 97, 1392-1397 (2006).

[5] Mustafa Balat, Havva Balat: Progress in biodiesel Processing. Applied Energy. 87, 1815-1835 (2010).

[6] Lin Lin, Zhou Cunshan, Saritporn Vittayapadung, Shen Xiangqian, Dong Mingdong: Opportunities & challenges for biodiesel fuel. Applied Energy. 88, 1020-1031(2011).

[7] L.C.Meher, S.N.Naik and L.M.Das: Methanolysis of Pongamia pinnata (karanja) oil for production of biodiesel. Journal of Scientific and Industrial Research. 63,913-918 (2004).

[8] Avinash kumar Agarwal, K.Rajamanoharan: Experimental investigations of performance and emissions of karanja oil and its blends in a single cylinder agriculture diesel engine. Applied Energy. 86, 106-112 (2009).

[9] P.K.Srivastava, Madumita Verma: Methyl ester of karanja oil as an alternative renewable source energy. Fuel. 87, 1673-1677 (2008).

[10] Atul Dhar, Avinash Kumar Agarwal: Performance, emission and combustion characteristics of karanja biodiesel in a transportation engine. Fuel. 119, 70-80 (2014).

[11] Saswat Rath, Sachin Kumar & R.K.Singh: Performance & emission analysis of blends of Karanja methyl ester with diesel in a compression ignition engine. International Journal of Ambient Energy. 32, 161-166 (2011).

[12] Ekrem Buyukkaya: Effect of biodiesel on a DI diesel engine performance, emission and combustion characteristics. Fuel. 89, 3099-3105 (2010).

View publication stats

Journal of Rock Mechanics and Geotechnical Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Rock Mechanics and Geotechnical Engineering

journal homepage: www.jrmge.cn

Full Length Article

Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique

Muhammad Nouman Amjad Raja [a,b,*], Sanjay Kumar Shukla [a,c]

[a] Geotechnical and Geoenvironmental Engineering Research Group, School of Engineering, Edith Cowan University, Joondalup, Perth, WA 6027, Australia
[b] Department of Civil Engineering, School of Engineering, University of Management and Technology, Lahore, Pakistan
[c] Civil Engineering Department, Delhi Technological University, Delhi, India

ARTICLE INFO

Article history:
Received 16 September 2020
Received in revised form
28 December 2020
Accepted 23 February 2021
Available online xxx

Keywords:
Geosynthetics
Repeated loads
Wraparound reinforcement technique
Model tests
Footing settlement
Bearing capacity
Reinforcement ratio
Strip footing

ABSTRACT

In the recent past, the potential benefits of wraparound geosynthetic reinforcement technique for constructing the reinforced soil foundations have been reported. This paper presents the experimental study on the behaviour of model strip footing resting on sandy soil bed reinforced with geosynthetic in wraparound and planar forms under monotonic and repeated loadings. The geosynthetic layers were laid according to the reinforcement ratio to minimise the scale effect. It is found that for the same amount of reinforcement material, the wraparound reinforced model resulted in less settlement in comparison to planar reinforced models. The efficiency of wraparound reinforced model increased with the increase in load amplitude and the rate of total cumulative settlement substantially decreased with the increase in number of load cycles. The wraparound reinforced model has shown about 45% lower average total settlement in comparison to unreinforced model, while the double-layer reinforced model has about 41% lower average total settlement at the cost of approximately twice the material and 1.5 times the occupied land width ratio. Moreover, wraparound models have shown much greater stability in comparison to their counterpart models when subjected to incremental repeated loading.

© 2021 Institute of Rock and Soil Mechanics, Chinese Academy of Sciences. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In the current construction practice, the use of geosynthetic-reinforced soil is increasing rapidly for making the projects, including geosynthetic-reinforced soil foundations, sustainable. During the past four decades, an extensive research has been carried out to evaluate the behaviour of shallow reinforced soil foundations subjected to monotonic loads (Khing et al., 1993; Omar et al., 1993; Yetimoglu et al., 1994; Adams and Collin, 1997; Basudhar et al., 2007; Abu-Farsakh et al., 2013; Kumar and Sahoo, 2013; Roy and Deb, 2017; Li et al., 2019; Raja and Shukla, 2020a, 2021). A limited number of studies were conducted on the reinforced soil bed under cyclic loading by simulating either the vehicle (tire) or train load or the combination of static and cyclic loads (Yeo et al., 1993; Leng and Gabr, 2002; Raymond, 2002; Das and Shin, 2009; Moghaddas Tafreshi and Dawson, 2010, 2012a; Boushehrian et al., 2011; Asakereh et al., 2013; Moghaddas Tafreshi et al., 2014; Wang et al., 2018; Ghotbi Siabil et al., 2020). The main objective of majority of these studies was to determine and optimise the most critical parameters which play a significant role in increasing the ultimate bearing capacity and reducing the settlement of the reinforced foundations, such as distance of the topmost reinforcement layer from the base of the footing (u), vertical spacing between the successive reinforcement layers (h), number of reinforcement layers (N), tensile modulus of the reinforcement (J), and width of the reinforcement layer (b). In most cases, the optimum value of the reinforcement width has been reported to be in the range of 4–8 times the footing width (B) (Fragaszy and Lawton, 1984; Omar et al., 1993; Shin and Das, 2000; Michalowski, 2004; Chakraborty and Kumar, 2014; Aria et al., 2020). However, often the field boundary conditions such as property line dispute and limited land space do not permit this optimum reinforcement width. In order to overcome this problem and to achieve the desired bearing capacity, multilayer reinforcements, 3D-geocells, grid-anchors, and horizontal-vertical reinforcements have been suggested as the possible alternative solutions (Latha and Murthy, 2007; Biswas et al., 2013; Mosallanezhad et al., 2016; Hou et al., 2017). In spite of these developments, the practical difficulties in manufacturing process,

* Corresponding author.
E-mail addresses: noumanamjad@live.com, m.raja@ecu.edu.au (M.N.A. Raja).
Peer review under responsibility of Institute of Rock and Soil Mechanics, Chinese Academy of Sciences.

https://doi.org/10.1016/j.jrmge.2021.02.001
1674-7755 © 2021 Institute of Rock and Soil Mechanics, Chinese Academy of Sciences. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

field installation and cost-effectiveness of these techniques are questionable especially when the budget is limited.

In the recent past, simple yet innovative and effective wraparound geosynthetic reinforcement technique has been proposed and implemented for enhancing the ultimate bearing capacity and reducing the settlement of reinforced foundations (Shukla, 2014, 2016). For this, the ends of the reinforcement layer are folded along the longitudinal axis. In this way, the occupied land area outside the footing width is reduced and soil above the reinforcement is confined, thus preventing its lateral movement. Kazi et al. (2015a,b, 2016) experimentally evaluated the ultimate bearing capacity of the strip footing resting on wraparound geotextile-reinforced sandy soil bed. They reported the increase in the ultimate bearing capacity of approximately 12%–30% and 65%–80% in comparison to those of model foundations reinforced with planar geotextile layer without wraparound ends and unreinforced models, respectively. The results of the numerical simulations conducted by Benmebarek et al. (2018) depicted that the wraparound geosynthetic-reinforced foundations offered the maximum improvement of bearing capacity for a strip footing under monotonic loading. Aria et al. (2019, 2021) experimentally and numerically investigated the behaviour of strip footing resting on soil bed strengthened by wraparound geotextile reinforcement technique. Test results and model simulations indicated that the bearing capacity ratio (BCR) of the model foundation generally increases with the increase in lap–length ratio and occupied land width ratio of wraparound reinforcement. Most recently, Raja and Shukla (2020b) developed an analytical expression to estimate the ultimate bearing capacity of wraparound geosynthetic-reinforced soil foundations. In spite of all these studies, the behaviour of the footing resting on soil bed strengthened by wraparound geosynthetic reinforcement technique subjected to repeated loading is yet to be investigated. The main focus of this study is to experimentally evaluate the load–settlement response of the strip footing on reinforced and unreinforced sand beds under slow repeated loads, and also exclusively, to demonstrate the usefulness of wraparound technique in comparison to unreinforced and planar geosynthetic-reinforced models. Moreover, the economic aspects are also considered by comparing the amount of geosynthetic material used in reinforced model with wraparound ends with those without wraparound ends. Further, in the paper, the term “planar reinforcement” will be used for geosynthetic arrangements without the wraparound ends.

2. Wraparound technique for foundations – Basic concept and mechanism

Fig. 1 shows the three-dimensional (3D) view of a footing resting on soil bed strengthened by wraparound geosynthetic reinforcement technique. The horizontal and vertical components of the wraparound are represented by l_x and l_y, respectively, which act as anchorage and provide additional confinement to the foundation soil in comparison to the planar geosynthetic arrangement. Fig. 2 illustrates the failure pattern of the wraparound geosynthetic-reinforced foundation soil in comparison to planar geosynthetic-reinforced model. As illustrated in Fig. 2a, four zones can be recognised for the increase in the ultimate bearing capacity of the strip footing resting on soil bed reinforced with wraparound geosynthetic reinforcement technique (Raja and Shukla, 2020b). Zone I is the active wedge zone analogous to the Terzaghi's failure mechanism for unreinforced soil foundation (Terzaghi et al., 1996) (Fig. 2a). The soil in this zone moves downwards, thus pushing the adjacent soil sideways. The area adjacent to active wedge, i.e. Zone II, is the area above the reinforced zone in which soil tends to move

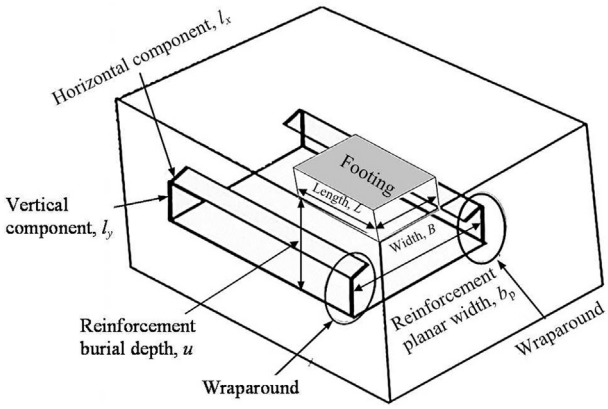


Fig. 1. Footing resting on soil bed strengthened by wraparound geosynthetic reinforcement technique.

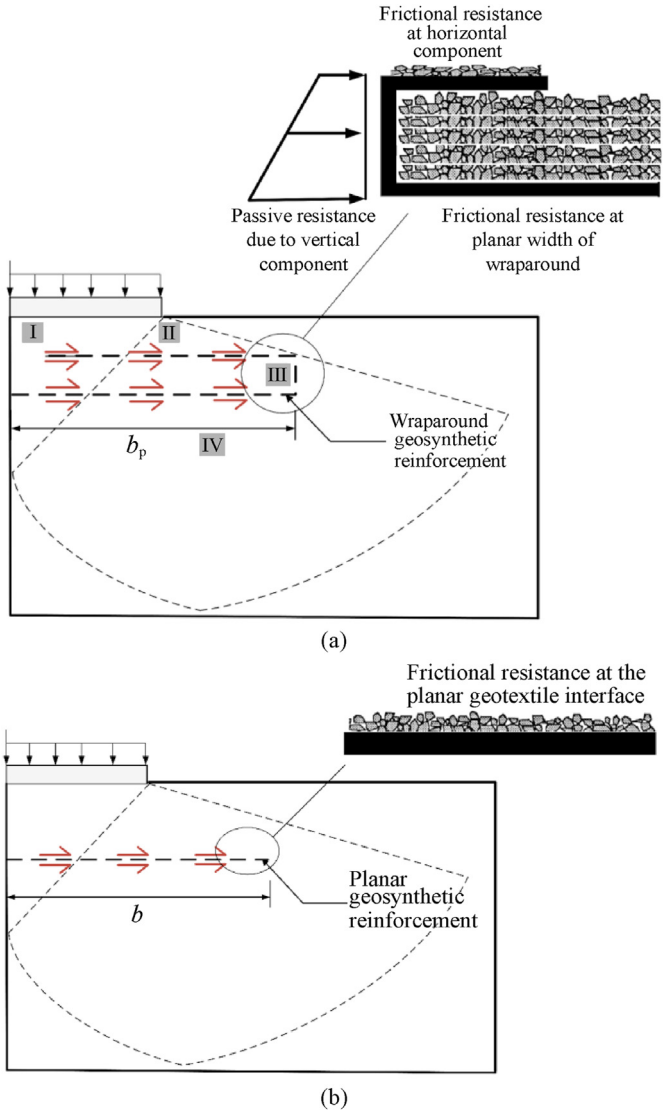


Fig. 2. A schematic mechanism of failure pattern of reinforced model for (a) wraparound and (b) planar reinforcement techniques (adapted from Raja and Shukla, 2020b).

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

horizontally to the sides due to induced shear stresses. This provides additional shear resistance to the horizontal displacement of foundation soil at soil–geosynthetic interface in comparison to planar geosynthetic-reinforced model (Fig. 2b). Zone III is also the area adjacent to Zone I but inside the wraparound ends, and provides the passive resistance due to confinement of soil. This is similar to the mechanism reported by Biswas et al. (2016) for geocell-reinforced foundations. Zone IV is the area beneath the geosynthetic layer and is unreinforced soil. Considering these mechanisms, the wraparound ends of the geosynthetic are considered to provide shear resistance at interface (soil–geosynthetic) and the passive resistance due to confinement of soil as presented in the zoomed view along with Fig. 2a, thus enhancing the ultimate bearing capacity of wraparound reinforced model foundations in comparison to planar reinforced foundation (Raja and Shukla, 2020b).

3. Experimental setup

A number of model footing load tests were conducted in the test cell located at the Geotechnical and Geoenvironmental Engineering Workstation, Edith Cowan University (ECU), Perth, Australia. The details about the materials, model test cell/tank and instrumentation, and experimental procedure adopted for this study are explained below.

3.1. Material description

The soil used in this study is a yellow brickie sand which has been extracted from the pit site located in the North Perth region, Australia. Physical properties of this sandy soil are measured according to the guidelines provided by AS 1289.0-2000 (2000). The soil is classified as the poorly graded sand (SP) according to the unified soil classification system (USCS). By applying moderate effort, this type of sand can be compacted to a relatively uniform density (Brachman et al., 2001). The non-woven type geotextile made of polypropylene with a tensile strength of 13.3 kN/m at the corresponding tensile strain of 47.5% is used in this study. Various other properties of the soil and the geotextile are also summarised in Table 1.

Table 1
Properties of soil and geosynthetic materials.

Material	Property	Magnitude
Soil	Specific gravity, G_s	2.64
	Effective grain size, D_{10} (mm)	0.27
	Median grain size, D_{50} (mm)	0.38
	Coefficient of uniformity, C_u	1.55
	Coefficient of curvature, C_c	0.96
	Percentage of finer less than 0.075 mm (%)	3
	Maximum dry unit weight, γ_{dmax} (kN/m^3)	15.97
	Minimum dry unit weight, γ_{dmin} (kN/m^3)	13.9
	Relative density, R_d (%)	70
	Angle of internal friction, ϕ (°)	35
	Cohesion, c (kPa)	2.5
	Soil classification (USCS)	Poorly graded sand (SP)
	Elastic modulus, E_s (kPa)	3920
Geosynthetic	Type	Non-woven (polypropylene)
	Tensile strength, MD/XMD (kN/m)	13.3
	Elongation strain, MD/XMD (%)	47.5
	Tensile stiffness, MD/XMD (kN/m)	28
	Apparent opening size (AOS) (mm)	0.15

Note: MD and XMD stand for the machine direction and cross machine direction, respectively.

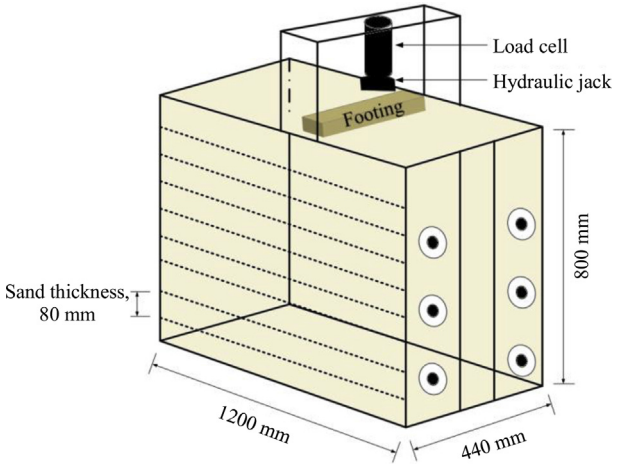


Fig. 3. Schematic view of the test cell used in the present study.

3.2. Laboratory test setup and instrumentation

The laboratory model tests were conducted in a metal steel tank with inside dimensions of 1200 mm × 440 mm × 800 mm (length × width × height). One long end of test cell was made of acrylic plexisheet of 25 mm in thickness. Following the approach of Shin and Das (2000) and Moghaddas Tafreshi and Dawson (2010), a light coat of petroleum jelly was applied at the ends to minimise the frictional resistance between the edge of the model footing and test cell walls. Moreover, test cell was braced (from outside) by the angle irons to prevent any potential yielding that may occur during the compaction of the sand and/or during the performance of footing load tests.

The rigid foundation has been modelled by well-seasoned sal wood footing. The measured dimensions of the footing for the model tests are 440 mm × 80 mm × 40 mm (length × width × thickness). Since the ratio of length to width is greater than 3, it will act as strip footing. The base of the footing was made rough by cementing the sand grains to it by epoxy glue. Fig. 3 represents the schematic view of the test cell used in the present study.

The sandy soil was compacted in a test tank in the form of seven layers with each having the thickness of 80 mm. The compaction was done manually by a wooden float to a relative density (R_d) of 70% which resulted in the average unit weight (γ) of 15.4 kN/m^3. The whole arrangement for the model footing load tests is represented in Fig. 4. Especial care was taken during the levelling of the sand bed at each layer, which was checked by digital auto-level to ensure the horizontal surface conditions (Fig. 4a). The vertical load was applied by hydraulic jack connected with low pressure pump system to adequately control the loading rate. The applied load was measured by the load cell (GEFRAN series) of 50 kN capacity. The load cell has 8 built-in sensors and is configured in a way that reduces any potential error caused by imperfect application of the load. The displacement transducers (accuracy of 0.001 mm) were used to measure the settlement of the footing (Fig. 4b). The load cell and displacement transducers were attached to the 16-channel Matest data acquisition system along with which the data logger software was used to record all the readings (Fig. 4c).

3.3. Testing programme

Four experimental test series for monotonic and seven for repeated loading conditions were performed. Test series for monotonic and repeated loading conditions were represented by letters A–D and E–K, respectively. The testing programme is

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

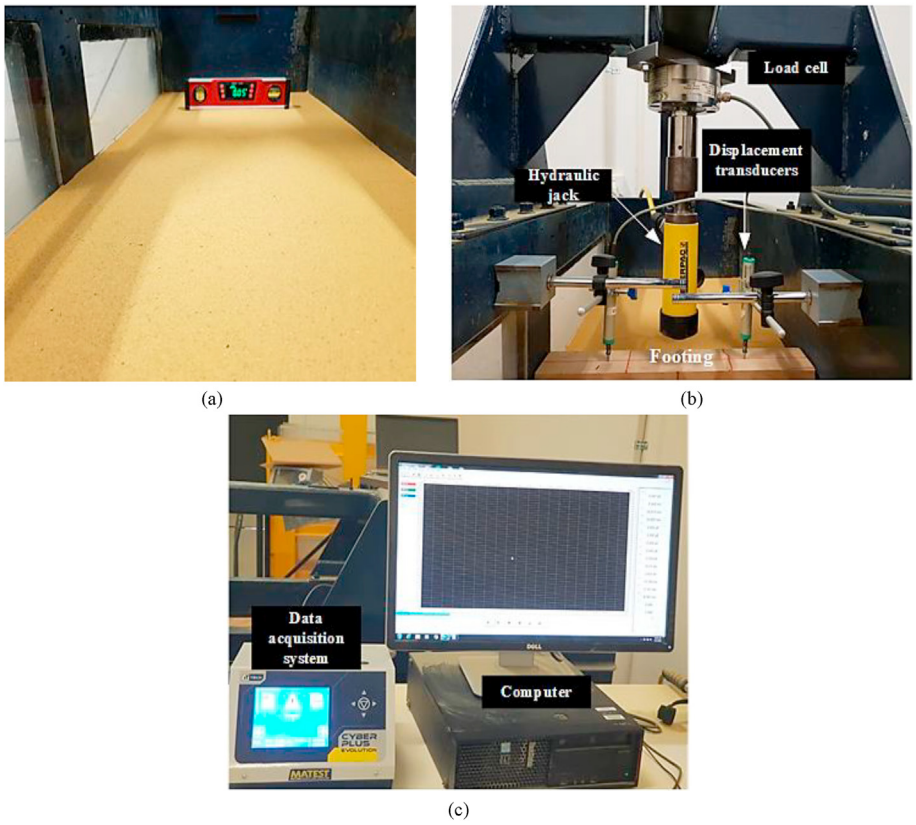


Fig. 4. Experimental model footing load test setup at workstation: (a) Levelling of sand bed; (b) Loading arrangement; and (c) Data acquisition system.

summarised in Table 2. The layout of the reinforcements (planar and wraparound) is presented in Fig. 5. The parameters involved are given below: u is the distance of topmost reinforcement layer from the base of footing, h is the distance between the subsequent layers of planar reinforcement, b is the reinforcement width of planar geosynthetic reinforcement, and b_p is the planar width of wraparound geosynthetic. All the parameters were normalised with the foundation width (B) and therefore represented as u/B, h/B, b/B, b_p/B, l_x/B, and l_y/B. Following the approach of Lal et al. (2017), h/B was kept equal to u/B when multiple planar reinforcements were used in the tests.

3.4. Loading scheme

For all the model footing load tests, the experimental procedure was performed according to the guidelines provided by ASTM D1196/D1196M-12 (2016). The load increment is applied and kept constant until the settlement of the footing is smaller than 0.03 mm/min for three successive minutes. Similar to the approach used by Kou et al. (2018), a vertical preload of approximately 0.05 kN was applied to surface footing before each trial. This ensures an even contact of surface footing with soil surface and thereafter readings were set to zero for consistency and uniformity. The settlement (s) is standardised with the footing width and expressed as the dimensionless factor (s/B). The ultimate bearing capacity values of the model foundations were estimated from their corresponding load–settlement curves.

For repeated loading tests, the model foundation was initially loaded with the portion of ultimate monotonic loads (e.g. 50%, 70%, and 90%), followed by the subsequent cycles of unloading and reloading. The repeated loading pattern is given in Fig. 6. The typical rate of loading and unloading was kept at 0.1 kN/s. A similar approach was utilised by El Sawwaf and Nazir (2010) for examining the behaviour of rectangular footings resting on reinforced soil bed under slow repeated loads. The total cumulative settlement due to repeated loading (s_t) was also standardised with the footing width and expressed as s_t/B.

It is noteworthy that the main aim of this study is to demonstrate the practical advantage of wraparound technique. However, in addition to this main objective, an attempt is made to curtail the scale effect of the model footing load tests by taking several measurements into account which are discussed in the following section.

4. Scale effect

The scale effect is one of the major concerns in the laboratory model footing load tests (1g tests). However, for these tests, if correctly performed, the scale effect can be minimised. This has been the key reason that 1g tests are still widely used in geotechnical engineering in a cost-effective manner (Al Heib et al., 2020). In the previous studies, researchers have given a valuable insight about the critical parameters influencing the scale effect in the reinforced model footing load tests and how it can be minimised (Das et al., 1996; Fakher and Jones, 1996; Chen, 2007; Chen and Abu-Farsakh, 2011; Tavakoli Mehrjardi and Khazaei, 2017). The most common physical factors influencing the ultimate bearing capacity (q_{uwr}) of wraparound geosynthetic-reinforced backfill systems can be given as follows: B, u, h, l_x, l_y, b_p, D_{50}, d_o, γ, E_r and E_s, where d_o is the aperture/opening size of reinforcement, and E_r is the elastic modulus of reinforcement. By utilising the dimension analysis rules, the function (g) governing the wraparound geosynthetic-reinforced backfill system can be written as follows:

$$q_{uwr} = g(B, u, b_p, l_x, l_y, D_{50}, h, E_s, \gamma, d_o, E_r) \quad (1)$$

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

Table 2
Summary of experimental series.

Loading scheme	Experimental series	Reinforcement technique	Parameters	Purpose
Monotonic loading	A	Unreinforced	–	To compare the performance and improvement
	B	Planar	$u/B = 0.3$; $N = 1$; $b/B = 4, 6, 8$	To evaluate the performance at various width ratios of reinforcement and to find the optimum ratio
	C	Planar	$u/B = h/B = 0.17$; $N = 2$; $b/B = 6$	To evaluate the effect of layer number on the optimum width ratio
	D	Wraparound	$u/B = 0.3$; $N = 1$; $b_p/B = 4, 3.5, 2.8$[a]; $l_x/B = 0.8, 1.05, 1.4$; $l_y/B = 0.2$; $B_{rt}/B = 6$	To evaluate the performance of wraparound technique
Repeated loading	E	Unreinforced	–	To compare the performance and improvement in repeated load tests
	F	Planar	$u/B = 0.3$; $N = 1, 2$[b]; $b/B = 6$; $q/q_u = 50\%$, 70%, 90%	To evaluate the settlement at various repeated load levels and to find the critical number of cycles for which footing settlement begins to stabilise
	G	Wraparound	$u/B = 0.3$; $N = 1,2$[b]; $b_p/B = 4$; $l_x/B = 0.8$; $l_y/B = 0.2$; $B_{rt}/B = 6$; $q/q_u = 50\%$, 70%, 90%	
	H	Wraparound	$u/B = 0.3$; $N = 1$; $b_p/B = 4$; $l_x/B = 0.8$; $l_y/B = 0.2$; $B_{rt}/B = 6$; $q/q_{awr} = 70\%$, 90%	To compare the performance of wraparound with multilayer reinforced soil foundations at higher load levels
	I	Planar	$u/B = h/B = 0.17$; $N = 2$; $b/B = 6$; $q/q_{awr} = 70\%$, 90%	
	J	Planar	$u/B = 0.3$; $N = 1, 2$[b]; $b/B = 6$; $q/q_u = 50\%$, 90% (Inc.)	To evaluate the performance of wraparound under incremental load
	K	Wraparound	$u/B = 0.3$; $N = 1$; $l_x/B = 0.8$; $l_y/B = 0.2$; $B_{rt}/B = 6$; $q/q_u = 50\%$, 90% (Inc.)	

Note: Inc. is for incremental loading.
[a] Corresponding to the full wraparound condition.
[b] For $N = 2$, $u/B = h/B = 0.17$.

Considering the length and force as the fundamental dimensions, 11 parameters in Eq. (1) can be reduced to 9 independent parameters which can be identified as $(\pi_1, \pi_2, \ldots, \pi_9)$ as depicted below:

$$\frac{q_{uwr}}{\gamma B} = g\left(\frac{u}{B}, \frac{h}{B}, \frac{l_x}{B}, \frac{l_y}{B}, \frac{b_p}{B}, \frac{D_{50}}{B}, \frac{\gamma D_{50}}{E_s}, \frac{d_o}{B}, \frac{E_s}{E_r}\right) \tag{2}$$

For ideal response, the real-life projects and model tests should be the same. However, it would be practically not feasible to maintain the complete similarity. Therefore, some of the parameters, which are significantly prone to "scale effect" (Fakher and Jones, 1996), can be listed. Considering the prototype footing (B_p) with dimension n times the model footing (B_m), the following equation can be expressed:

$$B_p / B_m = n \tag{3}$$

In order to maintain the similarity, the π terms for the prototype and model should be kept the same. For example, if the strip footing of 80 mm width is studied, then considering the value of n equal to 10, the parameters u, l_x, l_y, b_p, D_{50} and d_o should be considered 10 times those of the model. In this way, assuming that the soil in the prototype and model is the same, the following equation can be satisfied to obtain the bearing capacity of prototype system:

$$\left(\frac{q_{uwr}}{\gamma B}\right)_n = \left(\frac{q_{uwr}}{\gamma B}\right)_p \overset{\text{yields}}{\rightarrow} (q_{uwr})_p = n(q_{uwr})_m \tag{4}$$

The following sections are dedicated to the precautions taken in this study to minimise the scale effect according to the guidelines available in the well-established published literature related to the model footing load tests (unreinforced and reinforced).

4.1. Influence of ratio of median grain size to footing width (D_{50}/B)

In practice, the width of footing is usually very large compared to the median grain size of the soil particle. Several researchers have reported that the ratio of D_{50}/B has a significant impact on the bearing capacity of model foundations (Habib, 1974; Yamaguchi et al., 1977; Tatsuoka et al., 1991; Kusakabe, 1995; Hsieh and Mao, 2005). This is commonly called as the "particle size effect". Kusakabe (1995) recommended that the particle size effect can be neglected if D_{50}/B is kept smaller than 1/100. Cerato and Lutenegger (2003) suggested that the bearing capacity does not get affected if D_{50}/B is kept smaller than 1/200. Hsieh and Mao (2005) recommended this value to be smaller than 1/15 for geosynthetic-reinforced soils. Considering the above-mentioned recommendat ions, the particle size effect will be negligible as the ratio of D_{50}/B for this study is 0.00475.

4.2. Scale effect of the reinforcement (reinforcement ratio, R_r)

Reinforcement ratio is a key parameter affecting the bearing capacity of model footing resting on geosynthetic-reinforced soil bed (Chen, 2007; Chen and Abu-Farsakh, 2011, 2015; Ahmad et al., 2020). It incorporates the following parameters: u/B, h/B, E_s/E_r, and a_s/a_r, which is mathematically expressed as

$$R_r = \frac{a_r E_r}{a_s E_s} \tag{5}$$

where a_s is the area of reinforced soil, and a_r is area of reinforcement.

Mathematical forms of a_r, a_s and E_r are given as follows:

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

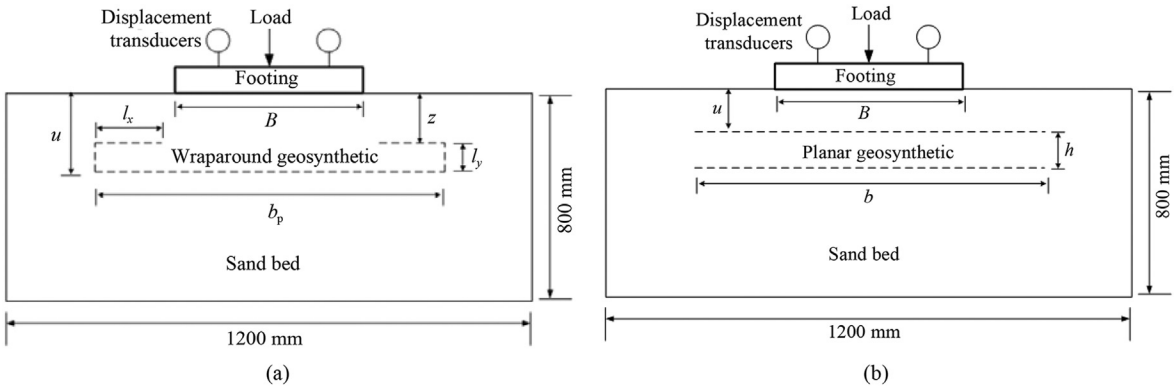


Fig. 5. Layout of (a) planar and (b) wraparound geosynthetic reinforcements.

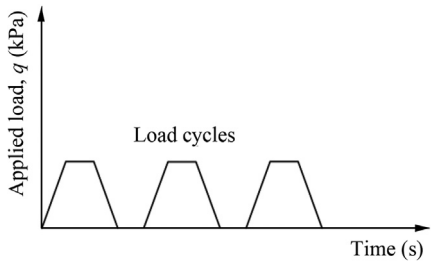


Fig. 6. Typical pattern of loading scheme in repeated load tests.

$$a_r = Nt_r \times 1 \tag{6}$$

$$E_r = J/t_t \tag{7}$$

$$a_s = [u + (N-1)h] \times 1 = d \times 1 \tag{8}$$

where N is the number of reinforcement layers, t_r is the reinforcement thickness, J is the tensile modulus of reinforcement, and d is the depth of reinforced zone.

Substituting Eqs. (6)–(8) into Eq. (5) results in

$$R_r = \frac{J}{uE_s} \tag{9}$$

Chen (2007) has carried out extensive experimental (small- and large-scale) and numerical analyses of reinforced soil foundations and concluded that if the value of R_r is kept the same as d/B, then the results of the laboratory reduced-scale model tests can be extrapolated to the full-scale reinforced soil foundations with minimal error.

In this study, the value of R_r is 0.297 which is very close to d/B of 0.3 for all the tests except for test series for multilayer reinforcements where the depth of reinforced zone slightly exceeds 0.3. Therefore, it can be admissible that the scale effect regarding R_r will have a low to negligible impact on the performed tests in this study.

4.3. Influence of ratio of aperture size to footing width (d_o/B)

The ratio of minimum opening size (d_o) of the reinforcement to the footing width has a significant effect on the BCR of the reinforced foundations. Cuelho et al. (2014) and Tavakoli Mehrjardi and Khazaei (2017) suggested that the d_o/B ratio should be restricted to 4 and 4–6, respectively, for achieving the highest BCR value with minimal scale effect. It is to be noted that this ratio has a significant impact when geogrid is used as the reinforcement. However, for geotextile-reinforced soil, this effect may not be applicable due to the negligible opening size of such reinforcements.

4.4. Test cell scale effect

Besides the above-mentioned parameters, another reason that can cause the scale effect issue is the insufficient dimensions of the test cell. Bransby and Smith (1975) suggested that for a wide enough test tank with smooth side walls, the boundary conditions and side frictional resistance do not have any noteworthy effect on the results of model footing load tests. Abdel-Baki and Raymond (1993) and Azzam and Nasr (2015) suggested that the length and height of tank should be kept minimum of 6 and 7 times the footing width, respectively, in order to minimise the scale effect. Lal et al. (2017) suggested that all the test cell dimensions should be minimum of 5 times the footing width. The test cell used in this study is 1200 mm long, 440 mm wide, and 800 mm deep, as shown in Fig. 4. All these dimensions are sufficient enough to minimise the scale effect on the performance of the model tests.

5. Results and discussion

A series of model tests was conducted on unreinforced, planar and wraparound reinforced models in a test cell. The details of the instrumentation used in the experimental procedure were given in Section 3. The ultimate bearing capacity of the unreinforced model foundation, q_u, was 96.6 kPa at a settlement, s, of 7.986 mm which gives the s/B value of approximately 0.1. The increase in the bearing capacity (at various s/B ratios) for planar and wraparound reinforced models is quantified as a dimensionless ratio, respectively termed as the improvement factors IF_{pg} and IF_{wg} in this study, which are mathematically expressed as

$$IF_{pg} = \frac{q_{(p)}}{q}, \; IF_{wg} = \frac{q_{(w)}}{q} \tag{10}$$

where $q_{(p)}$, $q_{(w)}$ and q are the bearing capacities of planar, wraparound geosynthetic-reinforced, and unreinforced model foundations, respectively.

It must be noted that the results of some of the model footing load tests were checked for repeatability by conducting them more than once. It has been observed that if the appropriate care is taken and conditions are kept the same, then there is no marked difference between the observed values.

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

5.1. Monotonic load tests

5.1.1. Planar reinforced models

The results of the test series A and B are presented in Fig. 7. For a medium dense type sand, the bearing pressure–settlement curve changes at a certain level indicating a shear failure in the soil and after that the curve becomes steeper (Shukla, 2014; Das and Sobhan, 2017). The reasonable agreement is achieved for the bearing pressure–settlement curve of unreinforced model foundation and the failure pattern of medium dense type sand. It is obvious from the figure that the bearing capacity of the reinforced model is greater than that of the unreinforced model. This is true even for the lower settlement levels; however, the effect is more pronounced after $s/B > 2.5\%$. This is in line with the previous studies on strip footing resting on geosynthetic-reinforced medium dense soil bed (e.g. Shin and Das, 2000; Cicek, 2011).

In the test series B, the model footing load tests were conducted at three b/B values of 4, 6, and 8. These values were chosen based on the previous studies conducted on geosynthetic-reinforced soil foundations. There is a lack of unanimity in the literature regarding the optimum width of geosynthetic reinforcement placed under the footing. Guido et al. (1986) reported the optimum width of 2 and 3 for geogrid- and geotextile-reinforced soil foundations, respectively. Omar et al. (1993) and Khing et al. (1993) suggested the optimum values of b/B equal to 8 and 6, respectively. Cicek et al. (2015) studied the effect of reinforcement length on the bearing capacity of strip footing on sandy soil and found an optimum value of b/B equal to 5 for their model study; however, they concluded that the increase in bearing capacity is highly linked with the geometry, material properties, and type of reinforcement and therefore, cannot be generalised. For a single layer of reinforcement, the optimum achievement in this study is obtained at $b/B = 6$. It is to be noted that the load-settlement curves are closer to each other for $b/B = 4$ and 6 at lower magnitudes ($s/B \leq 5\%$), but after that the trend begins to change, and once s/B reaches 7.5%, a significant improvement can be observed. Moreover, for a single reinforcement of $b/B = 4$, the failure point can be observed corresponding to the change in slope of load–settlement curve. However, the curves for $b/B = 6$ and 8 do not depict such failure points until higher settlement values. For $b/B = 6$ and 8, the curves follow almost the similar path until $s/B = 10\%$. This is indicated in Table 3 which reports the IF_{pg} values for planar reinforcement arrangement. The marked improvement can be observed especially for higher order and ultimate settlement values. For $b/B = 6$, the value of IF_{pg} is 1.553 for s/B at peak load level in comparison to IF_{pg} of 1.19 for $b/B = 4$; but when the reinforcement width ratio is further increased to 8, the improvement is not significant ($IF_{pg} = 1.584$).

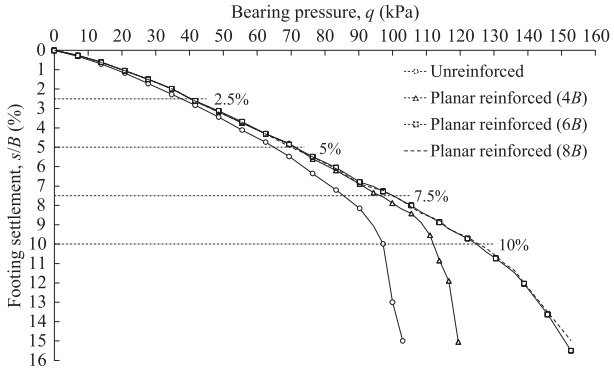


Fig. 7. Bearing pressure–settlement curve for test series A and B.

Table 3
Improvement factor for planar reinforcement arrangement at various settlement levels.

Footing settlement, s/B (%)	Improvement factor, IF_{pg}		
	$b/B = 4$	$b/B = 6$	$b/B = 8$
2.5	1.08	1.07	1.07
5	1.085	1.089	1.083
7.5	1.141	1.2	1.176
10	1.17	1.294	1.304
Ultimate[a]	1.19	1.553	1.584

[a] Corresponding to the peak load.

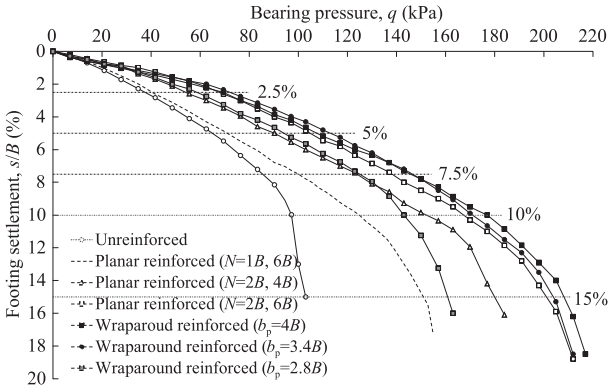


Fig. 8. Bearing pressure–settlement curve for test series C and D.

After the selection of optimum value of reinforcement width ratio (i.e. $b/B = 6$), the test series C with two planar layers of reinforcement were also conducted. For comparison purposes, the results of these load tests were discussed in the following section.

5.1.2. Wraparound reinforced models

The test series D represents the tests performed on the model footing resting on sand bed strengthened by wraparound reinforcement technique. For comparison purpose and to evaluate the benefits of wraparound technique, the results of the test series D are compared with those of series B and C. The test series C was conducted for $N = 2$ at $u/B = 0.17$ and $b/B = 4$ and 6. It should be noted that u/B was kept at 0.17 to minimise the scale effect due to the reinforcement ratio, as discussed in Section 4.2. For wraparound geotextile-reinforced tests, the parameters are given in Table 2. The total width of the wraparound geotextile (B_{rt}) reinforcement was kept equal to 6. However, in order to find the optimum planar width ratio (b_p/B) in the wraparound reinforcement technique, b_p was varied from $4B$ to $2.8B$ by changing the lap-length ratio (l_x/B) from 0.8 to 1.4. The load–settlement behaviour of test series C and D is presented in Fig. 8. It can be observed that the optimum performance is obtained when the planar width of the wraparound geotextile is 4. The curve for wraparound ($b_p/B = 3.5$) depicts the higher IF_{wg} values for $s/B \leq 7.5\%$ but there has been no discernible improvement for the s/B levels between 7.5% and 9% and after that the performance decreases in comparison to the wraparound with $b_p/B = 4$. Moreover, after the s/B value reaches approximately 12%, the curve has become steeper and much closer to the behaviour of planar reinforced model ($N = 2$ and $b/B = 6$). The performance of full wraparound (i.e. $b_p/B = 2.8$) is somewhat more similar to the planar reinforced foundation ($b_p/B = 4$ and $N = 2$) at lower s/B values, and for $s/B > 9\%$, the load–settlement behaviour was similar to that of reinforced model with single layer of geotextile ($b/B = 6$) but with additional thickness. This finding differs from the observation made by Aria et al. (2021): footing resting on

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

soil reinforced with geosynthetic layer with full wraparound ends performs better than those reinforced with partial wraparound ends. Moreover, the fully wrapped model will behave as the coherent body and deform as the centrally loaded slab. The authors believe that this is because the reinforcement used by Aria et al. (2021) was of extremely high stiffness (i.e. $J \approx 1000$ kN/m) which resulted in significantly higher BCR in model tests. Fakher and Jones (1996) indicated that the scale effect due to reinforcement stiffness is significant and may result in unrealistic high values of BCRs in laboratory reduced-scale model tests. Moreover, when fully wrapped geosynthetic reinforcement was utilised, this effect would have become more significant because of the horizontal component of wraparound (l_x) practically being equal to the planar width (b_p) of wraparound, as shown in Fig. 9. This geometry along with the very high tensile modulus of geosynthetic reinforcement may result in the deformation pattern of such models as centrally loaded beam or slab, as indicated by Aria et al. (2021). However, such mechanism is highly unlikely to occur in this study as the reinforcement properties and parameters were selected in a way to keep the scale effect at minimal level, as described in Section 4. In spite of this fact, even the fully wrapped model has performed better than the reinforced model with $N = 1$ and $b/B = 6$.

In order to evaluate the effectiveness of the wraparound technique, IF_{wg} values at various settlement levels are computed against the amount of material used and occupied land width. The mass of geotextile with width equal to that of base area of footing is standardised as x (in g) and occupied land width as y (in m). It is to be noted that the occupied land width means the planar width of geotextile which is the same as b for geotextile without wrapround ends and b_p for geotextile with wrapround ends. The comparison has been presented in Table 4. It can be easily observed that for the same amount of material, the reinforced model with wraparound ends (case 1, $IF_{wg} = 2.15$) has performed better than the model without wraparound ends (case 2, $IF_{pg} = 1.553$). For planar reinforcement, case 4 has shown the best performance and attained the IF_{pg} value of 2.07 at s_u/B, where s_u is the settlement at peak load. However, this is achieved at the cost of twice the amount of material and 1.5 times the occupied land width in comparison to wraparound arrangement (case 1). Fig. 10 represents the variation of IF_{wg} of wraparound reinforced model with l_x/b_p and l_x/l_y. It can be observed that if the total amount of the geotextile is kept constant, then the improvement in bearing capacity (IF_{wg}) increases with the increase in l_x/b_p and l_x/l_y from 0.2 to 0.3 and from 4 to 5.25, respectively. Further increase in l_x/b_p and l_x/l_y causes the decrease in IF_{wg}. The trend is especially true for settlement level (s/B) less than 10%. This can be attributed to the fact that at the lower settlement level, the frictional forces activated at the horizontal component along with the planar width (b_p) starts providing resistance to the foundation soil, thus causing an increase in the bearing capacity. However, as l_x is further increased, b_p becomes smaller, thus causing a decrease in optimal planar area of geosynthetic necessary for providing the resistance to the soil. However, it may be noted that for the same amount of geotextile, the increase in IF_{wg} for any settlement level is more than that in IF_{pg} (see Fig. 11). This is true for all the cases of l_x/b_p and l_x/l_y. The maximum improvement is achieved at ultimate load level for l_x/b_p and l_x/l_y values of 0.2 and 4, respectively. At higher settlement levels, it is expected that the shear resistance at soil–geosynthetic interface and the passive resistance due to confinement of soil as explained in Section 2 play their part in improving the bearing capacity and compression stiffness of the wraparound reinforced model foundations. However, for maximum benefits, it is important to place the wraparound geosynthetic with optimum planar width and lap-length ratio. It is important to note that the wraparound reinforced models have performed well even at the small settlement levels. Similar results were reported by Kazi et al. (2016) in examining the model foundations strengthened by wraparound geosynthetic reinforcement technique. This indicates the mobilisation of frictional and passive resistances in the wraparound zone even at smaller strain levels.

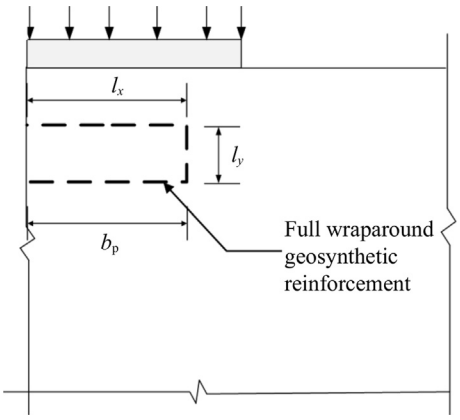


Fig. 9. Geometry of fully wrapped model.

5.2. Repeated load tests (unreinforced, planar and wraparound reinforced models)

After selecting the optimum width for planar and wraparound reinforced models, repeated load tests were conducted on unreinforced and reinforced model foundations (test series E–K). The repeated loads were applied according to the loading scheme as explained in Section 3.4. Three initial monotonic load levels, q/q_u of 50%, 70% and 90% were considered. For comparison purpose, the initial monotonic levels are kept according to the peak load level obtained for unreinforced models. This gives the values of applied load of approximately 48 kPa, 68 kPa and 87 kPa, respectively, for 50%, 70% and 90% of q/q_u. The results of the repeated load tests are discussed in the following sections.

5.2.1. Hysteresis loop analysis

Fig. 12 represents the hysteresis loops of model footing settlement versus the applied pressure for 10 load cycles (test series E–G). It is observed that with the application of initial monotonic load, the vertical settlement increases, and once the load is released (decreased), the elastic rebound can be observed. This is due to the elastoplastic nature of supporting soil and the reinforcement. For each cycle, the settlement attains a new maximum (peak) and new minimum value (residual) which is slightly greater than the previous one. The amount of plastic (permanent) settlement is more and gradually increases with the application of load cycles. Similar observations were made by Moghaddas Tafreshi et al. (2011) when analysing the response of circular footing resting on sand bed under cyclic loads. The trend is true for all the initial q/q_u levels. However, it can be observed that the settlement values for the wraparound reinforcement technique are always less than those for the unreinforced and planar reinforced models. This is even true for higher initial load levels, that is, the accumulated settlement of wraparound geotextile-reinforced model foundation at $q/q_u = 90\%$ is less than that of the planar single-layer reinforced model at $q/q_u = 70\%$. For double-layer reinforced model, the s/B values are much closer to that of wraparound case for higher initial load levels (i.e. $q/q_u = 70\%$ and 90%), but some difference can be observed for lower initial load level (i.e. $q/q_u = 50\%$). This is probably due to the fact that at lower load levels, the frictional forces at planar geosynthetic interface are not fully mobilised especially in the second layer of reinforcement. As the load started to increase, the involvement of the second layer

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

Table 4
Comparison of amount of material and occupied land width in terms of improvement factor at various settlement levels.

Arrangement	Case No.	Amount of material (g)	IF_{pg} or IF_{wg} at various s/B values					Occupied land width, y (m)
			$s/B = 2.5\%$	$s/B = 5\%$	$s/B = 7.5\%$	$s/B = 10\%$	Ultimate[a]	
Planar reinforcement	1 ($N = 1$)	$4x$	1.08	1.085	1.141	1.17	1.19	$4y$
	2 ($N = 1$)	$6x$	1.07	1.089	1.2	1.294	1.553	$6y$
	3 ($N = 2$)	$8x$	1.42	1.38	1.47	1.55	1.76	$4y$
	4 ($N = 2$)	$12x$	1.71	1.65	1.67	1.73	2.07	$6y$
Wraparound reinforcement	1 ($l_x/B = 0.8$)	$6x$	1.83	1.67	1.69	1.83	2.15	$4y$
	2 ($l_x/B = 1.05$)	$6x$	1.87	1.71	1.71	1.77	2.09	$3.5y$
	3 ($l_x/B = 1.4$)	$6x$	1.53	1.48	1.49	1.48	1.67	$2.8y$

[a] Corresponding to the peak load.

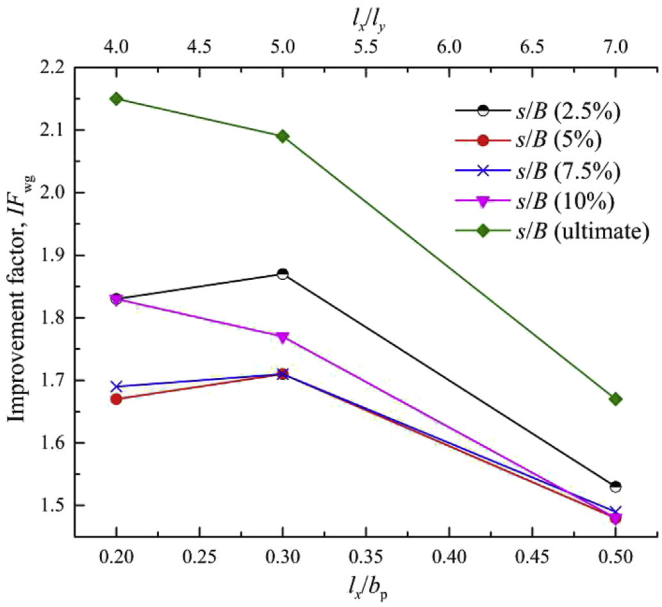


Fig. 10. Variation of IF_{wg} with l_x/b_p and l_x/l_y.

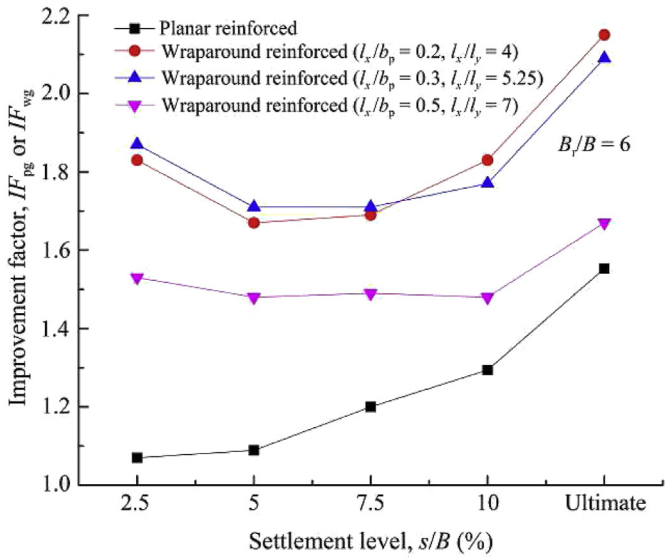


Fig. 11. Comparison of IF_{pg} for planar geosynthetic-reinforced model versus IF_{wg} for wraparound geosynthetic-reinforced modes at various settlement (s/B) levels.

has become prominent in resisting the soil movement. However, for the wraparound arrangement, in addition to the planar width, the resistance to the foundation soil is also provided at the horizontal and vertical components. This geometric configuration will also provide the interlocking of the particles in the wraparound zone (Fig. 2). This configuration along with the densification and confinement process in the wraparound zone reduces the settlement and increases the ultimate bearing capacity and compression stiffness of the reinforced soil. However, further evaluation is necessary to find the effect of number of load cycles on the performance of wraparound models.

5.2.2. Load cycle effect

In this section, the effect of number of load cycles is investigated on the performance of unreinforced and reinforced models. The maximum number of load cycles is restricted to 100 to observe the general trend. Fig. 13 represents the effect of number of load cycles (n) on the total settlement due to repeated loads of the models. It can be observed that for any initial load level, the settlement of footing is larger for first few load cycles, typically 1–20. It is evident that larger amount of total deformation occurs for unreinforced case in comparison to its counterpart models. Moghaddas Tafreshi and Dawson (2012b) reported the similar results in laboratory tests on footing resting on unreinforced and geotextile-reinforced sand bed under repeated loading. It is interesting to note that for unreinforced sand case ($q/q_u = 90\%$), the settlement values are much higher. This depicts that even the application of slow repeated load of magnitude slightly less than the ultimate load can lead to the unstable plastic shakedown state and possibly result in the incremental collapse after further load applications. Similar observation was made by Shajarati et al. (2012) in their study related to the behaviour of sandy soils under repeated loading. The planar single-layer reinforced model depicts the progressive increase in the settlement (with decreasing trend) with the increase in the load cycles. For double-layer reinforced models, the settlement increases at a much slower rate with an increase in number of load cycles, especially at higher load levels. This is in line with the observations reported by El Sawwaf and Nazir (2010) during the investigation of reinforced sand bed subjected to repeated loading. A similar behaviour is observed for the wraparound case; however, the settlement rate is extremely slow after the application of 30–40 cycles. This is the reason that for the repeated load level of lower magnitude (i.e. $q/q_u = 50\%$), the test is terminated after 70 cycles. For a more comprehended presentation, an improvement factor IF_s is introduced to compare the total settlement (s_t) for reinforced model in comparison to unreinforced model under repeated loading. Mathematically, the improvement factor is presented as follows:

$$IF_s = \frac{(s_t/B)_{U,l_n} - (s_t/B)_{R,l_n}}{(s_t/B)_{U,l_n}} \tag{11}$$

where $(s_t/B)_U$ is the total settlement of unreinforced model foundation, $(s_t/B)_R$ is the total settlement of reinforced model

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

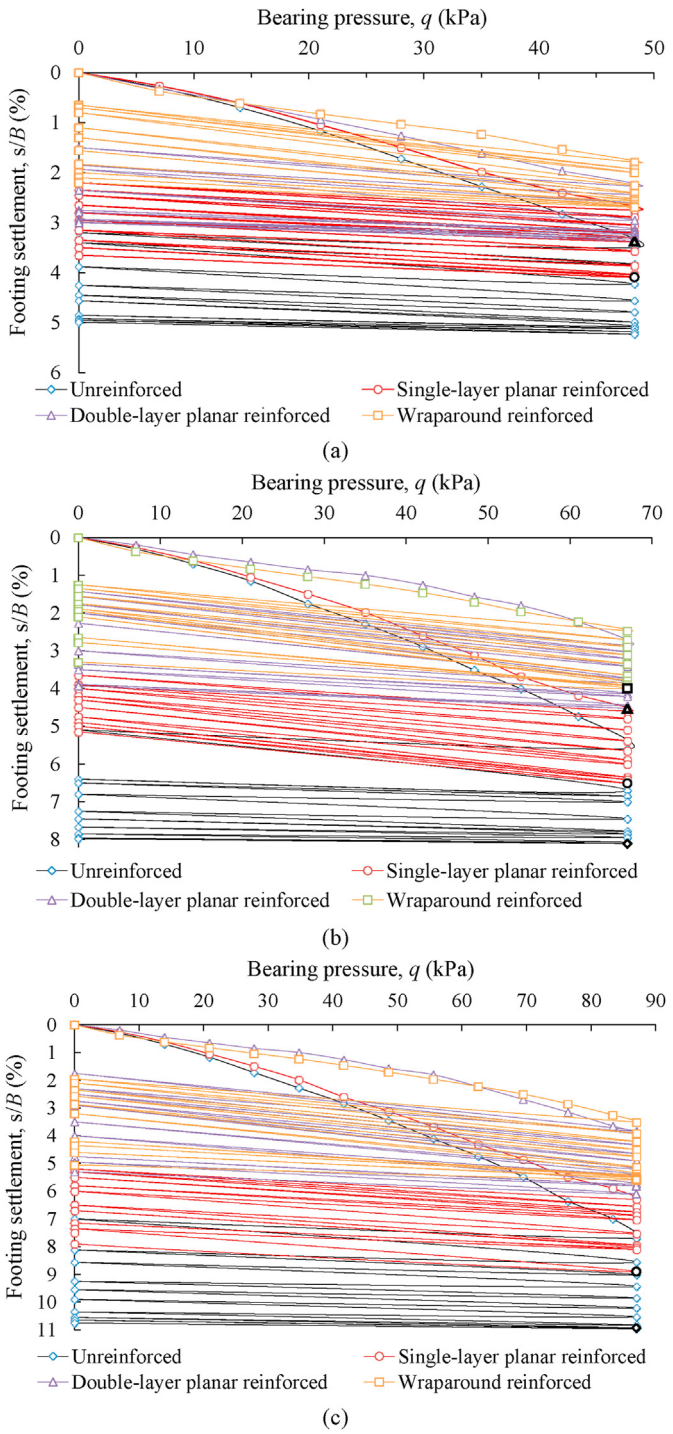


Fig. 12. Hysteresis loops of footing settlement for unreinforced and reinforced model foundations for: (a) $q/q_u = 50\%$, (b) $q/q_u = 70\%$, and (c) $q/q_u = 90\%$.

foundation, and l_n is the load level corresponding to n load cycle. Fig. 14 shows IF_s for planar and wraparound reinforced models for 50 load cycles. The values of IF_s for each load cycle are calculated using Eq. (8) and averaged over the interval of 10 cycles. It is easily identifiable that the wrapround model performed well in comparison to planar reinforced models. The improvement can be observed even for the initial 10 cycles in which most of the settlement occurs. However, the effect is more prominent at higher load levels. The average IF_s is 45% (all load levels) for wraparound reinforced models in comparison to 23% and 41% for single- and double-layer reinforced models, respectively. It is therefore admissible that the wraparound reinforcement technique can limit the accumulation of strains under repeated loading, thus providing greater stability to the foundation soil in comparison to planar reinforced models. This is true for any initial monotonic level and number of cycles under consideration.

It may be noted that the wraparound reinforced models were subjected to repeated loads that were less than their ultimate monotonic loads. Therefore, test series H and I were conducted to compare the behaviour of wraparound and multilayer planar reinforced models at higher levels of repeated loads. The fair comparison would not be possible for any load percentage as both models have different values of ultimate bearing capacities (Fig. 7). Therefore, for unbiased comparison, the load levels were kept at 70% and 90% load levels of average bearing capacity of wraparound and multilayer planar reinforced models at $s/B = 10\%$ as these values are closer to each other and of higher magnitude in comparison to other models. This gives the loads of 121.5 kPa and 156 kPa for $q/q_{awp} = 70\%$ and 90%, respectively, where q_{awp} is the average value of bearing pressure for wraparound and multilayer planar reinforced models. The results are shown in Fig. 15. It can be observed that the settlement values of multilayer reinforced foundations are much closer to those of wraparound reinforced model foundations. At the load level of $q/q_{awp} = 90\%$, the difference is almost negligible. This again highlights the fact about the higher mobilisation of interface friction at higher load levels and with the increase in number of load cycles. As a whole, for higher load levels, the total settlement has gradually increased

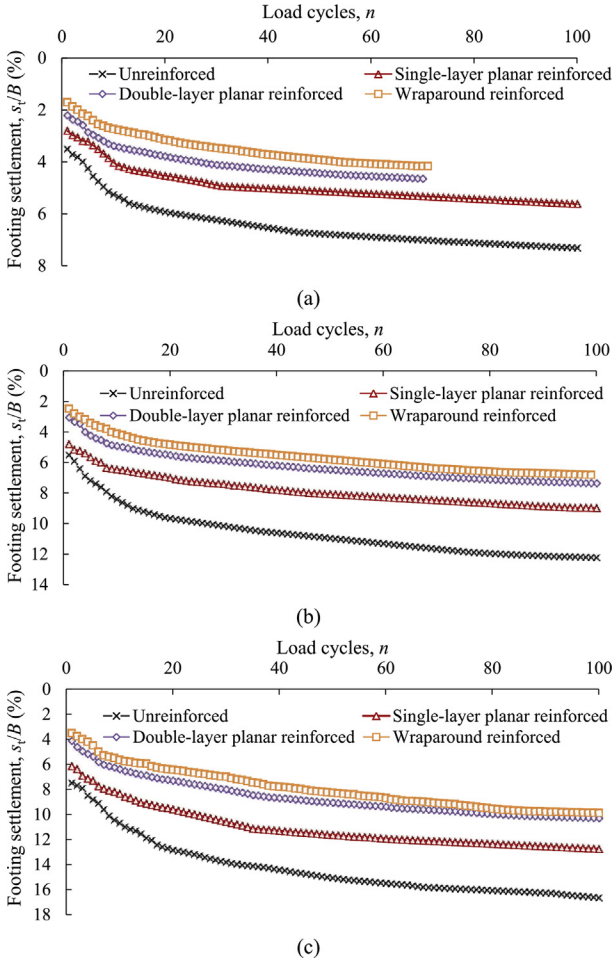
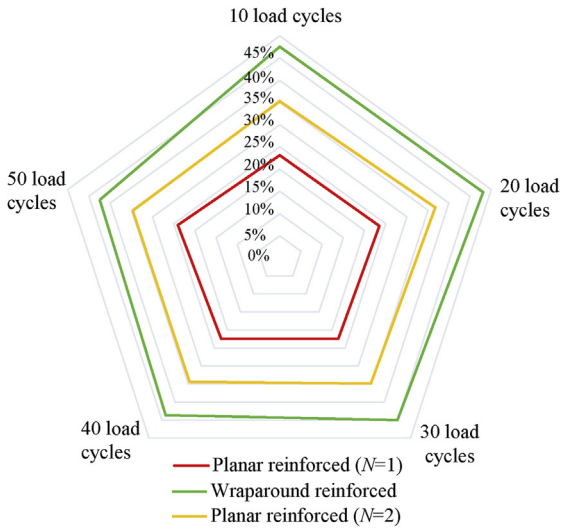
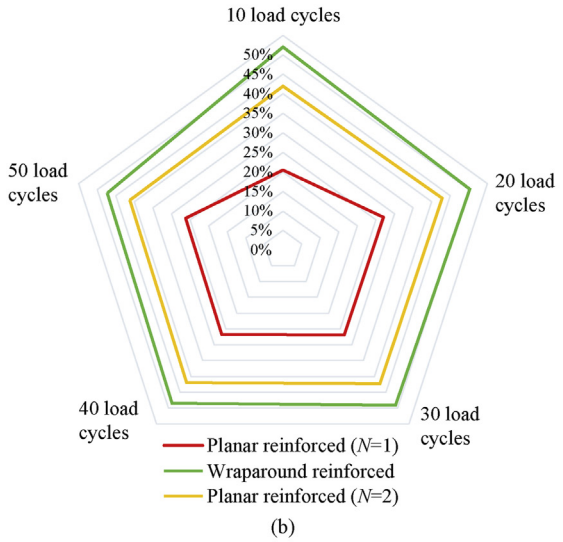


Fig. 13. Number of load cycles versus total settlement for unreinforced and reinforced model foundations for: (a) $q/q_u = 50\%$, (b) $q/q_u = 70\%$, and (c) $q/q_u = 90\%$.

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001





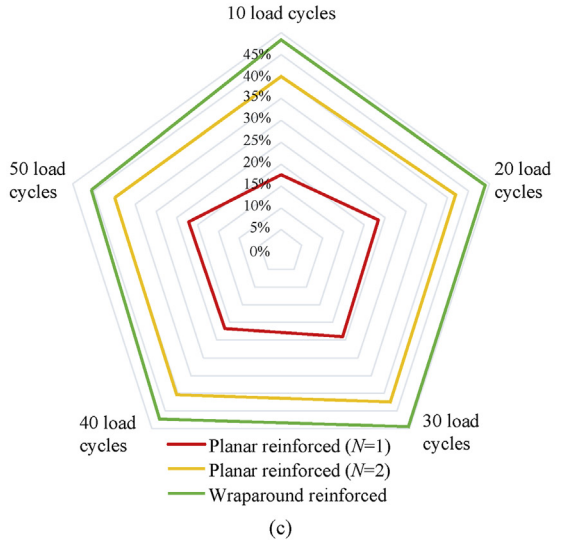


Fig. 14. Variation of improvement factor (IF_s) at various load cycles in planar and wraparound reinforced model foundations for: (a) $q/q_u = 50\%$, (b) $q/q_u = 70\%$, and (c) $q/q_u = 90\%$.

with the decreasing trend but no sign of failure is observed even after the application of 100 load cycles.

5.2.3. Effect of incremental increase of load

The test series J and K were performed to evaluate the effect of incremental repeated loads on the footing. The previous repeated load tests were performed by applying the cycles of same load levels (i.e. $q/q_u = 50\%$, 70% and 90%). However, for this series, the incremental (step-up) repeated loads are applied to evaluate the performance of wraparound geosynthetic-reinforced soil foundation. Ten cycles of each load level were applied and the corresponding settlement was measured. The major shift in the pattern for any load test has been observed at the point of change of loading amplitude, especially when the load was raised from 50% to 70% of the ultimate capacity (see Fig. 16). At higher loading shift (70%–90%), the wraparound reinforcement has shown the least change due to its densification mechanism and interlocking of soil particles within the wraparound zone. Moreover, the foundation soil reinforced with wraparound technique and planar reinforced layers ($N = 2$) has shown much greater stability than unreinforced and single-layer reinforced foundations.

6. Conclusions

This research presents an experimental investigation that analyses the behaviour of repeatedly loaded strip footing resting on sand bed strengthened by wraparound geosynthetic reinforcement technique. The reinforcement properties and test parameters were selected in a way to keep the scale effect to a minimal level. The effects of repeated load amplitude and number of cycles along with the effect of reinforcement parameters, such as number of layers and reinforcement width, were studied. The amounts of geosynthetic material used in wraparound and planar reinforced models were also compared. Based on the results and discussion presented in the previous section, the following general conclusions can be drawn:

(1) For the same amount of geotextile and less occupied land width, the model foundations strengthened with wraparound reinforcement technique have outperformed the planar reinforced models.
(2) For lower settlement levels ($s/B \leq 5\%$), the wraparound geotextile with smaller occupied land width ratio ($b_p/B = 3.5$) has performed well in comparison to the wraparound with slightly larger occupied land width ratio ($b_p/B = 4$). However, for higher settlement levels, the wraparound with occupied width ratio of 4 provides more stability to the foundation soil. The performance of the fully wrapped model ($b_p/B = 2.8$) is more similar to that of the planar double-layer reinforced model ($b/B = 4$); however, it should be noted that even the fully wrapped model outperforms the planar single-layer reinforced model with the same amount of geotextile and approximately 50% less occupied land width.
(3) For the same amount of geotextile, the total settlement due to vertical repeated loading is lower in wraparound geotextile-reinforced models than that in single-layer planar reinforced models. This is true for any initial load level (q/q_u) and number of load cycles. Moreover, for lower q/q_u (e.g. 50%), the wraparound geotextile-reinforced model foundation has shown negligible amount of settlement after the application of 30 load cycles.
(4) For any magnitude of q/q_u, the value of s_t/B for the wraparound geotextile-reinforced models tends to increase at a very slow rate in comparison to planar reinforced models.

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

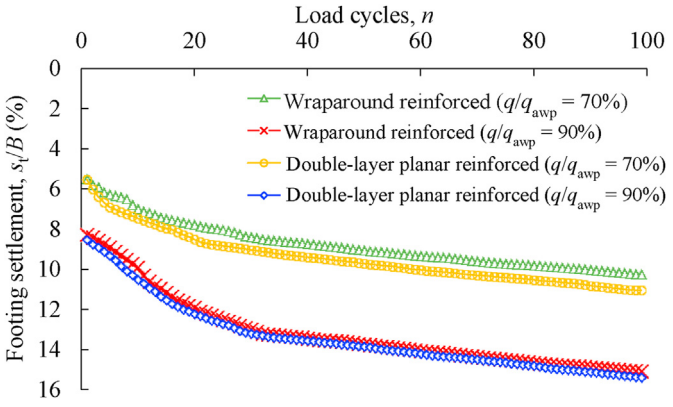


Fig. 15. Number of load cycles versus total settlement for wraparound and multilayer reinforced model foundations.

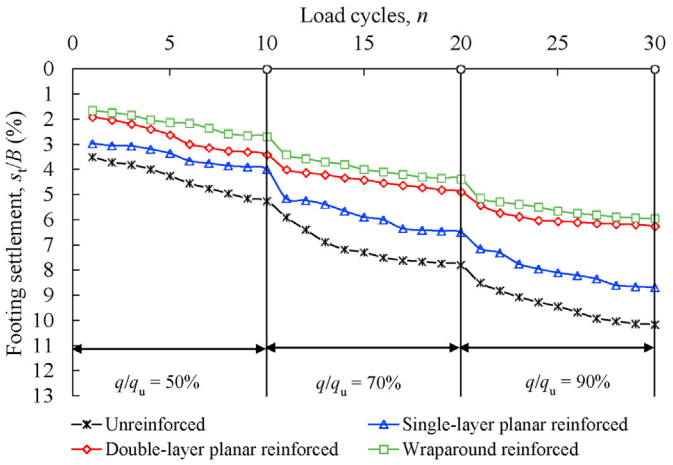


Fig. 16. Effect of incremental repeated loading on the settlement of unreinforced and reinforced model foundations.

This is even true for the initial few cycles in which most of the settlement occurs.

(5) The wraparound reinforced model has shown about 45% lower average total settlement in comparison to unreinforced model, while the double-layer planar reinforced model has about 41% lower average total settlement at the cost of twice the amount of material and 1.5 times the occupied land width ratio. Moreover, for an incremental repeated loading, the wraparound reinforced model performed at par with the double-layer models.

The present study has been carried out using the small size footing under 1*g* conditions. The load carrying capacity of the sand bed is dependent on the footing size. Therefore, for more comprehensive understanding, it is recommended to perform further tests using field-scale tests on large-sized footings. In future, particle chasing technology, image capturing and strain measuring devices can be employed to explicitly confirm the assumptions and considerations regarding the scale effect and to modify or update the proposed theoretical mechanism suggested for wraparound reinforcement.

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.

Acknowledgments

This research was jointly funded by the Higher Education Commission (HEC), Government of the Islamic Republic of Pakistan and Edith Cowan University, Perth, Australia.

References

Abdel-Baki, S., Raymond, G., 1993. Soil reinforcement for shallow foundation. In: Proceeding of the 2nd Geotechnical Engineering Conference. Cairo, Egypt, pp. 488–499.

Abu-Farsakh, M., Chen, Q., Sharma, R., 2013. An experimental evaluation of the behavior of footings on geosynthetic-reinforced sand. Soils Found. 53 (2), 335–348.

Adams, M.T., Collin, J.G., 1997. Large model spread footing load tests on geosynthetic reinforced soil foundations. J. Geotech. Geoenviron. Eng. 123 (1), 66–72.

Ahmad, H., Mahboubi, A., Noorzad, A., 2020. Scale effect study on the modulus of subgrade reaction of geogrid-reinforced soil. SN Applied Sciences 2 (3). https://doi.org/10.1007/s42452-020-2150-4.

Al Heib, M., Emeriault, F., Nghiem, H.L., 2020. On the use of 1g physical models for ground movements and soil-structure interaction problems. Journal of Rock Mechanics and Geotechnical Engineering 12 (1), 197–211.

Aria, S., Shukla, S.K., Mohyeddin, A., 2021. Behaviour of sandy soil reinforced with geotextile having partially and fully wrapped ends. Proceedings of the Institution of Civil Engineers – Ground Improvement 174 (1), 29–41.

Aria, S., Shukla, S.K., Mohyeddin, A., 2019. Numerical investigation of wraparound geotextile reinforcement technique for strengthening foundation soil. Int. J. GeoMech. 19 (4). https://doi.org/10.1061/(ASCE)GM.1943-5622.0001361.

Aria, S., Shukla, S.K., Mohyeddin, A., 2020. Optimum burial depth of geosynthetic reinforcement within sand bed based on numerical investigation. Int. J. Geotech. Eng. 14 (1), 71–79.

AS 1289.0-2000, 2000. Method of Testing Soils for Engineering Purposes – General Requirements and List of Methods. Standards Australian.

Asakereh, A., Ghazavi, M., Moghaddas Tafreshi, S.N., 2013. Cyclic response of footing on geogrid-reinforced sand with void. Soils Found. 53 (3), 363–374.

ASTM D1196/D1196M-12(2016), 2016. Standard Test Method for Nonrepetitive Static Plate Load Tests of Soils and Flexible Pavement Components, for Use in Evaluation and Design of Airport and Highway Pavements. ASTM International, West Conshohocken, USA.

Azzam, W.R., Nasr, A.M., 2015. Bearing capacity of shell strip footing on reinforced sand. J. Adv. Res. 6 (5), 727–737.

Basudhar, P.K., Saha, S., Deb, K., 2007. Circular footings resting on geotextile-reinforced sand bed. Geotext. Geomembranes 25 (6), 377–384.

Benmebarek, S., Djeridi, S., Benmebarek, N., Belounar, L., 2018. Improvement of bearing capacity of strip footing on reinforced sand. Int. J. Geotech. Eng. 12 (6), 537–545.

Biswas, A., Krishna, A.M., Dash, S.K., 2016. Behavior of geosynthetic reinforced soil foundation systems supported on stiff clay subgrade. Int. J. GeoMech. 16 (5), 04016007.

Biswas, A., Krishna, A.M., Dash, S.K., 2013. Influence of subgrade strength on the performance of geocell-reinforced foundation systems. Geosynth. Int. 20 (6), 376–388.

Boushehrian, A.H., Hataf, N., Ghahramani, A., 2011. Modeling of the cyclic behavior of shallow foundations resting on geomesh and grid-anchor reinforced sand. Geotext. Geomembranes 29 (3), 242–248.

Brachman, R.W.I., Moore, I.D., Rowe, R.K., 2001. The performance of a laboratory facility for evaluating the structural response of small-diameter buried pipes. Can. Geotech. J. 38 (2), 260–275.

Bransby, P.L., Smith, I.A.A., 1975. Side friction in model retaining-wall experiment. Journal of the Geotechnical Engineering Division, ASCE 101 (7), 615–632.

Cerato, A.B., Lutenegger, A.J., 2003. Scale effects of shallow foundation bearing capacity on granular material. J. Geotech. Geoenviron. Eng. 133 (10). https://doi.org/10.1061/(ASCE)1090-0241(2007)133:10(1192).

Chakraborty, M., Kumar, J., 2014. Bearing capacity of circular foundations reinforced with geogrid sheets. Soils Found. 54 (4), 820–832.

Chen, Q., Abu-Farsakh, M., 2011. Numerical analysis to study the scale effect of shallow foundation on reinforced soils. In: Han, J., Alzamora, D.E. (Eds.), Geo-Frontiers 2011: Advances in Geotechnical Engineering. American Society of Civil Engineers (ASCE), pp. 595–604.

Chen, Q., Abu-Farsakh, M., 2015. Ultimate bearing capacity analysis of strip footings on reinforced soil foundation. Soils Found. 55 (1), 74–85.

Chen, Q., 2007. An Experimental Study on Characteristics and Behavior of Reinforced Soil Foundation. PhD Thesis. Louisiana State University and Agricultural and Mechanical College, Baton Rouge, USA.

Cicek, E., Guler, E., Yetimoglu, T., 2015. Effect of reinforcement length for different geosynthetic reinforcements on strip footing on sand soil. Soils Found. 55 (4), 661–677.

Cicek, E., 2011. Analysis of Strip Footings on Geosynthetic Reinforced Sand. Institute of Science, Ataturk University, Erzurum, Turkey. PhD Thesis.

Cuelho, E., Perkins, S., Morris, Z., 2014. Relative Operational Performance of Geosynthetics Used as Subgrade Stabilization. Western Transportation Institute, Bozeman, USA.

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

Das, B.M., Puri, V.K., Omar, M.T., Evgin, E., 1996. Bearing capacity of strip foundation on geogrid-reinforced sand-scale effects in model tests. In: Proceedings of the 6th International Offshore and Polar Engineering Conference. Los Angeles, USA, pp. 527–530.

Das, B.M., Shin, E.C., 2009. Cyclic load-induced settlement of foundations on clay. In: Innovation in Civil and Construction Engineering. University of Missouri, pp. 241–246.

Das, B.M., Sobhan, K., 2017. Principles of Geotechnical Engineering, ninth ed. Cengage Learning.

El Sawwaf, M., Nazir, A.K., 2010. Behavior of repeatedly loaded rectangular footings resting on reinforced sand. Alexandria Engineering Journal 49 (4), 349–356.

Fakher, A., Jones, C.J.F.P., 1996. Discussion: bearing capacity of rectangular footings on geogrid-reinforced sand. Journal of Geotechnical Engineering 122 (4), 326–327.

Fragaszy, R.J., Lawton, E., 1984. Bearing capacity of reinforced sand subgrades. Journal of Geotechnical Engineering 110 (10), 1500–1507.

Ghotbi Siabil, S.M.A., Moghaddas Tafreshi, S.N.M., Dawson, A.R., 2020. Response of pavement foundations incorporating both geocells and expanded polystyrene (EPS) geofoam. Geotext. Geomembranes 48 (1), 1–23.

Guido, V.A., Chang, D.K., Sweeney, M.A., 1986. Comparison of geogrid and geotextile reinforced earth slabs. Can. Geotech. J. 23 (4), 435–440.

Habib, P.A., 1974. Scale effect for shallow footings on dense sand. Journal of the Geotechnical Engineering Division, ASCE 100 (1), 95–99.

Hou, J., Zhang, M.X., Dai, Z.H., Li, J.Z., Zeng, F.F., 2017. Bearing capacity of strip foundations in horizontal-vertical reinforced soils. Geotext. Geomembranes 45 (1), 29–34.

Hsieh, C., Mao, H.L., 2005. A bench-scale performance test for evaluation the geosynthetic reinforcement effects on granular base courses. In: Koerner, R.M., Koerner, G.R., Hsuan, Y.G., Ashley, M.V. (Eds.), Geo-Frontiers 2005: Geosynthetics Research and Development in Progress. ASCE, pp. 3963–3973.

Kazi, M., Shukla, S.K., Habibi, D., 2015a. An improved method to increase the load-bearing capacity of strip footing resting on geotextile-reinforced sand bed. Indian Geotech. J. 45 (1), 98–109.

Kazi, M., Shukla, S.K., Habibi, D., 2016. Behaviour of an embedded footing on geotextile-reinforced sand. Proceedings of the Institution of Civil Engineers – Ground Improvement 169 (2), 120–133.

Kazi, M., Shukla, S.K., Habibi, D., 2015b. Effect of submergence on settlement and bearing capacity of surface strip footing on geotextile-reinforced sand bed. International Journal of Geosynthetics and Ground Engineering 1 (1). https://doi.org/10.1007/s40891-014-0006-y.

Khing, K.H., Das, B.M., Puri, V.K., Cook, E.E., Yen, S.C., 1993. The bearing-capacity of a strip foundation on geogrid-reinforced sand. Geotext. Geomembranes 12 (4), 351–361.

Kou, Y., Shukla, S.K., Mohyeddin, A., 2018. Experimental investigation for pressure distribution on flexible conduit covered with sandy soil reinforced with geotextile reinforcement of varying widths. Tunn. Undergr. Space Technol. 80, 151–163.

Kumar, J., Sahoo, J.P., 2013. Bearing capacity of strip foundations reinforced with geogrid sheets by using upper bound finite-element limit analysis. Int. J. Numer. Anal. Methods GeoMech. 37 (18), 3258–3277.

Kusakabe, O., 1995. Chapter 6: foundations. In: Taylor, R.N. (Ed.), Geotechnical Centrifuge Technology. CRC Press, London, pp. 118–167.

Lal, D., Sankar, N., Chandrakaran, S., 2017. Effect of reinforcement form on the behaviour of coir geotextile reinforced sand beds. Soils Found. 57 (2), 227–236.

Latha, G.M., Murthy, V.S., 2007. Effects of reinforcement form on the behavior of geosynthetic reinforced sand. Geotext. Geomembranes 25 (1), 23–32.

Leng, J., Gabr, M.A., 2002. Characteristics of geogrid-reinforced aggregate under cyclic load. Transport. Res. Rec.: Journal of the Transportation Research Board 1786 (1), 29–35.

Li, L., Xiao, H., Gao, M., Liu, Y., Hu, Z., Yao, K., 2019. Experimental investigations on effect of geocell, waste tire chips, and geocell-tire chips on foundation reinforcement. J. Perform. Constr. Facil. 33 (6). https://doi.org/10.1061/(ASCE)CF.1943-5509.0001352.

Michalowski, R.L., 2004. Limit loads on reinforced foundation soils. J. Geotech. Geoenviron. Eng. 130 (4), 381–390.

Moghaddas Tafreshi, S.N., Dawson, A.R., 2012a. A comparison of static and cyclic loading responses of foundations on geocell-reinforced sand. Geotext. Geomembranes 32, 55–68.

Moghaddas Tafreshi, S.N., Dawson, A.R., 2012b. Laboratory tests of footing supported on geotextile-reinforced sand under repeated loading. In: Proceedings of the 15th World Conference on Earthquake Engineering (15WCEE). Curran Associates, Inc., pp. 20132–20138

Moghaddas Tafreshi, S.N., Tavakoli Mehrjardi, G., Ahmadi, M., 2011. Experimental and numerical investigation on circular footing subjected to incremental cyclic loads. Int. J. Civ. Eng. 9 (4), 265–274.

Moghaddas Tafreshi, S.N.M., Dawson, A.R., 2010. Behaviour of footings on reinforced sand subjected to repeated loading – comparing use of 3D and planar geotextile. Geotext. Geomembranes 28 (5), 434–447.

Moghaddas Tafreshi, S.N.M., Khalaj, O., Dawson, A.R., 2014. Repeated loading of soil containing granulated rubber and multiple geocell layers. Geotext. Geomembranes 42 (1), 25–38.

Mosallanezhad, M., Hataf, N., Sadat Taghavi, S.H., 2016. Experimental and large-scale field tests of grid-anchor system performance in increasing the ultimate bearing capacity of granular soils. Can. Geotech. J. 53 (7), 1047–1058.

Omar, M.T., Das, B.M., Puri, V.K., Yen, S.C., 1993. Ultimate bearing capacity of shallow foundations on sand with geogrid reinforcement. Can. Geotech. J. 30 (3), 545–549.

Raja, M.N.A., Shukla, S.K., 2020a. An extreme learning machine model for geosynthetic-reinforced sandy soil foundations. Proceedings of the Institution of Civil Engineers – Geotechnical Engineering. https://doi.org/10.1680/jgeen.19.00297.

Raja, M.N.A., Shukla, S.K., 2020b. Ultimate bearing capacity of strip footing resting on soil bed strengthened by wraparound geosynthetic reinforcement technique. Geotext. Geomembranes 48 (6), 867–874.

Raja, M.N.A., Shukla, S.K., 2021. Multivariate adaptive regression splines model for reinforced soil foundations. Geosynth. Int. https://doi.org/10.1680/jgein.20.00049.

Raymond, G.P., 2002. Reinforced ballast behaviour subjected to repeated load. Geotext. Geomembranes 20 (1), 39–61.

Roy, S.S., Deb, K., 2017. Bearing capacity of rectangular footings on multilayer geosynthetic-reinforced granular fill over soft soil. Int. J. GeoMech. 17 (9). https://doi.org/10.1061/(ASCE)GM.1943-5622.0000959.

Shajarati, A., Sørensen, K.W., Nielsen, S.D., Ibsen, L.B., 2012. Behaviour of cohesionless soils during cyclic loading. In: DCE Technical Memorandum No. 14. Aalborg. Aalborg University, Denmark.

Shin, E.C., Das, B.M., 2000. Experimental study of bearing capacity of a strip foundation on geogrid-reinforced sand. Geosynth. Int. 7 (1), 59–71.

Shukla, S.K., 2016. An Introduction to Geosynthetic Engineering. CRC Press/Taylor and Francis, London.

Shukla, S.K., 2014. Core Principles of Soil Mechanics. ICE Publishing, London.

Tatsuoka, F., Okahara, M., Tanaka, T., Tani, K., Morimoto, T., Siddiquee, M.S.A., 1991. Progressive failure and particle size effect in bearing capacity of a footing on sand. In: McLean, F.G., Campbell, D.A., Harris, D.W. (Eds.), Geotechnical Engineering Congress–1991, vol. 2. ASCE, pp. 788–802.

Tavakoli Mehrjardi, G., Khazaei, M., 2017. Scale effect on the behaviour of geogrid-reinforced soil under repeated loads. Geotext. Geomembranes 45 (6), 603–615.

Terzaghi, K., Peck, R.B., Mesri, G., 1996. Soil Mechanics in Engineering Practice. John Wiley & Sons.

Wang, J.Q., Zhang, L.L., Xue, J.F., Tang, Y., 2018. Load-settlement response of shallow square footings on geogrid-reinforced sand under cyclic loading. Geotext. Geomembranes 46 (5), 586–596.

Yamaguchi, H., Kimura, T., Fujii, N., 1977. On the scale effect of footings in dense sand. In: Proceedings of the 9th International Conference on Soil Mechanics and Foundation Engineering. Tokyo, Japan, pp. 795–798.

Yeo, B., Yen, S.C., Puri, V.K., Das, B.M., Wright, M.A., 1993. A laboratory investigation into the settlement of a foundation on geogrid-reinforced sand due to cyclic load. Geotech. Geol. Eng. 11 (1), 1–14.

Yetimoglu, T., Wu, J.T.H., Saglamer, A., 1994. Bearing capacity of rectangular footings on geogrid-reinforced sand. Journal of Geotechnical Engineering 120 (12), 2083–2099.

Muhammad Nouman Amjad Raja obtained his BSc degree from Department of Civil Engineering, University of Engineering and Technology Lahore, Pakistan in 2012 and his MSc degree in 2016 from Department of Civil, Geo and Environmental Engineering, Technische Universität München, Germany. He is currently a doctoral candidate at Geotechnical and Geoenvironmental Research Group, School of Engineering, Edith Cowan University, Australia. His current research focus is the investigation of geosynthetic-reinforced soil bed using experimental, analytical, and intelligent modelling techniques. He has already published several papers in reputed geotechnical engineering journals.

Please cite this article as: Raja MNA, Shukla SK, Experimental study on repeatedly loaded foundation soil strengthened by wraparound geosynthetic reinforcement technique, Journal of Rock Mechanics and Geotechnical Engineering, https://doi.org/10.1016/j.jrmge.2021.02.001

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397101

Fractional Order Sinusoidal Oscillator Using only a Single Operational Trans-resistance Amplifier

Manoj Kumar
Department of Electrical Engineering,
Delhi Technological University,
Delhi-110042 (India)
manoj.electro@gmail.com

D. R. Bhaskar
Department of Electrical Engineering,
Delhi Technological University,
Delhi-110042 (India)
drbhaskar@dtu.ac.in

Pragati Kumar
Department of Electrical Engineering,
Delhi Technological University,
Delhi-110042 (India)
pragatikumar@dtu.ac.in

Abstract— In this paper a fractional order sinusoidal oscillator (FSO) implemented with two fractional capacitors (FC), four resistors and a single operational trans-resistance amplifier (OTRA) has been presented. The characteristic equation of proposed fractional order oscillator has been derived using nodal analysis and the expressions for frequency of oscillation, condition of oscillation and phase difference between the output voltages are obtained. In one special case, when $\alpha = \beta = 1.0$, the circuit behaves as a classical single resistance controlled sinusoidal oscillator (SRCO) in which the frequency of oscillation and the condition of oscillation can be varied through separate resistors. These properties are retained even when the value of α and β lie in the range of $0 < \alpha, \beta < 1$. Since the input terminals of OTRA are at virtually ground, the proposed oscillator circuit is also insensitive to parasitic input capacitances and input resistances. The workability of proposed structure has been verified with PSPICE simulation results using a single OTRA (constructed with CMOS transistors using TMSC 0.18 µm technology) and fractional capacitors realized using the method proposed by Valsa and Dovrak and Friedl.

Index Terms*— Analog signals processing, Fractional capacitor, Fractional order analog circuit, Fractional order oscillator, Operational trans-resistance amplifier.

I. INTRODUCTION

Unlike the conventional harmonic oscillators which are realized with different types of amplifiers and conventional reactive elements, like capacitors/inductors and resistors, a fractional order oscillator is realized with amplifiers, resistors and fractance devices (fractional order capacitors/fractional order inductors). Though the concept of fractional order oscillator is not new [1], the research work in the area of fractional order oscillator has gathered momentum till 2001 when a fractional order Wein-bridge oscillator was proposed, wherein, the conventional capacitors in the classical Wein-bridge oscillator circuit were replaced by fractional order capacitors. It was shown that sustained oscillations were possible in this circuit the frequency of oscillation (FO) and the condition of oscillation (CO) were dependent not only on the values of resistors and capacitors, but also on the values of α, β, the parameters, defining the fractional order capacitors [2]. The general theoretical framework for the design of fractional order oscillator with n- fractance devices was developed by Radwan, Elwakil and Soliman [3]-[4] and since then a number of fractional order oscillator circuits [1-21], realized either by replacing the conventional capacitors with fractional order capacitors [4]-[6], [8], [14], [16]-[18], [20] or designed ab-initio [7], [9]-[13], [15], [19], [21] have appeared in open literature. Of the fractional order oscillators which were not designed by replacing the conventional capacitor(s) of an existing oscillator, current controlled current follower trans-conductance amplifier [9], second generation current conveyors and operational amplifier [10], differential voltage current conveyor [11], operational trans-resistance amplifiers (OTRA) [7], [12], operational trans-conductance amplifiers [13], [15], differential difference current conveyor [19] , multiplication mode current conveyor [21] have been used to propose fractional order oscillators with different properties.

Said, Radwan, Madian and Soliman, in [7], have presented a family of 1 and 2 OTRA based fractional order oscillators employing two fractional order capacitors and 2/3/4 resistors and discussed the operation of these circuits in detail, for various combinations of the fractional order parameters α and β. The single OTRA based circuits presented in Fig. 2a, and 2b, therein, employ two fractional order capacitors and two and three resistors respectively. In none of these circuits, there is any provision for varying the FO, once the circuit has been designed for a specified value of α and β, as all the passive components are present in expression for both CO and FO. This feature, if available, would increase the degree of freedom in designing the oscillator. The main aim of this paper, therefore, is to present a single OTRA based fractional order oscillator circuit employing two fractional order capacitors and four resistors in which, the FO, CO and the phase difference between the output voltages may be controlled by varying the value (s) of α and β.

Rest of paper is organised as follows: In section II, proposed structure of fractional order oscillator is described in details, while in section III, PSPICE simulation and theoretical results are given to validate the design perspective of fractional order oscillator circuit. Finally, in section IV conclusion of proposed fractional order oscillator is included.

II. PROPOSED STRUCTURE OF FRACTIONAL ORDER OSCILLATOR

An OTRA, whose circuit symbol is shown below in Fig.1, is the circuit implementation of a differential current controlled voltage source and belongs to the category of trans-resistance amplifier.

ISBN: 978-1-7281-8529-3/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:08:04 UTC from IEEE Xplore. Restrictions apply.

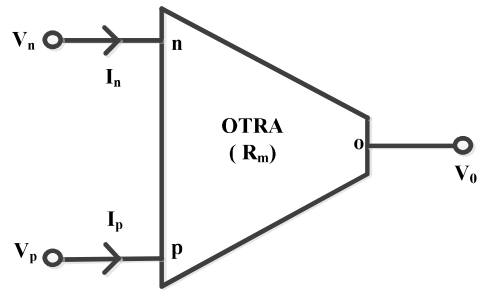


Fig.1. Symbol of OTRA

An ideal OTRA is characterized by the following terminal equations

$$V_o = R_m (I_p - I_n),\ V_p = 0,\ V_n = 0 \tag{1}$$

Where R_m is the trans-resistance gain of the OTRA should be infinity ideally. Under the conditions of negative feedback a very large value of R_m ($\rightarrow \infty$) forces the differential input (I_p-I_n) to be zero. The proposed OTRA based fractional order oscillator circuit is shown in Fig. 2.

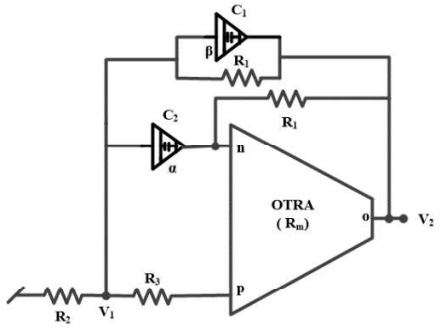


Fig.2. The proposed OTRA based fractional order oscillator

Applying KCL at node V_1, we obtain the following equation

$$V_1\left(\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}+C_2 s^{\alpha}+C_1 s^{\beta}\right) = V_0\left(\frac{1}{R_1}+C_1 s^{\beta}\right) \tag{2}$$

For an ideal OTRA, input current 'i_n' is forced to become equal to input current i_p, resulting in the following equation

$$V_1 = \frac{R_3}{R_1}\left(\frac{1}{1-R_3 C_2 s^{\alpha}}\right)V_2 \tag{3}$$

Solving equations (2) and (3), the characteristic equation (CE) can be expressed as:

$$s^{\alpha+\beta}+\frac{2}{R_1 C_1}s^{\alpha}+\frac{1}{C_2}\left(\frac{1}{R_1}-\frac{1}{R_3}\right)s^{\beta}+\frac{1}{R_1 C_1 C_2}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)=0 \tag{4}$$

The circuit represented by the above CE can support sustained oscillations, if, for the given values of the passive components and the fractional order parameters α and β, the equations (5) and (6) , obtained by equating the real part and imaginary part of the CE given in equation (4) admit a real solution.

$$\omega^{\alpha+\beta}\cos((\alpha+\beta)\pi/2)+\frac{2\omega^{\alpha}}{R_1 C_1}\cos(\alpha\pi/2)+\frac{1}{C_2}\left(\frac{1}{R_1}-\frac{1}{R_3}\right)\omega^{\beta}\cos(\beta\pi/2)+\frac{1}{R_1 C_1 C_2}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)=0 \tag{5}$$

$$\omega^{\alpha+\beta}\sin((\alpha+\beta)\pi/2)+\frac{2\omega^{\alpha}}{R_1 C_1}\sin(\alpha\pi/2)+\frac{1}{C_2}\left(\frac{1}{R_1}-\frac{1}{R_3}\right)\omega^{\beta}\sin(\beta\pi/2)=0 \tag{6}$$

The FO, and CO thus, obtained are given below in equations (7) and (8) respectively.

$$C_2\omega^{\alpha+\beta}(-\sin(\alpha\pi/2))+\frac{2\omega^{\alpha}C_2}{R_1 C_1}\sin((\beta-\alpha)\pi/2)+\frac{1}{R_1 C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)\sin(\beta\pi/2)=0 \tag{7}$$

The condition of oscillation (CO) is given by:

$$\frac{1}{R_3}=\frac{C_2\omega^{\alpha}\sin((\alpha+\beta)\pi/2)}{\sin((\beta)\pi/2)}+\frac{2C_2\omega^{\alpha-\beta}\sin((\alpha)\pi/2)}{R_1 C_1\sin((\beta)\pi/2)}+\frac{1}{R_1} \tag{8}$$

It is interesting to note that when α = β = 1 the circuit can operate like an integer order oscillator for which the FO and CO are given by

$$\text{FO: }\left(\frac{1}{R_1 C_2 C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)\right)^{\frac{1}{2}} \qquad \text{CO: }\frac{1}{R_3}=\frac{1}{R_1}\left(1+2\frac{C_2}{C_1}\right) \tag{9}$$

FO can be changed by varying the value of R_2 without disturbing the CO. Thus, the circuit can be operated as a single resistance controlled oscillator (SRCO). This SRCO has not been reported earlier in open literature.

Phase angle between V_1 and V_2 can be evaluated from equation (3) as

$$\varphi = \tan^{-1}\left(\frac{R_1 C_2\omega^{\alpha}\sin(0.5\alpha\pi)}{R_1 C_2\omega^{\alpha}\cos(0.5\alpha\pi)-\frac{R_1}{R_3}}\right) \tag{10}$$

III. Simulation and Numerical Results

The proposed structure of fractional order oscillator was simulated in PSPICE using CMOS OTRA structure [22] shown in Fig. 3. The circuit was designed for a nominal frequency of oscillation of 11.24Khz for (α = β = 1) by selecting the passive components as R_1= R_2= 20KΩ, and C_1= C_2= 1nF/$s^{(\alpha-1)}$. The workability of the circuit in fractional order domain was verified by selecting different combinations of α and β (0.8, 0.9) and measuring the frequency of oscillation and the corresponding phase difference between the available output voltages. The fractional order capacitors used in the simulation of proposed fractional order oscillator were designed using the method proposed by Valsa, Dovrak and Friedl [23]. The ladder structure (with a fifth order

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:08:04 UTC from IEEE Xplore. Restrictions apply.

approximation), along with the component values of the ladder elements are shown in Fig. 4 and Table I respectively. Nine different cases of fractional order oscillators (for different combinations of α and β) were simulated in PSPICE The condition of oscillation for each case was set by varying the value of the controlling resistor (R_3,) as given in equation (8). The transient responses and frequency spectrum of proposed fractional order oscillators are displayed in Fig.5. The simulation results (SM) and theoretical results (TH) showing frequency of oscillation (FO) and the phase difference for different values of α and β, along with the values of controlling resistor (R_3) used in simulation of fractional order oscillator circuit are shown in Table II. The theoretical values of FO and phase have been calculated using equation (9) and (10) respectively, and it is found that these values approximately match with simulation results.

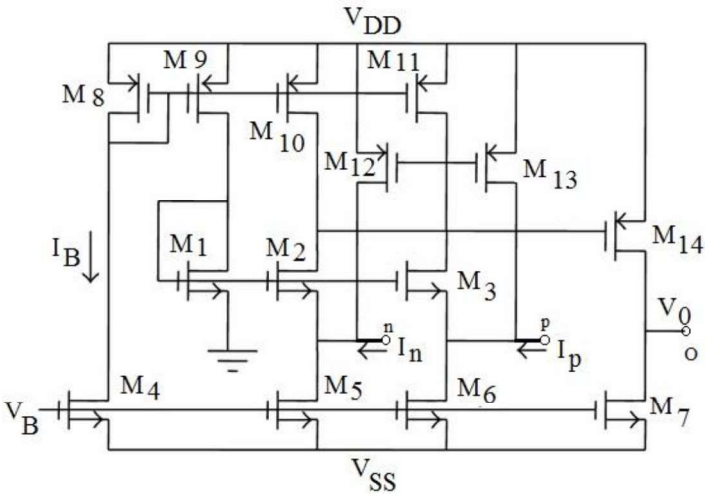
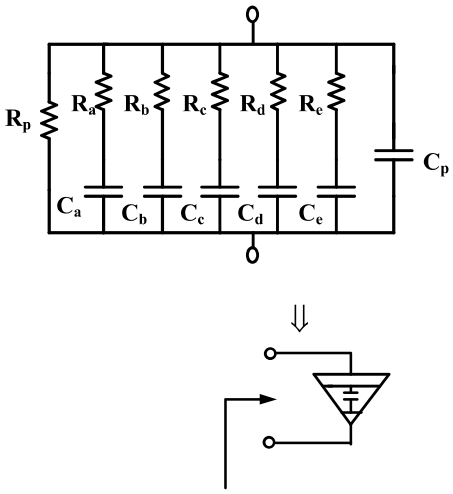


Fig.3. CMOS OTRA circuit proposed in [22]

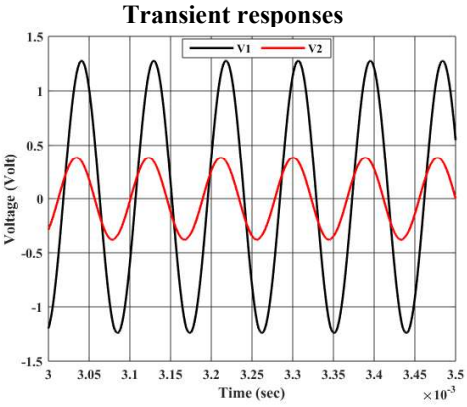


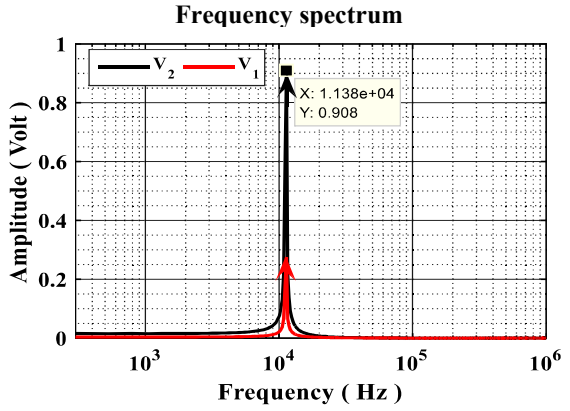
$$Z_{in} = \frac{1}{S^{\alpha} C_f}$$

Fig.4. RC-Ladder circuit to realize fractional order capacitors

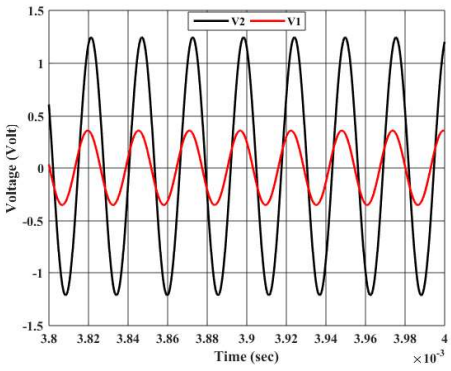
Cases

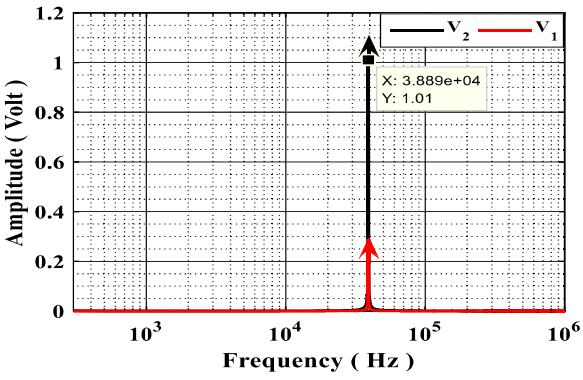
α = 1.0
β = 1.0





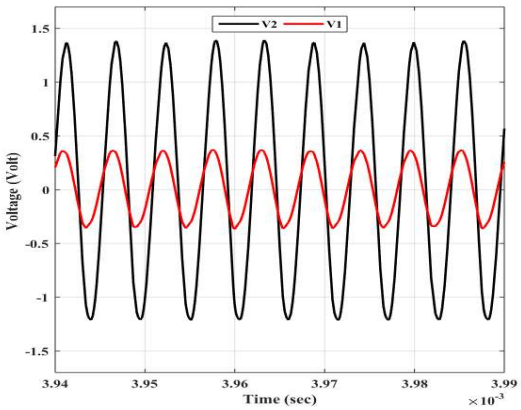
α = 0.9
β = 0.9

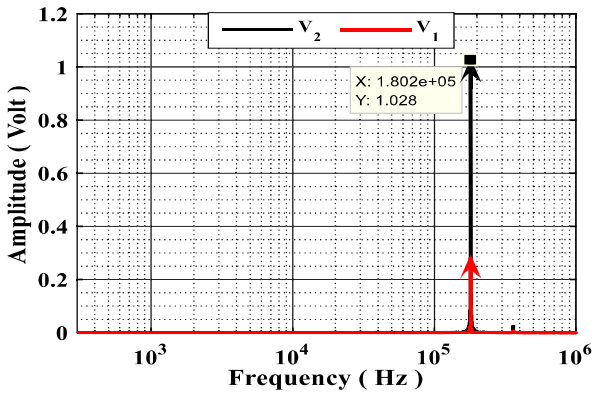




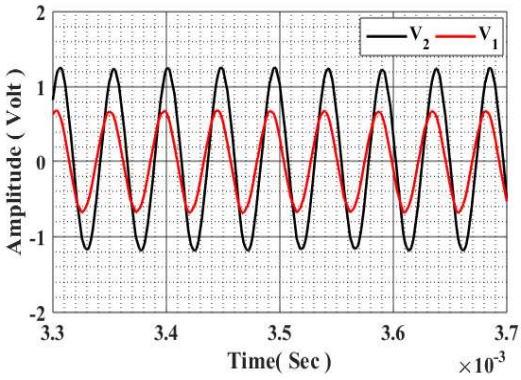
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:08:04 UTC from IEEE Xplore. Restrictions apply.

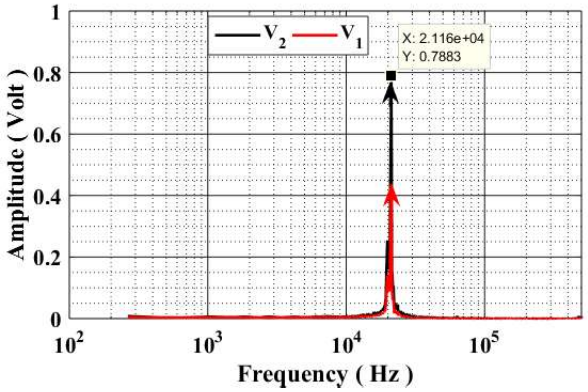
α = 0.8
β = 0.8



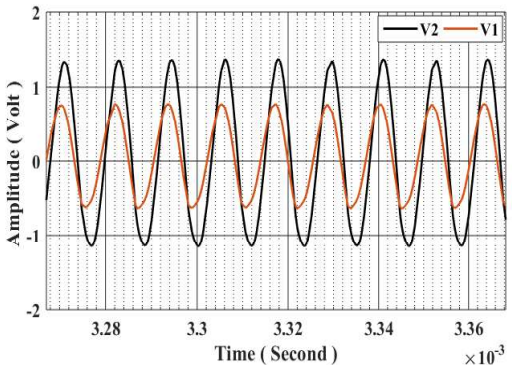


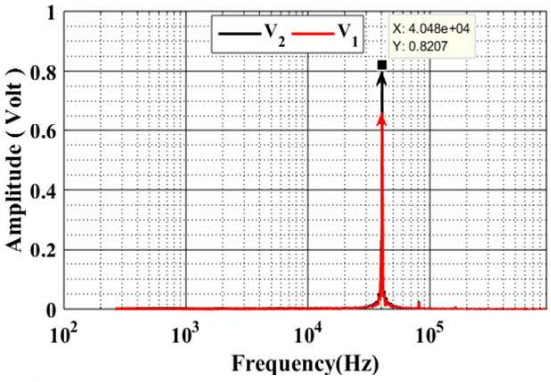
α = 0.8
β = 0.9



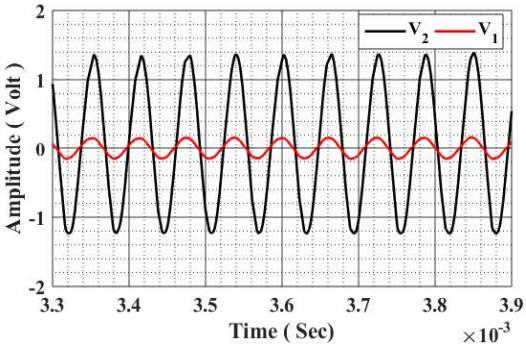


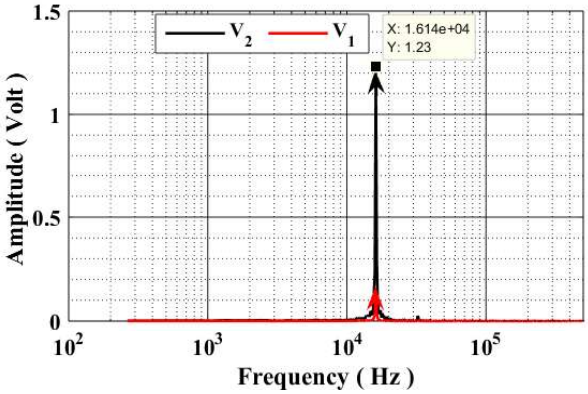
α = 0.8
β = 1.0





α = 0.9
β = 0.8





Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:08:04 UTC from IEEE Xplore. Restrictions apply.

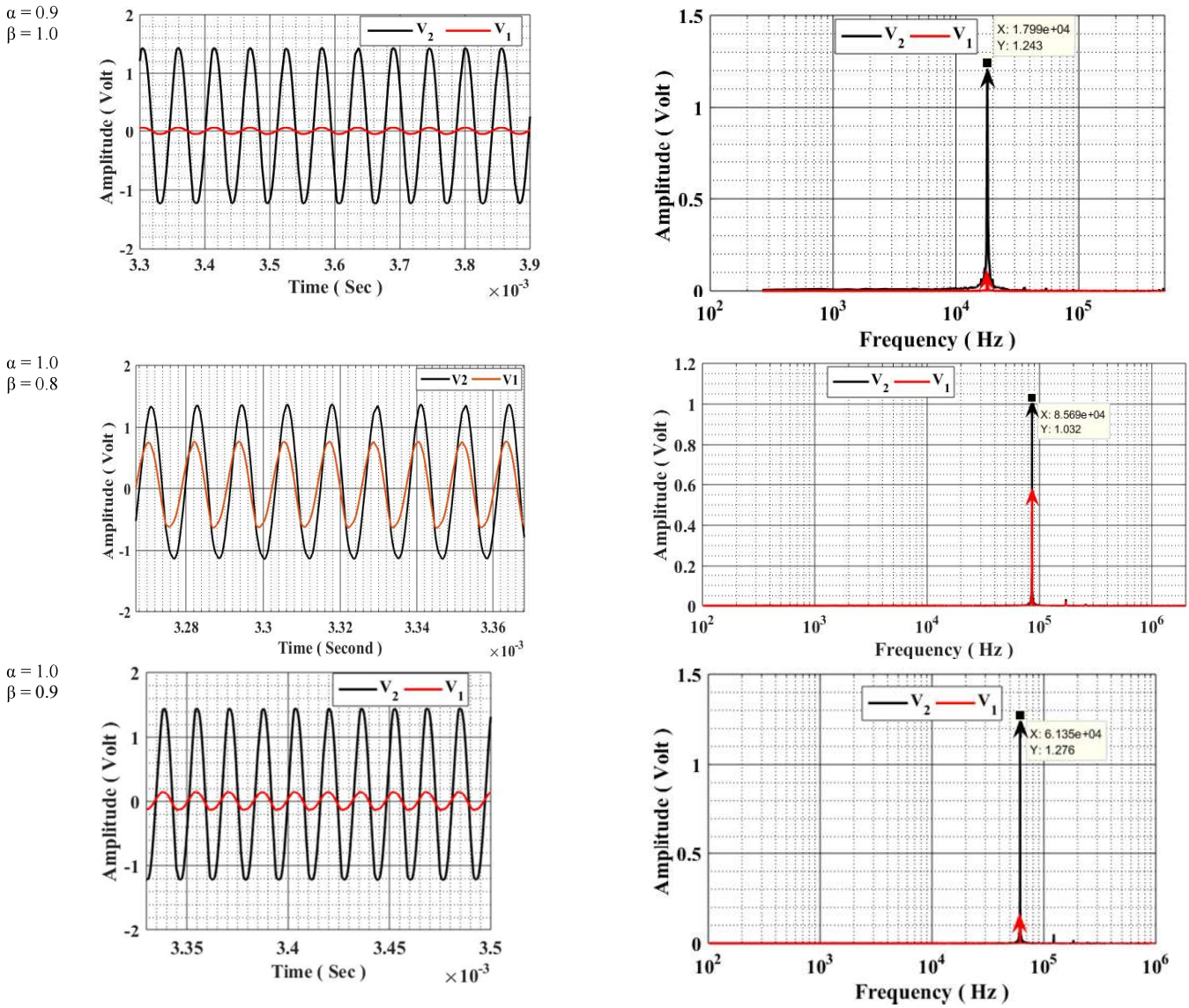


Fig.5. Transient responses and frequency spectrum of proposed fractional order oscillator for different values of α and β

TABLE I. COMPONENT VALUES OF 0.8 AND 0.9 ORDER FC

Order	$C_f = 1nF/s^{(\alpha-1)}$	
	Resistor	Capacitor
0.9	$R_a = 76.44M\Omega$	$C_a = 0.138nF$
	$R_b = 11.339M\Omega$	$C_b = 0.1058nF$
	$R_c = 1.682M\Omega$	$C_c = 0.856nF$
	$R_d = 0.25M\Omega$	$C_d = 0.0693nF$
	$R_e = 0.037M\Omega$	$C_e = 0.056nF$
	$R_P = 436.86M\ \Omega$	$C_P = 0.2372nF$
0.8	$R_a = 64.029M\Omega$	$C_a = 0.156nF$
	$R_b = 11.471M\Omega$	$C_b = 0.15058nF$
	$R_c = 2.153M\Omega$	$C_c = 0.0669nF$
	$R_d = 0.395M\Omega$	$C_d = 0.0438nF$
	$R_e = 0.072M\Omega$	$C_e = 0.0286nF$
	$R_P = 285.14M\ \Omega$	$C_P = 0.0542nF$

TABLE II. RESULTS OF FO, PHASE AND CONTROLLING RESISTOR (R_3)

Order	FO(SM/TH) (KHz)	Phase(SM/TH) (Degree)	R_3 (K Ω)
α = 1.0 β = 1.0	11.25/11.24	25.52/25.08	6.63
α = 0.9 β = 0.9	38.90/38.88	21.01/20.84	5.85
α = 0.8 β = 0.8	181.27/183.3	26.01/21.73	5.30
α = 0.8 β = 0.9	86.32/88.39	22.22/20.62	10.5
α = 0.8 β = 1.0	40.40/42.92	18.73/18.74	15.26
α = 0.9 β = 0.8	61.23/61.22	17.66/12.5	2.016
α = 0.9 β = 1.0	20.17/21.37	25.80/25.12	11.42

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:08:04 UTC from IEEE Xplore. Restrictions apply.

$\alpha = 1.0$ $\beta = 0.8$	18.00/18.02	7.72/5.39	0.824
$\alpha = 1.0$ $\beta = 0.9$	16.45/16.34	13.60/14.57	2.533

IV. Conclusion

A single OTRA-based FSO is presented, utilizing two fractional order capacitors and four resistors. The important parameters of FSO, like FO and CO and the phase difference between the output voltages can be controlled with fractional parameters 'α' and 'β'. In one of the special cases ($\alpha = 1$ and $\beta = 1.0$), the resultant oscillator circuit has the property of classical SRCO with independent control of frequency of oscillation and condition of oscillation, in which FO is controlled through R_2, while CO is controllable through R_3 without affecting each other, a property, which was not available in the existing single-OTRA based integer order oscillators. The SRCO property of the circuit will also hold for the value of α and β for any range between 0 and 1. PSPICE simulation results have been provided to validate the proposed FSO using CMOS OTRA with TMSC 0.18 μm technology.

Acknowledgment

The authors gratefully acknowledge Uttarakhand Technical University, Dehradun for funding this work under Competitive Research Fund of TEQIP-III program.

References

[1]. A. Oustaloup, "Fractional order sinusoidal oscillators: optimization and their use in highly linear FM modulation," IEEE Transactions on Circuits and Systems vol. 28, no. 3, pp. 1007-1009, 1981.

[2]. W. Ahmad, R. El-Khazali, and A. S. Elwakil, "Fractional-order Wien-bridge oscillator," Electronics Letters, vol. 37, no. 18, pp. 1110-1112,2001.

[3]. A. G. Radwan, A. S. Elwakil, & A. M. Soliman, "Fractional-order sinusoidal oscillators: design procedure and practical examples," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 55, no. 7, pp. 2051-2063, 2008

[4]. A. G. Radwan, A. M. Soliman, and A. S. Elwakil, "Design equations for fractional-order sinusoidal oscillators: four practical circuits examples," Int. J .Circ. Theory Appl., vol. 36, pp. 473-492, 2007.

[5]. A. L Said, A. G. Radwan, A.H. Madian, and A. M. Soliman, "Fractional order two port network oscillator with equal order," In 2014 26th International Conference on Microelectronics (ICM), pp. 156-159, 2014.

[6]. L. A. Said, A. H. Madian, A. G. Radwan and A. M. Soliman, "Current feedback operational amplifier (CFOA) based fractional order oscillators," 2014 21st IEEE International Conference on Electronics, Circuits and Systems (ICECS), Marseille, 2014, pp. 510-513, doi: 10.1109/ICECS.2014.7050034.

[7]. L. A. Said, A. G. Radwan A. H. Madian, and A. M. Soliman, "Fractional order Oscillators Based on Operational Transresistance Amplifiers," Int.J. Electron. Comm. (AEÜ), vol.69, pp. 988-1003, 2015.

[8]. L. A. Said, A. G. Radwan, A. H. Madian and A. M. Soliman, "Fractional-order oscillator based on single CCII," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), Vienna, 2016, pp. 603-606, doi: 10.1109/TSP.2016.7760952.

[9]. T. Comedang and P. Intani, "Current-Controlled CFTA Based Fractional Order Quadrature Oscillators," *Circuits and Systems*, vol. **7**, pp. 4201-4212, 2016, doi: 10.4236/cs.2016.713345.

[10]. L. A. Said, A. G. Radwan, A. H. Madian and A. M. Soliman, "Fractional order oscillator design based on two-port network," Circuits Syst. Signal Processing, vol. 35 pp. 3086–3112,2016.

[11]. D. Kubánek, F. Khateb, G. Tsirimokou, and C. Psychalinos, "Practical design and evaluation of fractional-order oscillator using differential voltage current conveyors," Circuits, Systems, and Signal Processing, vol. 35, no. 6, pp. 2003-2016, 2016.

[12]. L. A. Said, A. H. Madian, A. G. Radwan and A. M. Soliman, "Fractional order oscillator with independent control of phase and frequency," 2014 2nd International Conference on Electronic Design (ICED), Penang, 2014, pp. 224-229, doi: 10.1109/ICED.2014.7015803.

[13]. A. Kartci, N. Herencsar, J. Koton and C. Psychalinos, "Compact MOS-RC voltage-mode fractional-order oscillator design," 2017 European Conference on Circuit Theory and Design (ECCTD), Catania, 2017, pp. 1-4, doi: 10.1109/ECCTD.2017.8093281.

[14]. A. M. EL-Naggar, L. A. Said, A. G. Radwan, A. H. Madian, A. M. Soliman, "Fractional order four-phase oscillator based on double integrator topology," in IEEE Int. Conf. Modern Circ. and Syst. Technologies, pp. 1-4, 2017.

[15]. A. Kartci et al., "Fractional-order oscillator design using unity-gain voltage buffers and OTAs," 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, 2017, pp. 555-558, doi: 10.1109/MWSCAS.2017.8052983.

[16]. L. A. Said, A. G. Radwan, A. H. Madian, and A. M. Soliman, "Generalized family of fractional-order oscillators based on single CFOA and RC network," In 2017 6th International Conference on Modern Circuits and Systems Technologies (MOCAST), pp. 1-4, 2017.

[17]. L. A. Said, A. G. Radwan, A. H. Madian, and A. M. Soliman, "Three fractional-order-capacitors-based oscillators with controllable phase and frequency," Journal of Circuits, Systems and Computers, vol. 26, no. 10, 1750160, 2017.

[18]. S. K. Mishra, D. K. Upadhyay, and M. Gupta, "An approach to improve the performance of fractional-order sinusoidal oscillators," Chaos, Solitons & Fractals, Elsevier, vol. 116(C), pp. 126-135, 2018.

[19]. S. K. Mishra, M. Gupta and D. K. Upadhyay, "Design and implementation of DDCC-based fractional-order oscillator," International Journal of Electronics, vol. 106, no. 4, pp.581-598, 2019.

[20]. A. Pradhan, and R. K. Sharma, "Generalized Fractional-Order Oscillators using OTA," In 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 402-406, 2018.

[21]. A. Pradhan, K. S. Subhadhra, N. Atique, R. K. Sharma and S. S. Gupta, "MMCC-based current-mode fractional-order voltage-controlled oscillators," In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 763-768,2018.

[22]. H. Mostafa and A.M. Soliman, "A modified realization of the OTRA," Frequenz**, vol.** 60, pp. 3-4, 2006.

.

[23]. J. Valsa, P. Dvorak and M. Friedl, "Network model of the CPE," Radioengineering, vol. 20, no. 3, pp. 619–626, 2011.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:08:04 UTC from IEEE Xplore. Restrictions apply.

Nanoscale Advances

ROYAL SOCIETY OF CHEMISTRY

MINIREVIEW

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Nanoscale Adv.*, 2021, **3**, 2679

Gold nanomaterials for optical biosensing and bioimaging

Peng Si, [a] Nasrin Razmi,[b] Omer Nur, [b] Shipra Solanki,[cd] Chandra Mouli Pandey, [d] Rajinder K. Gupta,[d] Bansi D. Malhotra, [c] Magnus Willander*[b] and Adam de la Zerda*[a]

Gold nanoparticles (AuNPs) are highly compelling nanomaterials for biomedical studies due to their unique optical properties. By leveraging the versatile optical properties of different gold nanostructures, the performance of biosensing and biomedical imaging can be dramatically improved in terms of their sensitivity, specificity, speed, contrast, resolution and penetration depth. Here we review recent advances of optical biosensing and bioimaging techniques based on three major optical properties of AuNPs: surface plasmon resonance, surface enhanced Raman scattering and luminescence. We summarize the fabrication methods and optical properties of different types of AuNPs, highlight the emerging applications of these AuNPs for novel optical biosensors and biomedical imaging innovations, and discuss the future trends of AuNP-based optical biosensors and bioimaging as well as the challenges of implementing these techniques in preclinical and clinical investigations.

Received 16th November 2020
Accepted 12th March 2021

DOI: 10.1039/d0na00961j

rsc.li/nanoscale-advances

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

1. Introduction

Biological sensing and imaging are becoming increasingly important techniques for disease diagnostics and management due to their non-invasiveness, near real time feedback, high accuracy and reliability. Optical biosensing enables continuous monitoring of key health metabolites using wearable and hygienic devices. Optical imaging provides high-resolution images and does not require the use of radioactive contrast agents. These two techniques are further empowered by the development of nanotechnology, which significantly enhances their sensitivity, contrast, specificity and multiplexibility. Gold nanoparticles (AuNPs) are some of the most investigated nanotechnological tools for optical biosensing and bioimaging due to their easy fabrication, chemical stability, outstanding biocompatibility and versatile

[a]Department of Structural Biology, Stanford University, California, 94305, USA. E-mail: adlz@stanford.edu
[b]Department of Science and Technology, Physics and Electronics, Linköping University, SE-60174, Norrköping, Sweden. E-mail: magnus.willander@liu.se
[c]Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India
[d]Department of Applied Chemistry, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

Peng Si is currently a research scientist in Stanford University School of Medicine. He finished his postdoctoral training in Molecular Imaging Program at Stanford and received his PhD in Bioengineering from Nanyang Technological University, Singapore. His current research focuses on developing nanoscale contrast agents and molecular imaging methods for optical coherence tomography. He has published 16 peer-reviewed journal articles (12 first-authored) with over 1200 citations. He received World Molecular Imaging Congress Travel Award in 2017–2019.

Nasrin Razmi is an Aquasense Marie Curie Early Stage Researcher and a PhD student in Materials Science at Linkoping University, Sweden. Her main research interest centers around electrochemical sensors. She received her Master's degree in Biology-Biochemistry, and she holds a Bachelor's degree in Cell and Molecular Biology-Biochemistry.

© 2021 The Author(s). Published by the Royal Society of Chemistry

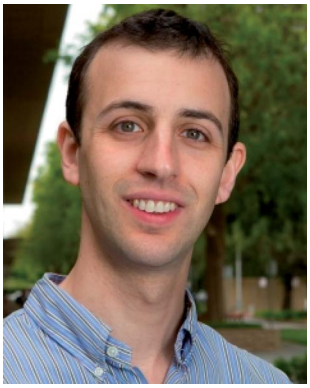
Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

optical properties. By engineering the shape and size of AuNPs, one can tune many optical properties of AuNPs including localized surface plasmon resonance (LSPR), ratio between light absorption and scattering coefficients, surface enhanced Raman scattering (SERS), fluorescence, *etc.* To date, researchers have fabricated dozens of Au nanostructures, including nanospheres (AuNS),[1] nanorods (AuNR),[2–4] nanoshells (AuNSh),[5–12] nanoprisms (AuNPr),[13–15] nanopyramids (AuNPy),[16] nanobipyramids (AuNBP),[17–19] nanocages (AuNC),[20–25] nanorings (AuNRg),[26–29] nanodisks (AuND),[22,30] nanostars (AuNSt),[31,32] nanorice,[33] nanobowls,[34] nanocrescents (AuNCr),[35,36] *etc.* (Fig. 1), each of which offers unique optical properties. In this article, we provide an overview of recent advances in optical biosensing and bioimaging techniques, which are enabled or enhanced by the unique optical properties of Au nanostructures. We focus on discussion of emerging biosensors and biomedical imaging which leverage the unique LSPR, SERS and luminescence properties of various gold nanomaterials (Fig. 2).

Chandra Mouli Pandey received his PhD in Chemistry from Banaras Hindu University, Varanasi, in collaboration with CSIR-National Physical Laboratory, New Delhi, in 2015. He is presently working as a DST-INSPIRE Faculty at the Department of Chemistry, Delhi Technological University, Delhi, on the development of biosensors for cancer biomarker detection. He has published more than 30 research papers in international journals and has recently authored a book entitled Biosensors: Fundamentals and Applications. His main research interest is the development of biosensors for health care based on nano-materials, self-assembled monolayers, conducting polymers and bio-nanocomposites.

Bansi D. Malhotra received his PhD from the University of Delhi, Delhi, in 1980. He has published more than 324 research papers in refereed international journals, has filed 12 patents, and has co-authored text books on "Nanomaterials for Biosensors: Fundamentals and Applications" and "Biosensors: Fundamentals and Applications". He is a recipient of the National Research Development Corporation Award 2005 for the invention of "Blood Glucose Biochemical Analyzer" and is a Fellow of the Indian National Science Academy, the National Academy of Sciences, India, and an Academician of the Asia Pacific Academy of Materials. His current activities include research and development of nanobiomaterials, biosensors, bio-fuel cells, ordered molecular assemblies, conducting polymers, Langmuir–Blodgett films, self-assembled monolayers, nano biotechnology, biomedical engineering, and biomolecular electronics. He is currently a SERB (Govt of India) Distinguished Fellow and Adjunct Professor at the Department of Biotechnology, Delhi Technological University, Delhi, India.

Magnus Willander is a professor (emeritus) in Linköping University and was earlier a professor in Gothenburg University, Sweden. He is an expert on nanomaterials and nanodevices including sensors. He has led several international projects, got several awards and published numerous scientific studies in his field.



Adam de la Zerda is an Associate Professor in the departments of Structural Biology and Electrical Engineering at Stanford University. He is working on the development of new cancer imaging technologies. He has received numerous awards including Forbes Magazine 30-under-30, Chan-Zuckerberg Biohub Investigator, Pew-Stewart Scholar, US Air Force Young Investigator Award, Dale F. Frey Award, NIH Director's Early Independence Award, Damon Runyon Fellowship, and Young Investigator Award at the World Molecular Imaging Congress. He has published over 35 papers in leading journals. He is the founder of a medical diagnostics start-up currently in stealth-mode called Visby Medical.

© 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

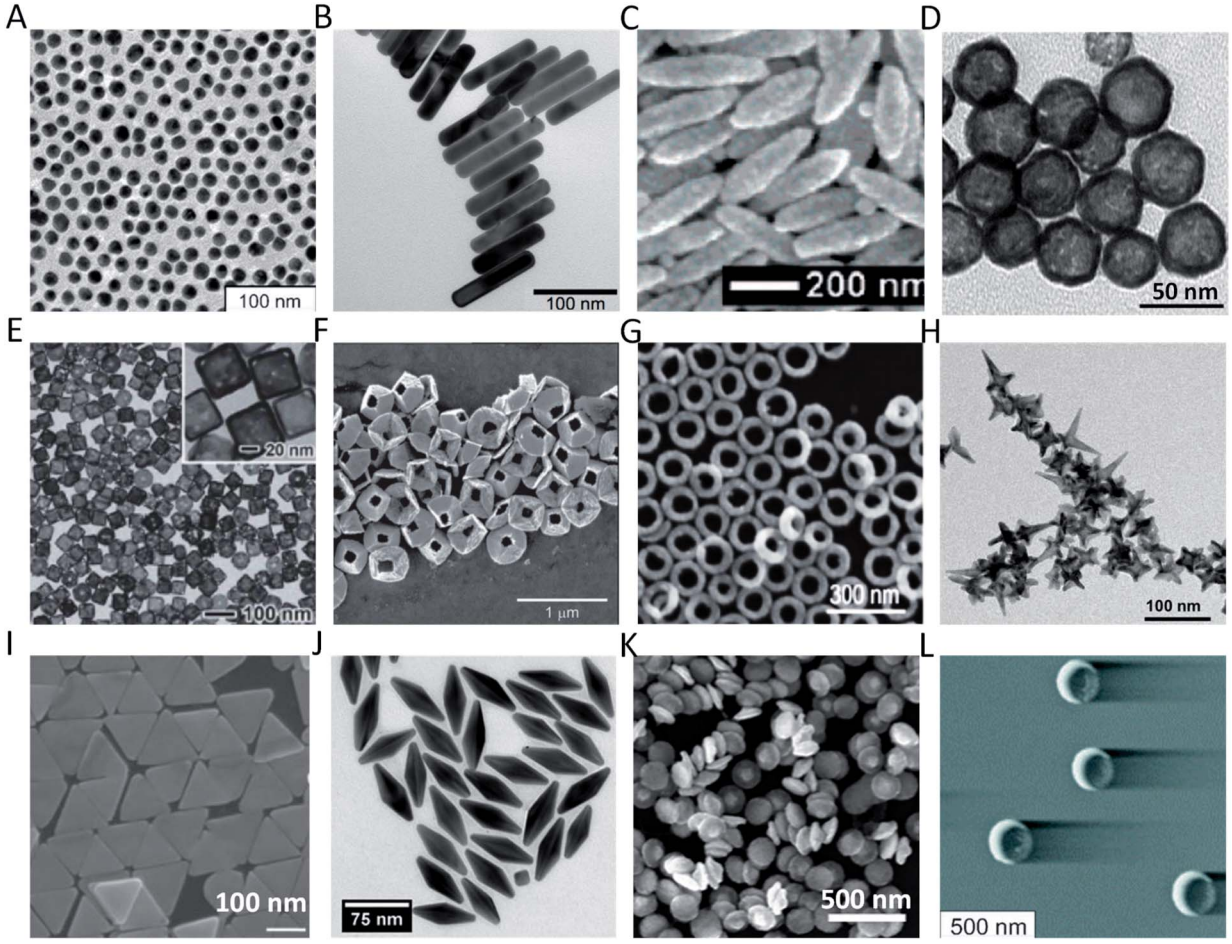


Fig. 1 Au nanostructures of different shapes and sizes with potential applications in optical biosensing and bioimaging: (A) 16 nm AuNS;[1] (B) AuNR;[3] (C) Au nanorice;[33] (D) AuNSh;[11] (E) AuNC;[25] (F) tipless AuNPy;[16] (G) AuNRg;[29] (H) AuNSt;[32] (I) AuNPr;[13] (J) AuNBP;[17] (K) AuND;[22] and (L) AuNCr.[35] Figures are adapted from references with permission. Copyright (2003, 2006, 2007, 2011, 2017) American Chemical Society. Copyright (2009) Elsevier Ltd.

2. Optical imaging and sensing based on the LSPR of AuNPs

2.1 Optical biosensing based on LSPR shift

Compared to other metal nanoparticles exhibiting LSPR, AuNPs have been found to be the most suitable due to their non-reactive nature and cost-effectiveness. AuNP-based biosensors designed by utilizing this property can be categorized into two different types: one relates to analysis of target analytes based on colorimetric methods developing color due to the aggregation of nanoparticles, and the other utilizes the change in the refractive index of AuNPs for detection.[37] The former assay requires a colloidal solution of AuNPs, whereas the latter method is based on a single AuNP which is usually mounted on a substrate.

AuNP-based colorimetric bioassays are some of the most successful optical biosensors as they involve fewer steps for detection and have also been commercialized. They are based on the wavelength shift caused by changes in interparticle distance.[38,39] The most common example is the pregnancy test kit in which specific antibody-coated nanoparticles show pink color on aggregation upon mixing with a solution containing the particular hormone.[40] Improved sensitivity and selectivity have been achieved by modifying AuNPs with other nanomaterials to make composites because the key objective in these assays is to control the inter-particle forces to manage aggregation. These types of assays are very simple and do not require any sophisticated instrumentation, and thus can be conveniently merged with smartphone imaging to make digital monitoring suitable for near-patient testing platforms.[41,42] A microfluidic colorimetric biosensor for *E. coli* detection based on AuNP aggregation was demonstrated by Zheng *et al.*[42] It was based on HRP + H_2O_2 + tyramine (Fig. 3). The device consisted of two mixing channels, one separation channel and a detection channel. Magnetic nanoparticle-modified *E. coli* capture antibodies, detection antibodies, catalase and a sample containing *E. coli* were first introduced into mixing channel 1. The magnetic nanoparticle-tagged bacterial conjugates containing the catalase enzyme were separated in the separation chamber. These bacterial conjugates were then mixed with hydrogen peroxide solution and then mixed with AuNP solution and

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

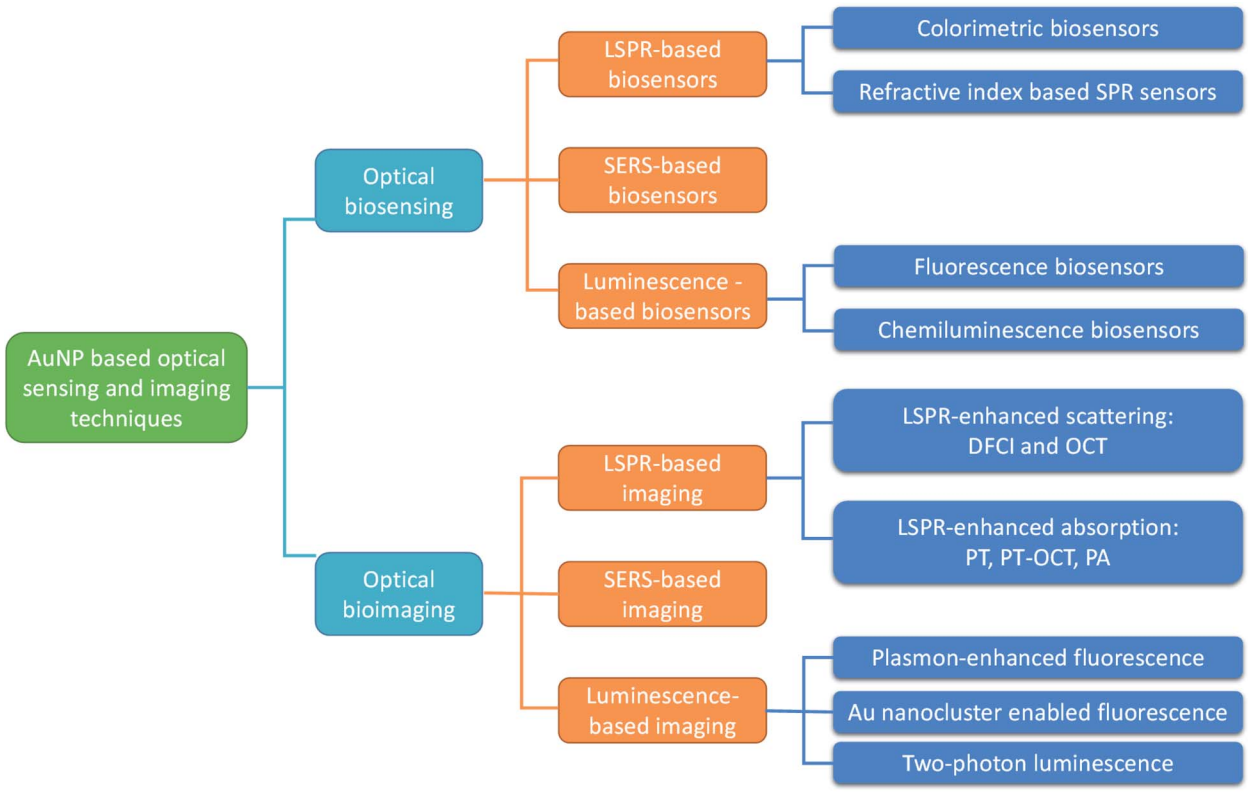


Fig. 2 Various optical biosensing and bioimaging techniques based on AuNPs.

crosslinking agents in the second mixing channel. The phenolic hydroxyl moieties resulted in the aggregation of AuNPs and color change from blue to red (Fig. 3B). Dynamic light scattering studies showed that the size of AuNPs changed from ~13 nm before aggregation to ~670 nm after aggregation and UV-visible studies showed that the aggregated AuNPs showed an absorption peak at 630 nm. This color change was detected using an app on a smart phone. Similarly, in another attempt to exploit the color changing property of AuNPs upon aggregation, a cysteine detection biosensor was designed in which the cysteine functionalized AuNPs did not form aggregates and thus no change in color occurred (Fig. 3C and D).[43]

Recent successes in the field of biosensors influenced by other fields have led to improved investment in colorimetric assays for point-of-care diagnostics.[44] Still, there remains the major challenge of avoiding non-specific aggregation triggering a color change in high ionic strength samples, including serum and urine. Other issues of concern include sample pretreatment requirements and lack of integration with other platforms. These limitations must be resolved and new strategies need to be developed to investigate different designs of AuNP-based biosensors.[45] Assembly-induced or aggregation-based colorimetric assays were developed and widely used for nucleic acid detection by the Mirkin research group.[46–48] They used AuNPs of size ~13 nm with an absorption band from 519–524 nm to tag single stranded DNA fragments and upon hybridization the decrease in distance between the AuNPs resulted in a color change. For example, the color of the solution containing unhybridized DNA was red and upon hybridization the color changed to blue due to the aggregation of AuNPs.[47] In another attempt, two different DNA sequences were identified on a single glass slide array.[48] Two different DNA fragments functionalized with ~50 nm and ~100 nm AuNPs were made for hybridization with probe DNA immobilized on the designed array and the color change occurring due to the aggregation of AuNPs was observed. The DNA fragments tagged with 50 nm AuNPs showed green color at a maximum absorption wavelength of 542 nm on the array after the hybridization event and the other tagged with 100 nm AuNPs showed orange color with a maximum absorption wavelength of 583 nm. Thus, the single array was able to differentiate between two different DNA fragments. In addition to nucleic acid detection, colorimetric sensors are among the most popular and the simplest methods designed for microbial detection. Bui and coworkers[49] reported a highly sensitive colorimetric immunoassay (without enzyme) based on plasmonic AuNPs of size 15 ± 2 nm conjugated to cysteine-loaded liposomes. A chemical cascade reaction due to the presence of a pathogen in the sample caused a visible color shift from red with an absorption peak at 520 nm to dark blue with two bands at 520 nm and 650 nm. The breakdown of cysteine-loaded nanoliposomes due to the presence of the microbes in the sample caused the aggregation of gold nanoparticles, resulting in a visible color change and pathogen detection. The fabricated sensing system was capable of visually detecting single cells of *Listeria*, *E. coli* 0157 and salmonella. In another study, Li and colleagues[50] proposed a visible

 © 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

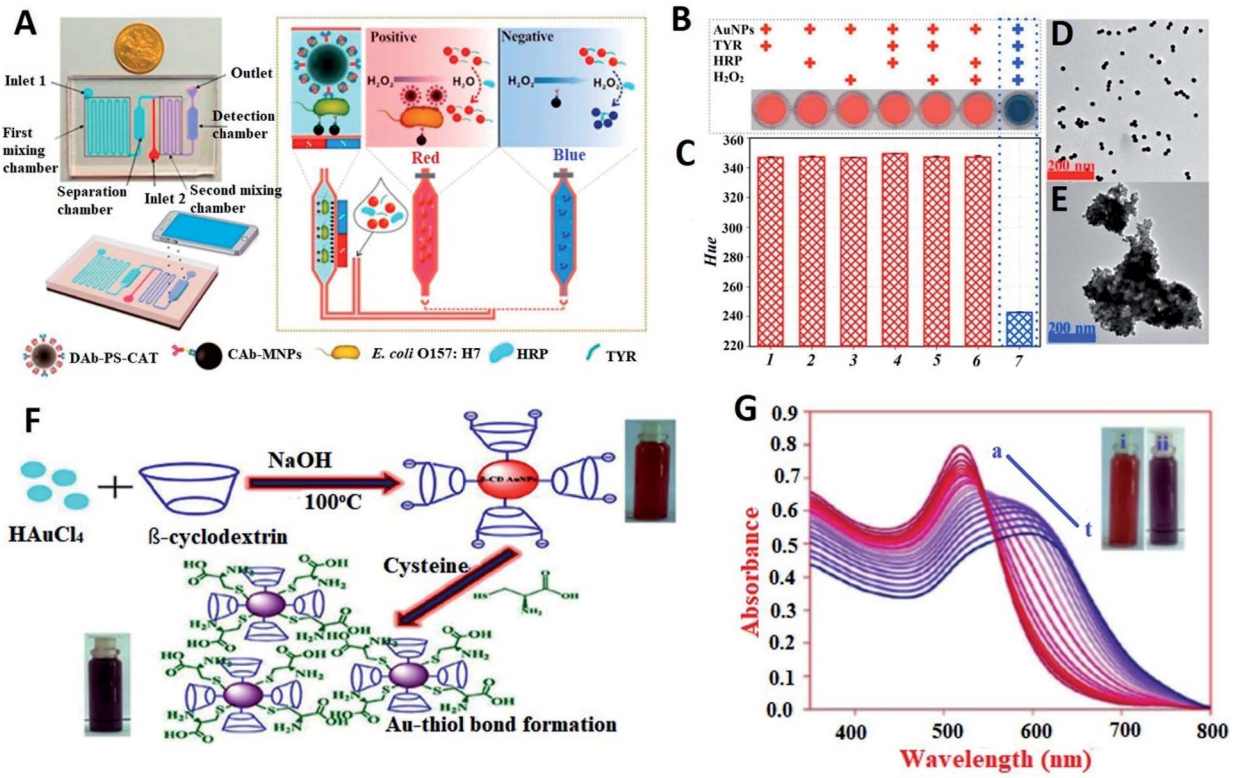


Fig. 3 AuNP-based optical biosensors. (A–E) A smartphone-based biosensor for colorimetric detection of *E. coli* O157:H7. (A) General schematic of the functioning of the biosensor; (B) results of mixing AuNPs with different combinations of cross-linking agents; (C) hue of the mixtures of the AuNPs with different combinations of cross-linking agents; (D) TEM image of the AuNPs before aggregation; (E) TEM image of the AuNPs after aggregation. Figures have been adapted from ref. 42 with permission from Elsevier, copyright 2019. (F and G) A AuNP-based Cys detection biosensor; (F) working principle of the biosensor and (G) SPR band intensity changes for β-CD AuNPs in the presence of incremental addition of Cys (the inset shows the digital images of β-CD AuNPs in the absence (i) and presence (ii) of Cys). Figures have been adapted from ref. 43 with permission from the Royal Society of Chemistry, copyright 2020.

colorimetric sensor using AuNPs with various surface charges allowing fast and differential response to 15 pathogens with different surface electronic properties in five seconds. A specific color shift pattern due to the interaction of gold nanoparticles and microorganisms was obtained for each of the microorganisms. The advantages of the proposed sensor are its simplicity, rapidness, and visual and label free detection. In addition, the proposed sensor is applicable in environmental monitoring and clinical diagnostics. Recently You *et al.*[51] fabricated a visual colorimetric nitrocellulose membrane strip-based sensing assay for Gram positive bacteria by using vancomycin-modified gold nanoparticles and one-pot reaction. They took advantage of the broad spectrum inhibiting activity against Gram-positive bacteria of the vancomycin antibiotic and 9.5 nm spherical AuNPs. The assay could differentiate between Gram-negative and Gram-positive bacteria. The obtained detection limits with the naked eye for all Gram-positive *S. aureus*, *M. luteus*, and *B. subtilis* were 1×10^9 cells per mL, and there was no interaction of the synthesized complex with Gram-negative bacteria such as *E. coli*. The proposed sensing array was able to detect *S. aureus* in tap water and orange juice.

Refractive index-based biosensors utilize the fact that the LSPR of AuNPs is sensitive to the changes near their surface, enabling them to monitor the biomolecular interactions with ultrahigh sensitivity. This is a label-free technique compared to the colorimetric assays. It has the advantage of preserving the structures of biomolecules and has been utilized to study the complete reaction kinetics and protein interactions. The plasmon resonances of different Au nanostructures have been evaluated to improve the performance and sensitivity of SPR assays. Even the amplification of SPR signals has been demonstrated by utilizing the coupling of LSPR of AuNPs with the propagating SPR signal of Au films leading to enhanced incident angle shift and higher sensitivity.[52] Active research is being conducted by different research groups to bring this technology to the market. However, challenges such as improvement in sensitivity, selectivity, detection limit, detection time and multiplexing capabilities need to be addressed. Using dark field microspectroscopy, Nusz and coworkers demonstrated that gold nanorods of dimensions 74 ± 9 nm (length) and 33 ± 6 nm (diameter) conjugated with biotin can detect streptavidin molecules at a sensitivity of 1 nM.[53] It is also possible for the system to detect serum proteins with nanomolar sensitivity.[54] The wavelengths studied in both the cases range from 700 to 800 nm. The readers can refer to the review article published by Hafner *et al.*[55] for more details regarding LSPR refractive index sensors. Here we focus on applications of this technique for bacteria detection. LSPR-based refractive index sensors have been used as a real time and rapid

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

methodology for bacteria detection by using different types of bioreceptors. However, in some measurements, interference of bacteria size can affect the detection limit, and in some cases, large size, high cost, complexity of the method and false result due to the effect of the composition of the sample or temperature, which causes refractive index fluctuations, are the disadvantages of SPR. Vaisocherová-Lísalová and coworkers constructed a rapid, sensitive and simultaneous SPR-based biosensor using streptavidin-coated gold nanoparticles and based on ultralow fouling and functionalizable poly(-carboxybetaine acrylamide) (pCBAA) brushes for multi-step detection of microorganisms in complex food samples. The proposed assay was composed of the following three steps: (i) crude food samples for sensor incubation, (ii) binding of secondary biotinylated antibodies, and (iii) binding of streptavidin-coated gold nanoparticles to the biotinylated antibodies. Limits of detection of 7.4×10^3 CFU per mL and 11.7×10^3 CFU per mL for *Salmonella* sp. and 57 CFU per mL and 17 CFU per mL for *E. coli* in cucumber and hamburger extracts were achieved.[56] Zhou *et al.* recently reported a biosensor based on nanogold enhanced surface plasmon resonance and colloidal gold test strips (ICTS) for *Vibrio parahaemolyticus* detection.[57] The response for the pathogen was improved dramatically after immunomagnetic separation.

2.2 Optical imaging based on LSPR-enhanced scattering

The scattered light intensity of AuNPs is about 10^5 to 10^6 times stronger than that of a fluorescein molecule and 100–1000 times stronger than that of a polymer bead.[58–60] Larger AuNPs show stronger scattering efficiencies than smaller AuNPs.[59] In contrast to fluorophores, light-scattering AuNPs are not photobleachable which makes them very attractive contrast agents for bioimaging. In this section, we focus on two imaging modalities which detect the backscattering of AuNPs: dark field confocal imaging (DFCI) and optical coherence tomography (OCT). DFCI is often used for *in vitro* imaging of cells and biomolecules and OCT is commonly used for *in vivo* imaging.

DFCI can detect AuNPs that are as small as 10 nm in diameter. Using a dark field microscope with a high NA lens, light scattered from AuNPs can be imaged as bright spots even though the dimensions of AuNPs are smaller than the diffraction limit of the lens. A major advantage of DFCI is that it allows high-contrast, true-color and multiplexed detection of gold nanoparticles with different sizes and shapes. Various types of AuNPs including AuNS, AuNR, AuNC and AuNSh have been implemented as DF optical contrast agents for cancer cell imaging by functionalizing their surface with ligand binding to cell surface receptors (Fig. 4).[61–65] AuNPs are also extensively utilized as DF probes to image a variety of intracellular biological processes such as cell division, endocytosis, virus infection, DNA damage, *etc.* For example, Qian *et al.* applied bioconjugated AuNS and DF imaging to track cell division in real time.[66] Wan and co-workers were able to label AuNPs on respiratory syncytial virus and visualize its infection of HEp-2 cells.[67] These studies have provided many new insights into intracellular events.

OCT is an interferometry-based imaging technique which illuminates samples with near-infrared (NIR) low-coherence light and detects the back-scattered light from samples. Intuitively, AuNPs are attractive OCT contrast agents because they

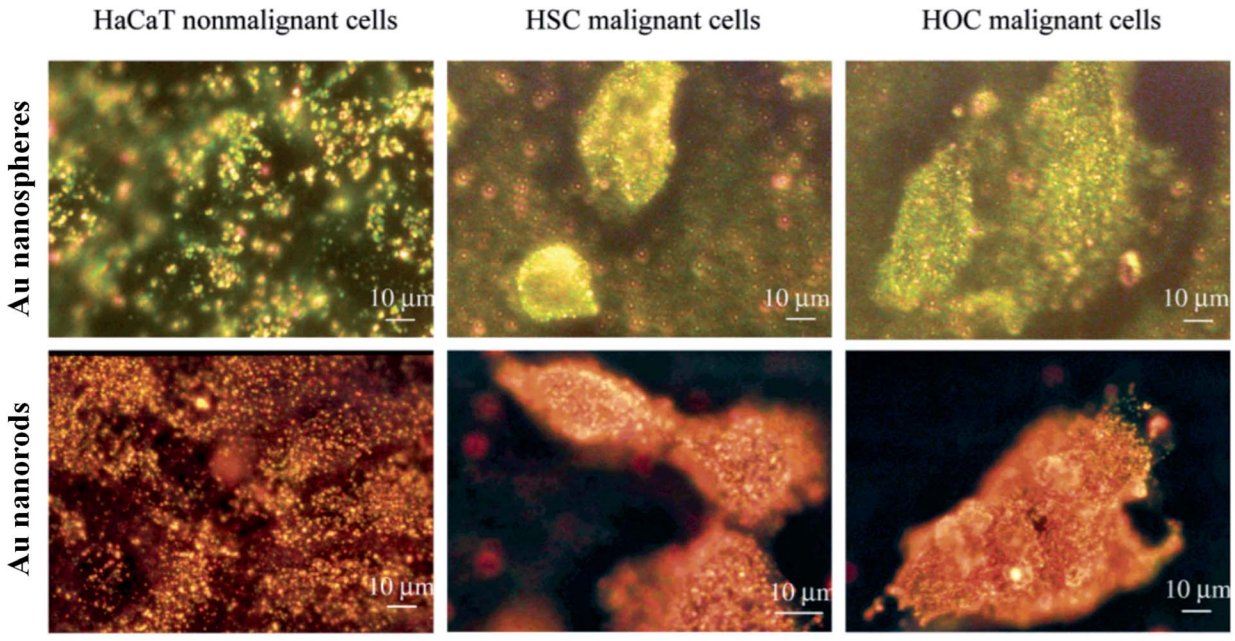


Fig. 4 Dark-field microscopy images of normal and two different types of cancer cells after incubating with AuNS/AuNR conjugated with anti-EGFR antibodies. AuNS show green to yellow color in the dark-field image (upper panel), indicating that their light scattering is mostly in the visible region. AuNR show orange to red color in the dark-field image (lower panel), indicating that their light scattering is mostly in the NIR region. Modified and reprinted with permission from ref. 156. Copyright (2006) American Chemical Society.

 © 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 1 Optical properties and optical imaging applications of different types of gold nanoparticles[a]

Au nanostructure	Fabrication method	LSPR (visible: 400–700 nm, NIR-I: 700–1000 nm, NIR-II: 1000–1700 nm)	Optical imaging modalities/ applications
Nanoshell	Overgrowth of core-bound particles[157]	NIR-I–NIR-II	OCT[5–10] PA[11,12]
Nanorod	Seeded growth with CTAB[158]	Visible–NIR-II	OCT[3,4,68–74] PA[99–104] TP fluorescence[147,148]
Nanostar	Seeded growth with CTAB or PVP; one-pot synthesis[159]	Visible–NIR-I	OCT[75] PA[105] SERS[31,32,118] TP fluorescence[149,150]
Nanocage	PVP-stabilized polyol, galvanic displacement[160]	Visible–NIR-I	OCT[20–22] PA[23,24]
Nanodisk	Nanoimprint lithography[22]	Visible–NIR-I	OCT[21,22]
Nanoring	Lithography using colloidal polystyrene nanoparticles as the template[27,28]	NIR-I–NIR-II	OCT[26,28]
Nanobipyramid	Seeded growth with CTAB and CTAC[17]	Visible–NIR-II	OCT[18,19] SERS[124,125]
Nanoprism	Seeded growth with CTAB[13]	NIR-I–NIR-II	OCT[14,77,161]
Nanopyramid	Lithography with patterned photoresist posts as the template[16,162]	Visible to NIR-I	SERS[16]
Nanocrescent	Lithography with a nanosphere template and vapor-phase deposition[35,36]	Visible to NIR-II	SERS[36,120]
Nanoclusters	Alkanethiol/phosphine-stabilized reduction[163]	Visible to NIR-I	Fluorescence[142–146]

[a] CTC: circulating tumor cells, OCT: optical coherence tomography, SERS: surface enhanced Raman scattering, TP: two-photon, NIR: near infrared, CTAB: cetyltrimethylammonium bromide, CTAC: cetyltrimethylammonium chloride; PVP: poly(vinylpyrrolidone).

can enhance the backscattering of light at a specific wavelength due to LSPR. LSPR also generates drastic difference in the intensity of backscattered light at different wavelengths, allowing different AuNPs to have unique spectral signatures. Due to their highly tunable plasmonic resonances and scattering coefficient, AuNPs of different shapes and dimensions have been extensively investigated as OCT contrast agents in recent years. AuNSh,[5–10] AuNR,[3,4,68–74] AuNSt,[75] AuNC,[20–22] AuNRg,[26,28] AuND,[21,22] AuNPr[14,15,76,77] and AuNBP[18,19] have been demonstrated to be promising contrast agents to improve OCT signals in tissue and blood vessels (Table 1). Among them, AuNSt, AuNC and AuND have plasmonic resonance peaks mainly in the first near infrared window (NIR-I, 700–1000 nm) so they are used as contrast agents for OCT with a light source in the NIR-I region; while AuNR, AuNSh, AuNRg, AuNPr and AuNBP can be employed as OCT contrast agents in both NIR-I and NIR-II windows. There are two commonly used methods to synthesize AuNPs as OCT contrast agents: nanolithography methods are used to fabricate gold nanodisks and nanorings, whereas seed-mediated growth is used to fabricate AuNPs of other geometries. The advantage of the nanolithography method is that it can fabricate AuNPs with precise dimensions with small variation in size, but it is expensive, time consuming and difficult to scale up the fabrication. In contrast, wet chemistry methods are low-cost, efficient and easy to scale up, but may have higher dimension variation and impurity.

AuNR are some of the most heavily investigated AuNPs as OCT contrast agents due to their highly tunable size and aspect ratio and facile preparation method. Their plasmonic resonance can be tuned from 600 to 1300 nm by adjusting their aspect ratio, and the scattering to absorption cross-section ratio can be changed by adjusting their size.[59,78] SoRelle *et al.*[3,4] showed that AuNR with larger dimensions (90 × 30 nm) exhibit 30-fold greater backscattering or OCT intensity than AuNR with a smaller size (50 × 15 nm). In addition, they show that two types of AuNR with different plasmonic resonances, 815 nm and 925 nm, can be used as spectral contrast agents and detected multiplexably by OCT. The AuNR spectral OCT contrast agents were then implemented for retinal imaging,[72] tracking of brain tumor-associated macrophages[79] and detecting circulating tumor cells.[80] Ratheesh *et al.*[71] demonstrated that AuNR with a higher aspect ratio of 8.8 and a plasmon-resonance peak at 1320 can be used as contrast agents to enhance OCT imaging in the 1300 nm range. Lippok and co-workers[73] demonstrated for the first time that AuNR can depolarize the backscattered OCT light and the depolarization signature of AuNR can be well detected in biological tissue.

AuNPr have been increasingly investigated as OCT contrast agents in recent years. Si *et al.* synthesized AuNPr with high scattering cross-section and LSPR in the NIR-II window, and demonstrated that intravenous injection of AuNPr can dramatically enhance the contrast of OCT vascular imaging,

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

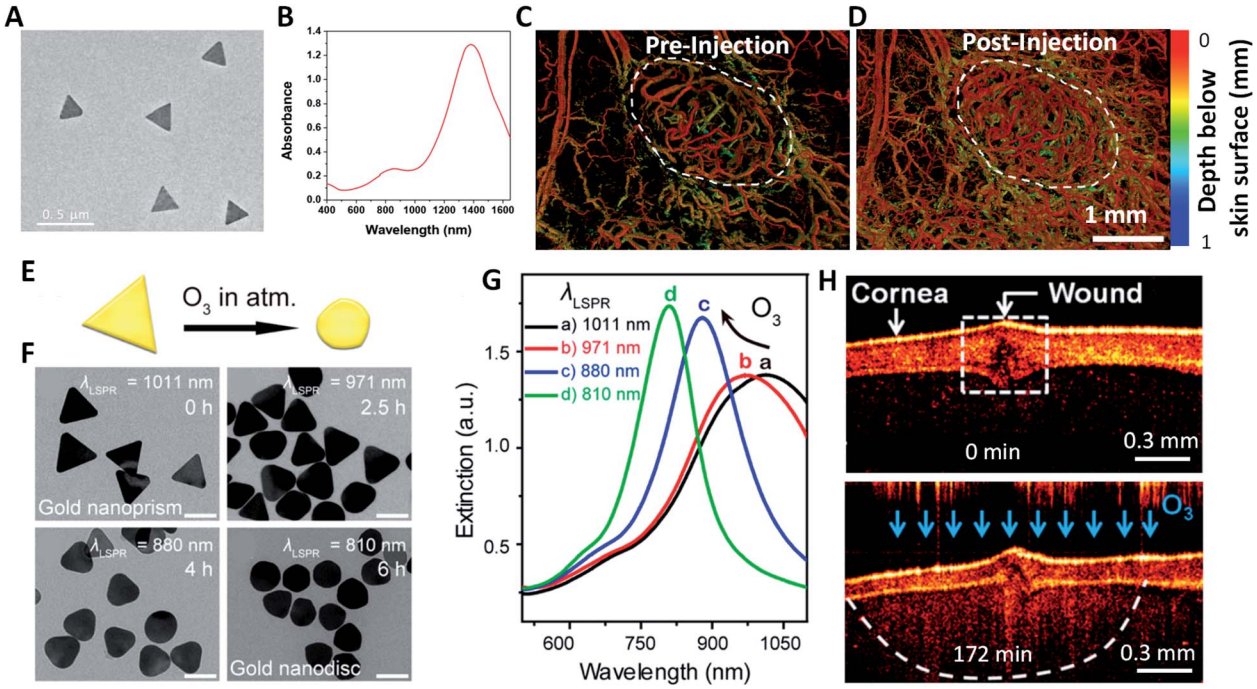


Fig. 5 AuNPr as OCT contrast agents for contrast enhanced vascular imaging and ozone sensing. (A and B) TEM image and extinction spectrum of AuNPr which have strong plasmonic scattering in the NIR-II region. (C and D) The OCT angiograms of a melanoma tumor implanted on a mouse ear before and after systematically injecting the NIR-II AuNPr. Reprinted with permission from ref. 14. Copyright (2018) American Chemical Society. (E–G) Geometric and plasmonic evolution of AuNPr upon exposure to ozone (O_3). Upon exposure to O_3, the corners of AuNPr are gradually rounded and the LSPR of AuNPr exhibits a blue shift accordingly. (H) OCT images of a AuNPr-infiltrated crucian carp eye before and at 172 min post exposure of 75 ppm O_3. Reprinted with permission from ref. 15. Copyright (2017) American Chemical Society.

especially for tumors (Fig. 5A–D).[14] Zhou and co-workers demonstrated that AuNPr can be used for three-dimensional (3D) ozone imaging in the anterior chamber of an isolated crucian carp eye.[15] Upon exposure to O_3, AuNPr transform their shape gradually into circular nanodiscs, resulting in a blue shift of their plasmon-resonance peak from 1050 nm to 830 nm (Fig. 5E–G). Using an OCT system with the center beam wavelength at 830 nm, O_3 can be sensitively detected due to the strong OCT intensity signals from gold nanodiscs (Fig. 5H). In another study, Zhou *et al.* synthesized a nanohybrid AuNPr@polyaniline core–shell structure, which can reversibly change its extinction and scattering properties at 830 nm under different acidic and basic conditions.[77] The AuNPr@polyaniline was employed as an OCT contrast agent to dynamically image the pH distribution in the anterior chamber of a fish's eye.

2.3 Optical imaging based on LSPR-enhanced absorption

Both photothermal (PT) and photoacoustic (PA) imaging take advantage of AuNPs' strong light absorption at their plasmon resonance peak. AuNPs produce heat upon absorbing laser light at their plasmon-resonance wavelength, which results in refractive index change of the surrounding medium and elastic expansion in the surrounding air. PT imaging relies on the detection of phase shift of the transmitted light beam caused by refractive index change, while PA imaging is based on detecting the acoustic waves generated by air expansion. In contrast to AuNP contrast agents, which usually have sizes above 30–40 nm, used for scattering-based imaging techniques, AuNPs suited for PT and PA imaging are much smaller in dimension, and have higher absorption-to-scattering cross-section.

PT imaging was first reported by Boyer and coworkers,[81] who used a green laser to focus on the sample where AuNPs absorb the light and produce heat. Another red laser beam is split into two orthogonally polarized beams, one of which focuses on the same point as the green laser. The red laser beams reflected from the sample are then recombined and the relative phase change is identified using a detector. They show that such a technique can sensitively image gold nanospheres as small as 10 nm in diameter. Berciaud *et al.* further demonstrated that PT imaging can detect even 1.4 nm AuNPs consisting of just 67 atoms, and measure the absorption spectra of individual AuNPs with diameter down to 5 nm.[82] Zharov *et al.* applied AuNPs[83–85] and Au-coated carbon nanotubes[86,87] as PT imaging contrast agents to visualize adherent and circulating cancer cells, circulating stem cells and lymphatic vessels. Further, these researchers demonstrated super-resolution PT microscopy based on non-linear dependence of signals on laser energy, with a 50 nm spatial resolution.[88]

Photothermal OCT (PT-OCT) combines PT-imaging and OCT to allow high-sensitivity and high-contrast *in vivo* imaging. The PT laser source is coupled to the sample arm of the OCT system, exciting the contrast agents that are within the optical path length of the OCT beam. Upon excitation, the photothermal effect of the contrast agents changes the

© 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

temperature of their surrounding environment and causes a variation of the local refractive index, which leads to an alteration of the optical path length which changes the phase of the backscattered OCT beam. The phase change is then detected by phase-sensitive OCT.[89] Many types of gold nanostructures, including AuNS,[90] AuNSh,[91] and AuNR,[92–97] have been investigated as PT-OCT contrast agents due to their strong absorption cross-section and the resulting photothermal effect. Among them, AuNR are most promising because they are highly efficient in absorbing light and releasing heat. Tucker-Schwartz *et al.*[93] showed that they could image AuNR accumulated in mouse tumor tissue using PT-OCT with very high signal-to-noise ratio. These AuNR were subcutaneously injected and then they accumulated in tumors through the leaky vasculature. Jung *et al.* demonstrated that they can image the concentration and time-dependent uptake of AuNR in sentinel lymph nodes using PT-OCT.[97] Recently, Lapierre-Landry *et al.*[94] and Gordon *et al.*[95] have demonstrated AuNR as PT-OCT contrast agents that can be specifically detected in the retina of laser-induced choroidal neovascularization (LCNV) mouse models after systematic injections. The former study shows that the PT-OCT signals in the LCNV lesion are associated with non-targeted AuNR passively accumulated in the eye, whereas the latter study applies PT-OCT to detect antibody-conjugated AuNR targeting the LCNV lesion. The latter study also investigated the effect of anti-angiogenic treatment in the retina of a LCNV mouse model using targeted AuNR as PT-OCT contrast agents. Notable reduction of PT-OCT signals is detected in the LCNV lesion after intravitreal delivery of neutralizing monoclonal anti-vascular endothelial growth factor (anti-VEGF) antibodies.

PA imaging takes advantage of the high resolution of optical imaging and the deeper tissue penetration of ultrasound imaging. It permits tissue penetration up to 5 cm and spatial resolution up to tens of micrometers. AuNPs that are strong absorbers of NIR light are ideal PA contrast agents as NIR light penetrates deeper in animal tissues. AuNP-based PA contrast agents have been intensively investigated in the past decade for a wide range of preclinical applications, including cancer detection, atherosclerotic plaque imaging, brain function measurements, image-guided therapy, *etc.* Because both PT and PA contrast agents are strong absorbers of NIR light, many AuNPs that are used for PT and PT-OCT imaging can be used for PA imaging, including AuNS,[98] AuNSh,[11,12] AuNR,[99–104] AuNC[23,24] and AuNSt.[105] Although AuNR have been utilized as PA contrast agents for a variety of applications including multiplexed cell tracking,[102] cardiovascular imaging,[100] stem cell imaging[103] and cancer cell detection,[104] regular AuNR have suboptimal thermal stability under a high-intensity pulsed laser used for PA irradiation, leading to plasmonic resonance shift and decay of PA signals over time. In a recent study, Chen *et al.*[2] report that miniature AuNR that are 5–11 times smaller than regular-sized AuNR, are 3-fold more thermally stable and produce 3.5-fold stronger PA signal intensity than their absorption-matched larger counterparts in the NIR-II window (Fig. 6A and B). Upon IV injection of these small targeted AuNR in a tumor-bearing mouse model, they improved the tumor delivery efficiency by 30% and PA contrast by 4.5 times (Fig. 6C and D).

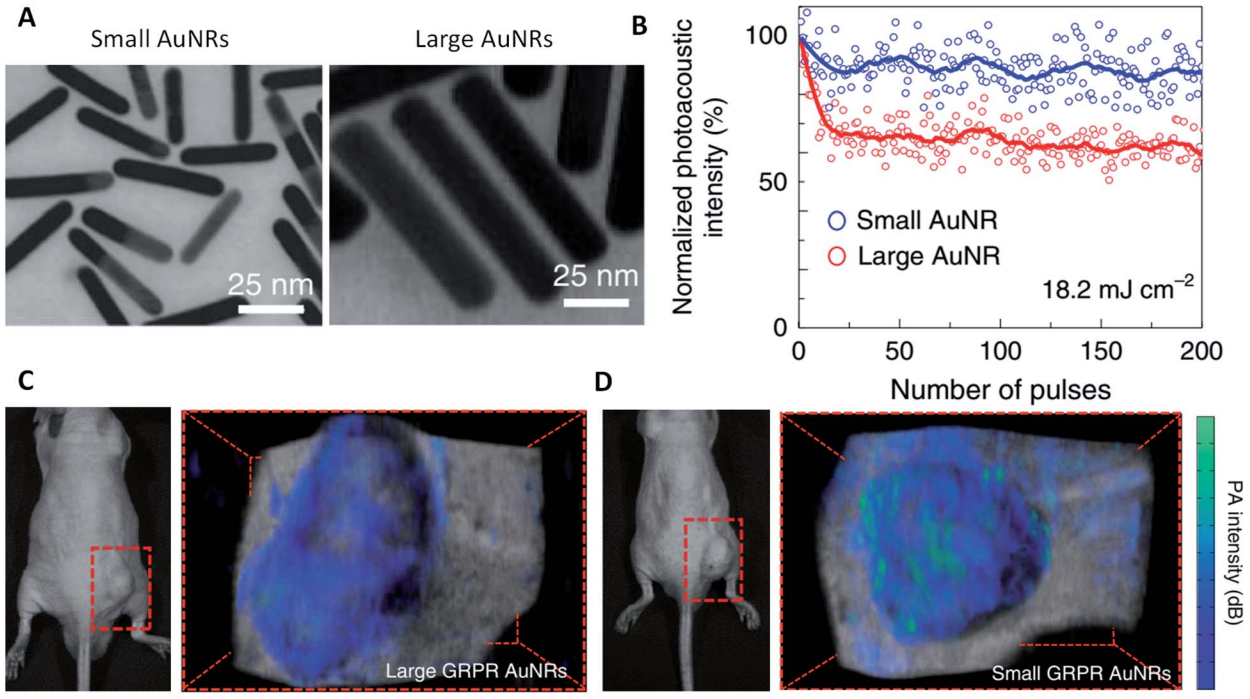


Fig. 6 Small AuNRs enhanced PA imaging. (A) TEM images of small and large AuNRs. (B) PA signal intensities of small and large AuNRs upon illumination with 200 laser pulses at 18.2 mJ cm^{-2}. (C and D) PA images of tumor-bearing mice after administration of prostate cancer targeting GRPR peptide-functionalized large and small AuNRs. Reprinted with permission from ref. 2. Copyright (2019) Springer Nature.

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

3. Optical sensing and imaging based on SERS

Though the Raman scattering effect is a very weak phenomenon, a drastic increase in the scattering intensity has been observed when it takes place in the vicinity of AuNPs, a phenomenon called SERS. SERS offers several advantages such as being non-destructive, having potential for multi-component analysis, yielding compound-specific information, and can be performed in the presence of water, making it suitable to study processes *in vivo*.[106]

3.1 SERS-based optical sensing

Raman scattering is gaining increased research attention in the field of biosensing because of its single-molecule sensitivity, intrinsically sharp fingerprints, and the availability of a wide range of photo-stable labels. To achieve high sensitivity the SERS signal has to be enhanced by either modifying the surface properties or by increasing the number of molecules generating Raman signal. Monolayer films of AuNPs have discrete physical and chemical properties due to quantum confinement and interparticle coupling.[107] These properties are highly useful in SERS applications. AuNP monolayers with nanogaps have shown an increase in the SERS signal but their surface uniformity hinders the performance.[108] Thus, numerous attempts have been made to fabricate AuNP monolayers with desired density and precisely controlled nanogaps.[109–111] For example, Shao *et al.* fabricated a multiple signal amplification sandwich type SERS biosensor for the detection of miRNA with a detection limit of 3 fM. A triple signal amplification was achieved by using AuNPs of selective morphology, hybridization chain reaction amplification and using silver nanoparticles in combination with AuNPs.[110]

Wang and co-workers[112] developed an interesting SERS-based biosensing platform that is able to differentiate between two different DNA targets simultaneously (Fig. 7). A lateral flow assay was able to detect DNA associated with Kaposi sarcoma-associated herpesvirus (KSHV) and bacillary angiomatosis (BA) (Fig. 7C). The Raman peak intensity centered at 1617 cm^{-1} was monitored, and its variation was used for the quantitative evaluation of KSHV and BA (Fig. 7D).

Currently, there are no SERS-based diagnostic devices in the market, but efforts are being made to reduce the complexity of software and instrumentation to commercialize the products. SERS is very promising for pathogen detection owing to its high sensitivity and selectivity.[113] Some of the commonly used platforms for SERS are nanomaterials such as gold nanostructured

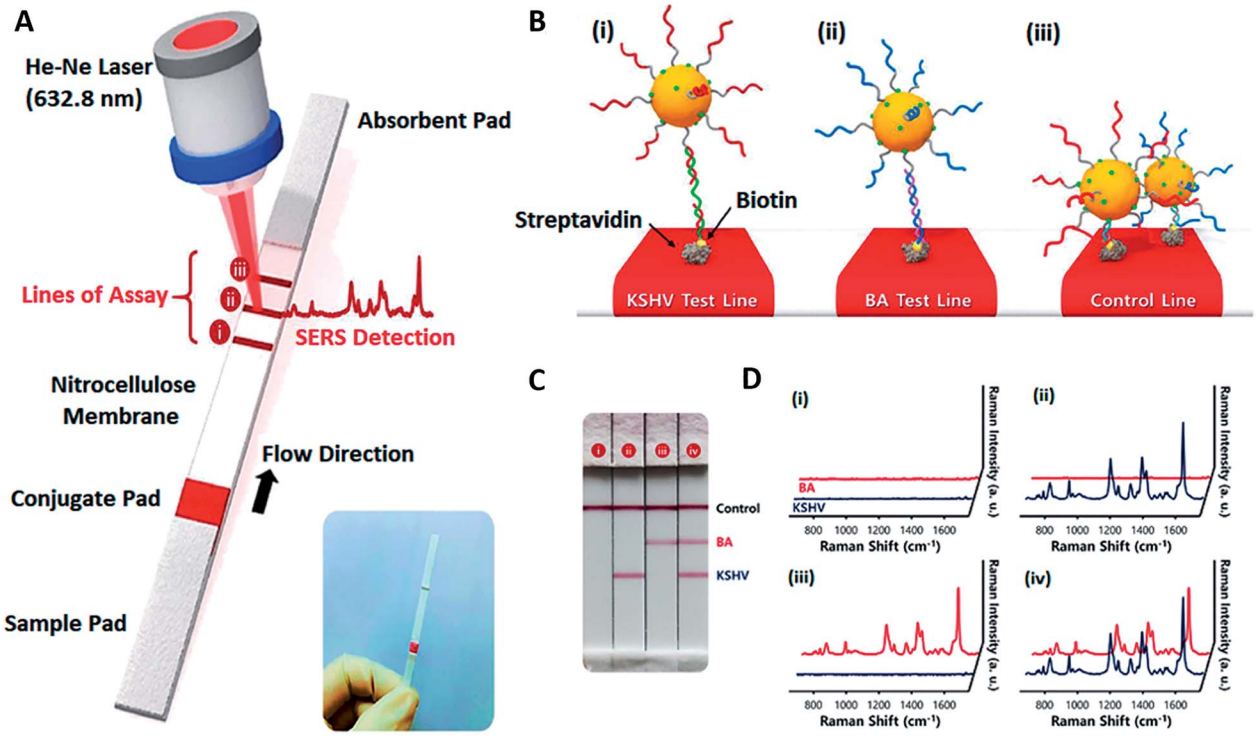


Fig. 7 SERS-based LFA for KSHV detection. (A) Scheme of the LFA biosensor for the simultaneous detection of two nucleic acids. The strip consists of two test lines and one control line. (B) (i) KSHV DNA–Au NP complexes were captured by the probe KSHV DNA on the first test line; (ii) BA DNA–Au NP complexes were captured by the probe BA DNA on the second test line, and (iii) excess KSHV and BA detection DNA attached to AuNPs were captured by control DNA through T20–A20 hybridization on the third control line. (C) Digital photographic images and (D) corresponding SERS spectra of the SERS-based LFA biosensor in the presence of (i) KSHV, 0 pM; BA, 0 pM; (ii) KSHV, 100 pM; BA, 0 pM; (iii) KSHV, 0 pM; BA, 100 pM; (iv) KSHV, 100 pM; BA, 100 pM. Assay time: 20 min. Reprinted with permission from ref. 112. Copyright (2017) American Chemical Society.

© 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

surfaces and gold colloidal particles. The Raman scattering from molecules in the vicinity of a gold nanomaterial surface due to the coupling effect of surface plasmon through the oscillating electric field can be described as electromagnetic enhancement. An enhancement factor of 10^6 to 10^{14} is enough to allow the detection of a single target molecule, which makes them highly suitable approaches for the detection of pathogens.[114,115] Mevold and coworkers[116] have conducted a SERS-based study for sensitive bacteria detection by using a graphene–gold nanoparticle nanohybrid. Cationic poly(diallyldimethylammonium chloride) (PDDA) functionalized graphene nanosheets were prepared. The traditional citrate thermal reduction method was applied to synthesize gold nanoparticles which were electrostatically immobilized onto graphene–PDDA nanohybrid sheets. The result of transmission electron microscopy (TEM) showed that the particle size was 15–20 nm. The zeta potential value of the nanohybrid ratio was 7.7–38.4 mV. The methodology was applied for the detection of *S. aureus* confirming its application in label-free and rapid biosensing of microorganisms and target biomolecules. In another study, Ma and colleagues[117] proposed a rapid and simple SERS-based biosensor for sensitive and selective *Salmonella typhimurium* (*S. typhimurium*) detection using spiny AuNPs functionalized with 4-mercaptobenzoic acid and specific thiolated *S. typhimurium* aptamers as SERS nanoprobes. Under optimal conditions, a low limit of detection of 4 CFU per mL was obtained for the detection of *S. typhimurium*.

3.2 SERS-based optical imaging

Compared with other optical imaging modalities, SERS has the advantage of high sensitivity, which allows contrast agents to be detected at a concentration as low as 1.5 fM.[118] In recent decades, SERS imaging agents have been increasingly studied by conjugating Raman active molecules on AuNPs with various structures. These AuNPs can dramatically enhance the Raman signals of dyes conjugated on them by an order of 10^{10}. AuNPs with rough geometrical features such as sharp tips and edges, commonly known as 'Raman hot spots', can even improve the Raman signals to the order of up to 10^{15} fold,[119] because the spatial confinement of the surface plasmons at these 'hot spots' produces a highly concentrated local secondary electromagnetic field. AuNPs with quite a few different morphologies can provide such 'Raman hot spots', including AuNSt,[31,32,118] AuNPy,[16] AuNCr,[36,120] nanourchins,[121–123] AuNBP,[124,125] *etc.* For example, Harmsen *et al.* demonstrated that an AuNSt-based SERS agent can be used to precisely detect macroscopic malignant lesions and microscopic tumor invasion with high sensitivity in animal models of pancreatic cancer, breast cancer, prostate cancer, and sarcoma, and in one human sarcoma xenograft model.[118] In recent studies, Bardhan and coworkers

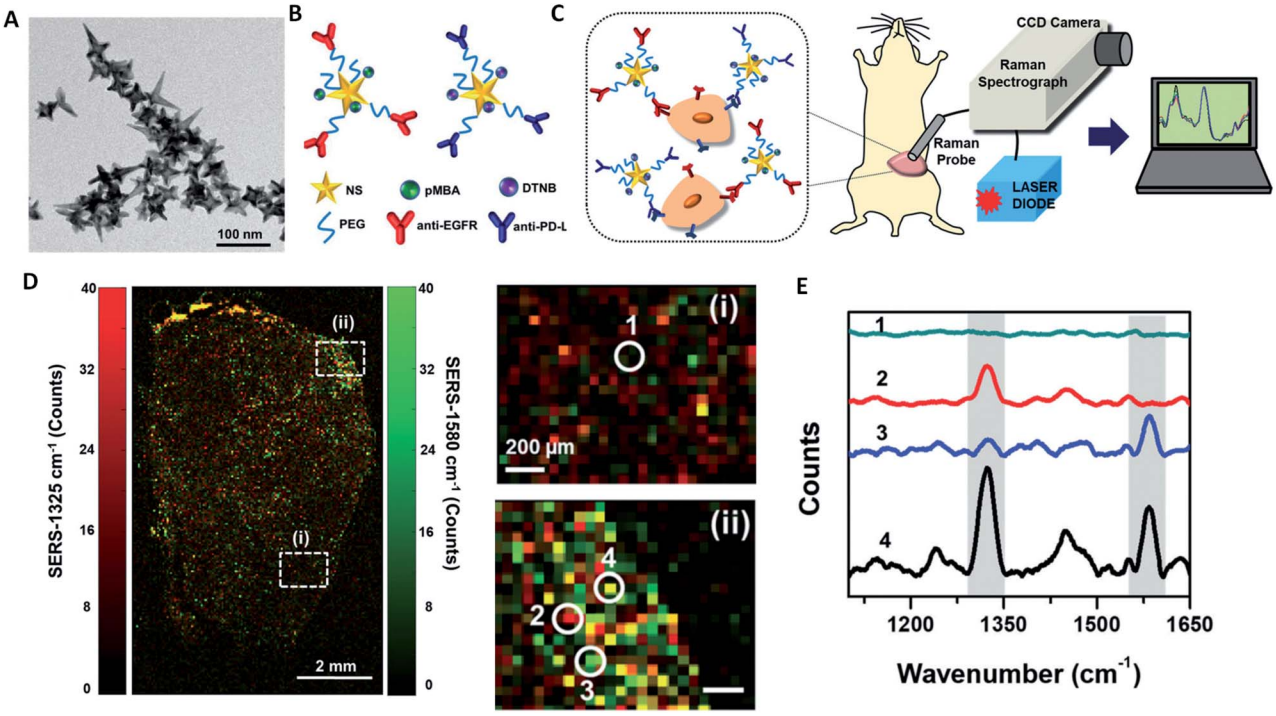


Fig. 8 Multiplexed SERS immunoimaging using AuNSt. (A) TEM images of AuNSt used for SERS. (B) Scheme of AuNSt functionalized with two different Raman-active tag/targeting antibody pairs. The Raman tag pMAB is paired with anti-EGFR, and DTNB is paired with anti-PD-L1. (C) Schematic illustration of the SERS setup for multiplexed tumor immunoimaging. (D) The SERS signals of both DTNB (1325 cm^{-1}) and pMBA (1580 cm^{-1}) in *ex vivo* mouse breast cancer tissue. Subsection (i) shows a region of interest (ROI) with little AuNSt accumulation, whereas subsection (ii) shows a ROI with rich AuNSt accumulation. (E) The corresponding Raman spectra of ROI (1)–(4) in (D), where (1) shows no AuNSt binding, (2) shows high PD-L1 expression, (3) shows high EGRF expression and (4) shows high expressions of both PD-L1 and EGFR. Reprinted with permission from ref. 32. Copyright (2018) the Royal Society of Chemistry.

© 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

reported multiplexed imaging of immuno-biomarkers using AuNSt-based SERS imaging (Fig. 8).[31,32]

4. Optical sensing and imaging based on AuNP-induced luminescence

4.1 Optical biosensors based on AuNP luminescence

Luminescence-based optical biosensors, in which the detection is carried out *via* either fluorescence or chemiluminescence, offer a variety of advantages such as high selectivity, excellent sensitivity and flexibility.[126] Fluorescence-based biosensors can be of colorimetric type giving visual results using AuNPs' individual fluorescence or fluorescence resonance energy transfer (FRET). Elahi and coworkers proposed a fluorescence biosensor for *Shigella* sp. detection by employing AuNPs.[127] As a signal reporter two DNA probes were immobilized on gold nanoparticles. A third DNA probe was immobilized on SMCC (sulfosuccinimidyl 4-*N*-maleimidomethyl cyclohexane-1-carboxylate) modified iron nanoparticles (MNPs). The DNA probe–AuNP–fluorescence DNA probe and the MNP–DNA probe were added to target DNA. By using a magnet, the MNP–DNA probe–target DNA–DNA probe–AuNP–fluorescence complex was isolated. Fluorescence intensity was studied by fluorescence spectrophotometry after the signal generation when a fluorescence DNA probe was released on the surface of AuNPs. By increasing the target DNA concentration, the fluorescence intensity was increased accordingly. A low detection limit of 102 CFU per mL for *Shigella* sp. was successfully achieved.

Plasmonic AuNPs can interact with different fluorophores resulting in either fluorophore quenching or enhancement also known as FRET. Quantum dots (QDs) and AuNPs are the most commonly used FRET pairs in fluorescence-based optical biosensors due to the overlapping of AuNPs' absorption spectrum with QDs' emission spectrum.[128–130] But the applicability of QD-based FRET biosensors is limited by their high cost, toxicity, the strong dependence of performance on surface states and chemical instability.[131] Metal nanoparticles or metal-doped nanoparticles are emerging as new alternatives of QDs in FRET-based biosensors.[131,132] Zhang *et al.* used amino capped silicon nanoparticles (SiNPs) and citrate capped AuNPs as FRET pairs to design an L-Csyteine (L-Cys) biosensor.[131] In the fabricated biosensor, fluorescence quenching of SiNPs with AuNPs and the restoration of fluorescence by the addition of L-Cys due to the release of AuNPs from the SiNP–AuNP complex have been demonstrated. The intensity of fluorescence restored is directly proportional to the concentration of L-Cys added. FRET-based biosensors possess many advantages such as real-time monitoring, visual monitoring, being capable of multiple analyte detection *etc.* Despite various merits, there are some demerits such as the cost of FRET reagents, selection of fluorophores, stability of functionalized nanoscale materials, low fluorescence resolution and low signal-to-noise ratio.

Electrochemiluminescence (ECL) or electron generated chemiluminescence based biosensors are emerging and less explored categories of biosensors which are gaining increasing attention. These are based on the phenomenon in which the luminescent species absorb energy from the electrochemical reaction occurring near their surface. AuNPs have been extensively used in ECL-based biosensors for signal amplification.[133–135] An ultrasensitive ECL biosensor for exosome detection was developed using recently discovered two-dimensional nanomaterials MXenes and AuNPs.[135] In this work, a glassy carbon electrode was modified with exosome specific aptamers and then attached to the target exosomes derived from HeLa cells. Then aptamer conjugated MXenes were attached to the modified electrode, followed by immersion in $HAuCl_4$ for *in situ* synthesis of AuNPs. The MXenes acted both as stabilizers and reductants, and the *in situ* formed AuNP–MXene hybrid not only demonstrated highly efficient recognition of exosomes but also provided a large catalytic surface with AuNPs. A 1000 times lower detection limit (*i.e.* 30 particles per μL) than that of the conventional ELISA was achieved. These ECL biosensors are highly advantageous in terms of specificity as the signal is due to the biorecognition reaction.

4.2 Optical imaging based on AuNP luminescence

AuNPs can significantly enhance the excitation of fluorescent dyes when the SPR of AuNPs overlaps with the absorption and emission spectra of fluorophores.[136] This phenomenon is called 'plasmon-enhanced fluorescence (PEF)', which was discovered shortly after the emergence of SERS.[137] The key to PEF is the distance between the AuNPs and fluorophores. AuNPs can enhance the fluorescence intensity of dyes only if the separation distance is appropriate (~5 nm). The fluorescence is unchanged if the separation distance is too large (>20 nm), whereas the fluorescence is quenched if the separation distance is very short (<2 nm).[138] PEF allows us to study weak fluorescence emission, even at the single molecule level. For example, Moerner *et al.* reported that the single-molecule fluorescence could be enhanced by gold bowtie nanoantennas by 1340-fold.[139] Khatua and coworkers used AuNR to achieve single-molecule fluorescence enhancement of a weak emitter, crystal violet, by a factor of 1000.[140] These findings are especially helpful to advance super-resolution fluorescence imaging. Jeynes *et al.* demonstrated AuNP-based DNA point accumulation for imaging in nanoscale topography (DNA-PAINT) to measure the nanoscale dimensions of individual telomeres in human cells.[141] The technique utilized an imaging probe composed of AuNPs and fluorescence capabilities, in combination with direct stochastic optical reconstruction microscopy (dSTORM) imaging and X-ray fluorescence microscopy, achieving an unprecedented spatial resolution of 5 nm.

Gold nanoclusters (AuNCls) were recently discovered to exhibit strong intrinsic fluorescence (in the visible and NIR region) that is highly photostable. AuNCls are AuNPs with diameters that are smaller than 2–3 nm, composed of several to hundreds of gold atoms. Because of their small sizes, AuNCls do not support the SPR effect. The intrinsic fluorescence of AuNCls can be directly used for both *in vitro* and *in vivo* imaging studies. Venkatesh and coworkers visualized the accumulation of 8-mercapto-9-propyladenine capped green AuNCls (emission at 510 nm) in cell nuclei.[142] Similarly, Wang *et al.* demonstrated

 © 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

nuclear staining with tripeptide (Lys–Cys-–Lys) stabilized red AuNCls (emission at 680 nm).[143] Further, the intracellular accumulation and immune responses of AuNCls with different ligands were studied by Fernández and coworkers.[144] Chandirasekar *et al.*[145] and Zhang *et al.*[146] have employed AuNCls for *in vivo* fluorescent imaging of zebrafish embryos and implanted tumor in a mouse model, respectively. Their studies show that AuNCls have no toxicity on developing embryos but have remarkable penetration and retention in tumors.

In addition to PEF and intrinsic fluorescence, AuNPs can be directly used for two-photon (TP) and multi-photon (MP) imaging, which allow for greater tissue penetration and are suitable for *in vivo* imaging. The excitation light sources of TP and MP imaging are usually high intensity lasers, and therefore AuNPs are advantageous than organic fluorophores due to their high photostability. In addition, the two-photon absorption coefficient of AuNPs can be as high as 100-fold greater than that of organic fluorophores.[147] *In vivo* imaging study found that the two-photon fluorescent signal produced by a single AuNR in mouse ear blood vessels is 58 times stronger than that from a single rhodamine molecule.[148] Another study shows that AuNR-labeled cancer cells exhibit three orders of magnitude higher TP fluorescence intensity than the autofluorescence intensity from unlabeled cancer cells.[147] Besides AuNR, AuNSt also show strong TP fluorescent signals both *in vitro* and *in vivo*, due to their SPR in the NIR region.[149] Gao compared the TP signals of five different types of AuNPs with similar sizes: nanospheres, nanocubes, nanotriangles, AuNR, and AuNSt and found that the TP signal intensity increases in the order of nanospheres, nanocubes, nanotriangles, AuNR and AuNSt. The TP fluorescence intensity of AuNSt is 50 000-fold greater than that of gold nanospheres.[150]

5. Conclusions and perspectives

The unique and highly tunable optical properties of AuNPs have shown enormous potential in the development of state-of-the-art biosensing and bioimaging techniques. AuNP-based optical biosensors have great potential to be used as point-of-care devices due to their non-invasiveness, high sensitivity and highly reliable analytical results. Various Au nanostructure-based contrast agents have emerged for many optical imaging techniques including DFIC, OCT, PA, PT, Raman spectroscopy and fluorescence microscopy (Table 1). These novel imaging techniques allow precise imaging of cellular and molecular biomarkers with high resolution. Although proof-of-concept research papers of AuNP-based optical biosensors and bioimaging are increasing every year, these techniques still have a long way to go towards commercial and clinical applications. For biosensing, efforts should be made for high throughput and multiplexed identification of biomarkers. Multi-technology combination is another approach that can be adopted to improve the performance of sensing devices. The integration of AuNP-based optical sensing platforms with smartphones, portable, wearable and implantable devices is a future trend of exploration. More research is needed in manufacturing AuNPs with large scale and long-term stability under various environmental conditions. In addition to this, full attention is required in improving the analytical performance of biosensors by preventing non-specific adsorption of biomolecules onto AuNPs and shortening the analysis time. For biomedical imaging, more thorough research needs to be conducted to study the biodistribution, clearance pathways and long-term toxicity of AuNPs with different geometries and dimensions. In addition, better targeting agents and targeting strategies are needed to improve the molecular imaging specificity. The development of AuNP composites/hybrids with other nanomaterials is an exciting area that needs to be further explored as it may open new avenues of research and applications. For instance, composites of AuNPs with graphene, metal oxide nanoparticles and other two-dimensional nanomaterials have been explored to enhance optical biosensing performances.[151–154] A nanoparticle with a gold core, Raman active layer, silica shell and Gd coating was developed to achieve triple-modality imaging, which combines MRI, PA and Raman imaging.[155] In the near future, we envision that there will be tremendous growth in developing multifunctional AuNPs and AuNP-based nanohybrids for multi-analyte sensing and multi-modality imaging.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

NR, MW, and ON gratefully acknowledge the financial support and funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. H2020-MSCA-ITN-2018-813680. NR wants to thank Prof. M. Hasanzadeh for valuable discussions. SS, CMP, RKG and BDM thank Prof. Yogesh Singh, Vice-Chancellor, Delhi Technological University, for his interest in this work. SS is thankful to CSIR, India, for the award of RA. CMP acknowledges the Department of Science and Technology, New Delhi, India, for the DST-INSPIRE Faculty Award (DST/INSPIRE/04/2015/000932). B. D. M. thanks the Science & Engineering Research Board (Government of India) for the award of a Distinguished Fellowship (SB/DF/011/2019). A. d. l. Z. would like to thank the Damon Runyon Cancer Research Foundation (DFS# 06-13), Claire Giannini Fund, the Susan G. Komen Breast Cancer Foundation (SAB15-00003), the Mary Kay Foundation (017-14), the Skippy Frank Foundation, the Donald E. and Delia B. Baxter Foundation, and a seed grant from the Center for Cancer Nanotechnology Excellence and Translation (CCNE-T; NIH-NCI U54CA151459). A. d. l. Z. is a Chan Zuckerberg Biohub investigator and a Pew-Stewart Scholar for Cancer Research supported by The Pew Charitable Trusts and The Alexander and Margaret Stewart Trust.

References

1 T. Shimizu, T. Teranishi, S. Hasegawa and M. Miyake, Size Evolution of Alkanethiol-Protected Gold Nanoparticles by

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Heat Treatment in the Solid State, *J. Phys. Chem. B*, 2003, **107**(12), 2719–2724.
2 Y.-S. Chen, Y. Zhao, S. J. Yoon, S. S. Gambhir and S. Emelianov, Miniature Gold Nanorods for Photoacoustic Molecular Imaging in the Second Near-Infrared Optical Window, *Nat. Nanotechnol.*, 2019, **14**(5), 465–472.
3 O. Liba, E. D. SoRelle, D. Sen and A. de la Zerda, Contrast-Enhanced Optical Coherence Tomography with Picomolar Sensitivity for Functional *In Vivo* Imaging, *Sci. Rep.*, 2016, **6**(1), 23337.
4 E. D. SoRelle, O. Liba, Z. Hussain, M. Gambhir and A. De La Zerda, Biofunctionalization of Large Gold Nanorods Realizes Ultrahigh-Sensitivity Optical Imaging Agents, *Langmuir*, 2015, **31**(45), 12339–12347.
5 A. Agrawal, S. Huang, A. Wei Haw Lin, M.-H. Lee, J. K. Barton, R. A. Drezek and T. J. Pfefer, Quantitative Evaluation of Optical Coherence Tomography Signal Enhancement with Gold Nanoshells, *J. Biomed. Opt.*, 2006, **11**(4), 041121.
6 A. M. Gobin, M. H. Lee, N. J. Halas, W. D. James, R. A. Drezek and J. L. West, Near-Infrared Resonant Nanoshells for Combined Optical Imaging and Photothermal Cancer Therapy, *Nano Lett.*, 2007, **7**(7), 1929–1934.
7 J. C. Y. Kah, M. Olivo, T. H. Chow, K. S. Song, K. Z. Y. Koh, S. Mhaisalkar and C. J. R. Sheppard, Control of Optical Contrast Using Gold Nanoshells for Optical Coherence Tomography Imaging of Mouse Xenograft Tumor Model *In Vivo*, *J. Biomed. Opt.*, 2009, **14**(5), 054015.
8 E. V. Zagaynova, M. V. Shirmanova, M. Y. Kirillin, B. N. Khlebtsov, A. G. Orlova, I. V. Balalaeva, M. A. Sirotkina, M. L. Bugrova, P. D. Agrba and V. A. Kamensky, Contrasting Properties of Gold Nanoparticles for Optical Coherence Tomography: Phantom, *In Vivo* Studies and Monte Carlo Simulation, *Phys. Med. Biol.*, 2008, **53**(18), 4995–5009.
9 M. Kirillin, M. Shirmanova, M. Sirotkina, M. Bugrova, B. Khlebtsov and E. Zagaynova, Contrasting Properties of Gold Nanoshells and Titanium Dioxide Nanoparticles for Optical Coherence Tomography Imaging of Skin: Monte Carlo Simulations and *In Vivo* Study, *J. Biomed. Opt.*, 2009, **14**(2), 021017.
10 J. Hu, F. Sanz-Rodríguez, F. Rivero, E. M. Rodríguez, R. A. Torres, D. H. Ortgies, J. G. Solé, F. Alfonso and D. Jaque, Gold Nanoshells: Contrast Agents for Cell Imaging by Cardiovascular Optical Coherence Tomography, *Nano Res.*, 2018, **11**(2), 676–685.
11 W. Lu, Q. Huang, G. Ku, X. Wen, M. Zhou, D. Guzatov, P. Brecht, R. Su, A. Oraevsky, L. V. Wang, *et al.*, Photoacoustic Imaging of Living Mouse Brain Vasculature Using Hollow Gold Nanospheres, *Biomaterials*, 2010, **31**(9), 2617–2626.
12 M.-L. Li, J. C. Wang, J. A. Schwartz, K. L. Gill-Sharp, G. Stoica and L. V. Wang, *In Vivo* Photoacoustic Microscopy of Nanoshell Extravasation from Solid Tumor Vasculature, *J. Biomed. Opt.*, 2009, **14**(1), 010507.
13 Y. Huang, A. R. Ferhan, Y. Gao, A. Dandapat and D.-H. Kim, High-Yield Synthesis of Triangular Gold Nanoplates with Improved Shape Uniformity, Tunable Edge Length and Thickness, *Nanoscale*, 2014, **6**(12), 6496–6500.
14 P. Si, E. Yuan, O. Liba, Y. Winetraub, S. Yousefi, E. D. SoRelle, D. W. Yecies, R. Dutta and A. de la Zerda, Gold Nanoprisms as Optical Coherence Tomography Contrast Agents in the Second Near-Infrared Window for Enhanced Angiography in Live Animals, *ACS Nano*, 2018, **12**(12), 11986–11994.
15 X. Jiang, P. Tang, P. Gao, Y. S. Zhang, C. Yi and J. Zhou, Gold Nanoprobe-Enabled Three-Dimensional Ozone Imaging by Optical Coherence Tomography, *Anal. Chem.*, 2017, **89**(4), 2561–2568.
16 C. M. Sweeney, C. L. Stender, C. L. Nehl, W. Hasan, K. L. Shuford and T. W. Odom, Optical Properties of Tipless Gold Nanopyramids, *Small*, 2011, **7**(14), 2032–2036.
17 D. Chateau, A. Liotta, F. Vadcard, J. R. G. Navarro, F. Chaput, J. Lermé, F. Lerouge and S. Parola, From Gold Nanobipyramids to Nanojavelins for a Precise Tuning of the Plasmon Resonance to the Infrared Wavelengths: Experimental and Theoretical Aspects, *Nanoscale*, 2015, **7**(5), 1934–1943.
18 P. Si, S. Shevidi, E. Yuan, K. Yuan, Z. Lautman, S. S. Jeffrey, G. W. Sledge and A. de la Zerda, Gold Nanobipyramids as Second Near Infrared Optical Coherence Tomography Contrast Agents for *In Vivo* Multiplexing Studies, *Nano Lett.*, 2020, **20**(1), 101–108.
19 E. Yuan, P. Si, Y. Winetraub, S. Shevidi and A. de la Zerda, A Spectral Demixing Model for Triplex *In Vivo* Imaging of Optical Coherence Tomography Contrast Agents, *ACS Photonics*, 2020, **7**(4), 893–900.
20 H. Cang, T. Sun, Z.-Y. Li, J. Chen, B. J. Wiley, Y. Xia and X. Li, Gold Nanocages as Contrast Agents for Spectroscopic Optical Coherence Tomography, *Opt. Lett.*, 2005, **30**(22), 3048.
21 H. B. Song, J. S. Wi, D. H. Jo, J. H. Kim, S. W. Lee, T. G. Lee and J. H. Kim, Intraocular Application of Gold Nanodisks Optically Tuned for Optical Coherence Tomography: Inhibitory Effect on Retinal Neovascularization without Unbearable Toxicity, *Nanomedicine*, 2017, **13**(6), 1901–1911.
22 J. S. Wi, J. Park, H. Kang, D. Jung, S. W. Lee and T. G. Lee, Stacked Gold Nanodisks for Bimodal Photoacoustic and Optical Coherence Imaging, *ACS Nano*, 2017, **11**(6), 6225–6232.
23 X. Yang, S. E. Skrabalak, Z. Y. Li, Y. Xia and L. V. Wang, Photoacoustic Tomography of a Rat Cerebral Cortex *In Vivo* with Au Nanocages as an Optical Contrast Agent, *Nano Lett.*, 2007, **7**(12), 3798–3802.
24 K. H. Song, C. Kim, C. M. Cobley, Y. Xia and L. V. Wang, Near-Infrared Gold Nanocages as a New Class of Tracers for Photoacoustic Sentinel Lymph Node Mapping on a Rat Model, *Nano Lett.*, 2009, **9**(1), 183–188.
25 Y. Xia, W. Li, C. M. Cobley, J. Chen, X. Xia, Q. Zhang, M. Yang, E. C. Cho and P. K. Brown, Gold Nanocages: From Synthesis to Theranostic Applications, *Acc. Chem. Res.*, 2011, **44**(10), 914–924.

 © 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

26 C.-K. Lee, H.-Y. Tseng, C.-Y. Lee, S.-Y. Wu, T.-T. Chi, K.-M. Yang, H.-Y. E. Chou, M.-T. Tsai, J.-Y. Wang, Y.-W. Kiang, *et al.*, Characterizing the Localized Surface Plasmon Resonance Behaviors of Au Nanorings and Tracking Their Diffusion in Bio-Tissue with Optical Coherence Tomography, *Biomed. Opt. Express*, 2010, **1**(4), 1060–1073.

27 E. M. Larsson, J. Alegret, M. Käll and D. S. Sutherland, Sensing Characteristics of NIR Localized Surface Plasmon Resonances in Gold Nanorings, *Nano Lett.*, 2007, **7**(5), 1256–1263.

28 H. Y. Tseng, C. K. Lee, S. Y. Wu, T. T. Chi, K. M. Yang, J. Y. Wang, Y. W. Kiang, C. C. Yang, M. T. Tsai, Y. C. Wu, *et al.*, Au Nanorings for Enhancing Absorption and Backscattering Monitored with Optical Coherence Tomography, *Nanotechnology*, 2010, **21**(29).

29 T. H. Chow, Y. Lai, X. Cui, W. Lu, X. Zhuo and J. Wang, Colloidal Gold Nanorings and Their Plasmon Coupling with Gold Nanospheres, *Small*, 2019, **15**(35), 1902608.

30 H. B. Song, J. S. Wi, D. H. Jo, J. H. Kim, S. W. Lee, T. G. Lee and J. H. Kim, Intraocular Application of Gold Nanodisks Optically Tuned for Optical Coherence Tomography: Inhibitory Effect on Retinal Neovascularization without Unbearable Toxicity, *Nanomedicine*, 2017, **13**(6), 1901–1911.

31 Y. C. Ou, X. Wen, C. A. Johnson, D. Shae, O. D. Ayala, J. A. Webb, E. C. Lin, R. C. Delapp, K. L. Boyd, A. Richmond, *et al.*, Multimodal Multiplexed Immunoimaging with Nanostars to Detect Multiple Immunomarkers and Monitor Response to Immunotherapies, *ACS Nano*, 2020, **14**(1), 651–663.

32 Y.-C. Ou, J. A. Webb, C. M. O'Brien, I. J. Pence, E. C. Lin, E. P. Paul, D. Cole, S.-H. Ou, M. Lapierre-Landry, R. C. DeLapp, *et al.*, Diagnosis of Immunomarkers *In Vivo* via Multiplexed Surface Enhanced Raman Spectroscopy with Gold Nanostars, *Nanoscale*, 2018, **10**(27), 13092–13105.

33 H. Wang, D. W. Brandl, F. Le, P. Nordlander and N. J. Halas, Nanorice: A Hybrid Plasmonic Nanostructure, *Nano Lett.*, 2006, **6**(4), 827–832.

34 J. Ye, P. Van Dorpe, W. Van Roy, G. Borghs and G. Maes, Fabrication, Characterization , and Optical Properties of Gold Nanobowl Submonolayer Structures, *Langmuir*, 2009, **13**, 1822–1827.

35 R. Bukasov and J. S. Shumaker-Parry, Highly Tunable Infrared Extinction Properties of Gold Nanocrescents, *Nano Lett.*, 2007, **7**(5), 1113–1118.

36 Y. Lu, G. L. Liu, J. Kim, Y. X. Mejia and L. P. Lee, Nanophotonic Crescent Moon Structures with Sharp Edge for Ultrasensitive Biomolecular Detection by Local Electromagnetic Field Enhancement Effect, *Nano Lett.*, 2005, **5**(1), 119–124.

37 P. Singh, SPR Biosensors: Historical Perspectives and Current Challenges, *Sens. Actuators, B*, 2016, **229**, 110–130.

38 C. Sönnichsen, B. M. Reinhard, J. Liphardt and A. P. Alivisatos, A Molecular Ruler Based on Plasmon Coupling of Single Gold and Silver Nanoparticles, *Nat. Biotechnol.*, 2005, **23**(6), 741–745.

39 P. K. Jain, S. Eustis and M. A. El-Sayed, Plasmon Coupling in Nanorod Assemblies: Optical Absorption, Discrete Dipole Approximation Simulation, and Exciton-Coupling Model, *J. Phys. Chem. B*, 2006, **110**(37), 18243–18253.

40 L. Qin, G. Zeng, C. Lai, D. Huang, P. Xu, C. Zhang, M. Cheng, X. Liu, S. Liu and B. Li, "Gold Rush" in Modern Science: Fabrication Strategies and Typical Advanced Applications of Gold Nanoparticles in Sensing, *Coord. Chem. Rev.*, 2018, **359**, 1–31.

41 M. Shellaiah, T. Simon, P. Venkatesan, K. W. Sun, F.-H. Ko and S.-P. Wu, Nanodiamonds Conjugated to Gold Nanoparticles for Colorimetric Detection of Clenbuterol and Chromium(III) in Urine, *Microchim. Acta*, 2018, **185**(1), 74.

42 L. Zheng, G. Cai, S. Wang, M. Liao, Y. Li and J. Lin, A Microfluidic Colorimetric Biosensor for Rapid Detection of Escherichia Coli O157: H7 Using Gold Nanoparticle Aggregation and Smart Phone Imaging, *Biosens. Bioelectron.*, 2019, **124**, 143–149.

43 R. Rajamanikandan, A. D. Lakshmi and M. Ilanchelian, Smart Phone Assisted, Rapid, Simplistic, Straightforward and Sensitive Biosensing of Cysteine over Other Essential Amino Acids by β-Cyclodextrin Functionalized Gold Nanoparticles as a Colorimetric Probe, *New J. Chem.*, 2020, **44**(28), 12169–12177.

44 H. Aldewachi, T. Chalati, M. N. Woodroofe, N. Bricklebank, B. Sharrack and P. Gardiner, Gold Nanoparticle-Based Colorimetric Biosensors, *Nanoscale*, 2018, **10**(1), 18–33.

45 T. Yang, Z. Luo, Y. Tian, C. Qian and Y. Duan, Design Strategies of AuNPs-Based Nucleic Acid Colorimetric Biosensors, *TrAC, Trends Anal. Chem.*, 2019, 115795.

46 R. Elghanian, J. J. Storhoff, R. C. Mucic, R. L. Letsinger and C. A. Mirkin, Selective Colorimetric Detection of Polynucleotides Based on the Distance-Dependent Optical Properties of Gold Nanoparticles, *Science*, 1997, **277**(5329), 1078–1081.

47 J. J. Storhoff, R. Elghanian, R. C. Mucic, C. A. Mirkin and R. L. Letsinger, One-Pot Colorimetric Differentiation of Polynucleotides with Single Base Imperfections Using Gold Nanoparticle Probes, *J. Am. Chem. Soc.*, 1998, **120**(9), 1959–1964.

48 T. A. Taton, G. Lu and C. A. Mirkin, Two-Color Labeling of Oligonucleotide Arrays via Size-Selective Scattering of Nanoparticle Probes, *J. Am. Chem. Soc.*, 2001, **123**(21), 5164–5165.

49 M.-P. N. Bui, S. Ahmed and A. Abbas, Single-Digit Pathogen and Attomolar Detection with the Naked Eye Using Liposome-Amplified Plasmonic Immunoassay, *Nano Lett.*, 2015, **15**(9), 6239–6246.

50 B. Li, X. Li, Y. Dong, B. Wang, D. Li, Y. Shi and Y. Wu, Colorimetric Sensor Array Based on Gold Nanoparticles with Diverse Surface Charges for Microorganisms Identification, *Anal. Chem.*, 2017, **89**(20), 10639–10643.

51 Q. You, X. Zhang, F.-G. Wu and Y. Chen, Colorimetric and Test Stripe-Based Assay of Bacteria by Using Vancomycin-Modified Gold Nanoparticles, *Sens. Actuators, B*, 2019, **281**, 408–414.

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

52 M. Matsishin, A. Rachkov, A. Lopatynskyi, V. Chegel, A. Soldatkin and A. El'Skaya, Selective Amplification of SPR Biosensor Signal for Recognition of Rpob Gene Fragments by Use of Gold Nanoparticles Modified by Thiolated DNA, *Nanoscale Res. Lett.*, 2017, **12**(1), 1–6.

53 G. J. Nusz, S. M. Marinakos, A. C. Curry, A. Dahlin, F. Hook, A. Wax and A. Chilkoti, Label-Free Plasmonic Detection of Biomolecular Binding by a Single Gold Nanorod, *Anal. Chem.*, 2008, **80**(4), 984–989.

54 S. M. Marinakos, S. Chen and A. Chilkoti, Plasmonic Detection of a Model Analyte in Serum by a Gold Nanorod Sensor, *Anal. Chem.*, 2007, **79**(14), 5278–5283.

55 K. M. Mayer and J. H. Hafner, Localized Surface Plasmon Resonance Sensors, *Chem. Rev.*, 2011, **111**(6), 3828–3857.

56 H. Vaisocherová-Lísalová, I. Víšová, M. L. Ermini, T. Špringer, X. C. Song, J. Mrázek, J. Lamačová, N. S. Lynn Jr, P. Šedivák and J. Homola, Low-Fouling Surface Plasmon Resonance Biosensor for Multi-Step Detection of Foodborne Bacterial Pathogens in Complex Food Samples, *Biosens. Bioelectron.*, 2016, **80**, 84–90.

57 J. Zhou, C. Zhang, X. Zhang, C. Lu, T. Ming, Y. Li and X. Su, Immunomagnetic Separation-Based Nanogold Enhanced Surface Plasmon Resonance and Colloidal Gold Test Strips for Rapid Detection of *Vibrio parahaemolyticus*, *Arch. Microbiol.*, 2020, 1–9.

58 J. Yguerabide and E. E. Yguerabide, Light-Scattering Submicroscopic Particles as Highly Fluorescent Analogs and Their Use as Tracer Labels in Clinical and Biological Applications I. Theory, *Anal. Biochem.*, 1998, **262**(2), 137–156.

59 P. K. Jain, K. S. Lee, I. H. El-Sayed and M. A. El-Sayed, Calculated Absorption and Scattering Properties of Gold Nanoparticles of Different Size, Shape, and Composition: Applications in Biological Imaging and Biomedicine, *J. Phys. Chem. B*, 2006, **110**(14), 7238–7248.

60 J. Yguerabide and E. E. Yguerabide, Light-Scattering Submicroscopic Particles as Highly Fluorescent Analogs and Their Use as Tracer Labels in Clinical and Biological Applications, *Anal. Biochem.*, 1998, **262**(2), 157–176.

61 C. Loo, L. Hirsch, M.-H. Lee, E. Chang, J. West, N. Halas and R. Drezek, Gold Nanoshell Bioconjugates for Molecular Imaging in Living Cells, *Opt. Lett.*, 2005, **30**(9), 1012–1014.

62 C. Loo, A. Lowery, N. Halas, J. West and R. Drezek, Immunotargeted Nanoshells for Integrated Cancer Imaging and Therapy, *Nano Lett.*, 2005, **5**(4), 709–711.

63 N. Chanda, R. Shukla, K. V. Katti and R. Kannan, Gastrin Releasing Protein Receptor Specific Gold Nanorods: Breast and Prostate Tumor Avid Nanovectors for Molecular Imaging, *Nano Lett.*, 2009, **9**(5), 1798–1805.

64 M. P. Melancon, W. Lu, Z. Yang, R. Zhang, Z. Cheng, A. M. Elliot, J. Stafford, T. Olson, J. Z. Zhang and C. Li, In Vitro and *In Vivo* Targeting of Hollow Gold Nanoshells Directed at Epidermal Growth Factor Receptor for Photothermal Ablation Therapy, *Mol. Cancer Ther.*, 2008, **7**(6), 1730–1709.

65 J. L. Li, L. Wang, X. Y. Liu, Z. P. Zhang, H. C. Guo, W. M. Liu and S. H. Tang, In Vitro Cancer Cell Imaging and Therapy Using Transferrin-Conjugated Gold Nanoparticles, *Cancer Lett.*, 2009, **274**(2), 319–326.

66 W. Qian, X. Huang, B. Kang and M. A. El-Sayed, Dark-Field Light Scattering Imaging of Living Cancer Cell Component from Birth through Division Using Bioconjugated Gold Nanoprobes, *J. Biomed. Opt.*, 2010, **15**(4), 046025.

67 X. Y. Wan, L. L. Zheng, P. F. Gao, X. X. Yang, C. M. Li, Y. F. Li and C. Z. Huang, Real-Time Light Scattering Tracking of Gold Nanoparticles-Bioconjugated Respiratory Syncytial Virus Infecting HEp-2 Cells, *Sci. Rep.*, 2014, **4**, 4529.

68 T. S. Troutman, J. K. Barton and M. Romanowski, Optical Coherence Tomography with Plasmon Resonant Nanorods of Gold, *Opt. Lett.*, 2007, **32**(11), 1438.

69 A. de la Zerda, S. Prabhulkar, V. L. Perez, M. Ruggeri, A. S. Paranjape, F. Habte, S. S. Gambhir and R. M. Awdeh, Optical Coherence Contrast Imaging Using Gold Nanorods in Living Mice Eyes, *Clin. Exp. Ophthalmol.*, 2015, **43**(4), 358–366.

70 Y. Winetraub, E. D. SoRelle, O. Liba and A. De La Zerda, Quantitative Contrast-Enhanced Optical Coherence Tomography, *Appl. Phys. Lett.*, 2016, **108**(2), 023702.

71 K. M. Ratheesh, P. Prabhathan, L. K. Seah and V. M. Murukeshan, Gold Nanorods with Higher Aspect Ratio as Potential Contrast Agent in Optical Coherence Tomography and for Photothermal Applications around 1300 nm Imaging Window, *Biomed. Phys. Eng. Express*, 2016, **2**(5), 055005.

72 D. Sen, E. D. SoRelle, O. Liba, R. Dalal, Y. M. Paulus, T.-W. Kim, D. M. Moshfeghi, A. de la Zerda, D. Sen, D. M. Moshfeghi, *et al.*, High-Resolution Contrast-Enhanced Optical Coherence Tomography in Mice Retinae, *J. Biomed. Opt.*, 2016, **21**(6), 066002.

73 N. Lippok, M. Villiger, A. Albanese, E. F. J. Meijer, K. Chung, T. P. Padera, S. N. Bhatia and B. E. Bouma, Depolarization Signatures Map Gold Nanorods within Biological Tissue, *Nat. Photonics*, 2017, **11**(9), 583–588.

74 E. D. Sorelle, D. W. Yecies, O. Liba, F. C. Bennett, C. M. Graef, R. Dutta, S. Mitra, L. M. Joubert, S. Cheshier, G. A. Grant, *et al.*, Spatiotemporal Tracking of Brain-Tumor-Associated Myeloid Cells *In Vivo* through Optical Coherence Tomography with Plasmonic Labeling and Speckle Modulation, *ACS Nano*, 2019, **13**(7), 7985–7995.

75 O. Bibikova, A. Popov, A. Bykov, A. Prilepskyii, M. Kinnunen, K. Kordas, V. Bogatyrev, N. Khlebtsov and V. Tuchin, Gold Nanostructures for OCT Imaging of Capillary Flow, *Proc. SPIE*, 2014, **9129**, 1–9.

76 X. Jiang, R. Liu, P. Tang, W. Li, H. Zhong, Z. Zhou and J. Zhou, Controllably Tuning the Near-Infrared Plasmonic Modes of Gold Nanoplates for Enhanced Optical Coherence Imaging and Photothermal Therapy, *RSC Adv.*, 2015, **5**, 80709–80718.

77 P. Tang, X. Jiang, Y. Wang, H. Chen, Y. S. Zhang, P. Gao, H. Wang, X. Li and J. Zhou, Plasmonic Nanoprobe of (Gold Triangular Nanoprism Core)/(Polyaniline Shell) for Real-Time Three-Dimensional pH Imaging of Anterior Chamber, *Anal. Chem.*, 2017, **89**(18), 9758–9766.

 © 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

78 X. Ye, L. Jin, H. Caglayan, J. Chen, G. Xing and C. Zheng, Improved Size-Tunable Synthesis of Monodisperse Gold Nanorods through the Use of Aromatic Additives, *ACS Nano*, 2012, **3**, 2804–2817.

79 E. D. SoRelle, D. W. Yecies, O. Liba, F. C. Bennett, C. M. Graef, R. Dutta, S. Mitra, L.-M. Joubert, S. Cheshier, G. A. Grant, *et al.*, Spatiotemporal Tracking of Brain-Tumor-Associated Myeloid Cells *In Vivo* through Optical Coherence Tomography with Plasmonic Labeling and Speckle Modulation, *ACS Nano*, 2019, **13**(7), 7985–7995.

80 R. Dutta, O. Liba, E. D. Sorelle, Y. Winetraub, V. C. Ramani, S. S. Jeffrey, G. W. Sledge and A. De La Zerda, Real-Time Detection of Circulating Tumor Cells in Living Animals Using Functionalized Large Gold Nanorods, *Nano Lett.*, 2019, **19**(4), 2334–2342.

81 D. Boyer, P. Tamarat, A. Maali, B. Lounis and M. Orrit, Photothermal Imaging of Nanometer-Sized Metal Particles among Scatterers, *Science*, 2002, **297**(5584), 1160–1163.

82 S. Berciaud, L. Cognet, P. Tamarat and B. Lounis, Observation of Intrinsic Size Effects in the Optical Response of Individual Gold Nanoparticles, *Nano Lett.*, 2005, **5**(3), 515–518.

83 V. P. Zharov and D. O. Lapotko, Photothermal Imaging of Nanoparticles and Cells, *IEEE J. Sel. Top. Quantum Electron.*, 2005, **11**(4), 733–751.

84 V. P. Zharov, E. I. Galanzha and V. V. Tuchin, Photothermal Flow Cytometry In Vitro for Detection and Imaging of Individual Moving Cells, *Cytometry, Part A*, 2007, **71**(4), 191–206.

85 V. P. Zharov, E. I. Galanzha and V. V. Tuchin, Photothermal Image Flow Cytometry *In Vivo*, *Opt. Lett.*, 2005, **30**(6), 628–630.

86 E. I. Galanzha, J. W. Kim and V. P. Zharov, Nanotechnology-Based Molecular Photoacoustic and Photothermal Flow Cytometry Platform for *In Vivo* Detection and Killing of Circulating Cancer Stem Cells, *J. Biophotonics*, 2009, **2**(12), 725–735.

87 J. W. Kim, E. I. Galanzha, E. V. Shashkov, H. M. Moon and V. P. Zharov, Golden Carbon Nanotubes as Multimodal Photoacoustic and Photothermal High-Contrast Molecular Agents, *Nat. Nanotechnol.*, 2009, **4**, 688–694.

88 D. A. Nedosekin, E. I. Galanzha, E. Dervishi, A. S. Biris and V. P. Zharov, Super-Resolution Nonlinear Photothermal Microscopy, *Small*, 2014, **10**(1), 135–142.

89 S. P. Mattison, W. Kim, J. Park and B. E. Applegate, Molecular Imaging in Optical Coherence Tomography, *Curr. Mol. Imaging*, 2015, **3**(2), 88–105.

90 M. C. Skala, M. J. Crow, A. Wax and J. A. Izatt, Photothermal Optical Coherence Tomography of Epidermal Growth Factor Receptor in Live Cells Using Immunotargeted Gold Nanospheres, *Nano Lett.*, 2008, **8**(10), 3461–3467.

91 C. Zhou, T.-H. Tsai, D. C. Adler, H.-C. Lee, D. W. Cohen, A. Mondelblatt, Y. Wang, J. L. Connolly and J. G. Fujimoto, Photothermal Optical Coherence Tomography in Ex Vivo Human Breast Tissues Using Gold Nanoshells, *Opt. Lett.*, 2010, **35**(5), 700.

92 J. M. Tucker-Schwartz, T. A. Meyer, C. A. Patil, C. L. Duvall and M. C. Skala, *In Vivo* Photothermal Optical Coherence Tomography of Gold Nanorod Contrast Agents, *Biomed. Opt. Express*, 2012, **3**(11), 2881.

93 J. M. Tucker-Schwartz, K. R. Beavers, W. W. Sit, A. T. Shah, C. L. Duvall and M. C. Skala, *In Vivo* Imaging of Nanoparticle Delivery and Tumor Microvasculature with Multimodal Optical Coherence Tomography, *Biomed. Opt. Express*, 2014, **5**(6), 1731.

94 M. Lapierre-Landry, A. Y. Gordon, J. S. Penn and M. C. Skala, *In Vivo* Photothermal Optical Coherence Tomography of Endogenous and Exogenous Contrast Agents in the Eye, *Sci. Rep.*, 2017, **7**(1), 1–9.

95 A. Y. Gordon, M. Lapierre-Landry, M. C. Skala and J. S. Penn, Photothermal Optical Coherence Tomography of Anti-Angiogenic Treatment in the Mouse Retina Using Gold Nanorods as Contrast Agents, *Transl. Vis. Sci. Technol.*, 2019, **8**(3).

96 V. Sebastián, S.-K. Lee, C. Zhou, M. F. Kraus, J. G. Fujimoto and K. F. Jensen, One-Step Continuous Synthesis of Biocompatible Gold Nanorods for Optical Coherence Tomography, *Chem. Commun.*, 2012, **48**(53), 6654.

97 Y. Jung, R. Reif, Y. Zeng and R. K. Wang, Three-Dimensional High-Resolution Imaging of Gold Nanorods Uptake in Sentinel Lymph Nodes, *Nano Lett.*, 2011, **11**(7), 2938–2943.

98 Q. Zhang, N. Iwakuma, P. Sharma, B. M. Moudgil, C. Wu, J. McNeill, H. Jiang and S. R. Grobmyer, Gold Nanoparticles as a Contrast Agent for *In Vivo* Tumor Imaging with Photoacoustic Tomography, *Nanotechnology*, 2009, **20**(39), 395102.

99 M. Eghtedari, A. Oraevsky, J. A. Copland, N. A. Kotov, A. Conjusteau and M. Motamedi, High Sensitivity of *In Vivo* Detection of Gold Nanorods Using a Laser Optoacoustic Imaging System, *Nano Lett.*, 2007, **7**(7), 1914–1918.

100 A. Taruttis, E. Herzog, D. Razansky and V. Ntziachristos, Real-Time Imaging of Cardiovascular Dynamics and Circulating Gold Nanorods with Multispectral Optoacoustic Tomography, *Opt. Express*, 2010, **18**(19), 19592–19602.

101 P.-C. Li, C.-R. C. Wang, D.-B. Shieh, C.-W. Wei, C.-K. Liao, C. Poe, S. Jhan, A.-A. Ding and Y.-N. Wu, *In Vivo* Photoacoustic Molecular Imaging with Simultaneous Multiple Selective Targeting Using Antibody-Conjugated Gold Nanorods, *Opt. Express*, 2008, **16**(23), 18605–18615.

102 C. L. Bayer, Y.-S. Chen, S. Kim, S. Mallidi, K. Sokolov and S. Emelianov, Multiplex Photoacoustic Molecular Imaging Using Targeted Silica-Coated Gold Nanorods, *Biomed. Opt. Express*, 2011, **2**(7), 1828–1835.

103 J. V. Jokerst, M. Thangaraj, P. J. Kempen, R. Sinclair and S. S. Gambhir, Photoacoustic Imaging of Mesenchymal Stem Cells in Living Mice via Silica-Coated Gold Nanorods, *ACS Nano*, 2012, **6**(7), 5920–5930.

104 S. Yang, F. Ye and D. Xing, Intracellular Label-Free Gold Nanorods Imaging with Photoacoustic Microscopy, *Opt. Express*, 2012, **20**(9), 10370–10375.

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

105 C. Kim, H. M. Song, X. Cai, J. Yao, A. Wei and L. V. Wang, *In Vivo* Photoacoustic Mapping of Lymphatic Systems with Plasmon-Resonant Nanostars, *J. Mater. Chem.*, 2011, **21**, 2841–2844.

106 S. Laing, L. E. Jamieson, K. Faulds and D. Graham, Surface-Enhanced Raman Spectroscopy for *In Vivo* Biosensing, *Nat. Rev. Chem.*, 2017, **1**(8), 1–19.

107 X. Lu, Y. Huang, B. Liu, L. Zhang, L. Song, J. Zhang, A. Zhang and T. Chen, Light-Controlled Shrinkage of Large-Area Gold Nanoparticle Monolayer Film for Tunable SERS Activity, *Chem. Mater.*, 2018, **30**(6), 1989–1997.

108 S. Si, W. Liang, Y. Sun, J. Huang, W. Ma, Z. Liang, Q. Bao and L. Jiang, Facile Fabrication of High-Density Sub-1 nm Gaps from Au Nanoparticle Monolayers as Reproducible SERS Substrates, *Adv. Funct. Mater.*, 2016, **26**(44), 8137–8145.

109 L. Song, Y. Huang, Z. Nie and T. Chen, Macroscopic Two-Dimensional Monolayer Films of Gold Nanoparticles: Fabrication Strategies, Surface Engineering and Functional Applications, *Nanoscale*, 2020, **12**, 7433–7460.

110 H. Shao, H. Lin, Z. Guo, J. Lu, Y. Jia, M. Ye, F. Su, L. Niu, W. Kang, S. Wang, *et al.*, A Multiple Signal Amplification Sandwich-Type SERS Biosensor for Femtomolar Detection of miRNA, *Biosens. Bioelectron.*, 2019, **143**, 111616.

111 Q. Cheng, L. Song, H. Lin, Y. Yang, Y. Huang, F. Su and T. Chen, Free-Standing 2D Janus Gold Nanoparticles Monolayer Film with Tunable Bifacial Morphologies via the Asymmetric Growth at Air–Liquid Interface, *Langmuir*, 2020, **36**(1), 250–256.

112 X. Wang, N. Choi, Z. Cheng, J. Ko, L. Chen and J. Choo, Simultaneous Detection of Dual Nucleic Acids Using a SERS-Based Lateral Flow Assay Biosensor, *Anal. Chem.*, 2017, **89**(2), 1163–1169.

113 M. Fleischmann, P. J. Hendra and A. J. McQuillan, Raman spectra of pyridine adsorbed at a silver electrode, *Chem. Phys. Lett.*, 1974, **26**(2), 163–166.

114 Y. Li, H. J. Schluesener and S. Xu, Gold Nanoparticle-Based Biosensors, *Gold Bull.*, 2010, **43**(1), 29–41.

115 D. S. Kopitsyn, M. V. Gorbachevskii, E. A. Botchkova, M. A. Bychenko and A. A. Novikov, Identification of Bacteria by Surface-Enhanced Raman Spectra after Peroxide Treatment, *Appl. Biochem. Microbiol.*, 2019, **55**(1), 78–82.

116 A. H. H. Mevold, W.-W. Hsu, A. Hardiansyah, L.-Y. Huang, M.-C. Yang, T.-Y. Liu, T.-Y. Chan, K.-S. Wang, Y.-A. Su and R.-J. Jeng, Fabrication of Gold Nanoparticles/Graphene–PDDA Nanohybrids for Bio-Detection by SERS Nanotechnology, *Nanoscale Res. Lett.*, 2015, **10**(1), 1–7.

117 X. Ma, X. Xu, Y. Xia and Z. Wang, SERS Aptasensor for *Salmonella typhimurium* Detection Based on Spiny Gold Nanoparticles, *Food Control*, 2018, **84**, 232–237.

118 S. Harmsen, R. Huang, M. A. Wall, H. Karabeber, J. M. Samii, M. Spaliviero, J. R. White, S. Monette, R. O'Connor, K. L. Pitter, *et al.*, Surface-Enhanced Resonance Raman Scattering Nanostars for High-Precision Cancer Imaging, *Sci. Transl. Med.*, 2015, **7**(271), 271ra7.

119 T. Vo-Dinh, H. N. Wang and H. T. Ngo *Biomedical Photonics Handbook*, ed. T. Vo-Dinh, CRC Press, 2014.

120 G. L. Liu, Y. Lu, J. Kim, J. C. Doll and L. P. Lee, Magnetic Nanocrescents as Controllable Surface-Enhanced Raman Scattering Nanoprobes for Biomolecular Imaging, *Adv. Mater.*, 2005, **17**(22), 2683–2688.

121 G. W. Kim and J. W. Ha, Polarization- and Wavelength-Dependent Defocused Scattering Imaging of Single Gold Nanostars with Multiple Long Branches, *Photochem. Photobiol. Sci.*, 2019, **18**, 1430–1435.

122 G. H. Jeong, Y. W. Lee, M. Kim and S. W. Han, High-Yield Synthesis of Multi-Branched Gold Nanoparticles and Their Surface-Enhanced Raman Scattering Properties, *J. Colloid Interface Sci.*, 2009, **329**(1), 97–102.

123 O. M. Bakr, B. H. Wunsch and F. Stellacci, High-Yield Synthesis of Multi-Branched Urchin-Like Gold Nanoparticles, *Chem. Mater.*, 2006, **18**(14), 3297–3301.

124 Y. Kang, H. X. Gu and X. Zhang, A Self-Referenced Method for Determination of Patulin by Surface-Enhanced Raman Scattering Using Gold Nanobipyramids as the Substrate, *Anal. Methods*, 2019, **11**, 5142–5149.

125 S. Y. Lee, Y. Han, J. W. Hong and J. W. Ha, Single Gold Bipyramids with Sharp Tips as Sensitive Single Particle Orientation Sensors in Biological Studies, *Nanoscale*, 2017, **9**, 12060–12067.

126 A. A. Saeed, J. L. A. Sánchez, C. K. O'Sullivan and M. N. Abbas, DNA Biosensors Based on Gold Nanoparticles-Modified Graphene Oxide for the Detection of Breast Cancer Biomarkers for Early Diagnosis, *Bioelectrochemistry*, 2017, **118**, 91–99.

127 N. Elahi, M. Kamali, M. H. Baghersad and B. Amini, A Fluorescence Nano-Biosensors Immobilization on Iron (MNPs) and Gold (AuNPs) Nanoparticles for Detection of *Shigella* spp., *Mater. Sci. Eng., C*, 2019, **105**, 110113.

128 S. Xu, F. Zhang, L. Xu, X. Liu, P. Ma, Y. Sun, X. Wang and D. Song, A Fluorescence Resonance Energy Transfer Biosensor Based on Carbon Dots and Gold Nanoparticles for the Detection of Trypsin, *Sens. Actuators, B*, 2018, **273**, 1015–1021.

129 X. Fu, X. Fu, Q. Wang, L. Sheng, X. Huang, M. Ma and Z. Cai, Fluorescence Switch Biosensor Based on Quantum Dots and Gold Nanoparticles for Discriminative Detection of Lysozyme, *Int. J. Biol. Macromol.*, 2017, **103**, 1155–1161.

130 S. Barua, S. Gogoi and R. Khan, Fluorescence Biosensor Based on Gold–Carbon Dot Probe for Efficient Detection of Cholesterol, *Synth. Met.*, 2018, **244**, 92–98.

131 H. Zhang and Z. Jia, Development of Fluorescent FRET Probes for "Off–On" Detection of L-Cysteine Based on Gold Nanoparticles and Porous Silicon Nanoparticles in Ethanol Solution, *Sensors*, 2017, **17**(3), 520.

132 B. Jin, S. Wang, M. Lin, Y. Jin, S. Zhang, X. Cui, Y. Gong, A. Li, F. Xu and T. J. Lu, Upconversion Nanoparticles Based FRET Aptasensor for Rapid and Ultrasensitive Bacteria Detection, *Biosens. Bioelectron.*, 2017, **90**, 525–533.

 © 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

133 W. Bai, Y. Wei, Y. Zhang, L. Bao and Y. Li, Label-Free and Amplified Electrogenerated Chemiluminescence Biosensing for the Detection of Thymine DNA Glycosylase Activity Using DNA-Functionalized Gold Nanoparticles Triggered Hybridization Chain Reaction, *Anal. Chim. Acta*, 2019, **1061**, 101–109.

134 A. Cui, J. Zhang, W. Bai, H. Sun, L. Bao, F. Ma and Y. Li, Signal-on Electrogenerated Chemiluminescence Biosensor for Ultrasensitive Detection of microRNA-21 Based on Isothermal Strand-Displacement Polymerase Reaction and Bridge DNA–Gold Nanoparticles, *Biosens. Bioelectron.*, 2019, **144**, 111664.

135 H. Zhang, Z. Wang, F. Wang, Y. Zhang, H. Wang and Y. Liu, In Situ Formation of Gold Nanoparticles Decorated Ti_3C_2 MXenes Nanoprobe for Highly Sensitive Electrogenerated Chemiluminescence Detection of Exosomes and Their Surface Proteins, *Anal. Chem.*, 2020, **92**(7), 5546–5553.

136 J. F. Li, C. Y. Li and R. F. Aroca, Plasmon-Enhanced Fluorescence Spectroscopy, *Chem. Soc. Rev.*, 2017, **46**, 3962–3979.

137 J. Gersten and A. Nitzan, Spectroscopic Properties of Molecules Interacting with Small Dielectric Particles, *J. Chem. Phys.*, 1981, **75**(3), 1139–1152.

138 Y. Wu, M. R. K. Ali, K. Chen, N. Fang and M. A. El-Sayed, Gold Nanoparticles in Biological Optical Imaging, *Nano Today*, 2019, 120–140.

139 A. Kinkhabwala, Z. Yu, S. Fan, Y. Avlasevich, K. Müllen and W. E. Moerner, Large Single-Molecule Fluorescence Enhancements Produced by a Bowtie Nanoantenna, *Nat. Photonics*, 2009, **3**(11), 654–657.

140 S. Khatua, P. M. R. Paulo, H. Yuan, A. Gupta, P. Zijlstra and M. Orrit, Resonant Plasmonic Enhancement of Single-Molecule Fluorescence by Individual Gold Nanorods, *ACS Nano*, 2014, **8**(5), 4440–4449.

141 J. C. G. Jeynes, K. Geraki, C. Jeynes, M. Zhaohong, A. A. Bettiol, E. Latorre, L. W. Harries and C. Soeller, Nanoscale Properties of Human Telomeres Measured with a Dual Purpose X-Ray Fluorescence and Super Resolution Microscopy Gold Nanoparticle Probe, *ACS Nano*, 2017, **11**(12), 12632–12640.

142 V. Venkatesh, A. Shukla, S. Sivakumar and S. Verma, Purine-Stabilized Green Fluorescent Gold Nanoclusters for Cell Nuclei Imaging Applications, *ACS Appl. Mater. Interfaces*, 2014, **6**(3), 2185–2191.

143 X. Wang, Y. Wang, H. He, X. Ma, Q. Chen, S. Zhang, B. Ge, S. Wang, W. M. Nau and F. Huang, Deep-Red Fluorescent Gold Nanoclusters for Nucleoli Staining: Real-Time Monitoring of the Nucleolar Dynamics in Reverse Transformation of Malignant Cells, *ACS Appl. Mater. Interfaces*, 2017, **9**(21), 17799–17806.

144 T. D. Fernández, J. R. Pearson, M. P. Leal, M. J. Torres, M. Blanca, C. Mayorga and X. Le Guével, Intracellular Accumulation and Immunological Properties of Fluorescent Gold Nanoclusters in Human Dendritic Cells, *Biomaterials*, 2015, **43**, 1–12.

145 S. Chandirasekar, C. Chandrasekaran, T. Muthukumara samyvel, G. Sudhandiran and N. Rajendiran, Biosurfactant Templated Quantum Sized Fluorescent Gold Nanoclusters for *In Vivo* Bioimaging in Zebrafish Embryos, *Colloids Surf., B*, 2016, **143**, 472–480.

146 C. Zhang, C. Li, Y. Liu, J. Zhang, C. Bao, S. Liang, Q. Wang, Y. Yang, H. Fu, K. Wang, *et al.*, Gold Nanoclusters-Based Nanoprobes for Simultaneous Fluorescence Imaging and Targeted Photodynamic Therapy with Superior Penetration and Retention Behavior in Tumors, *Adv. Funct. Mater.*, 2015, **25**(8), 1314–1325.

147 N. J. Durr, T. Larson, D. K. Smith, B. A. Korgel, K. Sokolov and A. Ben-Yakar, Two-Photon Luminescence Imaging of Cancer Cells Using Molecularly Targeted Gold Nanorods, *Nano Lett.*, 2007, **7**(4), 941–945.

148 H. Wang, T. B. Huff, D. A. Zweifel, W. He, P. S. Low, A. Wei and J. X. Cheng, In Vitro and *In Vivo* Two-Photon Luminescence Imaging of Single Gold Nanorods, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(44), 15752–15756.

149 H. Yuan, C. G. Khoury, H. Hwang, C. M. Wilson, G. A. Grant and T. Vo-Dinh, Gold Nanostars: Surfactant-Free Synthesis, 3D Modelling, and Two-Photon Photoluminescence Imaging, *Nanotechnology*, 2012, **23**(7), 075102.

150 N. Gao, Y. Chen, L. Li, Z. Guan, T. Zhao, N. Zhou, P. Yuan, S. Q. Yao and Q. H. Xu, Shape-Dependent Two-Photon Photoluminescence of Single Gold Nanoparticles, *J. Phys. Chem. C*, 2014, **118**(25), 13904–13911.

151 I. Khalil, W. A. Yehye, N. M. Julkapli, S. Rahmati, A. A. I. Sina, W. J. Basirun and M. R. Johan, Graphene Oxide and Gold Nanoparticle Based Dual Platform with Short DNA Probe for the PCR Free DNA Biosensing Using Surface-Enhanced Raman Scattering, *Biosens. Bioelectron.*, 2019, **131**, 214–223.

152 X. Cao, Z. Xia, W. Yan, S. He, X. Xu, Z. Wei, Y. Ye and H. Zheng, Colorimetric Biosensing of Nopaline Synthase Terminator Using Fe_3O_4@Au and Hemin-Functionalized Reduced Graphene Oxide, *Anal. Biochem.*, 2020, **602**, 113798.

153 L. Xiao, A. Zhu, Q. Xu, Y. Chen, J. Xu and J. Weng, Colorimetric Biosensor for Detection of Cancer Biomarker by Au Nanoparticle-Decorated Bi_2Se_3 Nanosheets, *ACS Appl. Mater. Interfaces*, 2017, **9**(8), 6931–6940.

154 Z.-M. Li, X. Zhang, T. Pi, J. Bu, R.-H. Deng, B.-Z. Chi and X.-J. Zheng, Colorimetric Determination of the Activity of Methyltransferase Based on Nicking Enzyme Amplification and the Use of Gold Nanoparticles Conjugated to Graphene Oxide, *Microchim. Acta*, 2019, **186**(8), 594.

155 M. F. Kircher, A. De La Zerda, J. V. Jokerst, C. L. Zavaleta, P. J. Kempen, E. Mittra, K. Pitter, R. Huang, C. Campos, F. Habte, *et al.*, A Brain Tumor Molecular Imaging Strategy Using a New Triple-Modality MRI-Photoacoustic-Raman Nanoparticle, *Nat. Med.*, 2012, **18**(5), 829–834.

156 X. Huang, I. H. El-Sayed, W. Qian and M. A. El-Sayed, Cancer Cell Imaging and Photothermal Therapy in the Near-Infrared Region by Using Gold Nanorods, *J. Am. Chem. Soc.*, 2006, **128**(6), 2115–2120.

Open Access Article. Published on 14 April 2021. Downloaded on 6/15/2021 11:16:51 AM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

157 S. Oldenburg, R. Averitt, S. Westcott and N. Halas, Nanoengineering of Optical Resonances, *Chem. Phys. Lett.*, 1998, **288**(2–4), 243–247.

158 X. Huang, S. Neretina and M. A. El-Sayed, Gold Nanorods: From Synthesis and Properties to Biological and Biomedical Applications, *Adv. Mater.*, 2009, **21**(48), 4880–4910.

159 A. Guerrero-Martínez, S. Barbosa, I. Pastoriza-Santos and L. M. Liz-Marzán, Nanostars Shine Bright for You, *Curr. Opin. Colloid Interface Sci.*, 2011, **16**(2), 118–127.

160 J. Chen, J. M. McLellan, A. Siekkinen, Y. Xiong, Z.-Y. Li and Y. Xia, Facile Synthesis of Gold–Silver Nanocages with Controllable Pores on the Surface, *J. Am. Chem. Soc.*, 2006, **128**(46), 14776–14777.

161 X. Jiang, P. Tang, P. Gao, Y. S. Zhang, C. Yi and J. Zhou, Gold Nanoprobe-Enabled Three-Dimensional Ozone Imaging by Optical Coherence Tomography, *Anal. Chem.*, 2017, **89**(4), 2561–2568.

162 J. Henzie, E.-S. Kwak and T. W. Odom, Mesoscale Metallic Pyramids with Nanoscale Tips, *Nano Lett.*, 2005, **5**(7), 1199–1202.

163 M. I. Halawa, J. Lai and G. Xu, Gold Nanoclusters: Synthetic Strategies and Recent Advances in Fluorescent Sensing, *Materials Today Nano*, 2018, 9–27.

© 2021 The Author(s). Published by the Royal Society of Chemistry

2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-7281-5791-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECT49130.2021.9392631

Harmonics Mitigation Techniques in Grid Integrated PV based Microgrid: A Comparative Analysis

Avdhesh Kumar, Rachana Garg and Priya Mahajan

Department of Electrical Engineering, Delhi Technological University, Delhi, India
E-mail : iesavd@gmail.com, rachana16100@yahoo.co.in, priyamahajan.eed@gmail.com

Abstract : **With the increased penetration of photovoltaic (PV) based distributed generation (DG), power quality (PQ) at the point of common coupling (PCC) becomes a major concern mainly due to harmonics. Harmonics, generated by nonlinear loads in the system, are multiples of fundamental frequency. This paper presents the comparative analysis and discussion of various control techniques for VSC of a grid connected PV based micro grid to mitigate the harmonics. The mitigation of harmonics is achieved using control techniques viz. conventional and adaptive of reference generation. Simulation of the developed model is carried out on MATLAB/Simulink. Performance of the system is evaluated under IEEE-519 standard.**

Keywords—Power Quality (PQ), Power Factor Correction (PFC), Total Harmonic Distortions (THD), Voltage Source Converter (VSC)

I. Introduction

The increasing penetration of solar photovoltaic (SPV) system in utility grids, degrades power quality (PQ) at the point of common coupling (PCC) is becoming a major concern. Harmonics are generated either by nonlinear local loads or power electronics converter used in the PV based microgrid system [1-4]. The integration of PV system requires voltage source converter (VSC), switching of these converters generates harmonics at the output of microgrid. In the present study, harmonics generated due to local loads only are mitigated using different control algorithm of VSC [5]– [7]. In [8] authors have reported that a phase locked loop (PLL) is required to transform three phases to synchronous frame in SRF based control but sometimes it shows computational delay and also it has to be tuned preceding to its operation. In [9] authors reported that, in the implementation of IRPT and PBT, reference current is estimated using voltage and current of the PCC, so fluctuation in voltage will reflect in the reference current. While unit template is simplified form of adaptive LMS algorithm. Adaptive LMS suffer from fluctuations and less accuracy in estimating mean square error (MSE). In [10] authors reported that in the past few decades, adaptive control algorithms have been proposed for many applications. Among them, the least mean square (LMS) algorithm is the distinguished one. Due to its simplicity, conventional LMS adaptive control has been implemented widely but the performance is often unsatisfactory because of poor dynamic performance, due to compromise between tracking capability and accuracy in fixed step size.

In this paper, PV based microgrid has been integrated to grid. VSC is controlled using both conventional techniques viz. synchronous reference frame (SRF), instantaneous reactive power theory (IRPT), unit template, power balance theory (PBT) and adaptive Least mean square (LMS) technique [11]–[16]. MATLAB/Simulink software has been used for modelling and simulation of these algorithms. The efficacy of control algorithm of VSC shows the efficient operation of the grid integrated PV system. VSC maintain unity power factor at grid by supplying reactive power demand of load along-with active power. Further, in case of unbalanced loads, VSC maintain the grid currents balanced by supplying current to unbalance loads as per each phase requirement and also eliminates the harmonics in the grid current.

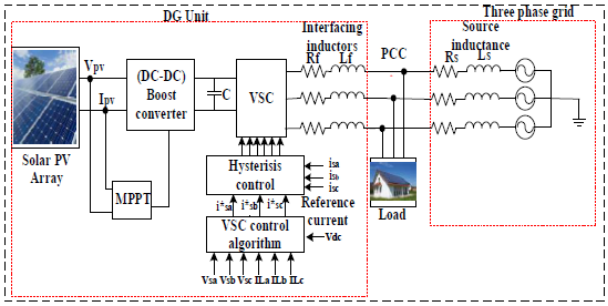


Fig.1 System configuration

II. System Configuration

Fig. 1 depicts proposed system configuration. A 10.25kW SPV system has been integrated to grid. Various parameters of the system are given in the appendix. The proposed system consists of a SPV array, dc-dc (boost) converter, VSC and consumer loads. As power from PV array is not constant and changes with environmental conditions, therefore to get the maximum efficiency of SPV array, Perturb and observe MPPT algorithm is used [17]–[20]. DC link voltage obtained after boost converter is converted to ac voltage of required magnitude and frequency using VSC.

III. VSC Control Algorithms

A. SRF Control algorithm

Fig. 2 represents the block diagram of reference current estimation using SRF control algorithm under UPF mode of operation. Park's transformation has been used to convert load current components from synchronous rotating reference frame to d-q reference frame.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:44 UTC from IEEE Xplore. Restrictions apply.

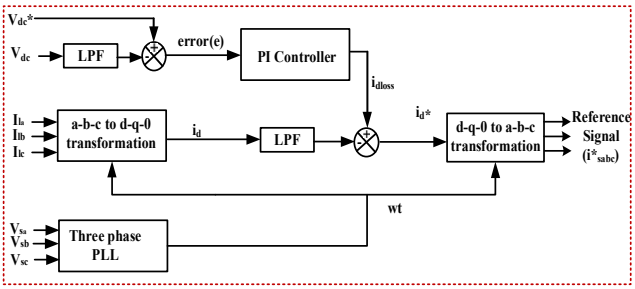


Fig. 2 SRF control algorithm

Three phase current in the a-b-c frame is transformed into d-q-0 frame by means of Park's conversion as given below:

$$\begin{bmatrix} i_d \\ i_q \\ i_0 \end{bmatrix} = \frac{2}{3}\begin{bmatrix} \cos w\,t & \cos(wt-120) & \cos(wt+120) \\ \sin wt & \sin(wt-120) & \sin(wt+120) \\ 1/2 & 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} i_{la} \\ i_{lb} \\ i_{lc} \end{bmatrix} \quad (1)$$

PI controller, is used to maintain the voltage across dc link. fundamental active (i_d) and reactive (i_q) component of load currents are extracted by means of low pass filter. VSC must supply the reactive power demand of the load to operate the SRF algorithm in unity power factor mode. Reference reactive component (i_q^*) must be zero $(i_q^* = 0)$, to compensate the reactive demand of the load. While (i_d) is added with output of PI controller (i_{dloss}) in order to regulate the dc link voltage.

$$i_d^* = i_d + i_{dloss} \quad (2)$$

Further converting the reference signal from d-q frame to a-b-c by means of inverse parks transformation, provides reference current ($i_{sa}^*, i_{sb}^*, i_{sc}^*$).

Generated reference current (i^*_{sabc}) must be in same phase with grid voltage, using reverse Park's conversion as represented below, i^*_{sabc} are obtained:

$$\begin{bmatrix} i_{sa}^* \\ i_{sb}^* \\ i_{sc}^* \end{bmatrix} = \begin{bmatrix} cos\ wt & \sin wt & 1 \\ \cos(wt-120) & \sin(wt-120) & 1 \\ \cos(wt+120) & \sin(wt+120) & 1 \end{bmatrix}\begin{bmatrix} i_d \\ i_q \\ i_0 \end{bmatrix} \quad (3)$$

Reference current are compared with sensed grid current $(i_{sa}, i_{sb,} i_{sa})$ in hysteresis control and generate switching signals to operate VSC [8-12].

B. IRPT Theory of Inverter Control

Fig. 3 depicts the block diagram representation of IRPT control algorithm of reference current. In this control Clark's transformations are used to converts the PCC voltage and load current in to α-β frame respectively as given below:

$$\begin{bmatrix} v_\alpha \\ v_\beta \end{bmatrix} = \sqrt{(2/3)}\begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix}\begin{bmatrix} v_{sa} \\ v_{sb} \\ v_{sc} \end{bmatrix} \quad (4)$$

$$\begin{bmatrix} i_{L\alpha} \\ i_{L\beta} \end{bmatrix} = \sqrt{(2/3)}\begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix}\begin{bmatrix} i_{La} \\ i_{Lb} \\ i_{Lc} \end{bmatrix} \quad (5)$$

Instantaneous magnitude of active power (p_L) and reactive power (q_L) are computed using the equations (6) and (7) as given below:

$$p_L = v_\alpha i_{L\alpha} + v_\beta i_{L\beta} \quad (6)$$

$$q_L = v_\alpha i_{L\beta} + v_\beta i_{L\alpha} \quad (7)$$

$$\begin{bmatrix} p_L \\ q_L \end{bmatrix} = \begin{bmatrix} \overline{p_L} & +\widetilde{p_L} \\ \overline{q_L} & +\widetilde{p_L} \end{bmatrix} \quad (8)$$

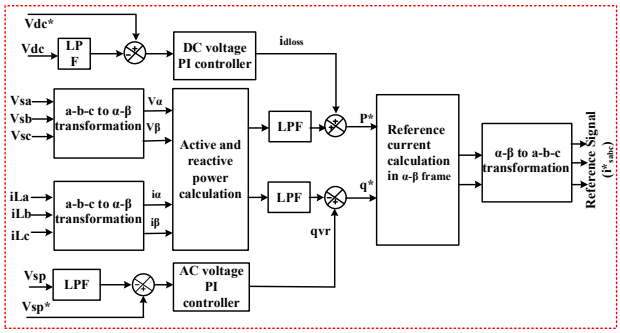


Fig. 3 IRPT control algorithm

Computed power contains both dc and ac component. To filter out the fundamental power components, low pass filters are used. Reference active and reactive component of power is estimated using Eq. given below:

$$p^* = \overline{p_L} + p_{loss} \quad (9)$$

$$q^* = \overline{q_L} + q_{vr} \quad (10)$$

Reference current ($i_{sa}^*, i_{sb}^*, i_{sc}^*$) is estimated as given below:

$$\begin{bmatrix} i_{sa}^* \\ i_{sb}^* \\ i_{sc}^* \end{bmatrix} = \sqrt{(2/3)}\begin{bmatrix} 1 & 0 \\ -1/2 & \sqrt{3}/2 \\ -1/2 & -\sqrt{3}/2 \end{bmatrix}\begin{pmatrix} v_\alpha & v_\beta \\ -v_\beta & v_\alpha \end{pmatrix}^{-1}\begin{bmatrix} p^* \\ q^* \end{bmatrix} \quad (11)$$

In PFC mode reference reactive power component (q^*) should be zero to maintain the grid at unity power factor. Reference current are compared with sensed grid current $(i_{sa}, i_{sb,} i_{sa})$ in hysteresis control and generate switching signals to operate VSC [13-15].

C. Unit template Control algorithm

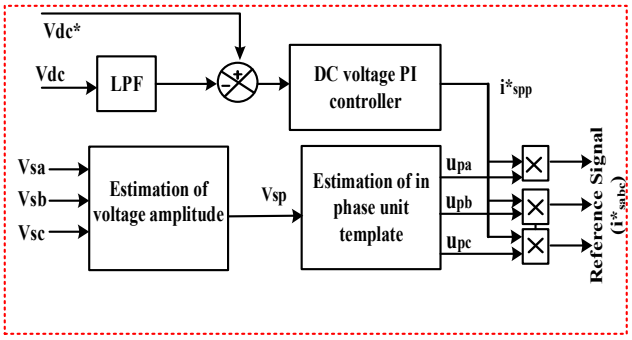


Fig.4 Unit template control algorithm

Unit template control algorithm of VSC for estimating the reference currents shown in Fig.4.The unit template is a simple VSC control technique for estimating the reference current. This control algorithm can be made flexible and it can be modified either for PFC or voltage regulation at PCC. The unit template control algorithm in PFC mode of operation is performed to estimate reference current as shown in the fig. 4. PCC voltage and DC bus voltage of VSC are used for implementing this control algorithm. A band-pass filter (BPF) can be used to remove distortion in sampled PCC voltages in real time implementation. Assume that after filtering the PCC voltage signal is (v_{sa}, v_{sb}, v_{sc}). Peak amplitude of PCC voltage is calculated as:

$$v_{sp} = \sqrt{\frac{2}{3}(v_{sa}^2 + v_{sb}^2 + v_{sc}^2)} \quad (12)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:44 UTC from IEEE Xplore. Restrictions apply.

The in-phase unit voltages template ($\mathcal{U}_{pa}, \mathcal{U}_{pb}, \mathcal{U}_{pc}$) can be calculated using phase voltages ($\mathcal{V}_{sa}, \mathcal{V}_{sb}, \mathcal{V}_{sc}$) with peak amplitude($\mathcal{V}_{sp}$) as follows:

$$\mathcal{U}_{pa} = \frac{\mathcal{V}_{sa}}{\mathcal{V}_{sp}}, \mathcal{U}_{pb} = \frac{\mathcal{V}_{sb}}{\mathcal{V}_{sp}}, \mathcal{U}_{pc} = \frac{\mathcal{V}_{sc}}{\mathcal{V}_{sp}} \quad (13)$$

Further, error in DC link voltage is compensated using PI controller. Output of PI controller is considered as amplitude of reference current (i^*_{spp}) . Amplitude of reference current(i^*_{spp}) are multiplied by in- phase unit template to get reference current ($i^*_{sa}, i^*_{sb}, i^*_{sc}$) as given below:

$$i^*_{sa} = i^*_{spp}\mathcal{U}_{sa}, i^*_{sb} = i^*_{spp}\mathcal{U}_{sb}, i^*_{sc} = i^*_{spp}\mathcal{U}_{sc} \quad (14)$$

Reference current ($i^*_{sa}, i^*_{sb}, i^*_{sc}$) are compared with sensed grid current (i_{sa}, i_{sb}, i_{sa}) in hysteresis control and generate switching signals to operate VSC.

D. Power balance theory control algorithm

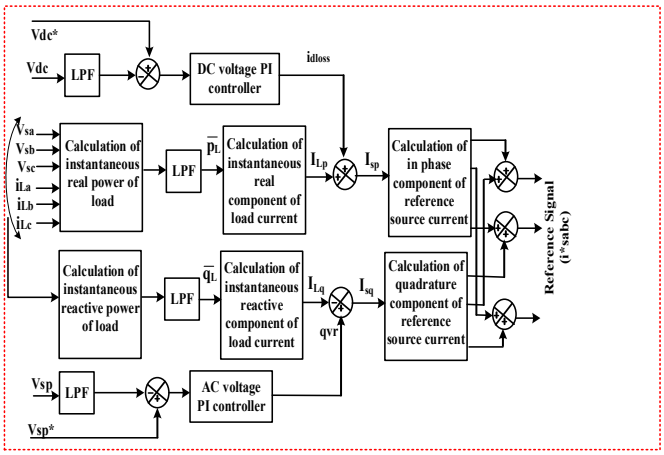


Fig.5 Power balance theory-based control algorithm

Fig. 5 depicts the block diagram representation of PBT control algorithm for generation of reference current. In this control, instantaneous power consumed by the loads can be utilized to extract of fundamental components of load currents. This algorithm sense the load currents, DC bus voltage, PCC voltage and grid current of proposed system. DC link PI controller output is added with fundamental active power component of load current as in eqn. (17) to regulate DC link voltage and to estimate the fundamental active power component of reference supply current. Eqn. (18) and (20) are used to compute instantaneous power (active and reactive) component of reference currents. Further to get reference current, in phase and quadrature reference current added together by taking respective phases.

The peak amplitude of voltage ($\mathcal{V}_t$)of grid voltage, can be calculated by:

$$\mathcal{V}_t = \sqrt{\frac{2}{3}(\mathcal{V}_{sa}^2 + \mathcal{V}_{sb}^2 + \mathcal{V}_{sc}^2)} \quad (15)$$

The in-phase unit template voltages ($\mathcal{U}_{pa}, \mathcal{U}_{pb}, \mathcal{U}_{pc}$) can be calculated from phase voltages ($\mathcal{V}_{sa}, \mathcal{V}_{sb}, \mathcal{V}_{sc}$) and peak amplitude ($\mathcal{V}_t$) as follows:

$$\mathcal{U}_{pa} = \frac{\mathcal{V}_{sa}}{\mathcal{V}_t}, \mathcal{U}_{pb} = \frac{\mathcal{V}_{sb}}{\mathcal{V}_t}, \mathcal{U}_{pc} = \frac{\mathcal{V}_{sc}}{\mathcal{V}_t} \quad (16)$$

The reactive/orthogonal/quadrature unit templates ($\mathcal{U}_{qa}, \mathcal{U}_{qb}, \mathcal{U}_{qc}$) can be obtained from in phase unit template voltages as follows:

$$u_{qa} = -\frac{u_{pb}}{\sqrt{3}} + \frac{u_{pc}}{\sqrt{3}}, u_{qb} = \frac{\sqrt{3}\,u_{pa}}{2} + \frac{(u_{pb}-u_{pc})}{2\sqrt{3}}, \quad u_{qc} = -\frac{\sqrt{3}\,u_{pa}}{2} + \frac{(u_{pb}-u_{pc})}{2\sqrt{3}} \quad (17)$$

The instantaneous active and reactive powers of the load are calculated as:

$$p_L = v_{sa}i_{La} + v_{sb}i_{Lb} + v_{sc}i_{Lc} = \overline{p_L} + \widetilde{p_L} \quad (18)$$

$$q_L = \sqrt{(1/3)[(v_{sa}-v_{sb})i_{Lc} + (v_{sb}-v_{sc})i_{La} + (v_{sc}-v_{sb})i_{Lb}]} = \overline{q_L} + \widetilde{q_L} \quad (19)$$

Fundamental power (active and reactive) component of the loads are extracted by means of LPFs from instantaneous powers.

Amplitude of active power fundamental component of load currents is estimated from average load power.

$$I_{Lp} = (2/3)(\frac{\overline{p_L}}{\mathcal{V}_{sp}}) \quad (20)$$

Moreover, amplitude of reactive power fundamental component of load currents is estimated from average load power:

$$I_{Lq} = (2/3)(\frac{\overline{q_L}}{\mathcal{V}_{sp}}) \quad (21)$$

Output of the DC voltage PI controller(i_{dloss}) of the proposed system is added with fundamental active power component of load currents (I_{Lp}) to estimate amplitude of the active power component of reference supply currents as given below:

$$I_{sp} = i_{dloss} + I_{Lp} \quad (22)$$

In-phase unit templates of respective phases is multiplied by amplitude component of the active power (I_{sp}) of reference supply current. The instantaneous fundamental active power component (in-phase) of reference supply currents($i^*_{spa}, i^*_{spb}i^*_{spc}$) are given as:

$$i^*_{sap} = I_{sp} * \mathcal{U}_{pa};\ i^*_{sbp} = I_{sp} * \mathcal{U}_{pb};\ i^*_{scp} = I_{sp} * \mathcal{U}_{pc} \quad (23)$$

Similarly, amplitude estimation of fundamental component of reactive power (I_{sq}) of reference supply currents is estimated is given in the eqn. below:

$$I_{sq} = q_{vr} - I_{Lq} \quad (24)$$

Further to get instantaneous values of the fundamental reactive power component of reference supply currents, amplitude of fundamental reactive component is multiplied by quadrature unit templates. Instantaneous quadrature reference current ($i^*_{sqa}, i^*_{sqb}, i^*_{sqc}$) are given as:

$$i^*_{saq} = I_{sq} * \mathcal{U}_{qa};\ i^*_{sbq} = I_{sq} * \mathcal{U}_{qb}; i^*_{scq} = I_{sq} * \mathcal{U}_{qc} \quad (25)$$

The adding in-phase and quadrature reference supply current will give the instantaneous fundamental reference supply currents as given:

$$i^*_{sa} = i^*_{sap} + i^*_{saq}, i^*_{sb} = i^*_{sbp} + i^*_{sbq}, i^*_{sc} = i^*_{scp} + i^*_{scq} \quad (26)$$

Further reference current ($i^*_{sa}, i^*_{sb}, i^*_{sc}$) are compared with sensed grid current (i_{sa}, i_{sb}, i_{sa}) in hysteresis control and generate switching signals to operate VSC.

E. Adaptive LMS Control algorithm

LMS control technique computes the reference currents to generate pulses for VSC conduction as demonstrated in Fig. 6. Control algorithm is described as follows:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:44 UTC from IEEE Xplore. Restrictions apply.

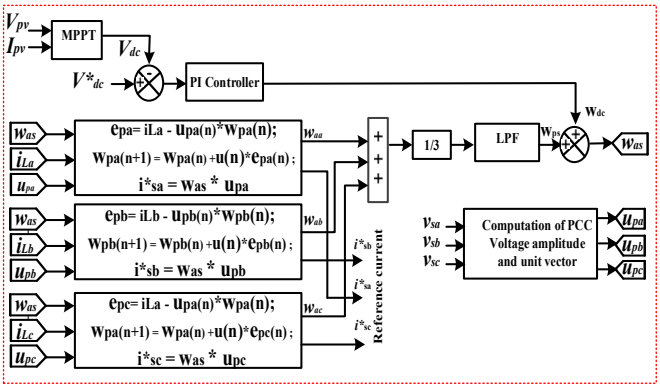


Fig. 6 Adaptive LMS Control algorithm

Estimation of in phase and unite templates of voltages

The peak amplitude of voltage ($\mathcal{V}_t$) of grid voltage, which can be calculated by:

$$\mathcal{V}_t = \sqrt{\frac{2}{3}(\mathcal{V}_{sa}^2 + \mathcal{V}_{sb}^2 + \mathcal{V}_{sc}^2)} \quad (27)$$

The in-phase unit voltages template ($\mathcal{U}_{pa}, \mathcal{U}_{pb}, \mathcal{U}_{pc}$) can be calculated by the relation of phase voltages ($\mathcal{V}_{sa}, \mathcal{V}_{sb}, \mathcal{V}_{sc}$) with peak amplitude ($\mathcal{V}_t$) voltages as follows [20]:

$$\mathcal{U}_{pa} = \frac{\mathcal{V}_{sa}}{\mathcal{V}_t}, \mathcal{U}_{pb} = \frac{\mathcal{V}_{sb}}{\mathcal{V}_t}, \mathcal{U}_{pc} = \frac{\mathcal{V}_{sc}}{\mathcal{V}_t} \quad (28)$$

Estimation of active weight using LMS control algorithm

Error in active component of load current in each phase ($e_{pa}; e_{pb}; e_{pc}$) at n_{th} instant is estimated as using in-phase unit template ($\mathcal{U}_{pa}, \mathcal{U}_{pb}, \mathcal{U}_{pc}$) and active load current component ($w_{pa}; w_{pc}; w_{pc}$)as follows:

$$e_{pa}(n) = i_{La}(n) - u_{pa} * w_{pa}(n) \quad (29)$$

$$e_{pb}(n) = i_{Lb}(n) - u_{pb} * w_{pb}(n) \quad (30)$$

$$e_{pc}(n) = i_{Lc}(n) - u_{pc} * w_{pc}(n) \quad (31)$$

The fundamental active weights w_{pa}, w_{pb}, w_{pc} at (n+1) instant be computed as:

$$w_{pa}(n+1) = w_{pa}(n) + \mu(n) * e_{pa}(n) \quad (32)$$

$$w_{pb}(n+1) = w_{pb}(n) + \mu(n) * e_{pb}(n) \quad (33)$$

$$w_{pc}(n+1) = w_{pc}(n) + \mu(n) * e_{pc}(n) \quad (34)$$

The average fundamental active weight components can be computed as:

$$w_{lp} = \frac{(w_{pa}+w_{pb}+w_{pc})}{3} \quad (35)$$

The sensed voltage of dc-link (V_{dc}) is compared with reference dc bus voltage (V_{dc}^*) to estimate error in DC link voltage and compensated using proportional integral (PI) controller. The output of the controller is dc loss weight (w_{dc}).

The total active weight (w_{ps}) component of the supply reference current is given by:

$$w_{ps} = w_{lp} + w_{dc} \quad (36)$$

The active in-phase reference supply current can be evaluated as:

$$i_{sa}^* = w_{ps}\mathcal{U}_{pa}, i_{sb}^* = w_{ps}\mathcal{U}_{pb}, i_{sc}^* = w_{ps}\mathcal{U}_{pc} \quad (37)$$

Generation of switching signal for VSC

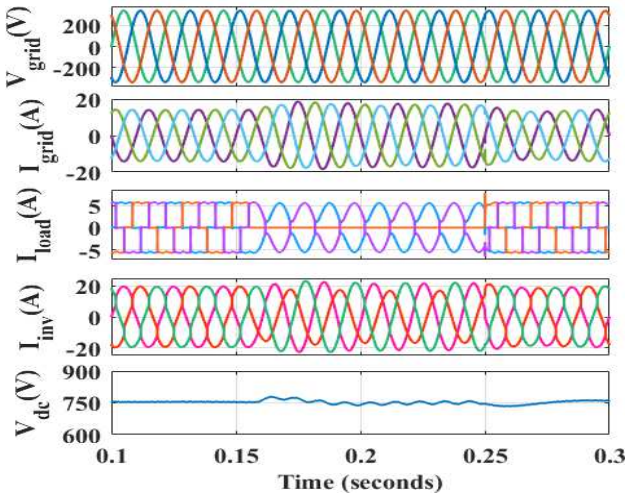
Reference current ($i_{sa}^*, i_{sb}^*, i_{sc}^*$) are compared with sensed grid current (i_{sa}, i_{sb}, i_{sa}) in hysteresis control and generate switching signals to operate VSC.

IV. RESULTS AND DISCUSSIONS

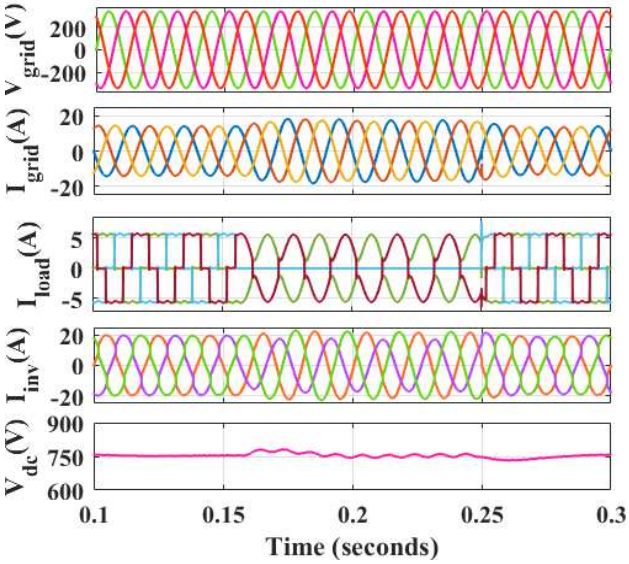
To analyze the efficacy of various control algorithms under nonlinear load, a bridge rectifier has been considered with RL load (R=100 Ω, L=100 mH). Further, to create the unbalance in the load, one phase is kept open from 0.15 to 0.25 seconds. Proposed system is supposed to be working at standard test condition (1000w/m^2, 25^{0}C). Proposed system is conducted in PFC mode of operation.

A. Performance of Various Control algorithm

Simulation results using various control algorithms under nonlinear(balance/unbalance) load are given in the Fig.7. To impose the unbalance between 0.15 to 0.25 second one phase is kept open. It can be seen from Fig. 7 that all the algorithms are efficient in maintaining the grid current sinusoidal and balanced for the unbalance load current between 0.15 to 0.25 sec. Fig 8 shows the grid current waveform and THD. It can be seen from Fig 8 (a) and table I, that load current THD is 29.28% while grid current THD is well within the IEEE standard-519.

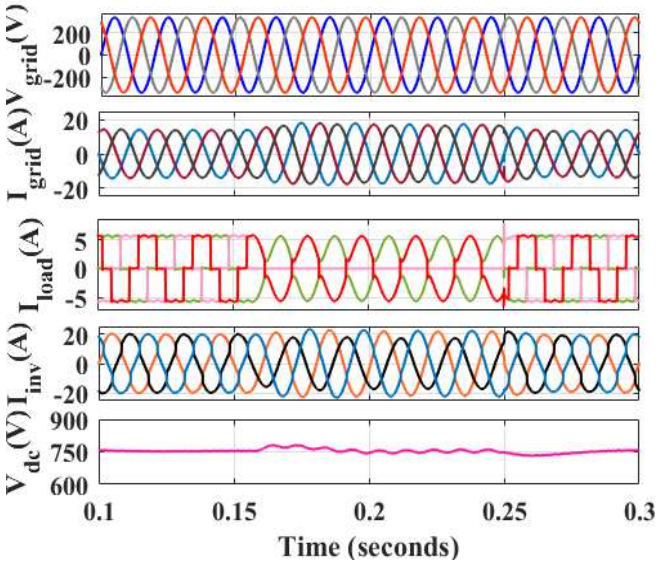


(a) Grid integrated microgrid under SRF control

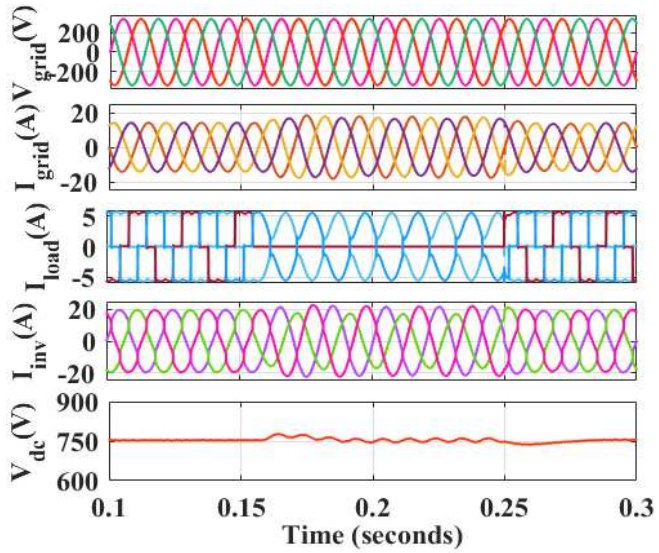


(b) grid integrated microgrid under IRPT control

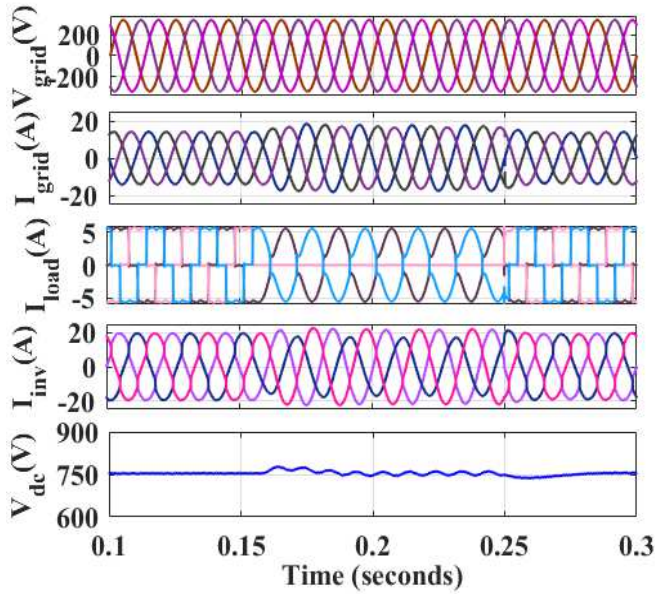
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:44 UTC from IEEE Xplore. Restrictions apply.



(c) Grid integrated microgrid under unit template control



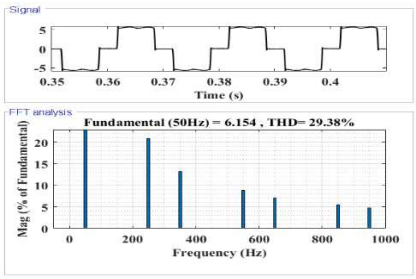
(d) Grid integrated microgrid under PBT control



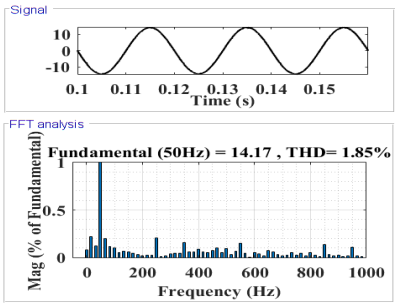
(e) Grid integrated microgrid under adaptive LMS control

Fig. 7(a), (b), (c), (d) and (e)

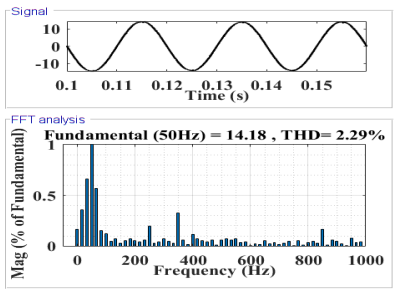
B. Grid curent waveform and THD under nonlinear load



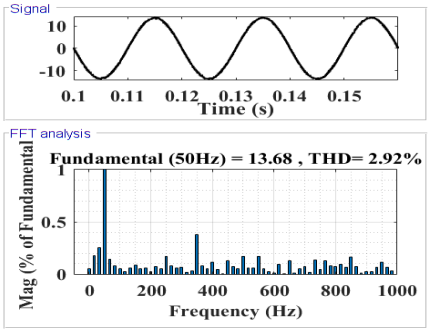
(a) Grid current waveform and THD before compensation



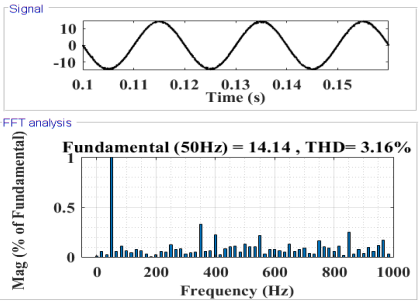
(b) Grid current waveform and THD after compensation using SRF control



(c) Grid current waveform and THD after compensation using IRPT control

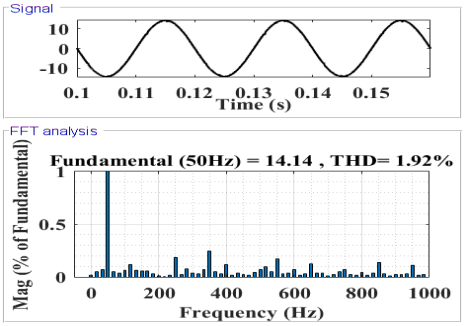


(d) Grid current waveform and THD after compensation using unit template control



(e) Grid current waveform and THD after compensation using PBT control

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:44 UTC from IEEE Xplore. Restrictions apply.



(f) Grid current waveform and THD after compensation using LMS control

Fig.8(a), (b), (c), (d), (e) and (f)

It can be seen from table I, that harmonics in grid current is minimum in case of SRF, while adaptive LMS provides slightly higher than SRF and less than other algorithm. Implementation of adaptive LMS is simplest than other, but advance type of adaptive LMS control can offer the best performance towards harmonics mitigation.

TABLE I.

S.N.	THD in grid current	
	Control algorithms	***THD***
1	SRF	1.85
2	IRPT	2.29
3	Unit template	2.92
4	PBT	3.16
5	LMS	1.92

Conclusion

In this paper, different conventional and adaptive LMS control algorithms have been implemented for controlling the PV base microgrid. The performance of the control algorithm is tested under nonlinear (balanced/unbalanced) load at STC. It has been observed that under different control, of VSC, microgrid operation is efficient and THD in grid current is maintained within limit along with UPF operation. Comparative analysis of different control algorithm under PV grid integration depicts that harmonics in grid current is minimum i.e. 1.85% in case of SRF while adaptive LMS offers minimum complexity in implementation and provides almost similar 1.92% THD in the grid current.

REFERENCES

[1] N. Beniwal, I. Hussain, and B. Singh, "A second-order volterra filter based control of SPV-DSTATCOM system to achieve Lyapunov's stability," *2016 IEEE 7th Power India Int. Conf.*, pp. 1–5, 2016, doi: 10.1109/POWERI.2016.8077292.

[2] B. Singh, P. Jayaprakash, D. P. Kothari, A. Chandra, and K. Al Haddad, "Comprehensive study of dstatcom configurations," *IEEE Trans. Ind. Informatics*, vol. 10, no. 2, pp. 854–870, 2014, doi: 10.1109/TII.2014.2308437.

[3] T. A. Youssef and O. Mohammed, "Adaptive SRF-PLL with reconfigurable controller for Microgrid in grid-connected and stand-alone modes," IEEE Power Energy Soc. Gen. Meet., pp. 1–5, 2013, doi: 10.1109/PESMG.2013.6673028.

[4] S. R. Arya, B. Singh, R. Niwas, A. Chandra, and K. Al-Haddad, "Power Quality Enhancement Using DSTATCOM in Distributed Power Generation System," *IEEE Trans. Ind. Appl.*, vol. 52, no. 6, pp. 5203–5212, 2016, doi: 10.1109/TIA.2016.2600644.

[5] B. Singh, A. Chandra, and K. Al-haddad, "Power Quality Problems and Mitigation Techniques" Wiley Online Library, vol. 9781118922. 2015.

[6] S. K. Patel, S. R. Arya, and R. Maurya, "Optimal Step LMS-Based Control Algorithm for DSTATCOM in Distribution System," *Electr. Power Components Syst.*, vol. 47, no. 8, pp. 704–720, 2019, doi: 10.1080/15325008.2019.1602797.

[7] S. Kumar, A. K. Verma, I. Hussain, and B. Singh, "Performance of grid interfaced solar PV system under variable solar intensity," *India Int. Conf. Power Electron. IICPE*, vol. 2015-May, pp. 1–6, 2015, doi: 10.1109/IICPE.2014.7115797.

[8] N. Beniwal, I. Hussain, and B. Singh, "Implementation of the DSTATCOM with an i-PNLMS-Based Control Algorithm under Abnormal Grid Conditions," *IEEE Trans. Ind. Appl.*, vol. 54, no. 6, pp. 5640–5648, 2018, doi: 10.1109/TIA.2018.2846739.

[9] N. Beniwal, I. Hussain, and B. Singh, "Second-order volterra-filter-based control of a solar PV-DSTATCOM system to achieve lyapunov's stability," *IEEE Trans. Ind. Appl.*, vol. 55, no. 1, pp. 670–679, 2019, doi: 10.1109/TIA.2018.2867324.

[10] Z. Li, D. Li, X. Xu, and J. Zhang, "New normalized LMS adaptive filter with a variable regularization factor," *J. Syst. Eng. Electron.*, vol. 30, no. 2, pp. 259–269, 2019, doi: 10.21629/JSEE.2019.02.05.

[11] B. Singh, A. Chandra, and K. Al-haddad, "*Power Quality Problems and Mitigation Techniques" Wiley Online Library*, vol. 9781118922. 2015.

[12] S. R. Arya, B. Singh, A. Chandra, and K. Ai-Haddad, "Control of DSTATCOM using adjustable step least mean square control algorithm," *2012 IEEE 5th Power India Conf. PICONF 2012*, vol. 2, no. 1, pp. 1–6, 2012, doi: 10.1109/PowerI.2012.6479508.

[13] R. Garg and P. Mahajan, "Performance Analysis of Grid Integrated PV System using SRF and IRPT Control", 1st International Conference on Signal Processing, VLSI and Communication Engineering (ICSPVCE), 28-30 March 2019, DOI: 10.1109/ICSPVCE46182.2019.9092869.

[14] B. Singh, D. T. Shahani, and A. K. Verma, "Power balance theory based control of grid interfaced solar photovoltaic power generating system with improved power quality," *PEDES 2012 - IEEE Int. Conf. Power Electron. Drives Energy Syst.*, 2012, doi: 10.1109/PEDES.2012.6484359.

[15] P. K. Barik, G. Shankar, and P. K. Sahoo, "Power quality assessment of microgrid using fuzzy controller aided modified SRF based designed SAPF," *Int. Trans. Electr. Energy Syst.*, 2019, doi: 10.1002/2050-7038.12289.

[16] M. Esmaeili, H. Shayeghi, K. Valipour, A. Safari, and F. Sedaghati, "Power quality improvement of multimicrogrid using improved custom power device called as distributed power condition controller," *Int. Trans. Electr. Energy Syst.*, vol. 30, no. 3, pp. 1–16, 2020, doi: 10.1002/2050-7038.12259.

[17] P. Verma, R. Garg, and P. Mahajan, "Asymmetrical interval type-2 fuzzy logic control based MPPT tuning for PV system under partial shading condition," *ISA Trans.*, no. January, 2020, doi: 10.1016/j.isatra.2020.01.009.

[18] S. K. Kollimalla and M. K. Mishra, "Variable perturbation size adaptive P&O MPPT algorithm for sudden changes in irradiance," *IEEE Trans. Sustain. Energy*, vol. 5, no. 3, pp. 718–728, 2014, doi: 10.1109/TSTE.2014.2300162.

[19] J. Mishra, "Performance Comparison of P & O and INC MPPT Algorithm for a Stand-alone PV System," no. 1, pp. 1–5, 2019.

[20] N. Femia, G. Petrone, G. Spagnuolo, and M. Vitelli, "A technique for improving P&O MPPT performances of double-stage grid-connected photovoltaic systems," *IEEE Trans. Ind. Electron.*, vol. 56, no. 11, pp. 4473–4482, 2009, doi: 10.1109/TIE.2009.2029589.

Appendix

Simulation Parameters of the system: Grid voltages: 415 V, frequency (f): 50 Hz, Line impedances (R_s, L_s): 0.01Ω, 0.1mH, Interfacing inductor (Lf): 7mH, $V_{dc(ref.)}$: 750 V; sampling time T_s = 5.5μs

Solar PV parameters: nominal power 10.25 kW, nominal voltage at MPP: 410V, nominal current: 25 A,

DC-DC (boost) converter parameters**:** Duty cycle (D): 0.43, Switching frequency (f_s): 10 kHz**,** Inductor (L): 0.5mH, Capacitor (C): 1000μF,

Nonlinear load: bridge rectifier R=100Ω, L=100mH (unbalancing imposed between 0.15 to 0.25sec)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:44 UTC from IEEE Xplore. Restrictions apply.

Hybrid Dynamic CML with Modified Current Source (H-MDyCML): A Low-Power Dynamic MCML Style

Riya JAIN[1], *Kirti GUPTA*[2], *Neeta PANDEY*[1]

[1]Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, 110042 Delhi, India
[2]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, Paschim Vihar, New Rohtak Road, 110063 Delhi, India

rj.riyajain26@gmail.com, kirtigupta22@gmail.com, n66pandey@rediffmail.com

DOI: 10.15598/aeee.v19i1.3944

Article history: Received Sep 15, 2020; Revised Dec 24, 2020; Accepted Feb 10, 2021; Published Mar 31, 2021.
This is an open access article under the BY-CC license.

Abstract. *With the growing demands of portable devices, it is necessary to pay attention to low-power digital integrated designs. This paper proposes a low-power MOS Current Mode Logic (MCML) design, termed as Hybrid Dynamic Current Mode Logic with modified current source (H-MDyCML). In H-MDyCML circuits, the functions are realized using complementary pass transistor logic which helps to overcome the problem of stacking of transistors in multiple levels. The dynamic current source has also been modified from an NMOS transistor to a PMOS transistor-driven current source which leads to the elimination of the use of CMOS inverter. H-MDyCML circuits are compared with other existing designing styles: Dynamic CML (DyCML), Hybrid DyCML (H-DyCML), and DyCML with modified current source (Dy-CML-NP). The proposed design (H-MDyCML) shows an overall improvement (in terms of Power Delay Product (PDP)) up to 94.77 % compared to DyCML, 52.17 % compared to Dy-CML-NP, and 91.40 % compared to H-DyCML, for single stage circuits.*

Keywords

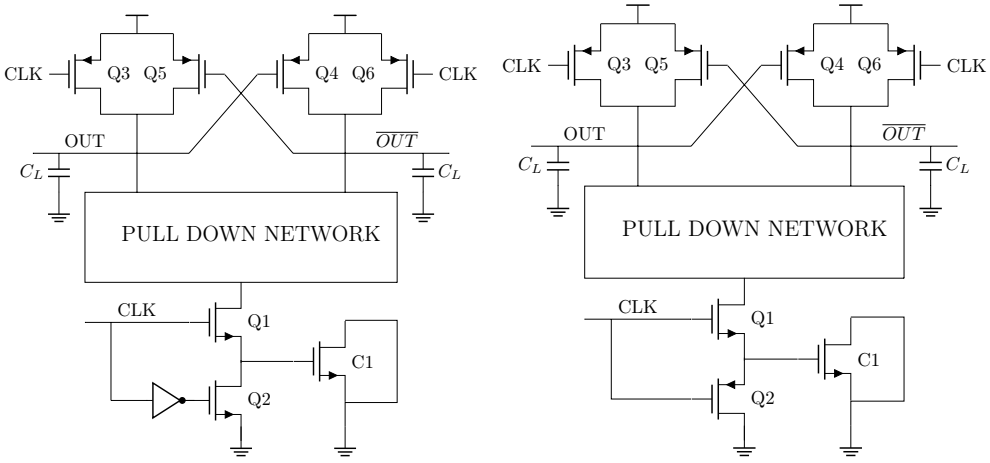
Complementary pass transistor logic, current mode logic, current source, low-power.

1. Introduction

The growth of digital integrated circuits has escalated rapidly in the recent past because of easily accessible Electronic Design Automation (EDA) tools and evolution in technology. This led to a drastic increase in demand for portable electronic goods. One of the major limitations faced by engineers while designing and developing such devices is battery life. Therefore, power dissipation has emerged as an important issue that results in the degradation of the performance of the overall system. To reduce power consumption, researchers are getting inclined towards low-power designs [1], [2], [3], [4] and [5].

Conventional CMOS logic is quite prominently used in digital design due to certain advantages such as high packaging density, negligible static power consumption, and well tested and documented design methodologies. However, while switching, digital circuits based on CMOS logic poses high dynamic power consumption. Switching power dominates the power dissipation in most digital systems which tends to increase at higher operating frequencies drastically. Therefore, there was a shift from CMOS-based digital designs to MCML-based digital designs [6], [7] and [8]. MCML designs have been proved to be a good replacement of CMOS-based designs due to reduced power consumption at high frequencies. However, because of the presence of a static current source, the circuits had static power dissipation accompanied by increased design complexity.

Therefore, there was a further shift from MCML-based circuits to DyCML-based circuits [9] and [10]. This is a newly proposed logic family in the literature that uses a dynamic current source in place of a static current source and works on the pre-charge and evaluation method [11] and [12]. Thus, DyCML circuits help to achieve high performance at lower power dissipation.



(a) Dynamic Current Mode Logic (DyCML). (b) DyCML with modified current mode logic (Dy-CML-NP).

Fig. 1: Basic Architecture of (a) DyCML style [10] and (b) DyCML style with modified current source (Dy-CML-NP) [15].

Nonetheless, the realization of functions with multiple inputs requires stacking of transistors which leads to a significant rise in delay. To overcome this problem, a new method, known as H-DyCML, was proposed recently [13] to avoid stacking of the transistors and hence reduce the delay. This method is based on Complementary Pass transistor Logic (CPL) to implement logic functions. However, there is still significantly high power consumption due to the use of an additional CMOS inverter which is required to obtain an inverted Clock (CLK) signal to drive the NMOS transistor of dynamic current source. Therefore, a method to eliminate the use of a CMOS inverter, namely Dy-CML-NP, has been recently proposed [14] and [15] where NMOS transistor in dynamic current source of a DyCML circuits has been replaced by a PMOS transistor which is driven by CLK signal instead of an inverted CLK signal. However, Dy-CML-NP also poses a limitation due to stacking of transistors. Therefore, this paper proposes a new method that overcomes the limitation of H-DyCML and Dy-CML-NP, and exploits the advantages of both methods by combining them.

The remaining article has been organized as follows: Sec. 2. traces the background of research emphasizing on importance, working, advantages and disadvantages of circuits designed using DyCML, H-DyCML, and Dy-CML-NP. Further, Sec. 3. introduces the proposed low-power design. Cascading of proposed H-MDyCML has been explained in Sec. 4. and Sec. 5. presents the simulation results and comparison of the performance of various other designing styles with proposed style. Then finally, the paper is concluded in Sec. 6.

2. Background

In MCML-based circuits, the use of a constant current source and reduced voltage swing results in lower dynamic power consumption but also brings with it certain drawbacks which limit its usage. Static power dissipation is the major drawback faced by engineers in designing systems using MCML logic. Therefore, to overcome this issue, there was a shift from MCML-based logic to DyCML-based logic [10] wherein a dynamic current source was used. A typical architecture of a DyCML circuit is shown in Fig. 1(a). DyCML circuits follow the principle of pre-charge and evaluation which exists in dynamic logic circuits.

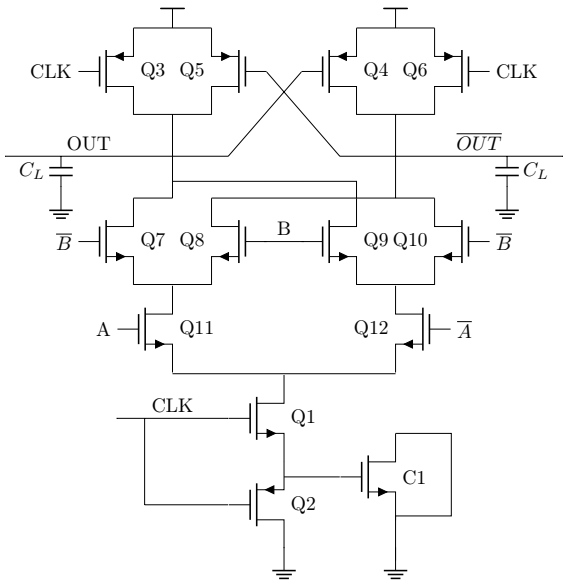


Fig. 2: Realization of XOR gate using DyCML with modified current source (Dy-CML-NP) [15].

One of the major changes made from MCML circuits was that the pull-up network was changed to active loads (Q3 and Q6) connected along with cross-coupled latches (Q4 and Q5) instead of load resistors only. Another major change made from MCML circuits was the replacement of static current source to dynamic current source built using NMOS (Q2) driven by inverted CLK signal. However, the logic function was evaluated similarly to that in MCML: using the 'Pull Down Network' block.

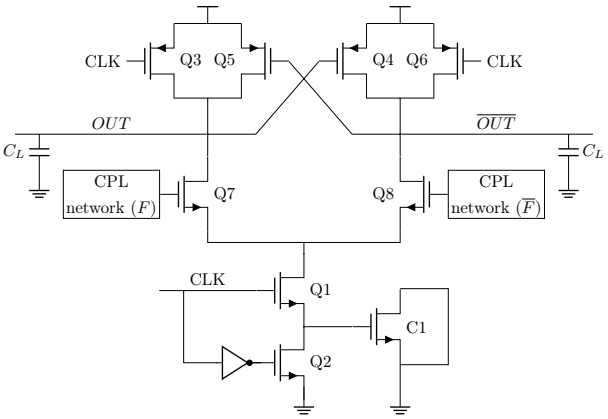


Fig. 3: Basic architecture of H-DyCML style [13].

Due to the presence of a dynamic current source in DyCML circuits, power was saved significantly along with increased performance as compared to MCML-based logic gates. However, the use of a CMOS inverter to obtain an inverted CLK signal that drives the NMOS transistor in the dynamic current source still contributed to high dynamic power dissipation. Therefore, to address this issue, a new method (Dy-CML-NP) was recently introduced in [15] which suggests to modify the dynamic current source in DyCML circuits by replacing the NMOS transistor with a PMOS transistor. Basic architecture of Dy-CML-NP is shown in Fig. 1(b), and it is evident that the requirement of the inverted CLK signal has been withdrawn. Although the issue of power dissipation has been addressed in Dy-CML-NP designing style, yet the delay remains critical due to multiple levels of stacking of transistors in the pull-down network which additionally increases the delay of a gate. This could be better understood by the design of 2-input XOR gate shown in Fig. 2. There are two levels of source-coupled transistors such that each input and its complement is applied to their gates. Therefore, the problem of stacking persists for both DyCML and Dy-CML-NP circuits. To overcome the problem of stacking of transistors in multiple levels, a new method was recently proposed in [13]. This method of designing, termed as H-DyCML, made use of complementary pass transistor logic to solve the issue. The working of H-DyCML-based logic can be explained with the help of Fig. 3. It also follows the pre-charge and evaluation method in a similar manner as in DyCML-based logic. The difference lies in the way a function was realized. The pull-down network in H-DyCML circuits consists of one level source-coupled NMOS transistor only such that their gate terminals are driven by the output of a CPL logic which is built using NMOS transistors.

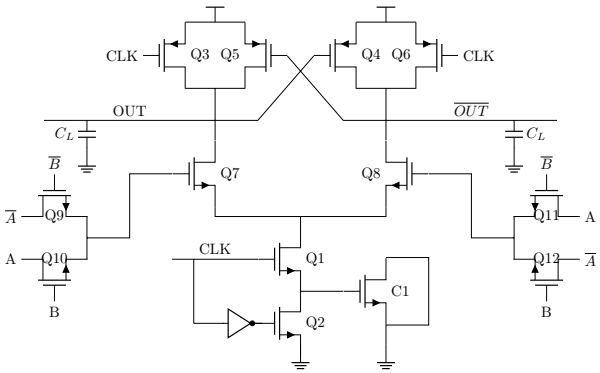


Fig. 4: Realization of XOR gate using H-DyCML style [13].

This is better enlightened in Fig. 4 which realizes 2-input XOR gate using H-DyCML style. Transistors Q9 and Q10 implement the function $\overline{A \oplus B}$ and Q11 and Q12 implement the function $A \oplus B$. If the output of the implemented logic $A \oplus B$ is HIGH, then Q7 is turned OFF and Q8 is turned ON. From this reason, OUT terminal is directly connected to the ground and OUT stays at V_{DD}. Since, NMOS transistors offer weak pull-up and strong pull-down, the maximum voltage level at the output of NMOS pass transistor (V_{OH}) is equal to gate voltage (V_{DD}) reduced by one threshold voltage (V_{tn}), that is, $V_{OH} = V_{DD} - V_{tn}$. Therefore, to avoid erroneous results, three techniques have been proposed in [13]:

- Use of level restorer circuit.
- Use of Multiple threshold voltage transistors.
- Use of transmission gates.

3. Proposed H-DyCML Circuits with Modified Current Source (H-MDyCML)

In this section, the proposed structure (H-MDyCML) for low-power, high-speed designing has been elaborated which jointly exploits the advantages of both H-DyCML and Dy-CML-NP designing styles. The proposed designing style has shown improvement in terms of:

- Delay when compared to Dy-CML-NP and DyCML.
- Power when compared to DyCML and H-DyCML.
- PDP when compared to DyCML, Dy-CML-NP, and H-DyCML.

Above improvements are supported by proper results and explanations in rest of the paper. Basic architecture of H-MDyCML circuits is shown in Fig. 5 and its working can be understood as explained further. The operation of the circuit still follows the pre-charge and evaluation mechanism as earlier. In the pre-charge phase, when CLK = LOW, transistors Q3, Q6, and Q2 are switched ON. Therefore, the load capacitors are charged to V_{OH}=V_{DD} through active loads (Q3 and Q6), and capacitor C1 is discharged through Q2. However, one thing to be noted here is that PMOS has a weak pull-down property which is why the capacitor discharges up to a level $V = V_{tp}$ because beyond that PMOS is switched OFF. Thus, capacitor C1 does not discharge to ground potential, but to one threshold voltage level above the ground potential. During the pre-charge phase, the output levels remain at V_{OH} and are unaffected by the inputs applied to the CPL logic because Q1 is switched OFF which disconnects the output nodes from capacitor C1. At this time, cross-coupled latches (Q4 and Q5) are also disconnected.

In the evaluation phase, when CLK = HIGH, transistors Q3, Q6, and Q2 are switched OFF. Q1 is switched ON and offers a direct path between output nodes and capacitor C1 which acts as a virtual short. However, C1 already contains some charge on it due to the weak pull-down nature of PMOS, as is explained above. During the evaluation phase, the values of inputs applied to CPL logic are evaluated. Since the two CPL functions implemented are the complement of each other, therefore, when the output of one CPL logic is HIGH, the output of another CPL logic is LOW which leads to discharging of one output node to $V_{DD} - V_{swing}$ and other remaining at V_{DD}. The output nodes are maintained at these levels with the help of latches built using cross-coupled PMOS transistors (Q4 and Q5). At the same time, the voltage across the capacitor goes from V_{tp} to $V_{DD} - V_{swing}$. The value of capacitor C1 is decided by the equation, as given in [15]:

$$C1 = \frac{V_{swing}}{V_{DD} - V_{swing} - |V_{tp}|} \cdot C_L. \qquad (1)$$

The reason for the subtraction of $|V_{tp}|$ in the denominator has been explained above; capacitor C1 is already charged to a potential of $|V_{tp}|$ after the pre-charged phase. For better understanding, 2-input XOR gate has been realized using H-MDyCML style as shown in Fig. 6.

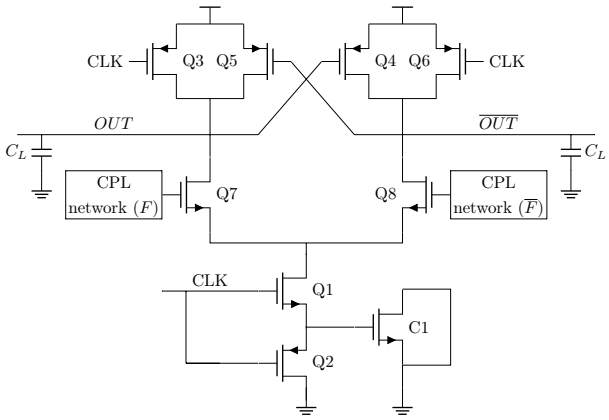


Fig. 5: Basic architecture of H-MDyCML style.

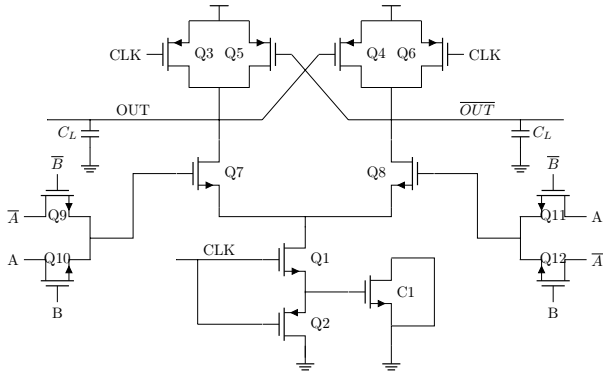


Fig. 6: Realization of XOR gate using H-MDyCML style.

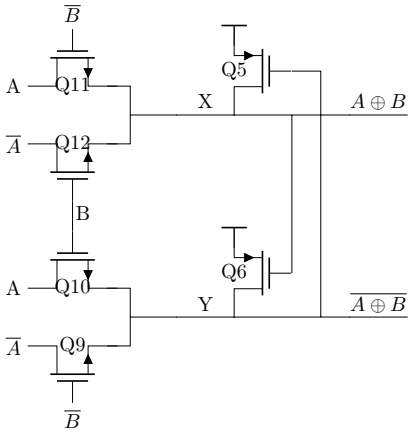


Fig. 7: Technique of level restoring [11] and [13].

Out of 3 techniques proposed in [13] to maintain the output of CPL logic at V_{DD} to avoid erroneous results, we choose the first technique of level restorer. It employs two cross-coupled PMOS transistors as shown in Fig. 7 which act as weak pull-up transistors. It implements XOR/XNOR functions. If applied inputs are such that the XOR function is set HIGH and XNOR is LOW, then Q5 will be turned ON, pulling up node

X and Q6 will be turned OFF, maintaining node Y at low potential. The reason for choosing this technique is the ease of choosing the voltage swing, that is, we can adjust the swing such that it is greater than the threshold voltage. However, dealing with other techniques increases the complexity and number of transistors utilized. Therefore, the technique of level restorer allows usage of a lesser number of transistors as compared to transmission gate logic without putting any constraint on threshold voltages of CPL logic transistors or level restorer transistors. Thus, the H-MDyCML style presents a good solution for low-power designs as it not only deals with the problem of increased delay due to stacking of transistors, but also helps remove CMOS inverter which is responsible for high dynamic power consumption.

4. Cascading of H-MDyCML Circuits

Simulation of individual gates is relatively simpler and the problem arises when two or more gates need to be cascaded together. The problem that arises while cascading is that the evaluation phase of consecutive stages occurs simultaneously. This results in erroneous results because the second stage starts evaluation before the first stage has finished evaluating inputs. Therefore, a certain delay must be incorporated after the first stage and before the second stage to separate the evaluation phase of two consecutive stages. To accomplish this task, additional circuitry, which act as a bridge, is added between the two stages. The circuitry has been proposed in [15] and looks as shown in Fig. 8 where EOE is the voltage drop across capacitor C1 and CLK is the clock signal of the previous stage. It is the Self-Timed Buffer (STB) similar to that proposed in [10], with a slight modification. Working of the shown STB is explained further.

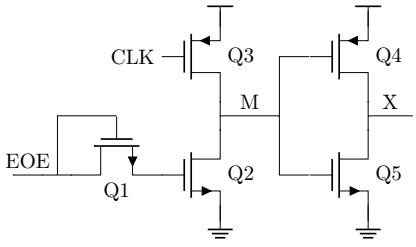


Fig. 8: Self-Timed Buffer [15].

In H-MDyCML circuits, in pre-charge phase, capacitor C1 is discharged to V_{tp} (when CLK = LOW). If we directly apply this signal to Q2, it will switch ON the transistor because $V_{tp} > V_{tn}$. Therefore, an additional NMOS is needed (Q1) which leads to one threshold voltage drop and thus switching OFF Q2. At the same time, Q3 is switched ON charging the node M to V_{DD}. This furthers activates Q5 and deactivates Q4 such that node X is subsequently pulled down to the ground. Similarly, in the evaluation phase, the total charge carried by load capacitors (C_L) and capacitor C1 is redistributed such that C1 is charged up to voltage equal to $V_{DD} - V_{swing}$ from V_{tp} (when CLK = HIGH). At this time, transistor Q3 is turned OFF while Q1 and consequently Q2 is turned ON. This pulls down the potential of point M to ground and subsequently, Q4 is turned ON while switching OFF Q5. Node X is charged up to V_{DD}. Therefore, CLK signal is transmitted to point X with some delay. One thing to be noted here is that this STB also eliminates the use of CMOS inverter to obtain the complement of CLK signal. This results in further reduction in dynamic power dissipation. An example of cascading is shown in Fig. 9. It shows a full adder in which two STBs are used. The first STB is used to cascade the first stage XOR gate to the second stage XOR and AND gates. The second STB is used to cascade the output of first and second stage AND gates to OR gate.

5. Simulation Results

Various gates designed using H-MDyCML style are simulated using Cadence Virtuoso. All the circuits are designed using 180 nm CMOS technology parameters. In order to avoid erroneous results at the output of CPL logic, the voltage swing needs to be kept greater than the threshold voltage. Therefore, V_{swing} is chosen to be 700 mV. The supply voltage is chosen as 1.8 V and since V_{swing} is 700 mV, $V_{DD} - V_{swing}$ becomes 1.1 V. Apart from this, C_L is taken to be 40 fF. Using Eq. (1), C1 comes out to be approximately 47 fF. For a fair comparison, various gates are also implemented using designing styles: DyCML, H-DyCML, and Dy-CML-NP in parallel with H-MDyCML. Various circuits used for comparison are: AND, OR, and XOR gates. For all the circuits, the parameters mentioned above are kept the same. However, the value of C1 differs in the case of DyCML and H-DyCML circuits since they use a current source driven by NMOS. The value of capacitor C1 for designing styles with NMOS driven current is source is given by [10]:

$$C1 = \frac{V_{swing}}{V_{DD} - V_{swing}} \cdot C_L. \tag{2}$$

According to Eq. (2) the value of capacitor C1 comes out to be approximately 26 fF for DyCML and H-DyCML. Different parameters compared for all the circuits are: Delay, Power, and PDP whose analysis for AND, OR, and XOR gates is summarized in Tab. 1.

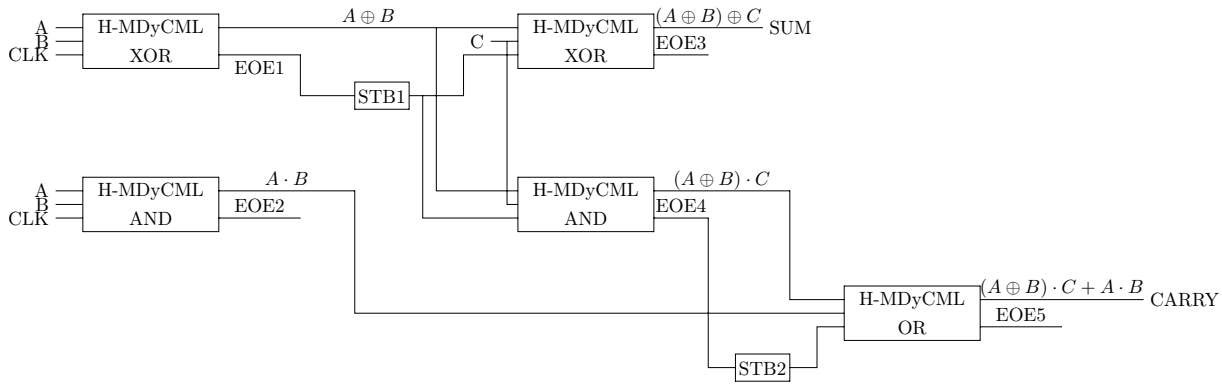


Fig. 9: Cascading required in Full-Adder circuit.

Tab. 1: Performance comparison.

Parameter	Style			
	DyCML	Dy-CML-NP	H-DyCML	H-MDyCML
AND gate				
Power (μW)	2.061	0.117	2.051	0.129
Delay (ps)	287.78	590.80	187.78	254.01
PDP (fJ)	0.593	0.069	0.384	0.033
OR gate				
Power (μW)	2.175	0.119	2.046	0.207
Delay (ps)	276.60	552.87	187.78	248.82
PDP (fJ)	0.602	0.066	0.384	0.052
XOR gate				
Power (μW)	2.073	0.111	2.047	0.129
Delay (ps)	304.92	591.10	182.83	253.09
PDP (fJ)	0.632	0.066	0.374	0.033
Full Adder				
Power (μW)	39.660	36.033	47.701	38.352
Delay (ps)	2646.39	2943.61	2615.65	2150.88
PDP (fJ)	104.956	106.068	124.769	82.492

Following observations can be made from the Tab. 1:

- H-MDyCML gates have less power consumption as compared to DyCML and H-DyCML gates due to the elimination of CMOS inverter. However, power consumption in H-MDyCML is greater than Dy-CML-NP due to the use of additional power supply in level storing circuit of CPL network.
- H-MDyCML gates have reduced delay as compared to DyCML and Dy-CML-NP gates due to replacement of stack of transistors with a CPL network to realize the logic. But, delay of H-MDyCML is greater than that of H-DyCML due to higher value of capacitor C1 in case of H-MDyCML, calculated using Eq. (1).
- PDP is the lowest for the proposed method due to combined effect of power and delay, which means that overall performance of the proposed designing style is better than others.

Apart from simulating individual gates, a full adder is designed, as shown in Fig. 9, and compared to show the effects of cascading on the performance of designs using different designing styles. The results for a full added for the same parameters as in the previous case are also recorded in Tab. 1 where it can be seen that the delay of the proposed H-MDyCML design is less than DyCML and Dy-CML-NP designs, and the power is less than DyCML and H-DyCML designs. Moreover, its PDP is the least. Therefore, it can be concluded that for single stage circuits, H-MDyCML shows an overall improvement (in terms of PDP) up to 94.77 %, 52.17 % and 91.40 % compared to DyCML, Dy-CML-NP and H-DyCML respectively. Also, H-MDyCML style shows an average improvement of 25.83 % from other designing styles in case of Full-Adder circuit. Decreased improvement with increasing stages in circuit is due to power and delay introduced by STBs.

Furthermore, the variation of performance of various designing styles has been observed with increasing voltage swing. Figure 10, Fig. 11 and Fig. 12 summarize the results for variations in delay, power, and PDP respectively with changing voltage swing for different designing styles. To analyze these variations, a 2-input

XOR gate is considered. Few things can be observed from the graphs in Fig. 10, Fig. 11 and Fig. 12:

- Delay shows an upward trend with increasing voltage swing (V_{swing}) for all designs because more charge needs to be transferred from load capacitors (C_L). The increase in delay across the domain of V_{swing} is significantly low, up to an average of 16.16 ps, which is evident from Fig. 10.

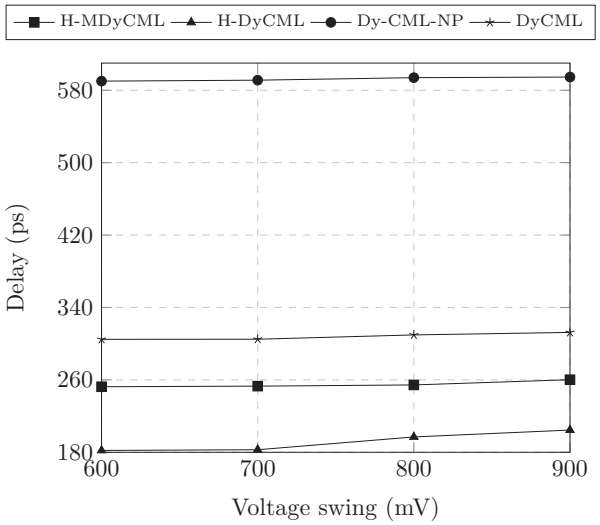


Fig. 10: Delay of a 2-input XOR gate realized using different styles with varying voltage swing (V_{swing}).

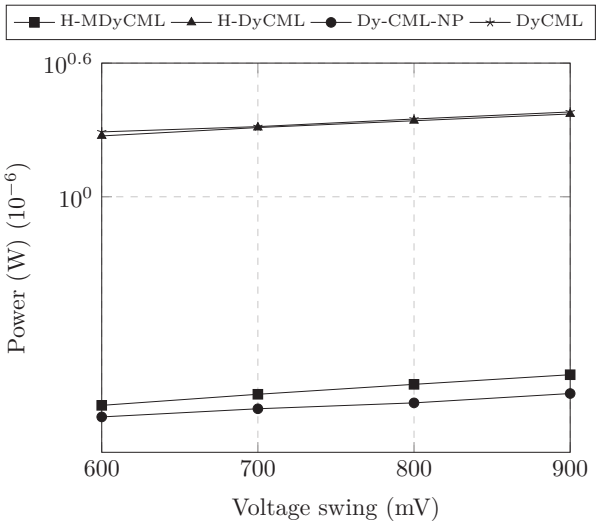


Fig. 11: Power of a 2-input XOR gate realized using different styles with varying voltage swing (V_{swing}).

- Power dissipation increases with increasing voltage swing for all designs because it is directly proportional to the V_{swing}.
- PDP increases with increase in voltage swing in case of all four designing styles due to the increase in power and delay. Moreover, PDP is the least for H-MDyCML style due to combined effect of improved power and delay.
- Difference between power dissipation of H-DyCML (or DyCML) and H-MDyCML increases with increasing voltage swing. This happens because power in H-DyCML and H-MDyCML circuits are of order 10^{-6} and 10^{-7} respectively. Since power is directly proportional to V_{swing}, absolute increase in power for H-DyCML is more than H-MDyCML. Similar trend is observed for PDP as it is directly proportional to power. Thus, improvement becomes more significant as V_{swing} is increased.

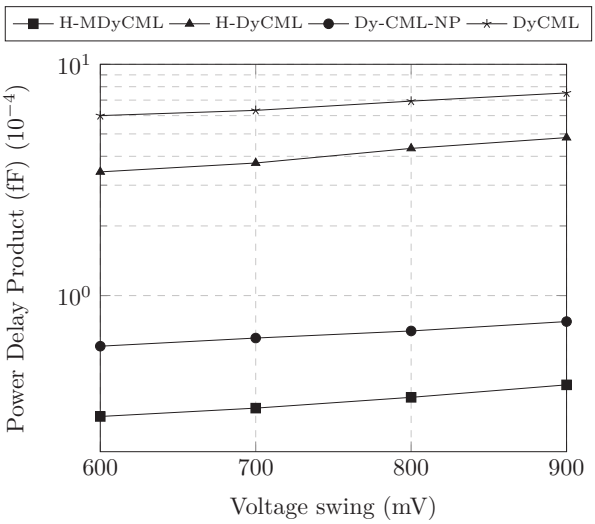


Fig. 12: PDP of a 2-input XOR gate realized using different styles with varying voltage swing (V_{swing}).

6. Conclusion

In this paper, a new method is proposed to design dynamic low-power, high-speed circuits. The new method combines the structure of H-DyCML and Dy-CML-NP to have reduced power consumption and delay. The reduced power consumption is possible due to the elimination of a CMOS inverter by using a dynamic current source built using a PMOS transistor as proposed in Dy-CML-NP circuits. The delay has been reduced due to use of the complementary pass transistor logic instead of a stack of transistors. Different gates are designed using DyCML, H-DyCML, Dy-CML-NP, and H-MDyCML styles and compared with each other in terms of power dissipation, delay, and power delay product. Therefore, it will be right to conclude that the circuits designed using the proposed method have better overall performance as compared to other methods.

References

[1] DAVARI, B., R. H. DENNARD and G. G. SHAHIDI. CMOS scaling for high performance and low power-the next ten years. *Proceedings of the IEEE*. 1995, vol. 83, iss. 4, pp. 595–606. ISSN 1558-2256. DOI: 10.1109/5.371968.

[2] CHANDRAKASAN, A. P., S. SHENG and R. W. BRODERSEN. Low-power CMOS digital design. *IEEE Journal of Solid-State Circuits*. 1992, vol. 27, iss. 4, pp. 473–484. ISSN 1558-173X. DOI: 10.1109/4.126534.

[3] CHANDRAKASAN, A. P. and R. W. BRODERSEN. *Low Power Digital CMOS Design*. 1st ed. New York: Springer, 1995. ISBN 978-1-4615-2325-3.

[4] KRAMBECK, R. H., C. M. LEE and H. F. S. LAW. High-speed compact circuits with CMOS. *IEEE Journal of Solid-State Circuits*. 1982, vol. 17, no. 3, pp. 614–619. ISSN 0018-9200. DOI: 10.1109/JSSC.1982.1051786.

[5] NG, P., P. T. BALSARA and D. STEISS. Performance of CMOS differential circuits. *IEEE Journal of Solid-State Circuits*. 1996, vol. 31, iss. 6, pp. 841–846. ISSN 1558-173X. DOI: 10.1109/4.509871.

[6] ISMAIL, A. H. and M. I. ELMASRY. A low power design approach for MOS current mode logic. In: *IEEE International [Systems-on-Chip] SOC Conference*. Portland: IEEE, 2003, pp. 143–146. ISBN 0-7803-8182-3. DOI: 10.1109/SOC.2003.1241480.

[7] ALIOTO, M. and G. PALUMBO. Feature - Power-aware design techniques for nanometer MOS current-mode logic gates: a design framework. *IEEE Circuits and Systems Magazine*. 2006, vol. 6, iss. 4, pp. 42–61. ISSN 1531-636X. DOI: 10.1109/MCAS.2006.264841.

[8] YAMASHINA, M. and H. YAMADA. An MOS Current Mode Logic (MCML) Circuit for Low-Power Sub-GHz Processors. *IEICE Transactions on Electronics*. 1992, vol. E75-C, no. 10, pp. 1181–1187. ISSN 0916-8516.

[9] ALLAM M. W. and M. I. ELMASRY. Dynamic current mode logic (DyCML): A new low-power high-performance logic style. In: *Proceedings of the IEEE 2000 Custom Integrated Circuits Conference*. Orlando: IEEE, 2000, pp. 421–424. ISBN 0-7803-5809-0. DOI: 110.1109/CICC.2000.852699.

[10] ALLAM, M. W. and M. I. ELMASRY. Dynamic current mode logic (DyCML): A new low-power high-performance logic style. *IEEE Journal of Solid-State Circuits*. 2001, vol. 36, iss. 3, pp. 550–558. ISSN 1558-173X. DOI: 10.1109/4.910495.

[11] KANG, S.-M. and Y. LEBLEBICI. *CMOS Digital Integrated Circuits: Analysis and Design*. 3rd ed. McGraw-Hill Higher Education, 2003. ISBN 978-0072460537.

[12] RABAEY, J. M., A. CHANDRAKASAN and B. NIKOLIC. *Digital Integrated Circuits: A Design Perspective*. 2nd ed. Pearson Education, 2003. ISBN 978-8-1203-22578.

[13] PANDEY, N., D. GARG, K. GUPTA and B. CHOUDHARY. Hybrid Dynamic MCML Style: A High Speed Dynamic MCML Style. *Journal of Engineering*. 2016, vol. 2016, iss. 1, pp. 1–10. ISSN 2314-4904. DOI: 10.1155/2016/8027150.

[14] PANDEY, N., K. GUPTA and B. CHOUDHARY. MCML Dynamic Register Design. In: *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*. Windsor: IEEE, 2018, pp. 582–586. ISBN 978-1-5386-7392-8. DOI: 10.1109/MWSCAS.2018.8623910.

[15] GUPTA, K., N. PANDEY and M. GUPTA. *Model and Design of Improved Current Mode Logic Gates: Differential and Single-ended*. 1st ed. Singapore: Springer, 2020. ISBN 978-981-15-0982-7.

About Authors

Riya JAIN received her B.Tech in the discipline Electronics and Communication Engineering from Delhi Technological University (DTU) in 2020. She is a research enthusiast who aims to contribute in the field of Electronics and Communication in coming future. Her areas of interest are: Digital Very-Large-Scale Integration (VLSI) design, Digital Design, and Computer Architecture. She has been contributing to the field of Electronics since 2016 and has been rigorously involved in research since 2018. Being a novice researcher, she is still exploring multiple domains in the field of Electronics and is motivated enough for further contribution to technological advancement in the same.

Kirti GUPTA received her B.Tech. in Electronics and Communication Engineering from Indira Gandhi Institute of Technology, Delhi in 2002, M.Tech. in Information Technology from School of Information Technology in 2006. She received her Ph.D. in Electronics and Communication Engineering from Delhi Technological University, in 2016. Since 2002, she is with Bharati Vidyapeeth's College of Engineering,

New Delhi and is presently serving as Professor in the same institute. A life member of International Society for Technology in Education (ISTE), and senior member of Institute of Electrical and Electronics Engineers (IEEE), she has published more than 100 research papers in international, national journals and conferences.

Neeta PANDEY received her M.E. in Microelectronics from Birla Institute of Technology and Sciences, Pilani in 1991 and Ph.D. from Guru Gobind Singh Indraprastha University, Delhi in 2009. She has served in Central Electronics Engineering Research Institute, Pilani, Indian Institute of Technology, Delhi, Priyadarshini College of Computer Science, Noida and Bharati Vidyapeeth's College of Engineering, Delhi in various capacities. At present, she is a professor in the Department of Electronics and Communication Engineering, Delhi Technological University. Her teaching and research interests include analog and digital VLSI design. A life member of ISTE, and senior member of IEEE, USA, she has coauthored over 100 papers in international, national journals of repute and conferences.

2021 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021

Improving Traditional Stock Market Prediction Algorithms using Covid-19 Analysis

Rajni Jindal
HOD-Computer Engineering Department
Delhi Technological University
New Delhi, India
rajnijindal@dce.ac.in

Nikhil Bansal
Computer Engineering Department
Delhi Technological University
New Delhi, India
nikhilbansal19.nb@gmail.com

Nitin Chawla
Computer Engineering Department
Delhi Technological University
New Delhi, India
nitinchawla.emails@gmail.com

Sanskriti Singhal
Computer Engineering Department
Delhi Technological University
New Delhi, India
sanskriti0512@gmail.com

2021 International Conference on Emerging Smart Computing and Informatics (ESCI) | 978-1-7281-8519-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ESCI50559.2021.9396887

***Abstract*—The stock market is an organized body where public companies offer their stocks through initial public offerings and traders buy/sell these stocks so as to obtain profits. It is dynamic and volatile in nature which makes the task of stock market trend prediction a complex problem. In recent times, the COVID-19 pandemic has made this task even harder. With the rising number of COVID-19 cases across the globe, the market has never been more volatile. This has resulted in the poor performance of various traditional trend prediction algorithms because these algorithms do not account for the impact of the pandemic on the stock market trends. The proposed work aims to enhance the stock market prediction ability of various common prediction models by taking into account the factors related to COVID-19. The forecasting techniques analysed are Decision Tree Regressor, Random Forest Regressor and Support Vector Regressor (SVR). Currently the most affected countries by COVID-19 are the United States of America, India and Russia. Therefore we have analysed the prediction performance of various approaches discussed in this paper on S&P 500 Index, Nifty50 Index and RTS Index using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Results obtained showcase that all the techniques used perform better when the COVID-19 features were included.**

Keywords—COVID-19, Stock Market Trend, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor (SVR), S&P 500 Index, Nifty50 Index, RTS Index.

I. INTRODUCTION

The daily stock market movement is extremely volatile and dynamic. It is influenced by many hidden factors and complex relationships which leads to a high degree of nonlinearity. Some theories on the nature of the stock market state that the stock market movement is completely random. One such theory is the Random walk theory [1] according to which stock market movement is independent of its past movement and hence cannot be predicted. While other assumptions like the Mean Reversion [2] argue that over a period of time the stock price is likely to move towards the average price. Another theory which is a development on the previous one called the Moving Average Reversion (MAR) argues that the average price is about the same as the mean of the prices in a previous time frame. Also various technical analysis show that past data can be analysed to make educated guesses that are quite close to the actual value.

COVID-19 outbreak is caused by SARS-CoV-2 and its first case was reported in China in the Wuhan region. WHO (World Health Organization) acknowledged it as a public health emergency internationally. COVID-19 has adversely affected more than 200 countries across the globe.

The lockdown in various countries has affected many lives, jobs and businesses. Businesses witnessed reduction in revenue generated which created cash flow challenges for the companies, thereby also affecting the economy. Some economists have declared COVID-19 as a Black-Swan Event for the stock market. It has had a catastrophic impact on the stock market. Numerous stocks have witnessed severe shifts. Some of them experienced drastic falls while the prices of others shot up. The investors are facing a hard time in investing their money as the prices are so unpredictable during the pandemic.

Various traditional prediction techniques have performed poorly and given unreliable results during the time of the pandemic since they do not take into account the various dimensions of the impact that the pandemic has had on the world. In this paper, we have compared the performance of such models by considering various COVID-19 related factors that have affected the stock market prediction.

II. RELATED WORK

Huang, Nakamori, and Wang, in their research paper [3] have proposed the widely used machine learning technique called Support Vector Machine to forecast the stock prices. In the paper they took into account the relationship that exists between various macroeconomic variables and the stock index. They have also highlighted the factor of how the US economy influences the Japan stock market flow therefore the S&P 500 Index is chosen as another input to be fed into the model. The authors have compared the performance of LDA,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:42:04 UTC from IEEE Xplore. Restrictions apply.

QDA and Elman Backpropagation Neural Networks (EBNN) and the proposed SVM. It was observed that SVM successfully forecasted the weekly stock market movement flow with the least amount of error.

Risul Islam Rasel et al. [4] used Artificial Neural Network (ANN) after pre-processing the data with the help of windowing techniques as an approach for predicting stock market trend to analyse financial instability. The paper proposed 3 efficient models predicting 1 day ahead, 5 days ahead and 10 days ahead. It was observed that the 1 day ahead model performed the best among the three models. They have also compared the performance of their approach with some other widely used algorithms - Support Vector Machine (SVM) and K-Nearest Neighbours (KNN). It was observed that ANN could predict the stock price more accurately than SVM and KNN.

Hyejung Chung et al. [5] proposed a "hybrid approach" combining the Genetic Algorithm with the LSTM neural network. This research suggests a method to decide the LSTM model's time window and configuration using Genetic Algorithm for stock market financial prediction. In all the error metrics, Genetic Algorithm improved LSTM Model showed enhanced results because of an improved training procedure on the LSTM network.

Kang Zhang et al. [6] presented a novel approach of applying generative adversarial networks to the problem of stock price prediction. The paper considered building the generator as an LSTM network and the discriminator as a Multilayer Perceptron (MLP). Both the generator and the discriminator indulge in a zero sum game. The generator tries to learn the distribution of the dataset and predict values as close as possible to the real data while the discriminator performs the task of distinguishing between real data and the data predicted by the generator. In this way both the generator and the discriminator improve over time resulting in a better learning process.The results showed that the proposed GAN architecture performed better than the classical methods for stock market trend prediction.

Anshul Mittal et al. [7] proposed a technique wherein the twitter dataset is used to perform sentiment analysis and the tweets are classified into 4 common feelings- Calm, Happy, Alert and Kind. These predicted moods from the sentiment analysis and the processed Dow Jones Industrial Average (DJIA) values are then given as an input to their Self Organizing Fuzzy Neural Networks (SOFNN) model to learn and predict future DJIA values.

It was observed that none of the approaches above could account for the impact of a pandemic on stock market trends.

III. RESEARCH METHODOLOGY

The following section elaborates upon the methodology and techniques adopted in our approach towards improving the traditional stock market prediction algorithms for the prediction task.

A. Our Approach

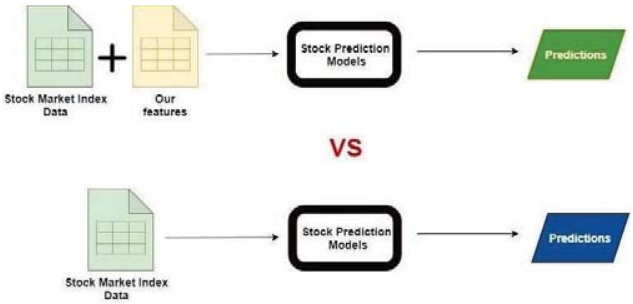


Fig. 1. Difference in Approach

With the rising number of COVID-19 cases across the globe, the market has never been more volatile. This has resulted in the poor performance of various traditional trend prediction algorithms because these algorithms do not account for the impact of the pandemic on the stock market trends. In our approach we have included various features to account for the pandemic so that our models are better able to learn the trend and give more accurate results. For the purpose of validation, we have tested our technique with stock market data of 3 different countries (The USA, India and Russia) while making use of 3 different traditional stock market prediction algorithms for the purpose of making predictions. The machine learning algorithms used include Decision Tree Regressor, Random Forest Regressor and Support Vector Regressor.

Before feeding the data into the model we have combined the standard stock market index attributes with the Covid-19 features. We then shifted the closing value column of our dataset by 1 field since we want to be able to predict the closing value of the index for the next day given the data of the current day. The dataset has been divided into a split of 80%- 20% i.e. 80% of the data is being used for the training purpose and 20% of the data is being used for testing purposes. We have then analyzed and compared the results obtained by using our features to those obtained by using the traditional features only.

B. Dataset Description

COVID-19 has adversely affected more than 200 countries across the globe. Currently the most affected countries by COVID-19 are the United States of America, India and Russia .Therefore we have analysed the prediction for S&P 500 Index, Nifty50 Index and RTS Index.

C. Data Collection

Historic data for each trading day of S&P 500 Index, Nifty50 Index and RTS Index : S&P 500 index is evaluated by taking a weighted average of 500 large companies registered in the United States. RTS Index is evaluated by taking a weighted average of 50 stocks in Russia listed on the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:42:04 UTC from IEEE Xplore. Restrictions apply.

Moscow Exchange. The data for the above was obtained from the Yahoo Finance website [8]. Nifty50 is evaluated by taking a weighted average of the largest 50 establishments in NSE India. The data for this was obtained from the website of National Stock Exchange (NSE) [9].

The statistical data for the COVID-19 was obtained from the website [9] of the World Health Organisation (WHO). The timeframe of the datasets is the year 2020 i.e. 3th January 2020 – 30th December 2020.

D. Data Preprocessing

Before feeding the data to various models to be analysed and compared we have performed some pre-processing. It was found that in the dataset of the stock market indexes that the values corresponding to dates on weekends and holidays were missing as the stock market is not active on those days. Therefore, in order to clean the data, we omitted the rows corresponding to those dates in both our datasets. The duplicate entries and other noises were carefully handled. For scaling of our data we have used the MinMaxScaler module of the scikit learn framework which adjusts the value of our input features. We have adjusted the value of each of our features between 0 and 1. Scaling the features brings the feature vectors to the same scale so that the model is not negatively impacted by large variations in the data and enables models to learn better. The following features have been chosen Date, Opening Price (OP) for the trading day, High Price (HP) for the trading day, Low Price (LP) for the trading day, Closing Price (CP) for the trading day, Shares Traded, Turnover, New Cases, Cumulative Cases, New deaths, Cumulative Deaths.

IV. RESULTS

We have used Root Mean Squared Error (RMSE) to evaluate our results. Another metric that is used to compare the results is the MAPE (Mean Absolute Percentage Error) since we are concerned about the variance of predicted value from the actual value i.e. relative error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}} \quad (1)$$

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad (2)$$

The best prediction results in the below table are in bold. The lower the value of RMSE and MAPE the better the model performs on the forecasting task. We observe that the RMSE values and MAPE values are less when we include the COVID-19 features in our dataset.

TABLE I. COMPARISON OF RESULTS

Country	Machine Learning Model Used	Root Mean Squared Error		Mean Absolute Percentage Error	
		With Covid Features	Without Covid Features	With Covid Features	Without Covid Features
India (Nifty50 Index)	Decision Tree Regressor	**212.672**	270.014	**1.569**	1.942
	Random Forest Regressor	**177.436**	190.426	**1.285**	1.355
	Support Vector Regressor	**128.883**	132.451	**0.902**	0.906
Russia (RTS Index)	Decision Tree Regressor	**57.517**	66.200	**1.574**	1.600
	Random Forest Regressor	**45.490**	50.762	**1.202**	1.265
	Support Vector Regressor	**33.924**	53.764	**0.922**	1.521
United States of America (S&P 500 Index)	Decision Tree Regressor	**53.409**	59.381	**1.273**	1.338
	Random Forest Regressor	**49.642**	53.272	**1.169**	1.281
	Support Vector Regressor	**45.572**	58.700	**1.134**	1.419

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:42:04 UTC from IEEE Xplore. Restrictions apply.

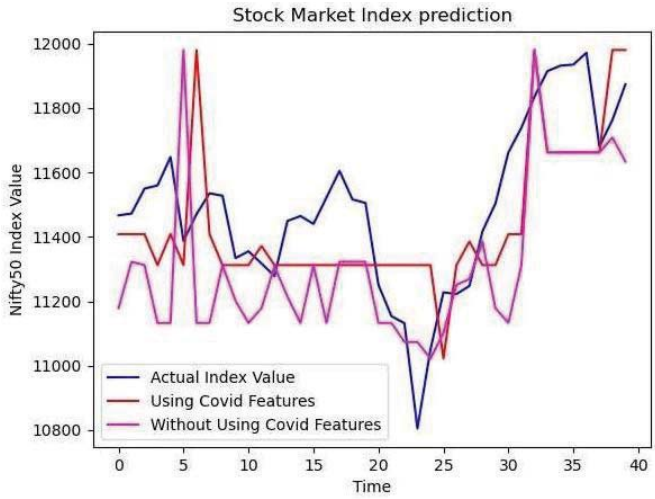


Fig. 2. Decision Tree Regressor (India)

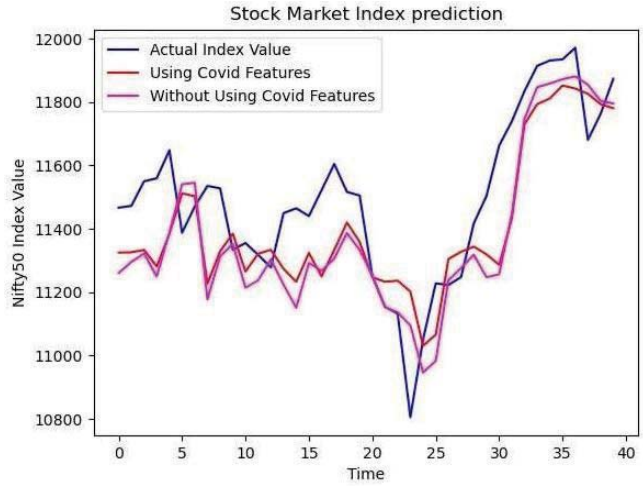


Fig. 3. Random Forest Regressor (India)

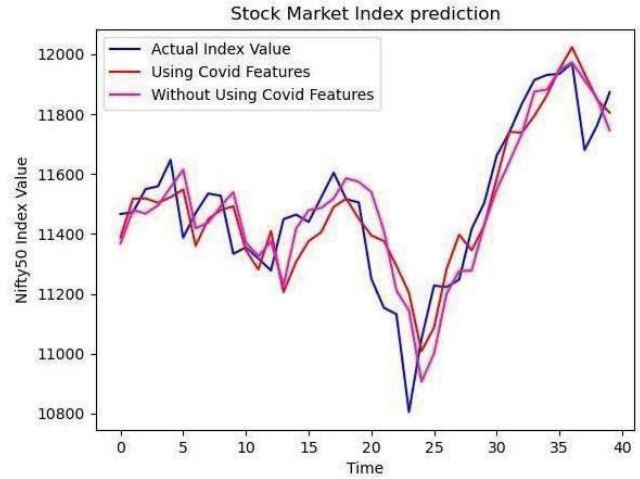


Fig. 4. Support Vector Regressor (India)

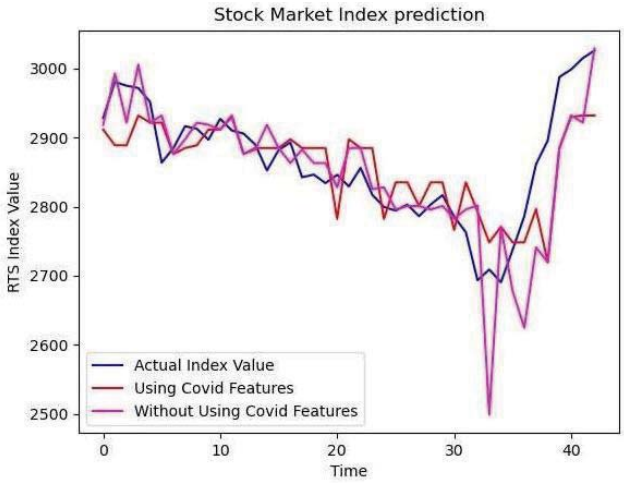


Fig. 5. Decision Tree Regressor (Russia)

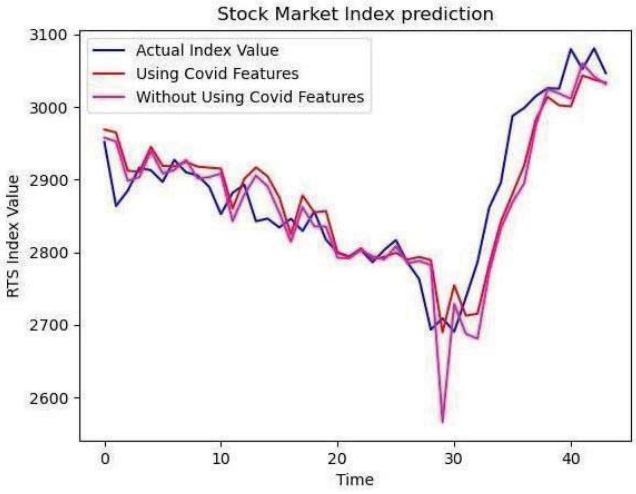


Fig. 6. Random Forest Regressor (Russia)

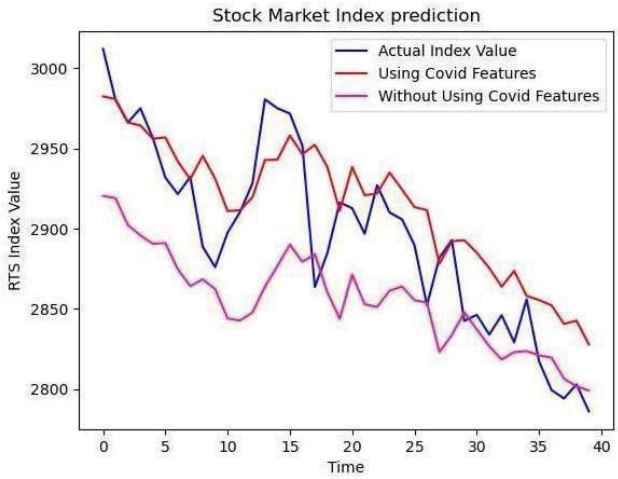


Fig. 7. Support Vector Regressor (Russia)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:42:04 UTC from IEEE Xplore. Restrictions apply.

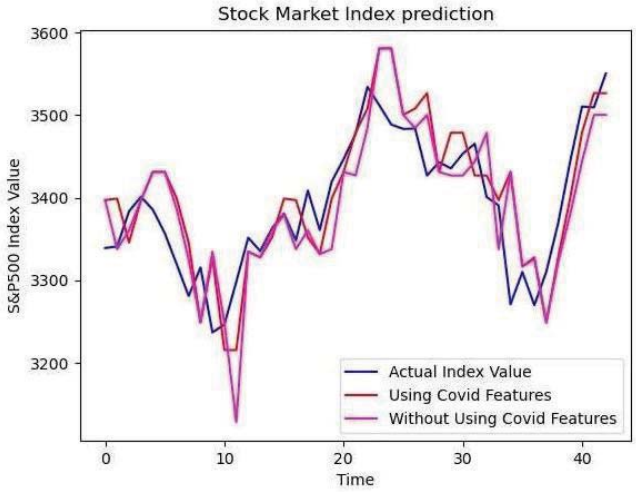


Fig. 8. Decision Tree Regressor (USA)

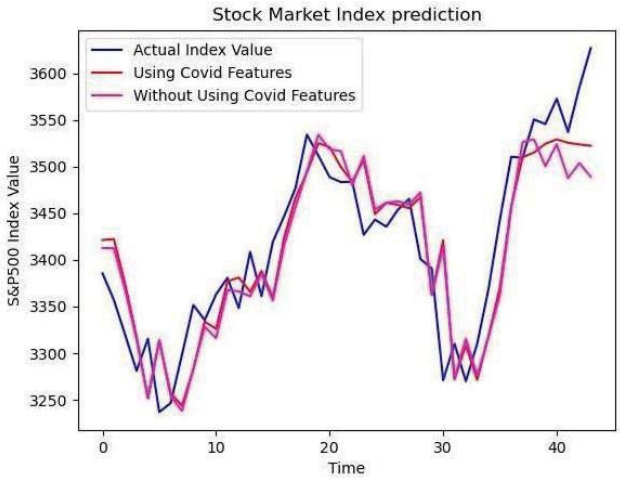


Fig. 9. Random Forest Regressor (USA)

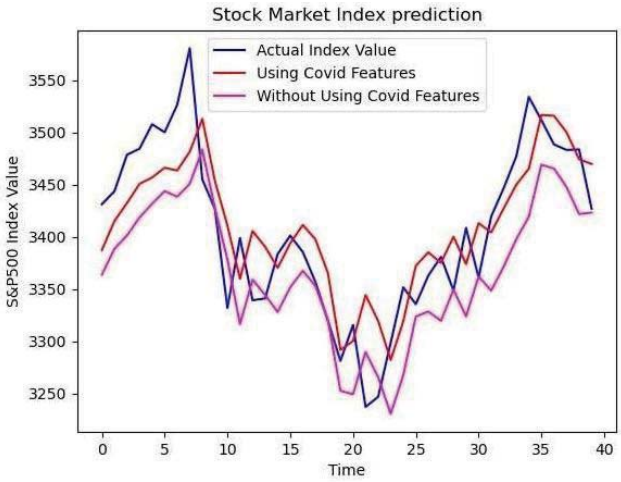


Fig. 10. Support Vector Regressor (USA)

V. CONCLUSION

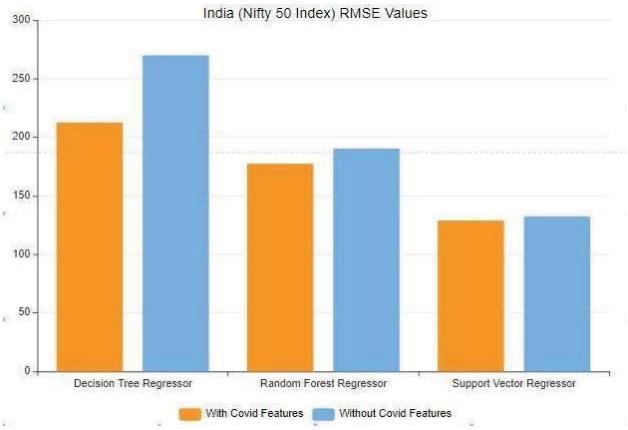


Fig. 11. RMSE Comparison Nifty50 Index (India)

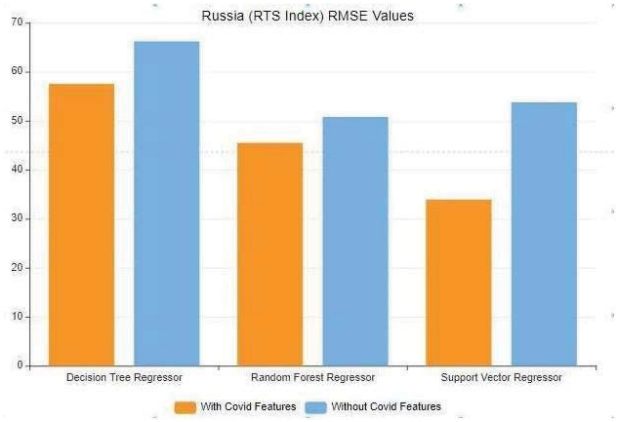


Fig. 12. RMSE Comparison RTS Index (Russia)

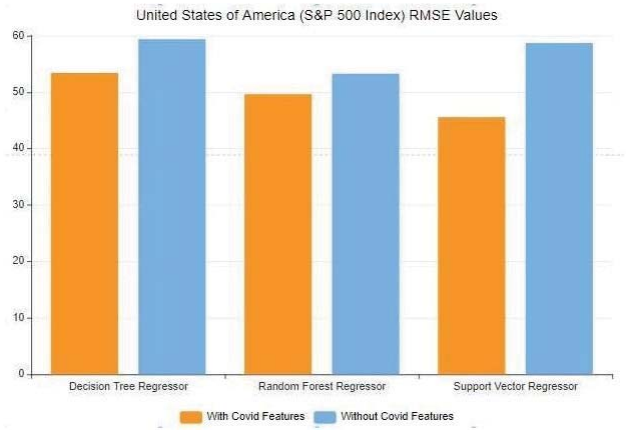


Fig. 13. RMSE Comparison S&P 500 Index (USA)

The following conclusions can be drawn from the above results :-

- There is a correlation between stockmarket trends and Covid-19 cases.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:42:04 UTC from IEEE Xplore. Restrictions apply.

- Making use of the number of cases andthe number of deaths in a country due to COVID-19 improves the prediction ability of the models.

VI. FUTURE SCOPE

As more data is being made available by the day there is a scope for predicting stock market trends during times of pandemic using neural network based approaches which are proven to be more robust in prediction tasks involving time series data like the LSTM (Long Short Term Memory) and GRU (Gated Recurrent Units) techniques.

REFERENCES

[1] Van Horne, J. C., & Parker, G. G. (1967). The random-walk theory: an empirical test. Financial Analysts Journal, 23(6), 87-92.

[2] Poterba, J. M., & Summers, L. H. (1988). Mean reversion in stock prices: Evidence and implications. Journal of financial economics, 22(1), 27-59.

[3] Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." Computers & operations research 32.10 (2005): 2513-2522.

[4] Rasel, R. I., Sultana, N., & Hasan, N. (2016). Financial instability analysis using ANN and feature selection technique:Application to stock market price

[5] Chung, H., & Shin, K. S. (2018). Genetic algorithm-optimized long short-term memory network for stock market prediction. Sustainability, 10(10), 3765.

[6] Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock market prediction based on generative adversarial networks. Procedia computer science, 147, 400-406.

[7] Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." Stanford University, CS229 (2011 http://cs229. stanford.edu/proj2011/GoelMittal-StockMarketPredic tionUsingTwitterSentimentAnalysis.pdf) 15 (2012).

[8] “Data for stock Market Index”[Online]. Available: https://in.finance.yahoo.com/ [Accessed Dec 20,2020].

[9] “Data for Nifty50 Index” [Online]

[10] Available: https://www.nseindia.com/ [Accessed Dec 20,2020]

[11] “Data for Covid” [Online]

[12] Available: https://www.who.int/data/collections/ [Accessed Dec 20, 2020]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:42:04 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/350513438

Influence of Walnut (Juglans L.) Shell Particles Addition on the Mechanical Properties of Epoxy Composites

Article *in* Applied Mechanics and Materials · March 2021

DOI: 10.4028/www.scientific.net/AMM.903.117

CITATIONS
0

READS
26

3 authors, including:

Mohit Mittal
Delhi Technological University
21 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Rajiv Chaudhary
Delhi Technological University
94 PUBLICATIONS **530** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Lignocellulosic fibers based composite material View project

Corrugation process parameters optimization View project

All content following this page was uploaded by Mohit Mittal on 31 March 2021.

The user has requested enhancement of the downloaded file.

Applied Mechanics and Materials
ISSN: 1662-7482, Vol. 903, pp 117-124
© 2021 Trans Tech Publications Ltd, Switzerland
Submitted: 2020-02-22
Revised: 2020-12-26
Accepted: 2020-12-30
Online: 2021-04-15

Influence of Walnut (*Juglans L.*) Shell Particles Addition on the Mechanical Properties of Epoxy Composites

Mohit Mittal[1,a *], Kanchan Phutela[2,b] and Rajiv Chaudhary[3,c]

[1]Department of Mechanical, Production, and Industrial Engineering, Delhi Technological University, Delhi – 110042 (India)

[2]Department of Food Technology, Kurukshetra University, Kurukshetra – 136119 (India)

[3]Department of Mechanical, Production, and Industrial Engineering, Delhi Technological University, Delhi – 110042 (India)

[a]mohit.30mittal@gmail.com, [b]phutelakanchan123@gmail.com, [c]rajivchaudhary@dce.ac.in

*Corresponding Author

Keywords: Walnut shell; Epoxy resin; Composite; Mechanical properties

Abstract. The primary objective of this research work is to explore the potential of walnut shell particles as a reinforcing agent in the thermoset matrix composite. Therefore, the number of composite samples was developed by varying the filler weight percentage (15%, 25%, 35%, and 45%) in an epoxy polymer. The composite samples were characterized in terms of mechanical properties by using tensile test, flexural or three-point bending test, Izod impact test, compression test, and Rockwell hardness test. Moreover, the effect of cellulosic filler content on the density and water absorptivity of epoxy-based composites was also investigated as per the ASTM standard. Agricultural waste in the form of walnut shell was characterized by the determination of grain size. The experimental results showed that the stiffness and hardness of epoxy thermoset were increased with the decrease in tensile and flexural strength after the addition of walnut shell fine particles. The composite of 35 wt% walnut filler exhibits a balanced combination of strength, stiffness, and toughness. It was found that the mechanical strength and modulus of filled composites were increased with the increase of filler content up to 35 wt%.

Introduction

The addition of cellulosic particles in polymer matrix composite leads to the development of low cost and lightweight material having improved thermo-mechanical properties. The biocomposite is the best alternative to wood, chipboard, plywood, and glass fiber reinforced composites [1,2,3]. As a substitute for glass, aramid, and carbon fibers; the lignocellulosic fillers have relatively good thermal insulation and specific mechanical properties, easy availability, low cost, sustainability, recyclability, and biodegradability [4,5]. Furthermore, the cellulosic fillers have required considerably lower energy for manufacturing and post-processing. According to the report presented by Bavan and Kumar, the consumption of natural fiber composites (NFCs) in India is 2075 thousand tons in 2010 [6]. With progressive concern over the environment, the biocomposites possess great importance in non-structural applications of an automobile, electronics, housing, and defense industries.

Currently, the non-wood natural fibers (hemp, kenaf, flax, sisal, pineapple, and coir) are used significantly in commercial applications. Among all types of natural fillers, the walnut shell (*Juglans regia L.*) has competitive advantages for thermosetting composites [7]. I.I Busuioceanu *et al.* [8] reported that the walnut shell reinforced composites have superior mechanical performance than the pine needle and green ash based composites. Several other studies also proved the positive attributes of walnut shell as reinforcement in polymer matrix composites [9-12]. According to the data presented by FAOSTAT of the United Nations in 2017, approx. 3.8 million tons of walnut shell is produced all over the world, with China contributing 51% of the total [13]. As an abundant resource of fruit waste, it becomes necessary to use it in an optimum manner. It was found that the price of 1 kg walnut shell particles is typically about 0.12 USD which is significantly less than the

All rights reserved. No part of contents of this paper may be reproduced or transmitted in any form or by any means without the written permission of Trans Tech Publications Ltd, www.scientific.net. (#556089967, Delhi Technological University, Bhiwani, India-25/03/21,16:40:40)

price of wood flour, 0.27 USD. Due to the above reasons of high availability, low cost, and better reinforcing effectiveness of walnut shell particles, it can be employed for the fabrication of low-cost and high strength composites which has potential to reduce the demand for wood-based material.
In order to develop a high-performance biocomposite material, the epoxy resin is used as a solvent which can cure over a wide range of temperature without any significant emission of volatile by-products. The epoxy thermoset is widely used in industrial applications because of their good mechanical, adhesive, and chemical resistant properties. Therefore, in this research work, the epoxy-based composite was developed and characterized for physical and mechanical properties. Moreover, the influence of walnut shell filler content (15, 25, 35, 45 wt.%) on the mechanical properties of epoxy composites was investigated.

Experimental

Materials: The medium viscosity epoxy resin [DGEBA (diglycidyl bisphenol-A); density: 1.17 g/cm^3; viscosity 500-700 mPa s; average molecular weight (M_n): 2300 g/mol] and hardener [TETA (triethylene tetraamine); density: 1.05 g/cm^3; viscosity 150-300 mPa s; average molecular weight (M_n): 146.23 g/mol] were procured from M/s Sakshi Dies and Chemicals, Delhi (India) and used as polymer matrices. The primary reasons to select epoxy thermoset as a viscoelastic matrix medium are: low viscosity and shrinkage; excellent chemical, thermal, and moisture resistant properties; and its ability to carry mechanical static and dynamic loads. The walnut particles obtained by grinding of the walnut shell (*Juglans* regia) were used as reinforcement fillers, having the composition shown in Table 1.

Table 1 Composition of walnut shell [14]

Material	Cellulose %	Holocellulose %	Lignin %	Ash %
Walnut shell	25.4	46.6	49.1	3.6

Preparation of biocomposites: Preparation of biocomposites involved the grinding of walnut shell on a laboratory vertical planetary ball mill machine (YKM-8L, China). After the milling process, the drying was performed at a temperature of 40 ± 5 ^{0}C for 4 days using a hot air oven (SV Scientific Laboratory, India). The dried walnut particles of different weight percentages (15%, 25%, 35%, and 45%) were mixed with epoxy resin by mechanical stirrer at 4000 RPM. Subsequently, the mixture was degassed in the furnace at a temperature of 60 ± 10 ^{0}C for 60 mins. Then the hardener (Triethylenetetramine, HY 951) was added in a proportion of 50 wt% for 100 wt% of epoxy resin and stirred at 7000 RPM for 3 mins to make a viscous solution. The solution was poured into the rubber molds coated with releasing gel. The casted samples were cured at room temperature (25 ^{0}C) for 3 days.

Methods: The characterization of milled walnut particles was done using vibratory sieve shaker, equipped with six sieves of following mesh sizes: 0.025, 0.037, 0.053, 0.075, 0.12, 0.3 mm. The total mass of 40 gm was taken for natural filler analysis and the measurements were carried out in a time of 1 hr.
Moisture absorption of composite specimens was determined according to ASTM D570 standard. The test was performed by immersing the samples in distilled water at room temperature. The weight gain (%) was calculated by using the formula

$$Weight\ gain\ (\%) = \frac{(W_t - W_O)}{W_O} \times 100 \quad (1)$$

where W_t – is the weight of the specimen after immersion time "t", W_0 - oven-dried weight.
Tensile tests were performed on biocomposite specimens using a Table Top Tinius Olsen Horizon H50KS, Universal Testing Machine in accordance with ASTM D638 standard. The tensile properties of the composite material were measured at room temperature with a crosshead speed of 2 mm/min. The standardized sample for tensile tests was in the dumbbell shape of 50 mm gauge length, 13 mm width, and 4 mm thickness.

The flexural properties were determined as per ASTM D790 standard. The composite specimen of dimensions 80 mm X 15 mm X 4 mm was subjected to a three-point bending test on UTM with a crosshead speed of 3 mm/min. The impact strength of notched composite specimens was determined using an Izod impact testing machine as per ASTM D256 standard. The dimensions of specimen for impact testing are 65 mm X 13mm X 4mm. The hardness of developed biocomposites was evaluated using the Rockwell Hardness Testing Machine on M scale with a ball diameter of 6.35 mm.

Results and Discussion

Natural filler analysis: The particle size distribution of the ground walnut shell was characterized by using a vibratory sieve shaker. Fig. 1 revealed the amount of material which remained on a sieve with a specified mesh size. It was observed that the majority of particles (94.7%) were in the size range of 37 µm to 120 µm after the grinding process.

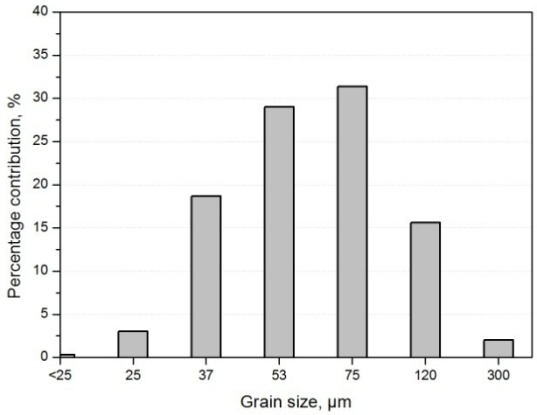


Fig. 1 Grain size distribution of walnut shell

Density: To fulfill the market demand for lightweight materials, the developed biocomposite was characterized for density measurement as a special function of filler weight fraction. Table 2 shows that the density of particle board material was reduced with the increase of filler weight fraction. This reduction was 5.5% for 45 wt% walnut particle composite in comparison to the pure epoxy matrix. It was due to the fact that the walnut particles are light in weight and occupy a large volume in the material. This kind of behavior is general and also found in the literature review.

Table 2 Density (gm/cm^3) of walnut particles reinforced composite

Sr. No.	EP	15% WP	25% WP	35% WP	45% WP
1	1.183	1.171	1.159	1.142	1.116
2	1.179	1.166	1.155	1.144	1.111
3	1.174	1.171	1.16	1.15	1.112
4	1.177	1.168	1.151	1.143	1.114
Mean	1.178	1.169	1.156	1.144	1.113
SD	0.0032	0.0021	0.0035	0.0032	0.0019

Water Absorption Capacity: The resistance of a material to the diffusion of water molecules is the primary requirement for those applications which are come in contact with water. Therefore, it becomes necessary to characterize the water absorption capacity of developed biocomposites under humid conditions. The water absorption tests were performed at room temperature as per ASTM D570 standard and the obtained results are shown in Table 3. The immersion time was up to the saturation (120 hr). It was observed that the addition of cellulosic particles in a polymer matrix led to an improvement in the diffusion of water molecules. The capacity of water absorption was found to be higher for 45% WP. This was attributed to the presence of a large amount of hydroxyl (-OH)

and other polar functional groups on the surface of walnut particles. It can be reported that the water absorption capacity of the epoxy polymer was substantially increased even after the addition of 15 wt% walnut particles. Therefore, the developed biocomposite cannot be employed effectively in structural applications.

Table 3 Water absorption capacity of the composite samples

Sr. no.	EP	15% WP	25% WP	35% WP	45% WP
1	2.01	4.34	5.62	7.46	9.22
2	1.94	3.98	5.33	7.23	8.73
3	2.00	4.27	5.03	7.68	9.05
4	1.89	4.08	5.42	7.44	9.16
Mean	1.96	4.17	5.35	7.45	9.04
SD	0.048	0.144	0.212	0.159	0.189

Mechanical Properties: To assess the influence of walnut particles incorporation in epoxy composite, the mechanical tests were performed as per ASTM standard.

Tensile properties: Tensile tests were carried out at a strain rate of 2 mm/min to analyze the mechanical behavior of walnut particles reinforced epoxy composites (WP). The results of the ultimate tensile strength and modulus of elasticity are presented in Table 4 and Fig. 2. It was observed that the loading of walnut particles caused the reduction in tensile strength of an epoxy thermoset. This kind of behavior was attributed to the poor stress transfer from polymeric resin to reinforcing fillers through the interfacial region. As the filler content was increased from 15 wt% to 35 wt%, the ultimate tensile strength was increased significantly. Moreover, the 35% WP shows almost the same strength in comparison to virgin epoxy polymer. Amongst all the developed samples, the 35% WP composite exhibits the highest stiffness and this was because of the close interaction and uniform distribution of walnut particles which leads to an easy and effective stress transfer from matrix to particles and particle to particle. The elastic modulus of epoxy resin was increased by 88.74% after the addition of 35 wt% walnut shell particles. It can also be noticed that the stiffness of polymer was increased considerably even after the loading of 15% walnut particle and it was due to the presence of hydroxyl (-OH) functional group on the surface of cellulosic fillers which leads to the creation of additional curing networks. Furthermore, the stiff characteristics of filler particles caused the increment in the resulting modulus of an epoxy polymer.

Table 4 Tensile properties of the composite samples

Properties	EP	15% WP	25% WP	35% WP	45% WP
Tensile strength (MPa)	30.47	19.85	24.18	30.09	22.18
% Elongation in length	6.83	7.11	5.76	4.08	5.22
Young's modulus of elasticity (GPa)	1.0025	1.2915	1.3017	1.8922	1.245

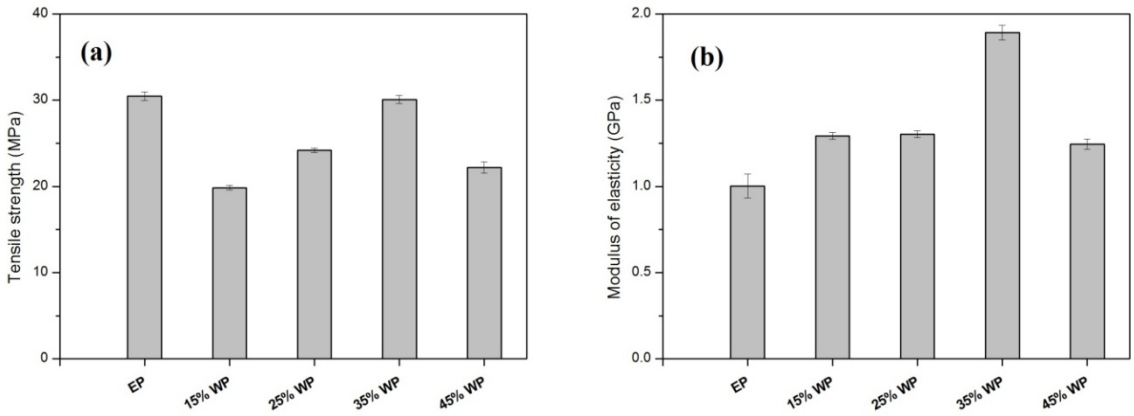


Fig. 2 Effect of filler content on the tensile properties of epoxy polymer composites

Flexural Properties: The flexural strength of developed composite samples was determined using a hydraulic operated universal testing machine at a rate of 3 mm/min under three-point bend test. The observations are reported in Table 5 and Fig. 3. Similar to tensile test results, the bending strength of an epoxy polymer was reduced after the incorporation of walnut shell particles. Amongst all the filled composite samples, the 35% WP exhibits highest flexural strength. In relation to the virgin polymer matrix, the 35% WP shows 44.17% higher flexural modulus. Beyond the limit of 35% filler content, the bending strength and modulus were reduced drastically and it was due to the aggregation of filler particles which results in the non-uniform stress distribution and localized stress concentration. It can be stated from the literature review that the upper end of a specimen faced compressive stress, lower-end experience tensile stress, and the axisymmetric plane is subjected to a shear load during the three-point bending test. The highest flexural modulus of 35% WP specimen may be due to the random arrangement of filler particles which leads to the generation of barrier against the mobility of viscoelastic medium.

Table 5 Flexural properties of the composite samples

Properties	EP	15% WP	25% WP	35% WP	45% WP
Flexural strength (MPa)	44.52	26.075	34.772	42.122	29.147
Flexural strain	0.059	0.046	0.044	0.021	0.026
Flexural modulus (GPa)	1.304	1.43	1.5449	1.88	1.8337

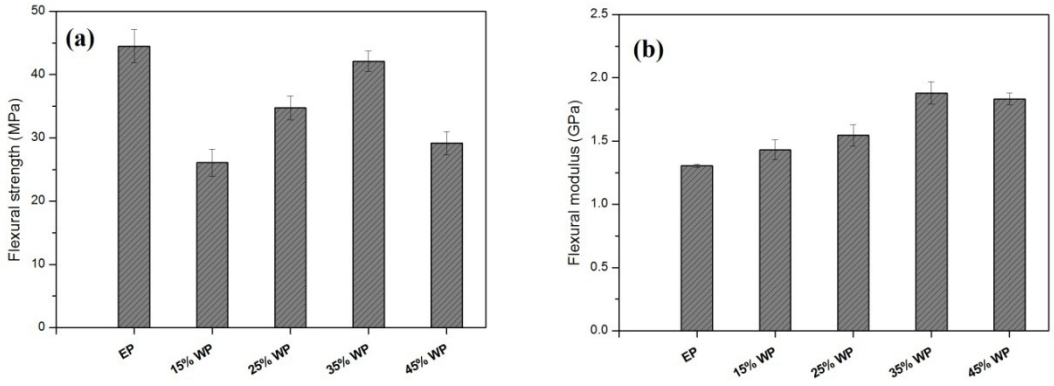


Fig. 3 Effect of filler content on the flexural properties of epoxy polymer composites

Compressive properties: The compressive properties of developed composites were determined by a universal testing machine at a rate of 2 mm/min under the displacement control mode. The results are shown in Table 6 and Fig. 4. A significant increase in the compressive strength of an epoxy thermoset was obtained after the incorporation of 35 wt% walnut shell particles. As the filler content was increased from 15 wt% to 35 wt%, both the strength and stiffness were increased by an amount of 21.18% and 35.96% respectively. Further addition of filler particle leads to the reduction in compressive strength and it was due to the improper particles distribution within the viscoelastic matrix medium. Similar to the variation in tensile and flexural moduli, the modulus of elasticity was also increased with the increase of walnut particle fraction. The composite revealed brittle behavior with less permanent deformation in comparison with the virgin polymer matrix.

Table 6 Compressive properties of the composite samples

Properties	EP	15% WP	25% WP	35% WP	45% WP
Compressive strength (MPa)	83.84	94.75	103.12	114.827	102.47
% Reduction in length	26.47	22.75	18.49	14.89	10.31
Modulus of elasticity (GPa)	1.335	1.546	1.779	2.102	1.875

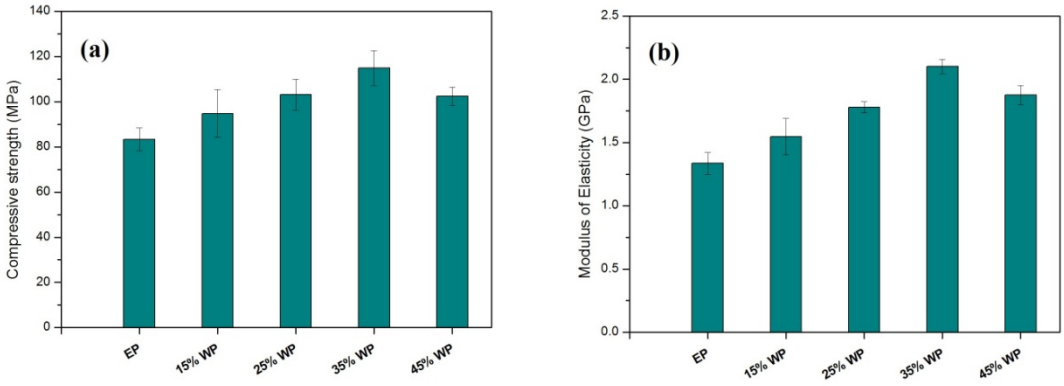


Fig. 4 Effect of filler content on the compressive properties of epoxy polymer composites

Hardness: The ability of a material to resist indentation, penetration and localized plastic deformation is defined by means of hardness. The hardness of biocomposites was measured on M scale of the Rockwell testing machine. The variation of hardness as a function of reinforced filler content is shown in Table 7 and Fig. 5 (a). It can be observed that the incorporation of walnut particles led to an increase in the hardness of an epoxy matrix. This results in the same path of tensile and compressive results which showed the poor ductility of filled composites. As the filler content was increased from 15 wt% to 45 wt% in epoxy polymer, the hardness was increased with the decrease of % elongation in length during the tensile test and % reduction in length during the compressive test. It depicts the brittle nature of highly filled epoxy composites.

Table 7 Rockwell hardness values of the composite samples

Sr. no.	EP	15% WP	25% WP	35% WP	45% WP
1	R-45	62	65	72	76
2	R-52	59	62	69	80
3	R-41	60	68	70	78
4	R-47	63	62	67	82
Mean	R-48.75	61	64.25	69.5	79
SD	2.86	1.58	2.49	1.8	2.24

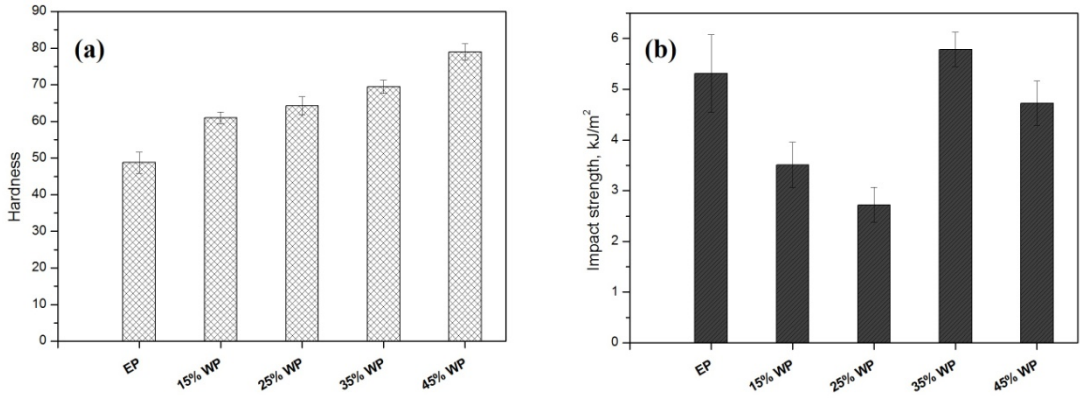


Fig. 5 Effect of filler content on (a) hardness and (b) Impact strength of epoxy composite

Impact strength: The impact strength of biocomposites was determined using an Izod Impact Tester as per ASTM 256 standard and the results are shown in Table 8 and Fig. 5 (b). It was observed that the impact toughness of epoxy thermoset was reduced after the incorporation of walnut shell particles. This was due to the presence of hydroxyl and residual moisture content in the filler providing the micropores that acted as a stress raiser. However, the composite of 35 wt% WP shows better impact strength in comparison to the virgin epoxy matrix. This may be due to the effective stress transfer from the viscoelastic medium to reinforcing particles. Generally, the

presence of macro fillers caused the generation of crack at the interfacial region under the condition of abrupt loading. The impact testing of samples proved the brittleness of walnut particle reinforced composites.

Table 8 Izod impact strength of various composite samples

Sr. no.	EP	15% WP	25% WP	35% WP	45% WP
1	5.83	4.04	2.64	5.94	4.72
2	5.21	3.86	2.19	6.24	4.08
3	6.09	3.12	2.98	5.55	5.31
4	4.11	3.01	3.08	5.42	4.79
Mean	5.31	3.51	2.72	5.787	4.725
SD	0.763	0.448	0.348	0.341	0.436

Conclusions

The present research work primarily explores the potential of walnut particles as reinforcing filler in epoxy matrix composites. Based on the experimental observations, the following conclusions can be made:

- The incorporation of walnut shell particles in epoxy thermoset resin led to the significant increase in stiffness and hardness of a material. However, the tensile and flexural strength was reduced after the addition of organic waste filler.
- Amongst all the developed composite samples, the composite of 35 wt% walnut filler exhibits the optimum set of mechanical properties. Moreover, it shows a considerably lower density than the virgin epoxy resin.
- The mechanical strength and modulus of composites were increased with the increase of filler content up to 35 wt%. The epoxy-based composite reinforced with 35% walnut fine particles showed a balanced combination of strength, stiffness, and toughness which proved it as a structural material.
- The walnut shell particles have significant potential to make low cost, stiff, and hard material that can be smoothly used for non-structural applications.

Overall, we can conclude that the addition of agricultural organic waste in thermosetting polymer results in the low-cost composite with acceptable mechanical properties.

References

[1] N. Saba, M.T. Paridah, M. Jawaid, Mechanical properties of kenaf fiber reinforced polymer composite: A Review, Constr. Build. Mater. 76 (2015) 87-96.

[2] S.J. Hans, H. Lee, H.M. Jeong, B.K. Kim, A.V. Raghu, K.R. Reddy, Graphene modified lipophilically by stearic acid and its composite with low density polyethylene, J. Macromol. Sci. 53 (2014) 1193-1204.

[3] K. Salasinska, M. Polka, M. Gloc, J. Ryszkowska, Composites with pistachio shell and sunflower husk: the effect of filler content and chemical constitution on the dimensional and fire stability, Polimery-W 61 (2016) 255-265.

[4] A. K. Mohanty, M. Misra, L.T. Drzal, Natural fibers, biopolymers, and biocomposites, CRC Press, New York, 2005.

[5] H.S. Kim, H.S. Yang, H.J. Kim, Biodegradability and mechanical properties of agro-flour-filled polybutylene succinate biocomposites, J. Appl. Polym. Sci. 97 (2005) 1513-1521.

[6] D. Bavan, G.C. Kumar, Potential use of natural fiber composite materials in India, J. Reinf. Plast. Compos. 29 (2010) 3600-3613.

[7] V.K. Singh, Mechanical behavior of walnut (Juglans L.) shell particles reinforced bio-composite, Sci. Eng. Compos. Mater. 22 (2014) 383-390.

[8] I.I. Busuioceanu, G. Enachescu, F. Baciu, G. Dinu, M. Stefanescu, Determination of mechanical behavior of new biocomposites, AIP Conference Proceedings, 2018, DOI: 10.1063/1.5024154.
[9] N. Ayrilmis, A. Kaymanakci, F. Ozdemir, Physical, mechanical, and thermal properties of polypropylene composites filled with walnut shell flour, J. Ind. Eng. Chem. 19 (2013) 908-914.
[10] K. Salasinska, M. Barczewski, R. Gorny, A. Klozinski, Evaluation of highly filled epoxy composites modified with walnut shell waste filler, Polym. Bull. 75 (2018) 2511-2528.
[11] M. Zahedi, H. Pirayesh, H. Khanjanzadeh, M.M. Tabar, Organo-modified montmorillonite reinforced walnut shell/polypropylene composites, Mater. Des. 51 (2013) 803-809.
[12] K. Acikalin, Thermogravimetric analysis of walnut shell as pyrolysis feedstock, J. Therm. Anal. Calorim. 105 (2011) 145-150.
[13] Information on http://www.fao.org
[14] D. K. Rao, A. Saxena, S. Kumar, N.K. Joshi, Poonam, Tensile and flexural behavior with characterization of hybrid bio-composite reinforced with walnut shell particles and coconut fibres, J. Adv. Sci. Tech. Res. 1 (2015) 94-102.

View publication stats

Journal of Atomic, Molecular, Condensed Matter & Nano Physics
Vol. 8, No. 1, pp. 63–69, 2021
ISSN 2582-8215
Published by RGN Publications

http://www.rgnpublications.com

DOI: 10.26713/jamcnp.v8i1.1496

Research Article

Interaction between Picric Acid and 5-aminoquinoline: A Theoretical Study

Nupur Pandey[1], Mohan Singh Mehata*[2] and Sanjay Pant[1]

[1] *Photophysics Laboratory, Department of Physics, Centre of Advanced Study, DSB Campus, Kumaun University, Nainital 263002, India*
[2] *Laser Spectroscopy Laboratory, Department of Applied Physics, Delhi Technological University, Delhi 110042, India*
***Corresponding author:** msmehata@gmail.com

Received: December 17, 2020 **Accepted:** December 29, 2020

Communicated by: Kriti Batra

Abstract. In the present work,density functional theory and time-dependent density functional theory were employed to investigate the hydrogen-bonded charge transfer complex of 5-aminoquinoline (5AQ) and picric acid (PA) at B3LYP/6-31G(d,p) level in the gas phase as well as $CHCl_3$. The geometrical parameters, optimization energies, interaction energies, frontier molecular orbitals (FMOs), and ground-state dipole moment were considered, which supported the stability of the complex. Hence, the theoretical study of the 5AQ-PA complex provides valuable information about its molecular structure, possible interaction sites and electronically excited states.

Keywords. 5AQ; DFT/TD-DFT; Hydrogen bonding; Charge transfer complex

PACS. 31.15.Ew; 31.15.ee; 31.10.+z

Copyright © 2021 Nupur Pandey, Mohan Singh Mehata and Sanjay Pant. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. Introduction

The association of the donor and the acceptor moiety via weak hydrogen bonding ($N-H^+\cdots O$) results in the formation of charge transfer (CT) complex, which was first proposed and discussed by Mulliken and Person [7].

CT complexes have been widely explored owing to their widespread applications in different fields, such as sensors, organic semiconductors, organic solar cells, medicinal chemistry, drug-receptor, antibacterial/antimicrobial studies, DNA binding, and second-order nonlinear optical activity [5, 6, 10].

A new absorption band is observed in the absorption spectrum of the complex, which neither assigns to the donor nor the acceptor. This new absorption peak assures the charge transfer from the *highest occupied molecular orbital* (HOMO) of the donor to the *lowest unoccupied molecular orbital* (LUMO) of the acceptor species. The complexation among the donor and the acceptor species gets affected by the functional group carrying sulfur, oxygen, or nitrogen atoms [1, 6].

Over the last few years, various organic donor-acceptor-based CT complexes have been explored using spectroscopic and theoretical techniques, including *density functional theory* (DFT) and *time-dependent density functional theory* (TD-DFT) and provided valuable information about the molecular structures, interaction sites and excited states [1, 4].

Numerous photophysical and photochemical studies on aminoquinolines have been carried out in the past few years as these molecules exhibit a wide spectrum of biological and chemical activities [9]. Aminoquinoline derivatives have a significant role in supramolecular chemistry, electron transfer/charge transfer processes, and sensing applications [8].

In the present study,the H-bonded charge-transfer complex of 5-aminoquinoline (5AQ) and (PA) in the gas phase and chloroform ($CHCl_3$) has been explored using DFT and TD-DFT.

2. Computational Method

The 5-aminoquinoline (5AQ), picric acid (PA), and their possible complex were optimized using the density functional theory (DFT) by employing the B3LYP level using 6-31G(d,p)basis set with the Gaussian-9 program's package [3]. Gauss-view 5.0 was used to draw the molecular structures. The geometrical parameters, optimization energies, zero-point energy corrections, interaction energies, and thermodynamic parameters were calculated from the optimized molecular structures and used to predict the stability of the complex. The polarizable continuum model (IEF-PCM) has been considered in the case of solvent.

3. Results and Discussion

Figure 1 displays the optimized molecular structures of the donor (5AQ), acceptor (PA), and the formed complex (5AQ-PA) in the gas phase. The most stable configuration of the formed complex favors the alignment of OH of PA towards the 5AQ's ring nitrogen atom. It can be seen that the complexation occurs mainly due to the proton transfer hydrogen bonding (H-bonding). The formation of H-bond has led to major changes in bond lengths, particularly for those near the H-bonding interactions sites in the complex.

In proton transfer, H-bonded complex O14-H15-N34, the increase in the length of OH bond from 1.672 Å to 1.819 Å and 1.801 Å was observed on switching from the gas phase to $CHCl_3$, while the length of N-H bond decreased from 1.054 Å to 1.038 Å and 1.036 Å. Table 1 displays the H-bonded parameters of the 5AQ-PA complex. The variations can also be observed in the bond lengths of the 5AQ part of the complex nearby the H-bond center. Considering the PA moiety of the complex, the bond length C(1)-C(6) increased from 1.420 to 1.458, C(1)-C(2) from 1.426 Å to 1.458 Å while C(1)-O(14) reduced from 1.315 Å to 1.25 Å compared with free PA in the gas phase. It is worth mentioning that similar results have been obtained in $CHCl_3$.

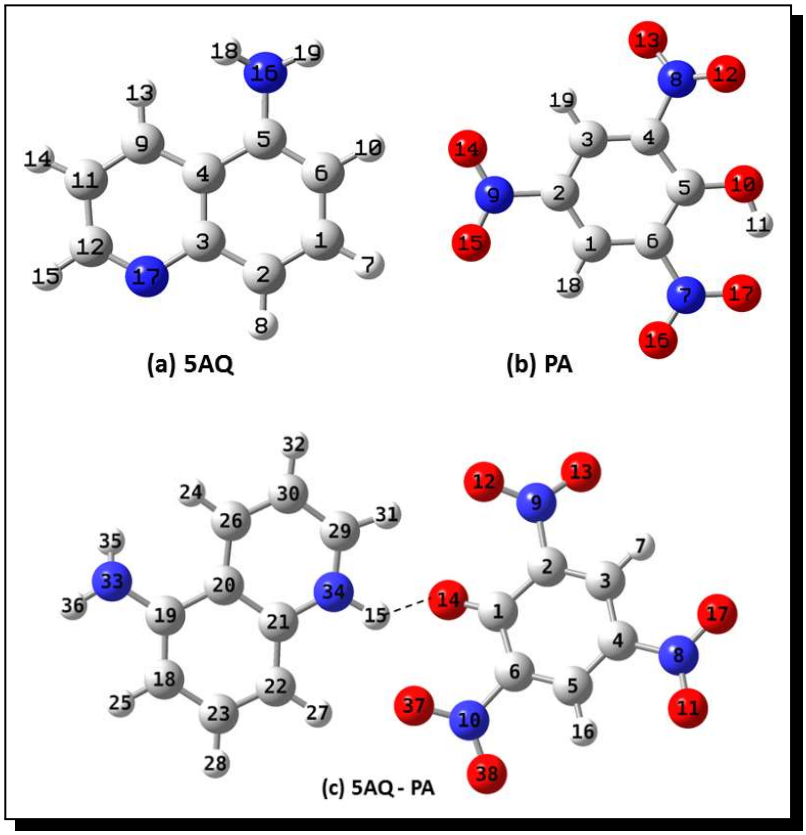


Figure 1. Optimized molecular structures of (a) 5AQ, (b) PA and (c) 5AQ–PA in the gas-phase with atomic numbering and H- bonds

Notable changes were seen in the optimized bond angles of 5AQ moiety near the H-bonding interaction region. Thus changes observed in the optimized bond lengths and bond angles endorse the presence of H-bonding as well as CT in the formed complex.

Table 1. Hydrogen bonding parameters of 5AQ–PA complex in the gas phase and $CHCl_3$

Medium	Distance (Å)			Angle (°)
	O14–H15	N34–H15	O14–N34	O14–H15–N34
Gas	1.668	1.054	2.561	139.210
$CHCl_3$	1.794	1.039	2.655	137.500

The interaction energy (ΔE) is an important parameter that defines the stability of the complex and can be calculated using eq. (1):

$$\Delta E = E_{\text{complex}} - (E_{\text{donor}} + E_{\text{acceptor}}). \tag{1}$$

The optimization energy (E_{total}), zero-point vibrational energy (E_{o}), corrected optimization energy ($E_{\text{corr.}}$), interaction energy (ΔE), and ground-state dipole moment of the complex are surmised in Table 2.

Table 2. Values of total optimization energy (E_{total}), interaction energy (ΔE), zero-point vibrational energy (E_0), corrected optimization energy (E_{corr}), and ground-state dipole moment (μ_g) of the complex using the B3LYP/631-G(d,p) method

Parameter	Gas	$CHCl_3$
E_{total}(kcal mol^{-1})	−864885.551	−864898.375
E_0 (kcal mol^{-1})	167.231	167.169
$E_{\text{corr.}}$ (kcal mol^{-1})	−864703.925	−864716.897
ΔE (kcal mol^{-1})	−12.262	−29.807
μ_g (Debye)	18.092	22.113

The interaction energy for 5AQ-PA recorded −12.262, and −29.807 kcal mol^{-1} in the gas phase and $CHCl_3$, respectively, confirming a stable complex formation.

The theoretically obtained dipole moment for 5AQ ranged between 3.4 D – 4.5 D in the gas phase and the chosen solvent, whereas it ranged from 1.5 D – 2.5 D for PA. Therefore, there is a huge difference between the dipole moment of the complex (5AQ-PA) compared with free 5AQ and PA, confirming electron density migration from 5AQ towards PA. The computed dipole moment of the complex showed higher values for $CHCl_3$ (22.11 D) than the gas phase (18.09 D).

The frontier molecular orbital diagram is a powerful tool in explaining the chemical reactivity and kinetic stability of the molecule [2]. Figure 2 represents the occupied orbital (HOMO) and virtual orbital (LUMO) of complex in the gas phase.

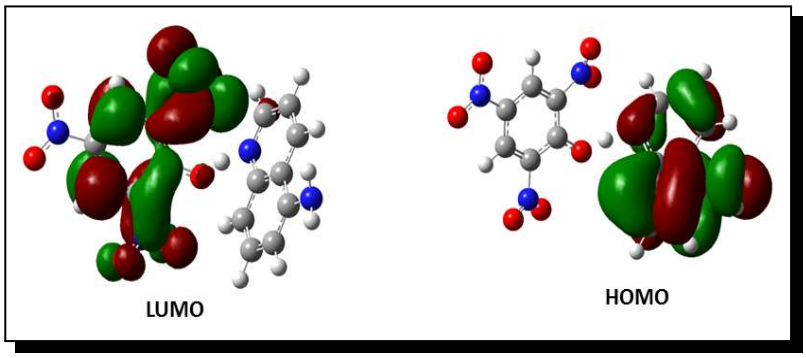


Figure 2. HOMO and LUMO of the complex in $CHCl_3$

The HOMO is mainly distributed over the 5AQ moiety, whereas LUMO is distributed over the PA moiety of the formed complex. The energy gap between the HOMO energy of the donor and the LUMO energy of the acceptor is comparatively small. In the present system, this energy gap is 2.13 eV, 1.86 eV in the gas phase and $CHCl_3$, which supports the possibility of CT from 5AQ towards PA.

The theoretically computed absorption spectrum of 5AQ, PA, and their complex in $CHCl_3$ is shown in Figure 3. In $CHCl_3$, the first three absorption bands along with the oscillator strength (f) were estimated at 431 nm ($f = 0.0770$), 398 nm ($f = 0.1498$), and 394 nm ($f = 0.003$), which corresponds to $S_0 \rightarrow S_1$, $S_0 \rightarrow S_2$, and $S_0 \rightarrow S_3$ transitions, respectively. The new band was detected in the visible region where neither the donor nor the acceptor absorbed. This new band originates possibly from the overlapping of LUMO of the acceptor with HOMO of the donor.

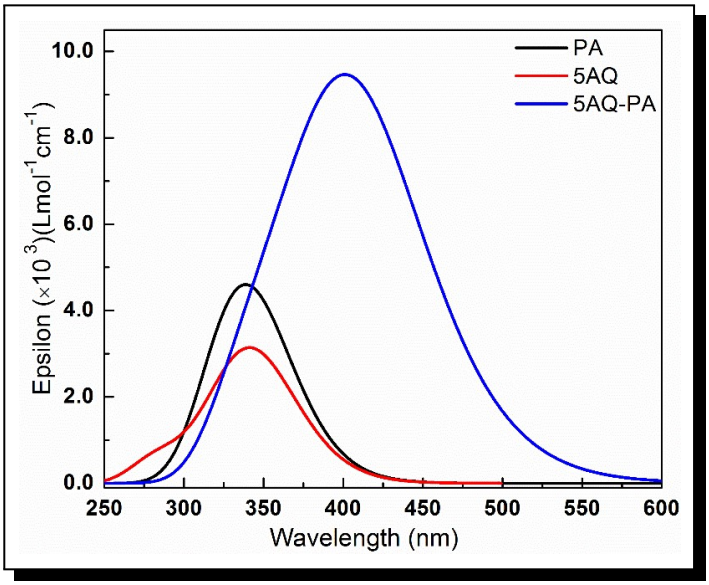


Figure 3. Estimated absorption spectra of 5AQ, PA, and their complex in $CHCl_3$ using IEFPCM/TD-DFT/B3LYP/631-G(d,p) method

4. Conclusion

The H-bonded CT complex of 5AQ (donor) with PA (π-acceptor) was studied based on the density functional theory (DFT) and time-dependent density functional theory (TD-DFT) at B3LYP level using 6-31G(d,p) basis set. The interaction energy of the studied complex was calculated in the gas phase and $CHCl_3$, supporting the formation of a stable complex. A new absorption band was noticed in the absorption spectrum where neither the donor nor the acceptor absorbed. Thus the theoretical study of the considered complex provides valuable information about its molecular structure, possible interaction sites and electronically excited states.

Acknowledgment

Author NP acknowledges the University Grants Commission (UGC), New Delhi, India for financial support under the research meritorious students in the sciences (RFSMS) scheme (F. No. 25-1/2014-15(BSR)5–53/2007. Author MSM is thankful to Prof. Keli Han, Dalian Institute of Chemical Physics, CAS, China for providing the Gaussian 09 facilities.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] K. M. Al-Ahmary, M. M. Habeeb and S. H. Aljahdali, Synthesis, spectroscopic studies and DFT/TD-DFT/PCM calculations of molecular structure, spectroscopic characterization and NBO of charge transfer complex between 5-amino-1,3-dimethylpyrazole (5-ADMP) with chloranilic acid (CLA) in different solvents, *Journal of Molecular Liquids* **277** (2019), 453 – 470, DOI: 10.1016/j.molliq.2018.12.072.

[2] S. Amirat, F. Madi, M. Bououdina, A. Gheid, A. Zaboub, L. Nouar and R. Merdes, Computational study on intermolecular charge transfer complex of 2,2′-bipyridine with picric acid: TD-DFT, NBO and QTAIM analysis, *Materials Research Express* **6** (2019), 75104, DOI: 10.1088/2053-1591/ab1507.

[3] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 09, Revision A.02*, Gaussian, Inc., Wallingford CT (2009), URL: `https://gaussian.com/g09citation/`.

[4] A. Karmakar, P. Bandyopadhyay, S. Banerjee, N. C. Mandal and B. Singh, Synthesis, spectroscopic, theoretical and antimicrobial studies on molecular charge-transfer complex of 4-(2-thiazolylazo)resorcinol (TAR) with 3,5-dinitrosalicylic acid, picric acid, and chloranilic acid, *Journal of Molecular Liquids* **299** (2020), 112217, DOI: 10.1016/j.molliq.2019.112217.

[5] R. Kavitha, S. Nirmala, R. Nithyabalaji and R. Sribalan, Biological evaluation, molecular docking and DFT studies of charge transfer complexes of quinaldic acid with heterocyclic carboxylic acid, *Journal of Molecular Structure* **1204** (2020), 127508, DOI: 10.1016/j.molstruc.2019.127508.

[6] L. Miyan and A. Ahmad, Spectroscopic and spectrophotometric studies on hydrogen bonded charge transfer complex of 2-amino-4-methylthiazole with chloranilic acid at different temperatures, *Journal of Molecular Liquids* **262** (2018), 514 – 526, DOI: 10.1016/j.molliq.2018.04.084.

[7] R. S. Mulliken and W. B. Person, Molecular compounds and their spectra. XXI. Some general considerations, *Journal of American Chemical Society* **91** (1969), 3409 – 3413, DOI: 10.1021/ja01041a001.

[8] N. Pandey, M. S. Mehata, N. Fatma and S. Pant, Efficient fluorescence quenching of 5-aminoquinoline: Silver ion recognition study, *Journal of Luminescence* **205** (2020), 475 – 481, DOI: 10.1016/j.jlumin.2018.09.062.

[9] N. Pandey, M. S. Mehata, N. Fatma and S. Pant, Modulation of Fluorescence properties of 5-Aminoquinoline by Ag^{+} in aqueous media via charge transfer, *Journal of Photochemistry and Photobiology A: Chemistry* **396** (2020), 112549, DOI: 10.1016/j.jphotochem.2020.112549.

[10] S. Shakya and I. M. Khan, Charge transfer complexes: Emerging and promising colorimetric real-time chemosensors for hazardous materials, *Journal of Hazardous Materials* **403** (2021), 123537, DOI: 10.1016/j.jhazmat.2020.123537.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

IoT based prediction of water quality index for farm irrigation

Rajesh Kumar Yadav
Department of Computer Science and Engineering
Delhi Technological University
New Delhi-110042, India
rkyadav@dtu.ac.in

Adarsh Jha
Department of Computer Science and Engineering
Delhi Technological University
New Delhi-110042, India
adarshjhaa100@gmail.com

Aditya Choudhary
Department of Computer Science and Engineering
Delhi Technological University
New Delhi-110042, India
adi20gamer@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395921

***Abstract*—Agriculture sector of Indian economy in which more than half of the population is involved, contributes to less than quarter of the GDP. With advancement in ICT, tools and techniques can be developed that can help in analyzing and automating various phases of farming for improving productivity. This work focuses on analysing the quality of irrigation water and developing a model for prediction of Irrigation Water Quality Index(IWQI) based on Salinity and Sodicity. Development of IWQI can save time and cost of lab tests for irrigation water. Five parameters of water Na^{+}, Cl^{-}, EC, HCO^{3-} and SAR are measured using which the IWQI is calculated. These five parameters are further reduced to three parameters using correlation analysis and a classification model for prediction of water quality class is developed using various classification techniques. Best result is obtained by Random Forest Classifier followed by Gradient Boosting and Neural Network Classifier. The classification model can be used in IoT based farming systems for preventing salinity based damage to the crops.**

***Index Terms*—Irrigation Water Quality Index, Precision Agriculture, Water salinity, Sensor Data**

I. Introduction

In a survey conducted in 2018, it was found that 50% of the workforce in India is involved in the agriculture sector, however its contribution is only 16% to the GDP [1].There exists a need to increase the efficiency of each stage in farming and at a cheap cost so that it is affordable. With improvement in technology, these needs can be addressed using innovative solutions like the Internet of Things(IoT). This technology allows us to use sensors which in turn can be used to obtain large datasets for physical and chemical characteristics of soil, water and weather. This data can be analysed and machine learning tools can be developed that can help in proper utilization of resources and increase the overall efficiency of crop production.

We require proper water quality standards for irrigation to improve crop yield and help maintain the soil productivity. Deficiency or excess of nutrients in soil also depends on water used for irrigation which effects the plant growth. However, most existing methods for water quality standards are for drinking water quality which is not suitable for plants and some methods which do exist for irrigation water quality use high number of parameters which ist't economically feasible for a farmer.

In this paper, we develop a model for prediction of Irrigation Water Quality Index(IWQI) class based on the degree of sodicity and salinity of water. The contribution of our work are as follows:

- We convert 5 parameters to quality measurement values which are used to calculate IWQI. We construct IWQI classes using IWQI values for irrigation water.
- We reduce 5 parameters to 3 parameters using correlation analyis to save costs of sensors required to implement the proposed work. These parameters are used to classify IWQI using seven classification techniques which are Support Vector Classifier, Neural Networks, Gradient Boosting , Random Forest, Decision Tree, Bagging and Naive Bayes classifier.
- We evaluate and choose the best performing classification algorithm to obtain IWQI class for water sample with 3 parameters.

The rest of the paper is organized as follows: Section II describes other works related to this paper, section III discusses the preliminary topics, Section IV describes the database used and algorithm of the model, Section V describes the results and finally Section VI gives the conclusion and future work.

II. Related Works

With advancement in IoT, many new tools and techniques have been developed for agriculture sector. The term Precision agriculture [2] was developed in which various types of data such as time series, geolocation, sensor etc. are gathered and analysed to improve the efficiency of agricultural practices. Irrigation is one of the important phases of agriculture which

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:35:18 UTC from IEEE Xplore. Restrictions apply.

involves adding water to soil after plantation which needs proper analysis of chemicals in water and quality of water for proper plant growth. Rajesh et al [3] proposed an automated system that measures the moisture, temperature and humidity of farms measured using sensors to determine the optimal water volume required for Irrigation. Mark et al. [4] studied the effect of saline water on plants and proposed methods for using them for irrigation while Shrivastav et al. [5] studied the effect of salinity on soil nutrients which decreases with increase in salinity due to its ability to reduce soil permeability.

Irrigation Water quality Indices(IWQIs) are one way to aggregate various parameters of water samples into a single value and determine their suitability for irrigation. Meireles et al [6] developed a WQI for irrigation water (which mainly focused on salinity and sodicity of water) using factor and principal component analysis to reduce thirteen parameters of water samples to just five parameters. Sayiter et al [7] used an Irrigation Water Quality Index(IWQI) using Sodium Adsorption Ratio (SAR), Kelly Index(KI), sodium percentage and Permeability Index(PI) and developed a model using Neural Networks for prediction of IWQI. Singh et al [8] uses index based approach for predicting water quality based on 12 parameters to classify water into 5 categories: excellent, good, medium, bad, and very bad. For performing this classification, they use multiple criteria decision analysis (MCDA) approach, indicating the importance of IWQI for managing water quality index. Hossam et al [9] used various dimensional reduction techniques such as Principal Component Analysis(PCA) and Factor Analysis(FA) to reduce 22 water quality parameters to 7 parameters:Sodium, Zinc, pH, Bicarbonate, Astantine, Boron and Nitrate and derived their weights using PCA. In the proposed work, we use the the IWQI developed by Meireles et al [6] and develop a classification model to predict the IWQI class.

III. Preliminaries

A. Salinity Based Irrigation Water Quality

Calculation of water quality for irrigation is necessary to study its effect on plant and soil health. Water Quality Index (WQI) is used for representing the overall quality of water. It is the aggregate of various chemical and physical parameters of water and is calculated by multiplying relative weights of parameters with parameter value. The parameters chosen should be able to be measured in all the water sources used for irrigation. In this work, we use the Irrigation Water Quality Index (IWQI) based on salinity and sodicity of water. The following subsections describe the water quality parameters, their weights and formula for calculating IWQI.

1) Parameters for IWQI: The IWQI used in this work was formulated by Meireles et al who reduced 13 parameters of water using Factor Analysis(FA) and Principal Component Analysis(PCA) to 5 parameters which are: Sodium, Chloride and Bicarbonate ion concentration, Sodium Absorption Ratio (SAR) and Electrical Conductivity(EC) [10]. Electrical conductivity is generally measured by passing current through solutions and the value of SAR is given by the combination of Calcium, Magnesium and Sodium ion concentration whose formula is given by Equation 1. It is the primary indicator of sodicity of water which affects the permeability of soil and water infiltration rate [11]. Both cases of water infiltration rate being too high or too low affects the growth of plants. The weights of the 5 parameters, which represents their contribution to water quality are converted to relative weights and are mentioned in Table I.

$$SAR = \frac{Na^{+}}{\sqrt{\frac{Ca^{2+}+Mg^{2+}}{2}}} \tag{1}$$

Here,
Na^{+}: Concentration of Sodium ion
Ca^{2+}: Concentration of Calcium ion
Mg^{2+}: Concentration of Magnesium ion

TABLE I
WEIGHTS FOR IWQI PARAMETERS

Parameters	Weights
Electrical Conductivity(EC)	0.211
Sodium(Na^{+})	0.204
Bicarbonate(HCO^{3-})	0.202
Chloride(Cl^{-})	0.194
Sodium Absorption Ratio (SAR)	0.189

2) Quality Measurement Values and IWQI: The value of parameters measured for water sample should neither be too high nor too low for optimal growth of plants due to which the parameters are normalized(between 0 to 100) to quality measurement values (q_i) according to limits shown in Table III [10] and Equation 2. The q_i values obtained are multiplied with relative weights of their respective parameters and finally added to obtain IWQI according to Equation 3 whose value is between 0 and 100. IWQI range is divided into various classes and each class represents the salinity characteristic.

$$q_i = q_{imax} - \frac{(x_{ij} - x_{inf}) * q_{iamp}}{x_{amp}} \tag{2}$$

Here,
q_i: quality measurement values,
q_{imax}: maximum value in particular quality measurement class,
q_{iamp}: amplitude of quality measurement class,
x_{ij}: value of parameter x,
x_{inf}: minimum value of parameter x in quality measurement class,
x_{amp}: It is the class amplitude to which the parameter belong

$$IWQI = \sum_{i=1}^{n} q_i w_i \tag{3}$$

Here,
q_i: quality measurement value ,
w_i: weight of parameter from Table I

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:35:18 UTC from IEEE Xplore. Restrictions apply.

TABLE II
CLASSIFICATION OF WATER SAMPLE BASED ON IWQI

IWQI	Soil	Plant
85–100	Can be used for any kind of soil	Most plants won't be affected
70–85	Can be used on soil with moderate permeability	Avoid use in plants with very low salt tolerance
55–70	Can be used on soils with moderate to high permeability	Avoid in plants with low salt tolerance
40–55	Can be used on soils with high permeability without dense layers	Used mainly in plants with high salt tolerance. Plants with moderate salt tolerance can be used with some control practices
0–40	Use for irrigation should be avoided	Avoid for all plants

TABLE III
LIMITING VALUES FOR QUALITY MEASUREMENT

q_i	EC	SAR	Na^+	Cl^-	HCO^{3-}
85-100	$0.2 \leq EC < 0.75$	$2 \leq SAR < 3$	$2 \leq Na^+ < 3$	$1 \leq Cl^- < 4$	$1 \leq HCO^{3-} < 1.5$
60-85	$0.75 \leq EC < 1.5$	$3 \leq SAR < 6$	$3 \leq Na^+ < 6$	$4 \leq Cl^- < 7$	$1.5 \leq HCO^{3-} < 4.5$
35-60	$1.50 \leq EC < 3$	$6 \leq SAR < 12$	$6 \leq Na^+ < 9$	$7 \leq Cl^- < 10$	$4.5 \leq HCO^{3-} < 8.5$
0-35	$EC <$ 0.20 or EC $\geq$ 3.00	$SAR <$ 2 or SAR$\geq$ 21	$Na^+ <$ 2 or $Na^+ \geq$ 9	$Cl^- <$ 1 or $Cl^- \geq$ 10	$HCO^{3-} <$ 1 or $HCO^{3-} \geq$ 8.5

B. Classification Algorithms

1) ANN: In ANN, we have an input layer, hidden layers and output layer each consisting of neurons in it [12]. Neurons receive some input and give an output after applying an activation function. The error is backpropagated to network for adjusting weights. This classifier is also known as MLP(Multilayer Perceptron Classifier).

2) SVM: This classification algorithm is very useful as it tries to maximize the margin between any two classes which makes the decision hyperplane more accurate [13].

3) Gradient Boost Classifier: It is a classifier which uses loss function, weak learners and an additive component to reduce the loss [14]. This algorithm uses decision trees as its weak learners. It is quite powerful but is prone to overfitting.

4) Random Forest Classifier: It is an ensemble learning method which uses multiple trees [15]. Here each classifier gives its own output and the output which is chosen by more number of classifiers becomes the final output.

5) Decision Tree: It is a classifier which uses a tree-like structure where at each node a decision is made according to the feature in the dataset and final output is given by the leaf node [16].

6) Naive Bayes: It is a simple classifier which uses Bayes theorem with an exception that features do not depend on each other [17]. This allows us to easily implement and use the algorithm for classification but its results are not reliable.

7) Bagging Classifier: This algorithm uses weak learners like decision trees on a random portion of its dataset, using the results from those trees to produce a final result [18].

IV. PROPOSED WORK

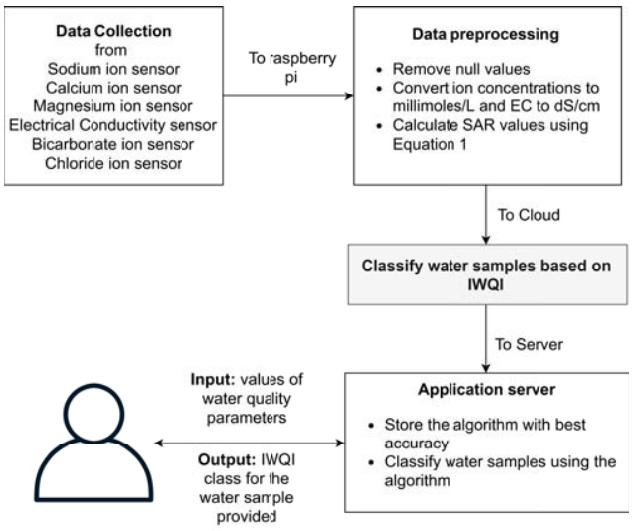


Fig. 1. Overall architecture

In this section, we propose a classification model for prediction based on the Irrigation Water Quality(IWQI) class described in section III-A2. The model can be incorporated into agricultural IoT systems for analysing water samples for irrigation purposes as shown in Fig. 1. The sensor network will be used to measure the values of various water quality parameters namely electrical conductivity and ion concentrations for Sodium, Magnesium, Calcium, Bicarbonate and Chloride. Data from sensor network is sent to a local machine like Raspberry pi that involves processing the data and sending it to the cloud. The cloud analyses the data by predicting the relationship between water quality parameters and the IWQI classes. Result of classification can be saved to an application server than can interact with users through web or mobile applications. In this work, we have focused on the data analysis part which involves developing the classification model described in section IV-B used the water quality data from major ions dataset [19] that is described in section IV-A.

A. Dataset - Major ions dataset

The dataset used in our work was obtained from US Geological Survey of Brackish Groundwater [19]. It consists of 66 thousand rows and contains values for concentration of dissolve solids, metals, nutrients, ions and physical properties such as conductivity out of which only EC, Cl^-, Na^+, HCO^{3-}, Ca^{2+} and Mg^{2+} are used for IWQI. EC stands for electrical conductivity which is measured in milliSiemens/cm. Cl^-, Na^+, HCO^{3-}, Ca^{2+} and Mg^{2+} represents Chloride, Sodium, Bicarbonate, Calcium and Magnesium ion concentration respectively. These concentrations were measured in mg/L which we converted to millimoles/L. EC needs to be in dS/cm so it

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:35:18 UTC from IEEE Xplore. Restrictions apply.

was divided by 100. We use Ca^{2+}, Mg^{2+} and Na^{+} to calculate SAR using Equation 1.

B. Irrigation Water Quality Index based classification

In this work, we develop a classification model and apply it on major ions dataset. This dataset consists of all the necessary features required to obtain the IWQI described in Section III-A1 and the model is built using these three steps: converting given values to quality measurement values, calculating irrigation water quality index(IWQI) using calculated values and relative weights of each parameter, assigning to a class according to ranges given in Table II and finally using different classification techniques and choosing the best of them.

1) Quality Measurement Values: The dataset consists of parameters with different units. Also, the value of any parameter being abnormally high or low decrease the quality of water. Hence, we need to normalize the values according to predefined limits of parameters. This is done by converting values of parameters to quality measurement values (q_i) that can be obtained using Equation 2 and Table 3. The quality measurement value (q_i) is a dimensionless quantity which can then be combined with other parameters without considering units. The Table IV indicates the number of values present in specific ranges of q_i for the dataset used. Here, we can see that majority samples have q_i values of EC in the range 85-100, bicarbonate in the range 35-85 and Cl^{-}, Na^{+} and SAR in the range 0-35. This indicates that most of the samples have optimal value of EC and either excess or deficit value of Cl^{-}, Na^{+} and SAR.

TABLE IV
NUMBER OF QUALITY MEASUREMENT VALUES IN EACH RANGE

q_i	Q_HCO^{3-}	Q_EC	Q_Cl^{-}	Q_Na^{+}	Q_SAR
0-35	5540	6580	32700	31400	32600
35-60	9950	0	0	0	794
60-85	20000	3500	0	2780	1950
85-100	2470	27900	5280	3810	2680

2) Calculating IWQI and developing a classification model: Using the quality measurement values calculated in section III-A2, we multiply them by their relative weights mentioned in Table I and add them to obtain IWQI according to Equation 3. The samples are further classified into quality classes according to quality ranges provided in Table II. Correlation matrix of quality parameters which was obtained using Pearson's Correlation is given in Table V.The following observations are made:

1) Cl^{-} is highly correlated with EC and Na^{+}.
2) Na^{+} is highly correlated with all the parameters.
3) EC is highly correlated with all the parameters except SAR.
4) HCO^{3-} is highly correlated with EC.
5) IWQI is highly correlated with Cl^{-}, Na^{+} and EC.

$$r = \frac{\sum(U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\sum(U_i - \bar{U})^2(V_i - \bar{V})^2}} \quad (4)$$

Here,
r: correlation coefficient
U_i: value of U-variable in the sample
$\bar{U}$: mean value of U-variable
V_i: value of V-variable in the sample
$\bar{V}$: mean value of V-variable

TABLE V
CORRELATION MATRIX FOR VARIOUS PARAMETERS FOR WATER SAMPLES

	Cl^{-}	Na^{+}	EC	HCO^{3-}	SAR	IWQI
Cl^{-}	1	0.467	0.506	0.159	0.083	0.44
Na^{+}	0.467	1	0.482	0.297	0.56	0.607
EC	0.506	0.482	1	0.835	0.018	0.423
HCO^{3-}	0.159	0.297	0.835	1	0.018	0.293
SAR	0.083	0.56	0.018	0.018	1	0.302
IWQI	0.44	0.607	0.423	0.293	0.302	1

We used different combination of parameters according to their correlation with IWQI value as features to the classification model in which Na^{+}, Cl^{-} and EC were found to be most suitable for prediction IWQI class. The classification result obtained using the three parameters are almost similar to that obtained using all of the five parameters. Therefore, correlation analysis is sufficient and other feature reduction techniques are not required. Choosing three parameters saves us the cost of measuring HCO^{3-} and SAR. The classification algorithms mentioned in Section III-B are used to predict quality classes and the method having best accuracy and F2 score is selected for further use. Multiple classification techniques have been used to cover different traditional, ensemble and deep learning based techniques whose performance can help in predicting the most suitable type of classification technique for the given dataset. The overall algorithm is mentioned in algorithm IV-B2 and flowchart in Fig. 2.

Classification of water samples based on IWQI

Input: Water quality dataset with 5 parameters Cl^{-}, Na^{+}, EC, HCO^{3-} and SAR
Output: algorithm with best score
1: Calculate quality measurement values(q_{param}) for each parameter.
2: Calculate IWQI using the Equation 3 and assign classes from Table II
3: Choose features of dataset as: features $\leftarrow Cl^{-}$,Na^{+},EC
4: Choose labels of dataset as:IWQI
5: Perform train test split on the dataset
6: Apply the 7 classifiers mentioned in Section III-B on the training set and evaluate on the testing set using accuracy and F1 score.
7: **return** algorithm with best score

V. EXPERIMENTS AND RESULTS

This section contains the simulation settings for implementing the classification model in V-A, the evaluation methods for the model V-C and the results of algorithms in section V-D.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:35:18 UTC from IEEE Xplore. Restrictions apply.

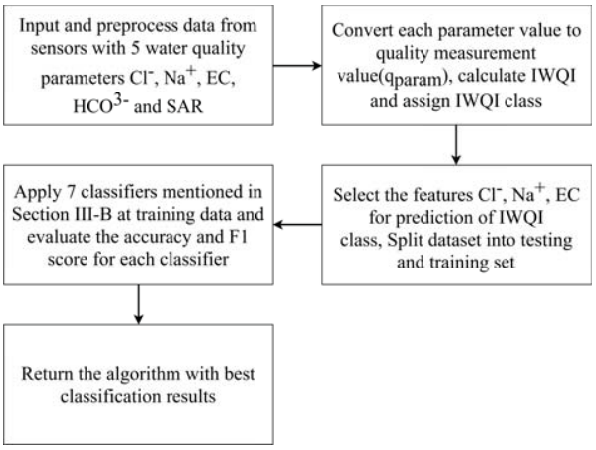


Fig. 2. Flowchart for classification of water sample based on IWQI

TABLE VI
PARAMETER SETTINGS FOR CLASSIFIERS

Classifier	Parameter Settings
Bagging Classifier	base_estimator:SVC, n_estimators:10
DecisionTree	criterion: gini, splitter: best, min_samples_split: 2
Naive Bayes	alpha: 1.0, class_prior: None, class_count: [7059. 15826. 5057. 2055. 402.]
Gradient Boosting	loss: deviance, learning rate: 0.1, n_estimators: 100
Random Forest	n_estimators: 100, criterion: gini, base_estimator_: DecisionTreeClassifier
SVM	C: 1.0, decision_function_shape: 'ovr', tol:0.001
MLP	hidden_layer_sizes:128, activation='relu', learning_rate=0.001

A. Tools and Simulation Settings

The IoT system for classifying water samples based on IWQI values that is shown in Fig. 1 requires measurement of various water quality parameters mentioned in section III-A1. As we are focused on the classification part, we obtain the parameters from the Major ions dataset described in section IV-A. The algorithm was implemented in Python 3 where data preprocessing is performed using Pandas [20] and Numpy [21]. Various classification algorithms are applied using Scikit-learn. The parameter settings for these algorithms are mentioned in section V-B.

B. Parameter Settings

The parameter settings for all the classification models are given in Table VI and the parameters are described below:

1) Decision Tree: criterion: function to evaluate split quality, splitter: method used to split at each node, min_sample_split: least amount of samples required to split a non-leaf node

2) Naive Bayes: alpha: the additive smoothing parameter, class_prior: estimated probability for new class prediction, class_count: number of samples under each class

3) Gradient Boosting: loss: function to optimize, learning_rate: for changing contribution of trees to the result, N_estimators: levels of boosting

4) Random Forest: n_estimators: Amount of tree before voting, criterion: function to evaluate split quality, base_estimator: base classifier to create sub estimator

5) Support vector classifier: C: Regularization, decision_shape: decision function's shape, tol: stopping tolerance

6) Bagging Classifier: base_estimator: base classifier used to fit on random subdomains, n_estimator: count of the base_estimator

7) MLP Classifier: hidden_layer_sizes: array representing sizes of hidden layers, activation: output function of hidden layer, learning_rate: step size of gradient descent

C. Evaluation Metrics For Classification Model

For testing our classifier, we use precision, accuracy, recall and F1 score. These metrics are calculated to test the reliability and robustness of classifiers used . These metrics are defined in Section V-C1, V-C2, V-C3 and V-C4 [22]. The equations for the same are given at 5, 6, 7 and 8.

1) Accuracy: It tell us how much data can be predicted accurately from a given classifier. This gives us an idea about the bulk of data which is correctly classified.

2) Precision: It tells us how many predictions are correct out of all observations saying output belongs to a particular class.

3) Recall: It tells us how many correct predictions were made for a class out of all the actual observations for a given class.

4) F1 Score: It can be defined as harmonic mean of precision and recall, can be used independently for evaluation of classifiers.

$$Accuracy = \frac{TPos + TNeg}{TPos + TNeg + FPos + FNeg} \tag{5}$$

$$Precision(P) = \frac{TPos}{TPos + FPos} \tag{6}$$

$$Recall(R) = \frac{TPos}{TPos + FNeg} \tag{7}$$

$$F1Score = 2 * \frac{P * R}{P + R} \tag{8}$$

Here,
TPos: Number of observations which are correctly classified for positive class in binary classification
TNeg: Number of observations which are correctly classified for Negative class in binary classification
FPos: Number of observations which are incorrectly classified for positive class in binary classification
FNeg: Number of observations which are incorrectly classified for negative class in binary classification

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:35:18 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
CLASSIFICATION RESULTS FOR ALGORITHM IV-B2

Methods	Accuracy	Precision	Recall	F1
DecisionTree	0.832	0.728	0.740	0.733
Naive Bayes	0.526	0.105	0.200	0.138
Gradient Boosting	0.858	0.762	0.757	0.760
Random Forest	**0.869**	**0.790**	**0.765**	**0.776**
SVM	0.845	0.762	0.711	0.726
Bagging	0.813	0.750	0.664	0.690
MLP	0.846	0.734	0.743	0.738

TABLE VIII
FIVE FOLD CROSS VALIDATION SCORES

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Bagging Classifier	0.829	0.823	0.814	0.830	0.813
Decision Tree	0.833	0.836	0.828	0.823	0.839
Naive Bayes	0.518	0.518	0.518	0.518	0.519
Gradient Boosting	0.862	0.866	0.869	0.859	0.862
Random Forest	**0.866**	**0.870**	**0.870**	**0.862**	**0.871**
SVM	0.853	0.849	0.846	0.841	0.848
MLP	0.847	0.853	0.857	0.847	0.849

D. Results of Classification based on IWQI

Algorithm IV-B2 is applied to major ions dataset with Electrical Conductivity, Cl^- and Na^+ as features and IWQI class as labels and results are evaluated using the metrics mentioned in Section V-C, details of which are mentioned in Section IV-B2. For the major ions dataset, the number of water samples in IWQI range 85-100 is around 1.3% of the total samples which affected the results of correctly classifying samples belonging to that range. This however does not adversely affects the overall result. Random Forest performed the best with an accuracy of 86.9% followed by Gradient Boosting and Neural Networks with accuracy of 85.8% and 84.6% respectively. Naive Bayes classifier performed the worst with accuracy of 52.6%. The results of all the algorithms are given in Table VII. We also performed five-fold cross validation and computed the accuracies whose results are in Table VIII. This concludes that all of the methods except Naive Bayes perform well in classifying the water samples.

VI. CONCLUSION AND FUTURE WORK

Analysis of irrigation water quality can help in increasing agricultural productivity and prevent damage to plants. In this work, we developed a classification model for prediction of Irrigation Water Quality Index (IWQI) class. The IWQI used here is an aggregate of five parameters that are SAR, EC, Chloride, Bicarbonate and Sodium ion concentration respectively. For classification, we used seven classification algorithms and selected three out of five parameters: EC, Cl^- and Na^+ for predicting IWQI class. Random Forest Classifier performed the best followed by Gradient Boosting and Neural Networks. The water quality index used here is mainly representation of salinity of water and the dataset used here is of groundwater. Other parameters such as acidity, oxygen demand etc. can be used in addition to the parameters used here to develop an index covering more factors. Also, dataset covering different types of water bodies can be used to improve the model. The proposed work used here can be incorporated into IoT based agricultural systems for analysing irrigation water quality that will be faster and economically feasible compared to manual lab tests. It can also help in deciding the water sample according to crop and soil properties.

REFERENCES

[1] S. Sunder, "India economic survey 2018: Farmers gain as agriculture mechanisation speeds up, but more r&d needed," *The Financial Express*, 2018.
[2] N. Zhang, M. Wang, and N. Wang, "Precision agriculture—a worldwide overview," *Computers and electronics in agriculture*, vol. 36, no. 2-3, pp. 113–132, 2002.
[3] R. K. Yadav, R. Dev, T. Mann, and T. Verma, "Iot based irrigation system," *International Journal of Engineering & Technology*, vol. 7, no. 4.5, pp. 130–133, 2018.
[4] M. E. Grismer and K. M. Bali, "Drought tip: Use of saline drain water for crop production," 2015.
[5] P. Shrivastava and R. Kumar, "Soil salinity: a serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation," *Saudi journal of biological sciences*, vol. 22, no. 2, pp. 123–131, 2015.
[6] D. La Mora-Orozco, H. Flores-Lopez, H. Rubio-Arias, A. Chavez-Duran, J. Ochoa-Rivero *et al.*, "Developing a water quality index (wqi) for an irrigation dam," *International journal of environmental research and public health*, vol. 14, no. 5, p. 439, 2017.
[7] S. Yıldız and C. B. Karakuş, "Estimation of irrigation water quality index with development of an optimum model: a case study," *Environment, Development and Sustainability*, pp. 1–16, 2019.
[8] S. Singh, N. Ghosh, S. Gurjar, G. Krishan, S. Kumar, and P. Berwal, "Index-based assessment of suitability of water quality for irrigation purpose under indian conditions," *Environmental monitoring and assessment*, vol. 190, no. 1, p. 29, 2018.
[9] H. S. Jahin, A. S. Abuzaid, and A. D. Abdellatif, "Using multivariate analysis to develop irrigation water quality index for surface water in kafr el-sheikh governorate, egypt," *Environmental Technology & Innovation*, vol. 17, p. 100532, 2020.
[10] A. C. M. Meireles, E. M. d. Andrade, L. C. G. Chaves, H. Frischkorn, and L. A. Crisostomo, "A new proposal of the classification of irrigation water," *Revista Ciencia Agronomica*, vol. 41, no. 3, pp. 349–357, 2010.
[11] R. S. Ayers, D. W. Westcot *et al.*, *Water quality for agriculture*. Food and Agriculture Organization of the United Nations Rome, 1985, vol. 29.
[12] M. H. Hassoun *et al.*, *Fundamentals of artificial neural networks*. MIT press, 1995.
[13] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
[14] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[16] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.
[17] K. P. Murphy *et al.*, "Naive bayes classifiers," *University of British Columbia*, vol. 18, p. 60, 2006.
[18] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
[19] S. Qi and A. Harris, "Geochemical database for the brackish groundwater assessment of the united states," *US Geological Survey data release*, 2017.
[20] W. McKinney *et al.*, "pandas: a foundational python library for data analysis and statistics," *Python for High Performance and Scientific Computing*, vol. 14, no. 9, pp. 1–9, 2011.
[21] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in science & engineering*, vol. 13, no. 2, pp. 22–30, 2011.
[22] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:35:18 UTC from IEEE Xplore. Restrictions apply.

Physics of Plasmas

Kinetic theory of effect of dust charge fluctuations on the parametric decay of lower hybrid wave instability by relativistic runaway electrons in tokamak

Cite as: Phys. Plasmas **28**, 043701 (2021); https://doi.org/10.1063/5.0041282
Submitted: 27 December 2020 . Accepted: 15 March 2021 . Published Online: 08 April 2021

Jyotsna Sharma, Suresh C. Sharma, and Ajay Gahlot

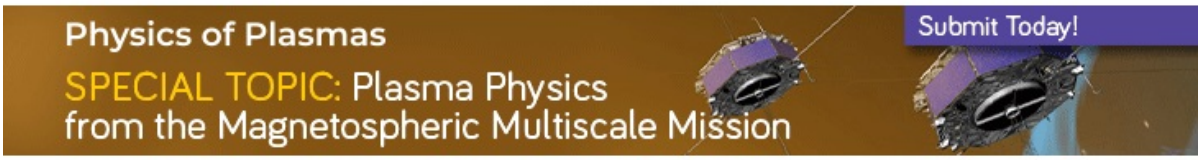
View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

A non-local fluid closure for modeling cyclotron resonance in collisionless magnetized plasmas
Physics of Plasmas **28**, 042105 (2021); https://doi.org/10.1063/5.0045335

On the fast waves in a cylindrical current-carrying plasma
Physics of Plasmas **28**, 042106 (2021); https://doi.org/10.1063/5.0022323

Wave trapping and E × B staircases
Physics of Plasmas **28**, 042302 (2021); https://doi.org/10.1063/5.0042930



AIP Publishing

© 2021 Author(s).

Physics of Plasmas ARTICLE scitation.org/journal/php

Kinetic theory of effect of dust charge fluctuations on the parametric decay of lower hybrid wave instability by relativistic runaway electrons in tokamak

Cite as: Phys. Plasmas **28**, 043701 (2021); doi: 10.1063/5.0041282
Submitted: 27 December 2020 · Accepted: 15 March 2021 ·
Published Online: 8 April 2021

View Online Export Citation CrossMark

Jyotsna Sharma,[1] Suresh C. Sharma,[2,a)] and Ajay Gahlot[3]

AFFILIATIONS

[1]Department of Physics, Amity School of Applied Sciences, Amity University Haryana, Manesar, Gurugram 122051, India
[2]Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi 110042, India
[3]Department of Applied Sciences, Maharaja Surajmal Institute of Technology, New Delhi 110058, India

[a)]Author to whom correspondence should be addressed: suresh321sharma@gmail.com

ABSTRACT

Dust charge fluctuation effect on the parametric decay of lower hybrid wave instability by relativistic runaway electrons is studied in a tokamak using kinetic treatment. Parametric upconversion of lower hybrid pump waves into relativistic runaway electron mode and upper sideband mode is described. A ponderomotive force is exerted on the runaway electrons by lower hybrid pump wave possessing large amplitude and upper sideband wave, which drives the runaway electron mode. The coupling of the oscillatory velocity of electrons with density perturbations produces nonlinear density perturbations on the upper sideband frequency. As a result, runaway generation is enhanced by a lower hybrid wave and the growth rate of the instability is measured as the square of the amplitude of the pump wave. Moreover, the presence of dust charge fluctuations and their number density in the tokamak have an appreciable effect on the growth rate of lower hybrid wave instabilities, which in turn affects the actual performance of the ITER due to the potential safety and operational issues.

Published under license by AIP Publishing. https://doi.org/10.1063/5.0041282

I. INTRODUCTION

Tokamak is a device that is being used to confine the plasma in the shape of a torus. In the near future, this confined plasma can be heated to utilize the energy released in producing electricity. However, the relativistic electron runaway (RE) having energies up to several MeV generates very severe trouble in tokamaks such as ITER (International Thermonuclear Experimental Reactor), FTU (Frascati Tokamak Upgrade), HT-7 (Heifei Tokamak-7), JET (Joint European Torus), and JT-60U (Japan Torus). REs can lead to damage of the first wall structure on impact. These REs can be generated via primary mechanism through Dreicer acceleration and secondary mechanism through an avalanche effect. For ITER, the avalanche effect dominates, which turns two-thirds of the pre-disruption current into the runaway current. When the condition of Čerenkov resonance is satisfied, lower hybrid waves increase the runaway by generating fast electrons through Landau damping. The plasma surface interactions result in the erosion of wall components of the tokamak, which produces some impurities in the plasma and degrades the wall. While due to the introduction of atomic impurities, dust particles of various shapes and sizes are also introduced in the plasma, which can severely affect the tokamak operation. Dust particles having high Z impurities like tungsten can become radioactive when they come into contact with tritium. Therefore, this type of dust present in the tokamak can spread serious health problems. Similarly, dust having large surface area can act as a catalyst or, due to its chemical activity, it increases the risk of explosion. Therefore, the study of produced dust particles and their sizes is very important for the fusion community in order to enhance the efficiency of tokamaks. Several researchers[1–28] have performed experiments to study the production of runaway electrons and dust charge fluctuations in tokamak. Helander *et al.*[1] have given the basic theory of runaway electron production, where the mechanisms of primary and secondary runaway are described and their dynamics has been analyzed during a tokamak disruption. Kuley and Tripathi[2] have studied the parametric decay of lower hybrid waves into two modes: (i) low

frequency electrostatic mode and (ii) lower hybrid upper sideband in tokamak. In their study, a kinetic theory of the process is described by the authors and the electron acceleration due to the parametric decay is discussed. Yoshino and Tokuda[10] have explained the behavior of runaway electrons in tokamak discharges, especially in the following three types of magnetic turbulence: the first is micromagnetic turbulence, the second is low-m/n magnetic islands in a sea of stochasticity, and the third is macroscale magnetic turbulence. Chen *et al.*[12] have studied the enhanced runaway production during ion Bernstein waves heating in the lower hybrid current driven plasmas as compared to that of lower hybrid current driven plasma in the Heifei Tokamak-7. Winter[21] has described the small particles (dust) existing in the magnetic confinement fusion devices. Krasheninnikov *et al.*[23] have described the role of dust in fusion plasma, and modeling of the dusty plasma in fusion is also well described. Zeng *et al.*[24] have studied the J-TEXT tokamak, where they explained the disruption-generated runaway electrons. They found that when intended disruption occurs with the use of argon injection, then the runaway electron plateau can be obtained more effortlessly by using high loop voltage as well as with the shorter beginning time of the high loop voltage. The authors also observed magnetic fluctuations during the disruptions in the beginning of the current quench. Liu *et al.*[25] have discussed the role of kinetic instabilities in elevated critical electric fields and runaway-electron avalanches. The authors found that by controlling the temperature and density of the background plasma, plasma waves can instinctively be energized during tokamak disruptions. In this way, the avalanche generation of runaway electrons can be cloaked. McDevitt and Tang[26] reported the runaway electron generation in axisymmetric tokamak geometry. Marco *et al.*[27] presented a sliding mode controller (SMC), a variable structure control scheme that is based upon the sliding surfaces to solve the plasma current control problem. Fulop *et al.*[28] found that lengthened plasmas do not produce such large runaway currents because of two reasons: first, because of the small induced electric fields coupled with larger plasmas, and second, because of the drop in the runaway avalanche gain due to the direct shaping effects.

In the present paper, the effect of dust charge fluctuations on the parametric decay of lower hybrid wave instability by relativistic runaway electrons in a tokamak using the kinetic approach is described. In Sec. II, the relativistic runaway electron's susceptibility is obtained using Vlasov equations. The non-linear coupling of lower hybrid pump waves with REs is explained in Sec. III. The pump wave imparts oscillatory velocity to the runaway electrons, which produces the non-linear current and drives the upper sideband. Electron acceleration process by the decay of lower hybrid waves and the dust charge fluctuation effect on the parametric decay of lower hybrid waves is discussed in Sec. IV.

II. RELATIVISTIC ELECTRON (RE) SUSCEPTABILITY

Consider a theoretical model of tokamak containing uniform plasma consisting of background electrons having density n_{0e} in a given magnetic field $\vec{B}_s(=B_0\hat{Z})$. Let us consider that the plasma contains a component of runaway electrons (REs) with a density perturbation ${}_0n^{0r}$, where the initial runaway electron density distribution[2] can be taken as

$$n^r = {}_0n^{0r} e^{[t(E-1)/(c_r\tau_{tr})]}, \tag{1}$$

where $E(=e\iota_{tr}E_z/m_ec)$ defines the parallel electric field (normalized), which is a constant of time, m_e defines the mass of runaway electrons, e is the electronic charge, c denotes the speed of the light, τ_{tr} $(=4\pi m_e^2c^3\varepsilon_0^2/e^4n_{0e}1n\Lambda)$ denotes the collision time of relativistic electrons, ε_0 is the permittivity of free space, $c_r=(\sqrt{(3Z+5)}/\pi)1n\Lambda$, $1n\Lambda$, is Coulomb's logarithm, the effective ion charge is denoted by Z, and ${}_0n^{0r}$ denotes the value of runaway density n^r at the time of launch of lower hybrid wave.

The distribution function for the relativistic tail of runaway electrons in a tokamak disruption[18] initially can be written as

$$\mathrm{f}_0(p_z,p_\perp)=\frac{{}_0n^{0r}\alpha_r}{2\pi c_r p_z}\exp\left(-\frac{\alpha_r p_\perp^2}{2p_z}-\frac{p_z}{c_r}\right), \tag{2}$$

where $p(=\frac{v\gamma_0}{c})$ is the normalized relativistic momentum of the runaway electrons, γ_0 is the relativistic gamma factor, $p_\perp$ and p_z are the perpendicular and axial components of the momentum, respectively, and $\alpha_r=\frac{E-1}{Z+1}$.

If the initial equilibrium is disturbed by an electrostatic perturbation,[2]

$$\varphi = ae^{i(\vec{k}.\vec{r}-\omega t)}, \tag{3}$$

where $\omega \ll \omega_c, \omega_c(=\frac{eB_0}{m_ec})$ is the electron cyclotron frequency.

The RE's response to electrostatic perturbation is given by the Vlasov equation,

$$\frac{\partial \mathrm{f}}{\partial t}+\mathrm{v}\cdot\nabla\mathrm{f}=-\frac{e}{m_e}\left[\nabla\varphi-\frac{1}{c}\mathrm{v}\times\mathrm{B}_s\right]\nabla_p\mathrm{f}. \tag{4}$$

For $\omega \ll \omega_c$, the perturbed density[2] of runaway electrons can be written as

$$n_{1b}=\frac{k^2\varepsilon_0}{e}(\chi_{rb}+i\chi_{ib})\varphi, \tag{5}$$

where χ_{rb} and χ_{ib} are given in Ref. 2.

III. NON-LINEAR COUPLING AND GROWTH RATE

Consider a parametric coupling of lower hybrid pump wave with potential[2] φ_0, which is given as

$$\varphi_0 = a_0e^{i(k_{z0}.z+k_{0\perp}.\vec{r}-\omega_0t)} \tag{6}$$

with a RE mode given by Eq. (3) and the upper sideband mode

$$\varphi_0 = a_2e^{i(\vec{k}_z.\vec{r}-\omega_2t)}, \tag{7}$$

where $\omega_2=\omega+\omega_0, \vec{k}_2=\vec{k}+\vec{k}_0$.

The dispersion relation[2] is given as: $\omega_x^2=\omega_L^2(1+\frac{m_ik_{zx}^2}{m_ek_x^2})$, m_i is the ion mass and $\omega_L^2=\omega_{pi}^2/(1+\frac{\omega_{ep}^2}{\omega_c^2})$, $\omega_{pi}=(z4\pi n_{0i}e^2/m_i)^{1/2}$ is the ion plasma frequency, $\omega_{ep}=(4\pi n_{0e}e^2/m_e)^{1/2}$ is the electron plasma frequency.

The oscillatory velocity of REs as described by Kuley and Tripathi[2] can be written as

$$\mathrm{v}_{bxz}=-\frac{ek_{zx}\varphi_x}{m_e\gamma_0^3(\omega_x-k_{zx}\mathrm{v}_{bz0}^0)}.$$

The upper sideband wave and pump wave on coupling exert a ponderomotive force on the electrons, which is denoted by f_p. Here, f_p can have two components, one is perpendicular to the magnetic field

Published under license by AIP Publishing

($f_{\perp p}$) and other one is parallel to the magnetic field (f_{zp}). Here, the response of electrons to $f_{\perp p}$ is suppressed by the applied magnetic field strongly. However, the electrons effectively respond to f_{zp} in the parallel direction. Thus, the nonlinearity in frequency arises due to $f_{zp} = -m_e v.\nabla v_z$.

So, the parallel ponderomotive force for the background electrons is

$$f_{zp} = ie\varphi_{pp}k_{z1} = -\left(\frac{m_e}{2}\right)\left[v^*_{\perp 0}.\nabla_\perp v_{2z} + v_{2\perp}.\nabla_\perp v^*_{z0}\right].$$

Following Kuley and Tripathi,[2] the ponderomotive potential can be written as

$$\varphi_{ppb} = \frac{e\varphi_0\varphi_2 k_{\perp 2}k_{\perp 1}\times\omega_c}{2m_e\omega_c^2 ik_{z0}\gamma_0^2}\left[-\frac{k_{z2}}{(\omega_2 - k_{z2}v_{bz0})} + \frac{k_{z0}}{(\omega_0 - k_{z0}v_{bz0})}\right]. \tag{8}$$

So, the response of REs can be rewritten as

$$n_{1b} = (\chi_{rb} + i\chi_{ib})\left(\frac{\varepsilon_0 k^2}{e}\right)(\varphi + \varphi_{ppb}). \tag{9}$$

The dust density perturbation[22] can be written as

$$n_{1d} = -\frac{n_{d0}k_\perp^2\phi_1 Q_{od}}{m_d\omega^2}, \tag{10}$$

where the charge on dust grains[3] is balanced by plasma currents existing on the surface of grains and is given by the equation

$$Q_{1d} = \frac{|I_{0e}|}{i(\omega + i\eta)}\left(\frac{n_{1i}}{n_{0e}} - \frac{n_{1e}}{n_{0e}}\right) \tag{11}$$

where $\eta \simeq 10^{-2}\,\omega_{ep}\left(\frac{a}{\lambda_d}\right)\frac{n_{0e}}{n_{0i}}$, λ_d is the Debye length, a is the dust grain size, and electron perturbed density can be written as

$$n_{1e} = -\frac{en_{0e}}{m_e}\left(\frac{\nabla_\perp^2\varphi}{\omega_c^2} + \frac{k_z^2\varphi}{\omega^2}\right). \tag{12}$$

Similarly, the ion perturbed density is

$$n_{1i} = -\frac{en_{0i}}{m_i}\left(\frac{\nabla_\perp^2\varphi}{\omega_c^2 - w_{ic}^2} + \frac{k_z^2\varphi}{\omega^2}\right), \tag{13}$$

$w_{ic}(= eB_0/m_i c)$ is the ion cyclotron frequency.

Now putting the values from (12) and (13) in Eq. (11), we obtain

$$Q_{1d} = -\frac{|I_{0e}|}{i(\omega + i\eta)}\frac{e}{m_i}\left(\frac{\nabla_\perp^2\varphi}{\omega^2 - \omega_{ic}^2} - \frac{k_z^2\varphi}{\omega^2}\right) + \frac{|I_{0e}|}{i(\omega + i\eta)}\frac{e}{m_e}\left(\frac{\nabla^2_{\perp\varphi}}{\omega_c^2 - \omega^2} - \frac{k_z^2\varphi}{\omega^2}\right). \tag{14}$$

Using all Eqs. (9), (10), (12)–(14) in Poisson's equation,

$$\nabla^2\varphi = \frac{1}{\varepsilon}(n_{1e}e + n_{1b}e - n_{1i}e - n_{0d}Q_{1d} - Q_{0d}n_{1d}), \tag{15}$$

and solving the equation, we get

$$-k^2\varphi = \frac{1}{\varepsilon}\left[e\left(\frac{k^2\varepsilon}{e}\right)\chi_e\varphi - e\left(\frac{{}_0n^0ek_z^2}{m_e\omega^2}\right)\varphi_{pp} + e\left(\frac{k^2\varepsilon}{e}\right)\chi_b(\varphi + \varphi_{ppb}) + e\left(\frac{k^2\varepsilon}{e}\right)\chi_i\varphi - n_{0d}Q_{1d} + \left(\frac{k^2\varepsilon}{Q_{0d}}\right)\chi_d\varphi Q_{0d}\right] \tag{16}$$

or

$$-k^2\varphi = \frac{k^2\varepsilon}{\varepsilon}\left[\left\{\chi_e\varphi + \varphi\frac{i\beta}{\omega + i\eta}\chi_e\right\} + \left\{\varphi_0\chi_i + \chi_i\frac{i\beta}{\omega + i\eta}\varphi_0\frac{n_{0e}}{n_{0i}}\right\} + \chi_d\varphi + \chi_b\varphi + \chi_b\varphi_{ppb} + \chi_e\varphi_{pp}\right], \tag{17}$$

where $\beta = \left(\frac{|I_{0e}|}{e}\right)\frac{n_{0d}}{n_{0e}}$ is the coupling parameter.

On solving Eq. (17) further, we get

$$\left[1 + \chi_e\left(1 + \frac{i\beta}{\omega + i\eta}\right) + \chi_i\frac{1}{\delta}\left(\frac{i\beta}{\omega + i\eta} + 1\right) + \chi_d + \chi_b\right]k^2\varphi = \frac{k^2\varepsilon}{\varepsilon}[-\chi_b\varphi_{ppd} - \chi_e\varphi_{pp}] \tag{18}$$

or

$$\left(1 + \chi_e' + \chi_i'\frac{1}{\delta} + \chi_b + \chi_d\right)\varphi = [-\chi_b\varphi_{ppd} - \chi_e\phi_{pp}], \tag{19}$$

where

$$\chi_e' = \left(\frac{i\beta}{\omega + i\eta} + 1\right)\chi_e,$$

and

$$\chi_i'\frac{1}{\delta} = \frac{1}{\delta}\left(\frac{i\beta}{\omega + i\eta} + 1\right)\chi_i,$$

Equation (19) can be rewritten as

$$\varepsilon^1\varphi = -\chi_b\varphi_{ppd} - \chi_e\phi_{pp}, \tag{20}$$

$$\varepsilon^1 = 1 + \chi_e' + \chi_d + \frac{1}{\delta}\chi_i' + \chi_b. \tag{21}$$

Also,

$$\left(\frac{i\beta}{\omega + i\eta} + 1\right)\chi_e = \chi_e\left(\frac{i\beta\omega + \omega^2 + \beta\eta + \eta^2}{\eta^2 + \omega^2}\right),$$

$$\chi_i\frac{1}{\delta}\left(\frac{i\beta}{\omega + i\eta} + 1\right) = \frac{1}{\delta}\left(\frac{\omega^2 + i\beta\omega + \eta^2 + \beta\eta}{\eta^2 + \omega^2}\right)\chi_i,$$

and

$$\varepsilon^1 = 1 + \chi_e\left[1 + \frac{\beta\eta}{\omega^2 + \eta^2}\right] + \chi_i\left[1 + \frac{\beta\eta}{\omega^2 + \eta^2}\right]\frac{1}{\delta} + \chi_{rb} + \chi_d + i\left[\frac{\chi_e\beta\omega}{\omega^2 + \eta^2} + \frac{\chi_i\beta\omega}{\omega^2 + \eta^2}\frac{1}{\delta} + \chi_{ib}\right] \tag{22}$$

or

Published under license by AIP Publishing

$$\varepsilon^1 = 1 + \chi_e'' + \chi_i'' \frac{1}{\delta} + \chi_{rb} + \chi_d + i\left[\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right], \tag{23}$$

where

$$\chi_e'' = \chi_e\left[1 + \frac{\beta\eta}{\omega^2+\eta^2}\right],$$

and

$$\chi_i'' = \chi_i\left[1 + \frac{\beta\eta}{\omega^2+\eta^2}\right]\frac{1}{\delta}, \quad \chi_e^{ii} = \chi_e\frac{\beta\omega}{\omega^2+\eta^2},$$
$$\chi_i^{ii} = \chi_i\left[\frac{\beta\omega}{\omega^2+\eta^2}\right]\frac{1}{\delta}, \quad \delta = \frac{n_{0i}}{n_{0e}},$$
$$\chi_b = \chi_{rb} + i\chi_{ib}, \quad \chi_e = \frac{\omega_{ep}^2 k_\perp^2}{\omega_c^2 k^2} - \frac{\omega_{ep}^2 k_z^2}{\omega^2 k^2}, \quad \text{and}$$
$$\chi_i = \frac{\omega_{pi}^2 k_\perp^2}{(\omega^2 - \omega_{ic}^2)k^2} - \frac{\omega_{pi}^2 k_z^2}{\omega^2 k^2}.$$

The density perturbation couples with the electron velocity (oscillatory) and produces non-linear density perturbations at a frequency equal to the upper sideband frequency. Therefore, solving the equation of continuity (for the background electron), we get

$$\frac{\partial}{\partial t} n_{nl2} + \nabla\left(\frac{n}{2} v_0^*\right) = 0; \tag{24}$$

we obtain

$$n_{nl2} = \frac{n}{2\omega_2\left(\vec{k}_2.\vec{v}_0^*\right)}. \tag{25}$$

and for REs,

$$n_{2n}^{n1} = \frac{n_b}{2\omega_2}\left(\vec{k}_2.\vec{v}_0^*\right). \tag{26}$$

Now, for the upper sideband wave, using Eqs. (25) and (26) in Poisson equation, one obtains

$$\varepsilon_2\varphi_2 = \frac{k_2 e\varphi_0}{k_2^2 m_e \omega_c^2}(1+\chi_i)\frac{k_2.\omega_c \times \nabla_\perp \varphi_1}{2\omega_2}, \tag{27}$$

where

$$\varepsilon_2 = 1 + \frac{\omega_{ep}^2}{\omega_c^2} + \gamma_0\frac{\omega_{ep}^2(\omega_2 - k_{z2}v_{bz0})}{\omega_2} - \frac{\omega_{pi}^2}{\omega_2^2} \tag{28}$$

denotes the dielectric function with (ω_2, k_2). The given Eqs. (20) and (27) are called nonlinear coupled equations, from where we can obtain a nonlinear dispersion relation as

$$\varepsilon^1\varepsilon_2 = \zeta, \tag{29}$$

where the coefficient of coupling can be written as

$$\zeta = \frac{U^2k^2\sin^2\theta}{4}\left[\frac{\chi_e}{\omega_2^2}\left(\frac{\omega k_{z0}}{\omega_0 k_z} - 1\right) + \frac{\chi_b}{k_z\gamma_0^2\omega_2} \times \left(\frac{k_{z0}}{(\omega_0 - k_{z0}v_{bz0})} - \frac{k_{2z}}{(\omega_2 - k_{z2}v_{bz0})}\right)\right], \tag{30}$$

where U denotes the magnitude of the velocity of electron, whereas θ measures the angle between k_{z2} and k_{z0}.

Using the Taylor series expansion, we get

$$\varepsilon^1(\omega_r + i\Gamma) = \varepsilon^1(\omega_r) + \frac{\partial\varepsilon^1}{\partial\omega_r} i\Gamma \tag{31}$$

and

$$\varepsilon_2(\omega_r + i\Gamma) = \varepsilon_2(\omega_{r2}) + i\Gamma\frac{\partial\varepsilon_2}{\partial\omega_{r2}}. \tag{32}$$

Also

$$\varepsilon^1\varepsilon_2 = \varepsilon^1 \mathrm{x} i\Gamma\frac{\partial\varepsilon_2}{\partial\omega_2}, i\Gamma = \frac{\varepsilon^1\varepsilon_2}{\frac{\partial\varepsilon_2}{\partial\omega_2}\varepsilon^1}. \tag{33}$$

Finally, Eq. (33) gives the growth rate of the unstable mode as

$$i\Gamma = \frac{\left[\zeta\left(1 + \chi_e'' + \chi_{rb} + \chi_d + \chi_i''\frac{1}{\delta}\right) - i\zeta\left(\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right)\right]}{\left[\left(1 + \chi_e'' + \chi_{rb} + \chi_d + \chi_i''\frac{1}{\delta}\right)^2 + \left(\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right)^2\right]\frac{\partial\varepsilon_2}{\partial\omega_2}},$$

or

$$\Gamma = \frac{-i\zeta\left(\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right)}{\left[\left(1 + \chi_e' + \chi_i''\frac{1}{\delta} + \chi_{rb} + \chi_d\right)^2 + \left(\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right)^2\right]\frac{\partial\varepsilon_2}{\partial\omega_2}}, \tag{34}$$

where

$$\frac{\partial\varepsilon_2}{\partial\omega_2} = \frac{\gamma\omega_{ep}^2}{\omega_c^2}\left[\frac{k_{z2}v_{bz0}}{\omega_2^2}\right] + 2\frac{\omega_{pi}^2}{\omega_2^3},$$

and

$$\Gamma = \frac{\mathrm{Im}\left[\zeta\left(1 + \chi_e'' + \chi_i''\frac{1}{\delta} + \chi_{rb} + \chi_d\right) - i\zeta\left(\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right)\right]}{\left[\left(1 + \chi_e' + \chi_i''\frac{1}{\delta} + \chi_{rb} + \chi_d\right)^2 + \left(\chi_e^{ii} + \chi_i^{ii}\frac{1}{\delta} + \chi_{ib}\right)^2\right]\frac{\partial\varepsilon_2}{\partial\omega_2}}. \tag{35}$$

Equation (35) gives the growth rate of the unstable wave in the presence of dust grains. However, in the absence of the dust grains, Eq. (35) reduces to Eq. (35), p. 062507–4 of Kuley and Tripathi.[2]

IV. RESULTS AND DISCUSSION

We have used the typical parameters corresponding to tokamak,[12] e.g., electron density $\sim 4 \times 10^{13}\,\mathrm{cm}^{-3}$, electron temperature $\sim 3 \times 10^3$ eV, ion temperature $\sim 1.5 \times 10^3$ eV, energy of fast electrons $\sim 10^5$ eV, magnetic field $\sim 2.5 \times 10^4$ Gauss, frequency of the lower hybrid pump wave $\sim 2.45 \times 10^9$ Hz, $\gamma_0 = 1.2$, and the density of fast electrons $\sim 2 \times 10^{10}\,\mathrm{cm}^{-3}$ for numerical calculations. By using these parameters, we have plotted the growth rate $\Gamma(\mathrm{s}^{-1})$ with the relative density of dust grains $\delta(= n_{0i}/n_{0e})$, as shown in Fig. 1, where the growth rate shows an enhancement with the increase in the relative density of negatively charged dust grains. In tokamak, REs contain

Published under license by AIP Publishing

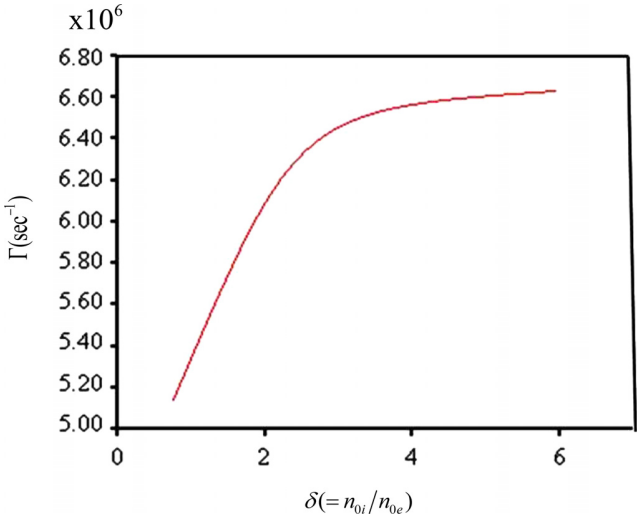


FIG. 1. Growth rate $\Gamma(s^{-1})$ with $\delta(=n_{0i}/n_{0e})$, the relative density of negatively charged dust grains.

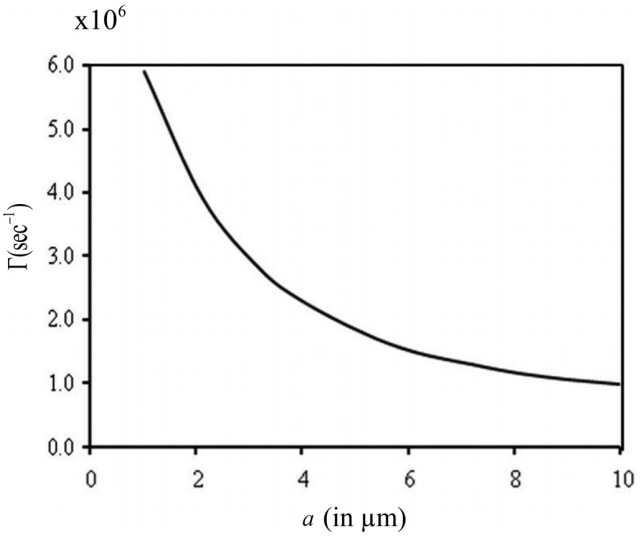


FIG. 2. Growth rate $\Gamma(s^{-1})$ with dust grain size a (in μm) for the same parameters as used in Fig. 1.

free energy and this energy adds up to the pump photons, which produces the upper sideband photons. These REs provide the upconversion of the frequency of lower hybrid pump waves. Now, the phase velocity (parallel) of these waves (upconverted by the REs) becomes comparable to that of the velocity of light in vacuum. Thus, as a result, the upper sideband wave attains an amplitude equivalent to that of the pump wave and then accelerates the electrons to the energy nearly around 2 MeV. On the account of this process, the growth rate increases and becomes maximum for $\theta = 90^{\circ}$, where θ is the angle between $k_{\perp 2}$ and $k_{0\perp}$. The growth rate is slower for other values of θ. We have compared our theoretical results with the experimental observations[12] existing for the heating of lower hybrid waves in tokamak and it has been observed that our results show the consistency with these experimental results. Both the studies show the existence of electrons of MeV energy in tokamak, and the REs with this much energy can lead to severe damage in tokamak. Dust plays a very significant role in tokamak. Dust particles having a high atomic number can become radioactive and damage the tokamak operation. Our results show that the presence of dust grains, their size, and density affects and modifies the growth rate. The dust grains, usually found in tokamaks, have appreciable effect on the growth rate of lower hybrid waves and range in size from 100 nm to 100 μm; that is why we have considered the effect of dust grain size and it is plotted in Fig. 2. Here, it can be seen that, as the size of dust grain increases, the growth rate decreases. More is the size of dust grains, more electrons become attached to the dust grain surface and hence the electron current is reduced, whereas the ion current is increased, which in turn decreases the growth rate. The effect of dust charge fluctuations on the growth rate can also be explained from Eqs. (16)–(35). These equations involve two parameters, one is β called the coupling parameter, i.e., the coupling between the plasma and dust grains, whereas the other parameter is η, the dust charging rate. The dust charging rate is directly proportional to the dust grain radius. Small dust grains require large charging time.

Variation of the growth rate $\Gamma(s^{-1})$ with the relative density of the negatively charged dust grains $\delta(=n_{0i}/n_{0e})$ for distinct values of the dust grain number density $n_{d0}(\text{in cm}^{-3})$ is shown in Fig. 3, which displays the increase in growth rate with the dust grain number density and the relative density of negatively charged dust grains. Figure 4 disports the variation of the growth rate $\Gamma(s^{-1})$ with the relative density of negatively charged dust grains $\delta(=n_{0i}/n_{0e})$ for different values of dust grain size a (in μm) and shows the increase in growth rate with the size of dust grains. The major safety issues in a tokamak are due to the radiological and chemical hazards. The chemical hazards are dependent on the dust grain distribution and sizes. If the size of the

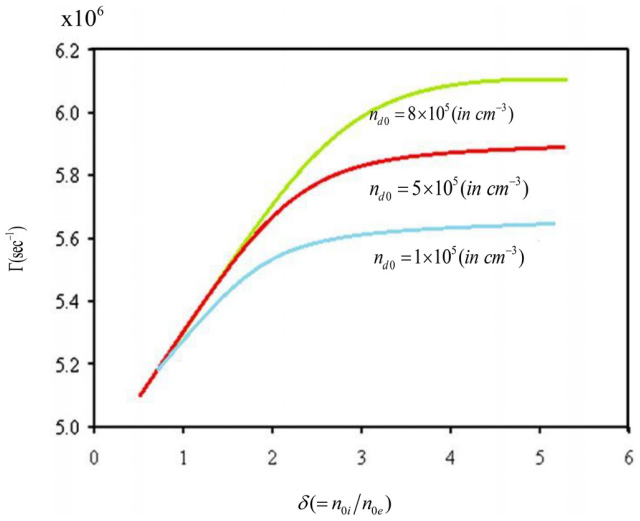


FIG. 3. Growth rate $\Gamma(s^{-1})$ vs relative dust grain density $\delta(=n_{0i}/n_{0}e)$ for distinct values of $n_{d0}(\text{in cm}^{-3})$, dust grain density.

Published under license by AIP Publishing

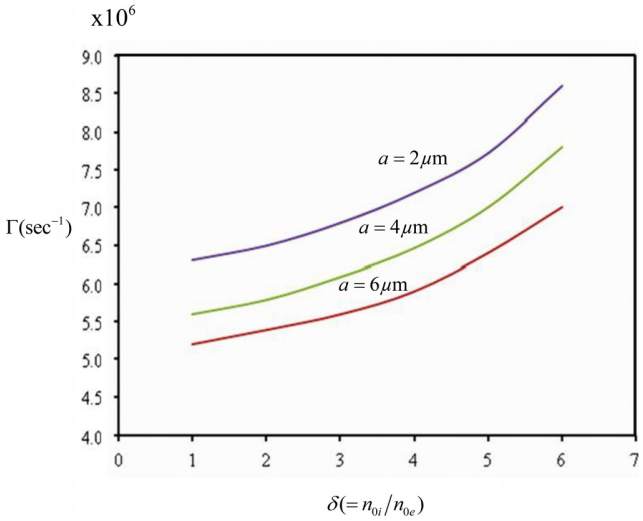


FIG. 4. Growth rate $\Gamma(s^{-1})$ vs relative density of negatively charged dust grains $\delta(= n_{0i}/n_{0e})$ for distinct values of dust grain size a (in μm).

dust grains is smaller, they can probably get into the air easily and thus contain a little hazardous material as compared to the large sized dust grains.

Temperature has also a significant effect on the growth rate $\Gamma(s^{-1})$ due to the fact that the thermal velocity of electrons increases with increase in temperature, which in turn increases the growth rate, as shown in Fig. 5. In Fig. 6, a graph is plotted between the growth rate $\Gamma(s^{-1})$ and the negatively charged dust grain density n_{d0} (in cm^{-3}), where the growth rate shows a linear increase with the increase in the dust grain density. The literature existing so far does not report any theoretical results for the consideration of the effect of dust charge

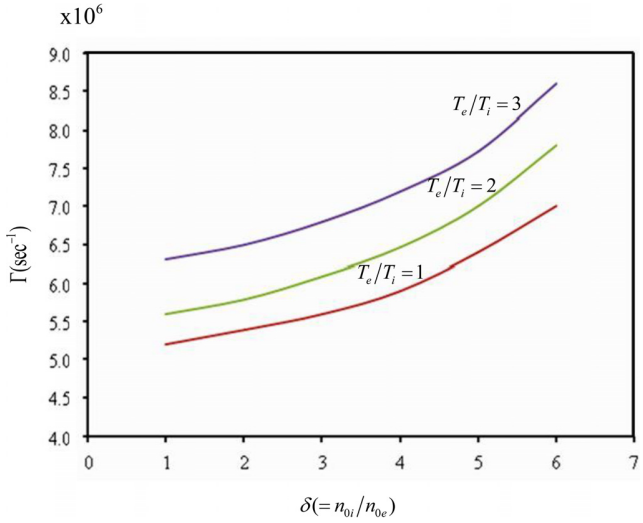


FIG. 5. Growth rate $\Gamma(s^{-1})$ vs relative density of negatively charged dust grains $\delta(= n_{0i}/n_{0e})$ for distinct values of temperature ratio.

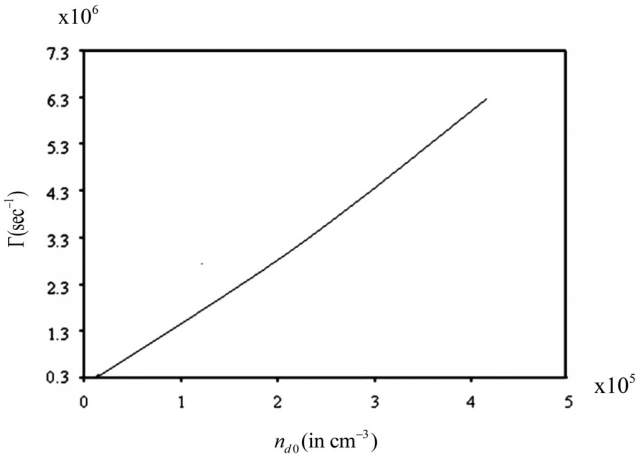


FIG. 6. Growth rate $\Gamma(s^{-1})$ as a function of dust grain density n_{d0}(in cm^{-3}).

fluctuations on the parametric decay of lower hybrid waves in tokamaks. Therefore, our theoretical results are very significant in understanding and explaining the role of dust grain size and number density of dust particles present in the tokamak during different experiments.

V. CONCLUSION

In the present manuscript, the kinetic theory of parametric upconversion of lower hybrid waves by relativistic runaway electrons (REs) in a tokamak is described. The REs having energies up to several MeV generate very severe trouble in tokamaks by damaging the first wall structure on impact. The production of these current carrying fast electrons of such high energy with high residual electric field is significantly enhanced by the lower hybrid waves present in the tokamak through electron Landau damping (parallel) when the condition of the Čerenkov resonance is satisfied. Thus, the growth of the lower hybrid wave instability is a serious issue for the operation of a tokamak. Moreover, the presence of dust in a tokamak also becomes a reason for the enhancement of the growth rate of lower hybrid wave instability. Dust is produced in a tokamak by the process of sputtering and large disruptions during a shot by power flux to surfaces, in the form of hydrocarbons and tungsten compounds. The growth rate of the lower hybrid wave instability is found to be affected due to the presence of dust population. The growth rate of the instability increases with the increase in the relative density of dust grains, whereas it decreases with the increase in dust grain size. The electron to ion temperature ratio is also found to be one of the parameters that increases the growth rate of the instability.

ACKNOWLEDGMENTS

The author J.S. is thankful to SERB-DST, Government of India (No. EMR/2016/002699), for partial support.

DATA AVAILABILITY

The data that support the findings of this manuscript are available within the article. In addition, the data that support the findings

Published under license by AIP Publishing

of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]P. Helander, L.-G. Eriksson, and F. Anderson, Plasma Phys. Controlled Fusion **44**, B247 (2002).
[2]A. Kuley and V. K. Tripathi, Phys. Plasmas **17**, 062507 (2010).
[3]M. R. Jana, A. Sen, and P. K. Kaw, Phys. Rev. E **48**, 3930 (1993).
[4]M. Shimada, D. J. Campbell, V. Mukhovato, M. Fujiwara, N. Kirneva, K. Lackner, M. Nagami, V. D. Pustovitov, N. Uckan, J. Wesley, N. Asakura, A. E. Costley, A. J. H. Donn, E. J. Doyle, A. Fasoli, C. Gormezano, Y. Griboy, O. Gruber, T. C. Hender, W. Houlberg, S. Ide, Y. Kamada, A. Leonard, B. Lipschultz, A. Loarte, K. Miyamoto, V. Mukhovtov, T. H. Osborne, A. Polevoi, and A. C. C. Sips, Nucl. Fusion **47**, S1 (2007).
[5]J. A. Wesson, R. D. Gill, M. Hugon, F. C. Schuller, J. A. Snipes, D. J. Ward, D. V. Bartlett, D. J. Campbell, P. A. Duperrex, A. W. Edwards, R. S. Granetz, N. A. O. Gottardi, T. C. Hender, E. Lazzaro, P. J. Lomas, N. Lopes Cardozo, K. F. Mast, M. F. F. Nave, N. A. Salmon, P. Smeulders, P. R. Thomas, B. J. D. Tubbing, M. F. Turner, and A. Weller, Nucl. Fusion **29**, 641 (1989).
[6]V. Riccardo, and JET EFDA Contributors. Plasma Phys. Controlled Fusion **45**, A269 (2003).
[7]V. V. Plyusin, V. Riccardo, R. Jaspers, B. Alper, V. G. Kiptily, J. Mlynar, S. Popovichev, E. d L. Luna, and F. Andersson, and JET EFD Contibutors, Nucl. Fusion **46**, 277 (2006).
[8]J. R. Martin-Solis, B. Espito, R. Sanchez, F. M. Poli, and L. Panaccione, Phys. Rev. Lett. **97**, 165002 (2006).
[9]J. R. Martin-Solis, B. Espito, R. Sanchez, and G. Granucci, Nucl. Fusion **44**, 974 (2004).
[10]R. Yoshino and S. Tokuda, Nucl. Fusion **40**, 1293 (2000).
[11]Y. Kawano, R. Yoshino, Y. Neyatani, Y. Nakamura, S. Tokuda, and H. Tamai, Fusion Sci. Technol. **42**, 298 (2002).
[12]Z. Y. Chen, B. N. Wan, S. Y. Lin, Y. J. Shi, L. Q. Hu, J. Younis, X. Z. Gong, J. F. Shan, F. K. Liu, B. J. Ding, and X. Gao, and HT-7 Team, Plasma Phys. Controlled Fusion **48**, 1489 (2006).
[13]Z. Y. Chen, B. N. Wan, B. L. ling, X. Gao, A. Ti, Q. Du, S. Sajjad, S. Y. Lin, Y. J. Shi, and HT-7 Team, Plasma Phys. Controlled Fusion **50**, 015001 (2008).
[14]H. Dreicer, Phys. Rev. **117**, 329 (1960).
[15]M. N. Rosenbluth and S. V. Putvinski, Nucl. Fusion **37**, 1355 (1997).
[16]L.-G. Eriksson, P. Helander, F. Anderson, and M. Lisak, Phys. Rev. Lett. **92**, 205004 (2004).
[17]C. S. Liu and Y. Mok, Phys. Rev. Lett. **38**, 162 (1977).
[18]G. Pokol, T. Fulop, and M. Lisak, Plasma Phys. Controlled Fusion **50**, 045003 (2008).
[19]T. Fulop, G. Pokol, P. Helander, and M. Lisak, Phys. Plasmas **13**, 062506 (2006).
[20]L.-G. Eriksson and P. Helander, Comput. Phys. Commun. **154**, 175 (2003).
[21]J. Winter, Phys. Plasmas **7**, 3862 (2000).
[22]S. C. Sharma and J. Sharma, Phys. Plasmas **17**, 043704 (2010).
[23]S. I. Krasheninnikov, R. D. Smirnov, A. Y. Pigarov, T. K. Soboleva, and D. A. Mendis, J. Plasma Phys. **76**, 377 (2010).
[24]L. Zeng, Z. Y. Chen, Y. B. Dong, H. R. Koslowski, Y. Liang, Y. P. Zhang, H. D. Zhuang, D. W. Huang, and X. Gao, Nucl. Fusion **57**, 046001 (2017).
[25]C. Liu, E. Hirvijoki, G. Y. Fu, D. P. Brennan, A. Bhattacharjee, and C. Paz-Soldan, Phys. Rev. Lett **120**, 265001 (2018).
[26]C. J. McDevitt and X.-Z. Tang, Europhys. Lett. **127**, 45001 (2019).
[27]A. Marco, A. J. Garrido, S. Coda, and I. Garrido, and TCV Team, *Advanced Controls in Complex Clean Energy Devices, Subsystems, and Processes* (Wiley, 2019), p. 2319560.
[28]T. Fülöp, P. Helander, O. Vallhagen, O. Embreus, L. Hesslow, P. Svensson, A. J. Creely, N. T. Howard, and P. Rodriguez-Fernandez, J. Plasma Phys. **86**, 474860101 (2020).

Published under license by AIP Publishing

AIP Advances

Learning of classification tasks with an array of uniform-mode spin Hall nano-oscillators

Cite as: AIP Advances **11**, 045117 (2021); https://doi.org/10.1063/9.0000192
Submitted: 15 October 2020 . Accepted: 02 April 2021 . Published Online: 14 April 2021

Utkarsh Singh, Neha Garg, Saurabh Kumar, Pranaba Kishor Muduli, and Debanjan Bhowmik

COLLECTIONS

Paper published as part of the special topic on 65th Annual Conference on Magnetism and Magnetic Materials, 65th Annual Conference on Magnetism and Magnetic Materials, 65th Annual Conference on Magnetism and Magnetic Materials, 65th Annual Conference on Magnetism and Magnetic Materials, 65th Annual Conference on Magnetism and Magnetic Materials and 65th Annual Conference on Magnetism and Magnetic Materials

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Ultrafast Ising Machines using spin torque nano-oscillators
Applied Physics Letters **118**, 112404 (2021); https://doi.org/10.1063/5.0041575

Acoustical Fatigue Test Procedures
Noise Control **6**, 11 (1960); https://doi.org/10.1121/1.2369402

Range and Energy Deposition Enhancement of a Fast Electron Beam by External Electric Fields
Journal of Vacuum Science and Technology **10**, 1000 (1973); https://doi.org/10.1116/1.1318452

Call For Papers!
AIP Advances
SPECIAL TOPIC: Advances in Low Dimensional and 2D Materials

AIP Publishing

© 2021 Author(s).

AIP Advances ARTICLE scitation.org/journal/adv

Learning of classification tasks with an array of uniform-mode spin Hall nano-oscillators

Cite as: AIP Advances 11, 045117 (2021); doi: 10.1063/9.0000192
Presented: 6 November 2020 • Submitted: 15 October 2020 •
Accepted: 2 April 2021 • Published Online: 14 April 2021

View Online Export Citation CrossMark

Utkarsh Singh,[1] Neha Garg,[2,3] Saurabh Kumar,[3] Pranaba Kishor Muduli,[2] and Debanjan Bhowmik[3,a)]

AFFILIATIONS

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India
[2] Department of Physics, Indian Institute of Technology Delhi, New Delhi 110016, India
[3] Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India

Note: This paper was presented at the 65th Annual Conference on Magnetism and Magnetic Materials.
[a)] **Author to whom correspondence should be addressed:** debanjan@ee.iitd.ac.in. **URL:** http://web.iitd.ac.in/~debanjan/

ABSTRACT

Recently, a system of spintronic vortex oscillators has been experimentally trained to classify vowel sounds. In this paper, we have carried out a combination of device-level and system-level simulations to train a system of spin Hall nano oscillators (SHNOs) of smaller size (25X lower in area compared to those vortex oscillators) for such data classification tasks. Magnetic moments precess in an uniform mode as opposed to the vortex mode in our oscillators. We have trained our system to classify inputs in various popular machine learning data sets like Fisher's Iris data set of flowers, Wisconsin Breast Cancer (WBC) data set, and MNIST data set of handwritten digits. We have employed a new technique for input dimensionality reduction here so that the clustering/target synchronization pattern changes based on the nature of the data in the different data sets. Our demonstration of learning in a system of such small SHNOs for a wide range of data sets is promising for scaling up the oscillator-based neuromorphic system for complex data classification tasks.

© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/9.0000192

I. INTRODUCTION

A. Motivation

Spintronics-based neuromorphic/in-memory computing hardware has been considered to be a low energy alternative to transistor-based digital hardware, which has memory-computing separation, for data classification tasks.[1] One approach for spintronic neuromorphic computing is designing an analog crossbar array of domain-wall-based or skyrmion-based synaptic devices.[2–6] Fully Connected Neural Network (FCNN) algorithms, of spiking or non-spiking type, are then implemented on this crossbar.[2,5–11] The properties/parameters of the activation function, or neuron, are static in this case. A large number of parameters needs to be tuned to achieve learning in such FCNN.[12]

Recently, in an alternative approach, Romera *et al.*[12] use the dynamic properties of synchronized spintronic oscillators in an experimentally fabricated network with far less parameters, which need to be tuned, to achieve on-chip learning for classification of seven vowel sounds. Magnetic Tunnel Junction (MTJ) cells of quite large diameter ($\approx$ 375 nm) are used as vortex mode spin torque oscillators for this purpose. But Romera *et al.*[12] also propose that in order to scale up the system for more complicated classification tasks and to reduce the power consumption, a spin torque oscillator of much lower diameter needs to be used.[13,14]

At this lower length scale, the nanomagnet may not oscillate in the vortex mode but rather oscillate close to the uniform mode. In the uniform mode, most individual magnetic moments inside the ferromagnetic layer of the spin torque oscillator always remain parallel to each other and precess in unison about an axis.[13,15,20] So it is important that considering this uniform mode of oscillation, classification is shown for a system of such coupled spin torque oscillators.

B. Our contributions here

In this paper, we have carried out simulation-based device-system co-study of an array of uniform-mode spin-orbit torque nano oscillators, also known as spin Hall nano oscillators (SHNOs),[20] of

© Author(s) 2021

diameter 75 nm (5X lower in diameter compared to the vortex oscillator in Romera *et al.*,[12] 25X lower in area) to demonstrate learning of classification tasks in such a system. We model such uniform-mode SHNOs through micromagnetics (the 'mumax3' package[16]) and show that they can be locked to the external radio frequency (RF) magnetic field, which mimic the input values in reduced dimensions much like the vortex mode oscillators in Romera *et al.*[12] (see Sec. II below for this device-level study). Thus, classification of the input can also be achieved through an array of such uniform-mode oscillators much like in the vortex oscillators in Romera *et al.*,[12] as we show in Sec. III through our system-level study.

Also Romera *et al.* show classification in only one data set (vowel sounds) experimentally.[12] But in our simulations, we show learning on several popular machine learning data sets like Fisher's Iris data set of flowers,[17] Wisconsin Breast Cancer (WBC) data set,[18] and the MNIST data set of handwritten digits[19] (Sec. 1 of supplementary material). The MNIST data set is much larger than the vowel data set with each input being a 784-dimensional vector (28-by-28-pixel image) as opposed to 12-dimensional vector in the vowel data set.[12] There are also many more samples in the MNIST data set than in the vowel data set.[12,19] Thus, we have trained our system on more diverse, and sometimes more complex, data classification tasks, compared to Romera *et al.*[12]

In order to make the spin-oscillator-based classification scheme in Romera *et al.*[12] applicable for all these different data sets, in our paper, we have changed the input dimensionality reduction technique from what was followed in Romera *et al.*[12] and made it more versatile. We explain this in details in Sec. III below. Since we also report high classification accuracy (above 80 % for MNIST, above 90 % for other data sets) through our system-level study in Sec. III, it shows that our learning scheme is quite effective for a wide range of classification tasks.

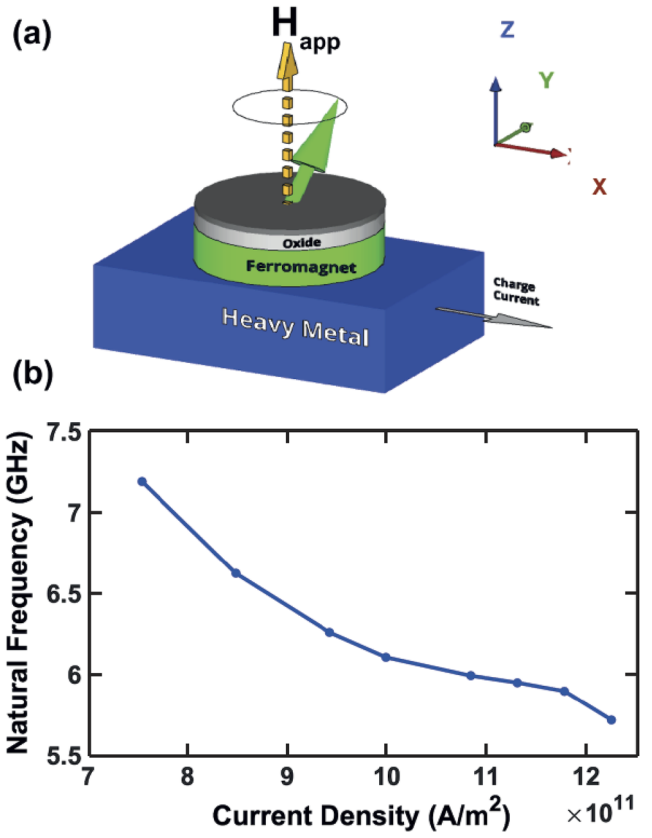


FIG. 1. (a) Schematic of the spin Hall nano oscillator (SHNO), which we simulate micromagnetically. The direction of charge current and that of the applied dc magnetic field (H_{app}) are also shown. z is the vertical/out-of-plane direction. (b) Variation of the natural/auto-oscillation frequency of the SHNO with the magnitude of the charge current density, as obtained from the micromagnetic simulation.

II. MICROMAGNETIC SIMULATION OF UNIFORM-MODE SHNO

A. Natural frequency of SHNO

First, we model a heavy metal/ferromagnetic metal heterostructure-based spin Hall nano oscillator (SHNO) using micromagnetics.[20–22] In-plane charge current flowing through the heavy metal layer results in vertical spin current due to the spin Hall effect. The spin current consequently exerts a spin orbit torque on the magnetization in the ferromagnetic metal layer.[23,24] The ratio of spin current to in-plane charge current density is determined by the spin Hall angle of the heavy metal.[23] If the ferromagnet exhibits perpendicular magnetic anisotropy (PMA),[25] as in the case considered here, the magnetization is found to precess about the vertical axis for an appropriate range of values of in-plane current density due to the spin-orbit torque (Figure 1 (a)), as reported through simulations and experiments previously.[20,25–27]

Here, the ferromagnetic metal layer of diameter 75 nm and thickness 2 nm (in the SHNO) is modeled as a circular grid of magnetic moments in the micromagnetic simulation package 'mumax3.'[16] The package essentially solves Landau Lifschitz Gilbert (LLG) equation (modified by an expression for spin orbit torque) to solve for the dynamics of each magnetic moment under the influence of spin orbit torque due to in-plane charge current flowing through the heavy metal layer below the ferromagnetic metal layer. Dipole coupling and exchange interaction between the magnetic moments are taken into account. The following parameters are used for the magnetic moments in the ferromagnetic layer: exchange correlation constant $(A_{ex}) = 3 \times 10^{-11}$ J/m, saturation magnetization $M_{sat} = 1313 \times 10^3$ A/m, Perpendicular Magnetic Anisotropy (PMA) field $H_k = 1.79$ T (Figure 1 (a)), vertically applied, external dc magnetic field $H_{app} = 0.1$ T, gyromagnetic ratio $\gamma = 17.6 \times 10^{10}$ Hz/T, and damping constant $\alpha = 0.005$.[27–30] To carry out the 'mumax3' simulation, we take the cell size to be 2 nm × 2 nm × 1 nm. 'Mumax3'[16] uses a Runge Kutta method with an adaptive time step to solve the modified LLG equation for the magnetic moments.

In order to simulate the influence of the spin orbit torque on the magnetic moments, spin current, calculated from the charge current using the value of spin Hall angle for the heavy metal (Pt in this case), has been used in our micromagnetic simulation.[16] Thickness of the heavy metal (Pt) layer is taken to be 10 nm, which is greater than the spin diffusion length in Pt.[33,34] Hence, we can consider the vertical spin current density injected by the heavy

© Author(s) 2021

metal layer on the ferromagnetic layer above it (J_s) = in-plane charge current density (J_c) × spin Hall angle (0.07 here, considering Pt).[23,31,32] The field-like torque term has not been considered here because the field-like torque has been found to be very low in such heavy metal/ferromagnetic metal system like the one we have been considered here.[35] Taniguchi *et al.*[27] show that when the ferromagnetic layer exhibits PMA and the spin polarization ($\vec{\sigma}$), due to the charge current flow, is in in-plane, the magnetization ($\vec{m}$) exhibits oscillation (like we observe here next) even in the presence of just Slonczweski-like torque, which has the form: $\vec{m} \times (\vec{m} \times \vec{\sigma})$. The field-like torque term, of the form $\vec{m} \times \vec{\sigma}$, has not been included in Taniguchi *et al.*[27] and yet oscillation is observed. So we do not consider the field-like torque term here and only consider the Slonczweski-like spin-orbit torque term. Micromagnetic simulations on a similar heavy metal (Pt)/ferromagnet system reported elsewhere[8,10,11] also do not consider the field-like torque.

In our micromagnetic simulation, for the value of current density greater than 7.5×10^{11} A/m^2, all the magnetic moments, apart from the ones in the periphery, precess around the z-axis (Figure 1 (a)) while being parallel to each other. This can be seen from the directions of the magnetic moments in the SHNO at different snapshots of time within one time period of precession of the moments Fig. 2. Thus our system is very close to the uniform mode of magnetic oscillation, as mentioned in Sec. II.

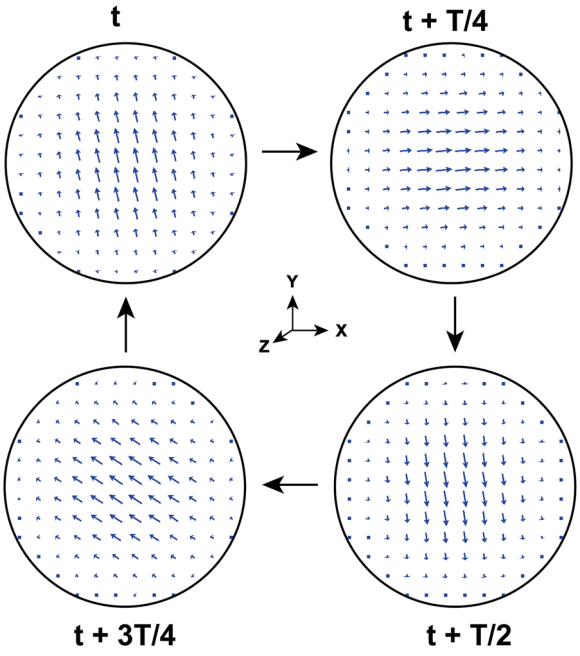


FIG. 2. Different snapshots (in time) of the ferromagnetic layer of SHNO show that its magnetic moments oscillating at a natural frequency of 6.7 GHz, where t is any arbitrary time and *T* is the period of oscillation (inverse of frequency of SHNO) = 1.5×10^{-10} s. Blue dot indicates that magnetic moment is totally out of the plane/vertical (direction: +z), while blue arrow indicates the moment is partially out of the plane (+z) and there is also an in-plane component. The direction of the in-plane component is given by the head of the arrow. 360 degree rotation of the head of the arrow over time shows the precession.

The 'mumax' software provides us with the spatial average of the magnetic moments in x, y, and z directions as functions of time. Since the moments precess about the z-axis here (Fig. 1(a)), we plot the spatial average of the moments in x direction as a function of time to get a sinusoidally varying function of time (refer to the magnetic moment (in x direction) vs time plot in Sec. III of supplementary material). As shown in that plot, the difference in time between two consecutive data points in our simulation is ≈1 ps here. When natural frequency of the oscillator is 6.5 GHz (Fig. 3(c)) and hence time period of each oscillation is ≈153.8 ps, there are about 154 simulated points within each oscillation cycle.

Carrying out Fast Fourier Transform (FFT) of the variation of magnetization with time, the natural frequency/auto-oscillation frequency of the oscillator is obtained. The natural/auto-oscillation

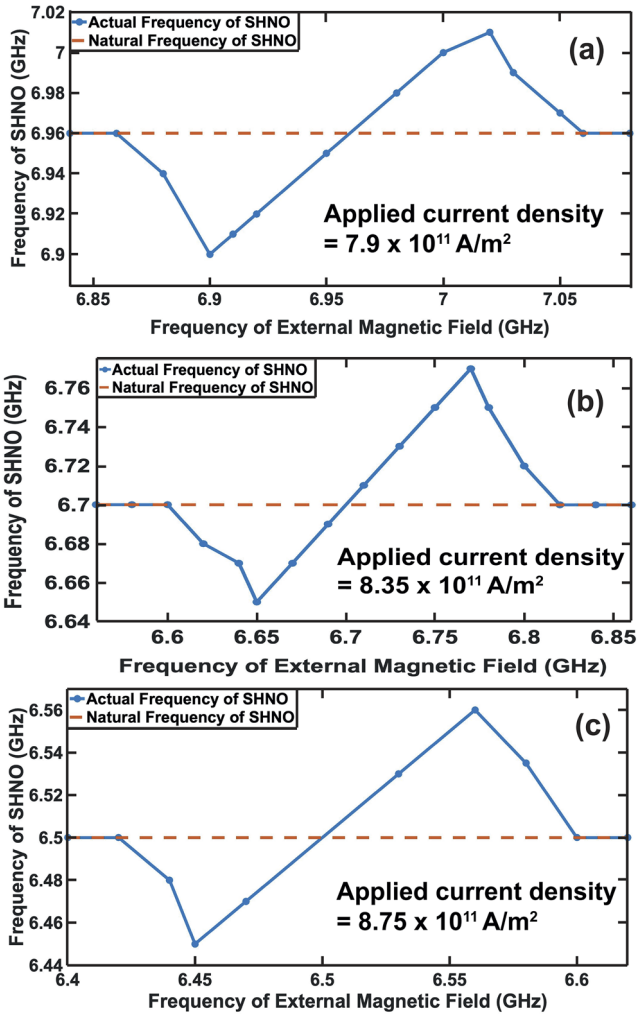


FIG. 3. Synchronization of the oscillator with vertically applied, external RF magnetic field for three different current densities, which result in three different natural frequencies of the oscillator based on Fig. 1(b): (a) applied current density = 7.9×10^{11} A/m^2 (b) applied current density = 8.35×10^{11} A/m^2 (c) applied current density = 8.75×10^{11} A/m^2.

© Author(s) 2021

frequency is defined as the frequency of oscillation in the absence of an external RF magnetic field. From our micromagnetic simulations, we observe that as the current density increases, the natural frequency of oscillation decreases (the precession angle increases), in accordance with Taniguchi *et al.*[27] (Figure 1 (b)). The range in which natural frequency of our SHNO can be modulated with current matches with that observed experimentally on nanoconstriction SHNO with PMA, based on Pt (heavy metal)/CoNi (ferromagnet) bilayer.[26]

While taking FFT of the variation of magnetization with time, we have eliminated the effect of transient oscillations. To do this, while taking FFT, we have not considered the first 70 ns and only considered the next 30 ns. For the range of the natural frequency of the oscillator we consider here, i.e., 6.2–7.2 GHz (as we see in Sec. II B next), this implies that the first 434–504 cycles are ignored (to remove transient effects) and only the next 186–216 cycles are considered for the FFT calculation.

B. Synchronization of SHNO to external RF magnetic field

Similar to Romera *et al.*,[12] the oscillators here have to be synchronized to the frequencies of the external RF magnetic field, corresponding to the inputs in reduced dimension. So we use micromagnetic simulation to determine the synchronizing/locking characteristic of our uniform-mode oscillator with respect to the frequency of external RF magnetic field: Fig. 3(a), (b), and (c). From Fig. 3(b), when the natural frequency of the oscillator is 6.7 GHz (corresponding applied current density = 8.35×10^{11} A/m^2) and an external RF magnetic field of magnitude 1 mT is applied vertically/out of the plane, we observe that within the external field's frequency range of around 6.65–6.77 GHz, the frequency of oscillation of our oscillator is equal to the frequency of the external field, and not the natural frequency of the oscillator. Thus, when natural frequency is 6.7 GHz, the locking range/synchronization bandwidth is about ± 0.06 GHz around the natural frequency (0.12 GHz in total).

Similarly, from all the micromagnetic simulations we carry out of this oscillator on 'mumax3,' we find out that for any other value of natural frequency between 6.2 and 7.2 GHz, the locking range/synchronization bandwidth is also approximately 0.12 GHz (± 0.06 GHz) around that value of natural frequency. For example, Fig. 3(c) shows that when the applied current density is 8.75×10^{11} A/m^2 and hence the natural frequency of the spin oscillator is 6.5 GHz (Fig. 1 (b)), within ± 0.06 GHz of 6.5 GHz, the oscillator is locked/synchronized to the external RF magnetic field. The same bandwidth is observed when the current density is 7.9×10^{11} A/m^2 (natural frequency = 6.96 GHz, consistent with Fig. 1 (b)) (Fig. 3(a)). The magnitude of the external RF magnetic field is 1 mT in all these cases (Fig. 3 (a), (b), and (c)).

This information about the locking range/synchronization bandwidth is useful for tuning the frequencies of the oscillators according to the synchronization pattern for the different input categories by adjusting the applied current on the oscillators, as we see in Sec. III. In this paper, we restrict ourselves to the current-density range of 7.5×10^{11} A/m^2 to 9.6×10^{11} A/m^2 (and hence, the natural-frequency range of 7.2 GHz to 6.2 GHz) such that the synchronization bandwidth is always 0.12 GHz (± 0.06 GHz) around the natural frequency value, as we observe in Fig. 3 (a), (b), and (c).

III. CLASSIFICATION WITH AN ARRAY OF UNIFORM-MODE SHNO-s

A. Dimensionality reduction of input to design the target synchronication pattern

The input in any data set has to be transformed from a much higher dimensional space to a low-dimensional space for classification with a few SHNOs, as shown in the schematic of the system we design for training here (Fig. 4). In Romera *et al.*,[12] the synchronization pattern (2-dimensional cluster plot) of the four oscillators for the different categories of input (7 vowel sounds) is first decided upon. Accordingly, using least-square regression in software, the matrix coefficients are calculated such that each higher dimensional input vector, belonging to an input category (vowel sound), is linearly transformed, through the matrix multiplication, to a 2-dimensional vector, corresponding to the cluster assigned for

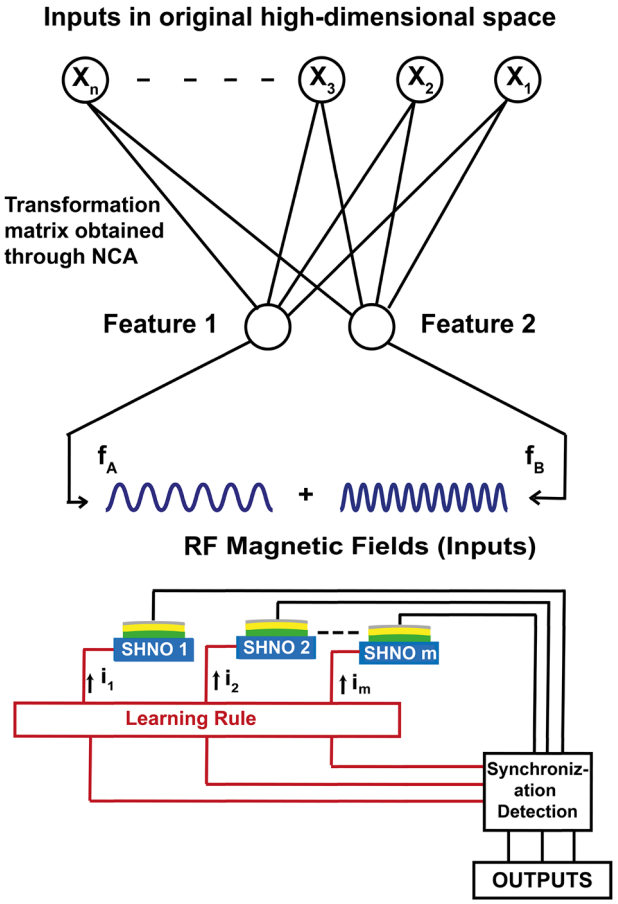


FIG. 4. Schematic of the system of SHNOs used here for data classification. The input is transformed from the original high-dimensional space to a two-dimensional space (Feature 1, Feature 2) using the NCA technique as shown in Fig. 5. RF magnetic field with frequencies f_A, proportional to Feature 1, and f_B, proportional to Feature 2, is applied on m SHNOs. The SHNOs are selectively synchronized with the input frequencies to determine the category/label of the input, as shown in Figure 6. Currents entering the m SHNOs ($i_1, i_2, \ldots, i_m$) are modulated for this purpose.

© Author(s) 2021

that category in the synchronization pattern. After that, the frequencies of the four oscillators are mapped to the synchronization pattern following a learning rule.[12]

But if this method in Romera *et al.*[12] is employed for not just the vowel data set but for many other data sets, then the clusters in the lower dimensional space (of the input) do not change from one data set to another. Only the transformation matrix coefficients will change from one data set to another, while the cluster configuration remains fixed. Thus, this method does not take advantage of the inherent nature of the data in different data sets for cluster formation and subsequent classification.

But in this paper, since we have shown classification on different data sets as opposed to one data set in Romera *et al.*,[12] we have used a more versatile technique—Neighbouring Components Analysis (NCA)[36]—to reduce the input dimensions and determine the clusters/synchronization pattern. Here in the NCA method also, the transformation from higher to lower dimensional space of the input is linear (can be represented by a matrix multiplication) (Fig. 4) just as in the dimensionality reduction method in Romera *et al.*[12] Also, supervised learning is followed in both cases. But in Romera *et al.*,[12] the data-independent oscillator synchronization pattern is used for supervision to calculate the coefficients of the transformation matrix, making the clusters independent of data sets.

In our NCA technique, as described in Goldberger *et al.*,[36] we iteratively calculate the coefficients of the transformation matrix A such that the linear transformation of a high dimensional data point x_i i.e., Ax_i gives a point in lower dimension. NCA exploits the presence of labels in a supervised manner to arrive at this matrix A such that if two points x_i and x_j belong to same class C, they are closer in the transformed space; otherwise, they are far apart.

While the reduction of the input from the higher-dimensional space to the lower-dimensional space is linear both here and in Romera *et al.*[12] (both given by a transformation matrix), the formation of the clusters in the reduced two-dimensional space depends on the nature of the input data and varies from data set to data set here, as shown in Fig. 5. This is not the case in Romera *et al.*[12] since a fixed synchronization pattern is used there. Thus our technique is more versatile than that followed in Romera *et al.*[12]

In our case, for the WBC data set, the number of input features is 30. So for WBC, we carry out a 30-to-2 reduction. For the Iris data set, the number of input features is 4 (sepal length, sepal width, petal length, petal width). So for Iris, we carry out a 4-to-2 reduction. For the MNIST data set, the number of input features is 784 since each image is a 28-by-28-pixel image. So for MNIST, we carry out a 784-to-2 reduction (see Sec. I of supplementary material). The coefficients of the final matrix for input-dimension reduction, obtained through the aforementioned NCA technique mentioned above, have been provided in Sec. VI of supplementary material for all the data sets: WBC, Iris, and MNIST.

For the different data sets, the two features in the reduced two-dimensional space in Fig. 5 (Feature 1, Feature 2) are then transformed to two input RF magnetic field frequencies f_A and f_B (Fig. 4). Please refer to Sec. VII of supplementary material for details on this transformation. These two transformed frequencies must be in that range of natural frequencies (6.2–7.2 GHz) of our uniform-mode spin oscillators where the synchronization bandwidth is nearly constant, as determined in Sec. II B. This leads to the generation of the target synchronization patterns for our spin oscillators.

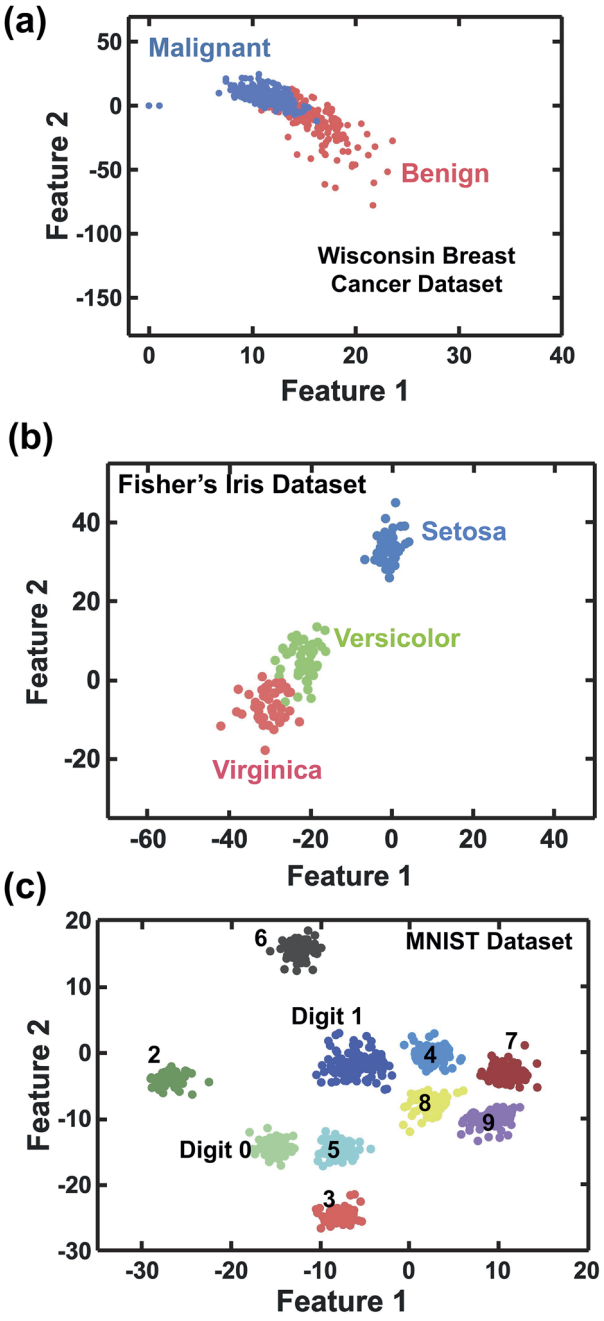


FIG. 5. Through the application of the versatile NCA technique, each input in (a) Wisconsin Breast Cancer (WBC) (b) Fisher's Iris (c) MNIST data sets is transformed into two dimensions: Feature 1, Feature 2. Then Feature 2 is plotted against Feature 1 for each of the data sets to show clustering of the data in reduced dimensions.

B. Tuning the SHNOs' frequencies to the target synchronization pattern to achieve learning

Though our method for the generation of the clusters/target synchronization pattern is different from that followed in Romera

© Author(s) 2021

et al.,[12] our method for modulating the frequencies of the oscillators to match with the synchronization pattern is similar to that used in Romera *et al.*[12] Much like in Romera *et al.*,[12] we also modulate the currents going into the oscillators to change their natural frequencies, as shown in our system-level schematic (Fig. 4). For this purpose, the system of oscillators is modeled in the numerical package Matlab. Each oscillator's natural frequency modulation characteristic (Fig. 1(b)) and synchronization (to external RF magnetic field) characteristic (Fig. 3) are obtained from micromagnetic simulations of the SHNO device, as shown in Sec. I, and incorporated in the Matlab code. Thus this becomes a device-system co-study.

For all the data sets we use here, the natural frequencies of the oscillators are initialized to random values within the chosen frequency range of 6.2–7.2 GHz since the synchornization bandwidth is almost constant in this range (± 0.06 GHz around the natural frequency value, as mentioned in Sec. II B). Then during the training process, the natural frequencies of the oscillators have been updated by modulating their current densities, using a current density step of around 1×10^{10} A/m^2, but again restricting the natural frequencies within that range of 6.2–7.2 GHz. Since our target synchronization pattern is also in this range, this restriction works fine for our training process.

The learning rule we use for updating the natural frequency of each oscillator after every iteration in our training process is described in Sec. IV of supplementary material. Our learning rule ignores the existence of mutual coupling between the oscillators.

Romera *et al.* use 4 vortex oscillators to classify the seven vowel sounds. In our paper, since we use different data sets for classification, the number of oscillators (m in Fig. 4) used varies from one data set to data set. The number of oscillators needed increases with an increase in the number of categories to be classified into, or output labels, in a data set. So we use 2 oscillators for WBC data set (2 output labels: 'malignant' or 'benign') (Fig. 5(a)), 2 oscillators for Iris data set (3 output labels for the 3 flower types: setosa, virginica, and versicolor) (Fig. 5(b)), and 5 oscillators for MNIST data set (10 output labels for the 10 digits: 0 to 9) (Fig. 5(c)).

For the WBC data set, we classify the input as 'malignant' when the oscillator 2 couples to f_B and 'benign' when the oscillator 1 couples to f_A [Fig. 6(a)]. Thus the output of the synchronization detector in Fig. 4 determines the category/label of the input. Although the training set and the test set must be taken from the same data set, we must ensure that the samples in the test set are different from the ones in the training set. So we have used the first 455 samples in the WBC data set for training and the next 114 samples for testing. The 30 input features corresponding to some of the training and test samples in the WBC data set are tabulated in Sec. I of supplementary material. After 150 iterations, training accuracy is 96 % (Fig. 7) and test accuracy is 92 % (Table I). Intermediate stages of the training are shown in Fig. 8.

We verify that the final synchronization pattern, after training, for the WBC data set is correct (Fig. 6(a)) by carrying out further micromagnetic simulations of the oscillators on 'mumax3,' in the presence of two external RF magnetic fields with frequencies (f_A and f_B) simultaneously applied. As observed from Fig. 6(a), the two-oscillator system is trained on the WBC data set when oscillator 1's natural frequency is 6.45 GHz and oscillator 2's natural frequency is 6.7 GHz. In Sec. V of supplementary material, we show through micromagnetic simulations that when current densities through the

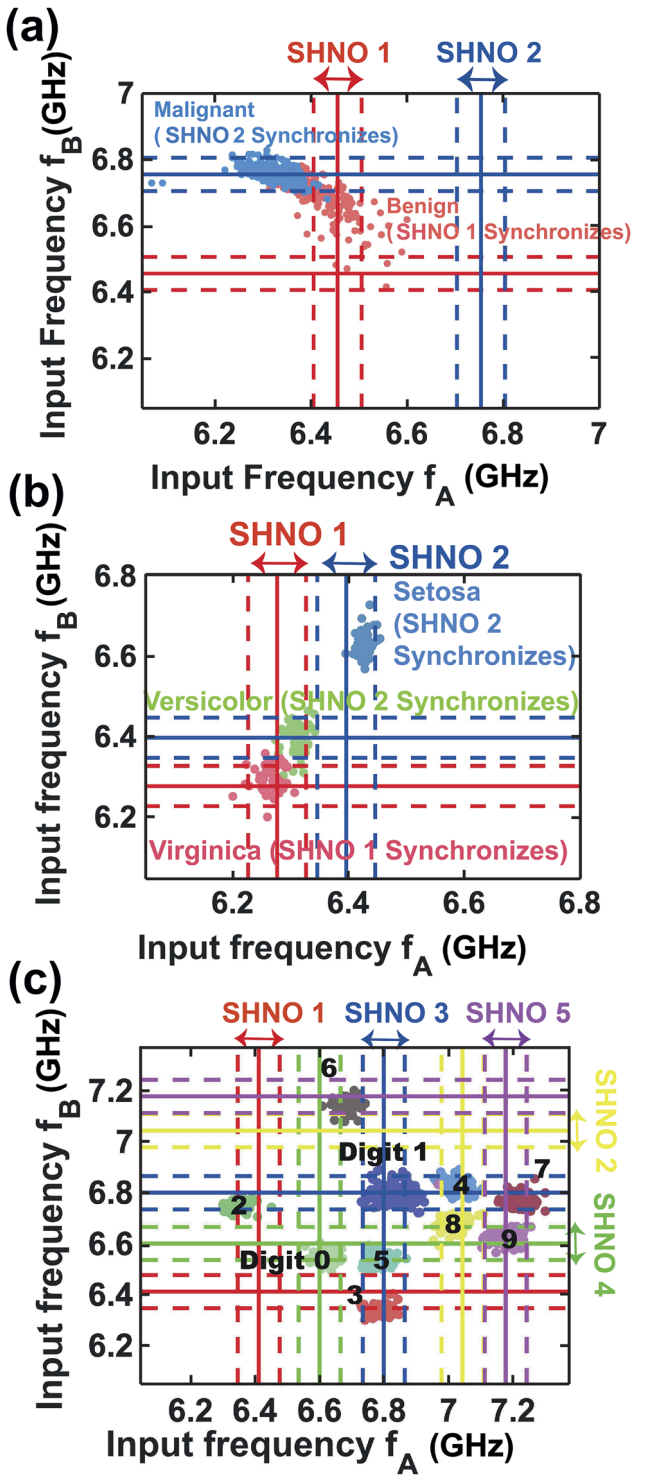


FIG. 6. The figure shows the final synchronization pattern obtained after training. The oscillators (SHNOs) are suitably synchronized with the input frequencies (of RF magnetic field) to achieve learning for the data sets (a) WBC (b) Fisher's iris (c) MNIST. The number of steps needed to achieve the final training/learning stage, shown here, is 150 for all these data sets.

© Author(s) 2021

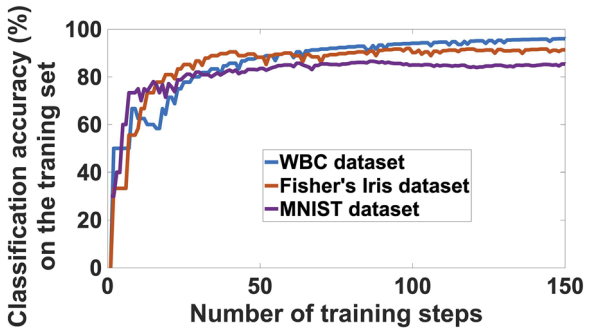


FIG. 7. Classification accuracy, for different training data sets, obtained from our uniform-mode-SHNO-based network.

oscillators are such that the natural frequencies of the two oscillators take the above-mentioned values and the f_A and f_B values correspond to that of the different samples in the 'malignant' cluster, then oscillator 2's frequency changes with f_B value while oscillator 1's frequency neither follows f_A or f_B. But when the f_A and f_B values correspond to that of the different samples in the 'benign' cluster, then oscillator 1's frequency changes with the f_A value while oscillator 2's frequency neither follows f_A or f_B. This is what is expected from Fig. 6(a): for 'malignant' cluster, oscillator 2 synchronizes with the second external RF magnetic field (of frequency f_B) and for 'benign' cluster, oscillator 1 synchronizes with the first external RF magnetic field (of frequency f_A). Thus our micromagnetic simulation results under two external RF magnetic fields agree with the synchronization pattern (after training) obtained from the Matlab code.

For the Iris data set, we classify the input as of the 'setosa' flower type when oscillator 2 synchronizes with f_A, as 'versicolor' type when oscillator 1 synchronizes with f_A and oscillator 2 synchronizes with f_B at the same time, and as 'virginica' type when oscillator 1 synchronizes with f_A (Fig. 6(b)). Similar to what we have done for the WBC data set, here we have used the first 120 samples of the Iris data set for training and the next 30 samples for testing. The 4 input features corresponding to some of the training and test samples in the Iris data set are tabulated in Sec. I of supplementary material. After 150 iterations, training accuracy is 91 % (Fig. 7) and test accuracy is 94 % (Table I). Intermediate stages of the training are shown in Fig. 9.

For the MNIST data set, we have similar combinations of synchronization of different oscillators with the input frequencies to classify the data set (Fig. 6(c)). The MNIST data base, by default, provides us with different images of handwritten digits (0–9) for training and different images for the same handwritten digits (0–9) for testing. From this MNIST data base, we took 800 images for training and 200 images for testing. Some of the input images (28-by-28-pixel) for the training set and the test set are shown in Sec. I of supplementary material. After 150 iterations, training accuracy is 85 % (Fig. 7) and test accuracy is 83 % (Table I).

Thus, we obtain fairly high classification accuracy numbers for all the data sets, showing that we have achieved successful learning with our system of uniform-mode oscillators. For comparison, we train in software (Matlab) a standard single-layer perceptron/Fully Connected Neural Network (FCNN) without a hidden layer.[38] Such an FCNN has a similar number of parameters as our oscillator-based system since our system includes the parameters in the matrix which transforms the input data from higher dimensional space to two-dimensional space (Fig. 4). We use the same three data sets as before for the single-layer perceptron/FCNN. The non-linear activation function used at the output layer here is the very popular 'softmax' function.[37] Stochastic Gradient Descent method is used to optimize the mean-squared error over every iteration.[38] The structure and underlying non-linear function/equations for the FCNN, which we use here for comparison, are provided in Sec. II of supplementary material.

While the final train accuracy numbers are comparable or higher for the single-layer perceptron compared to our SHNO network (Table I), it is the performance on the test set after training (test accuracy) that matters more.[37] We observe that both for Iris and MNIST data sets, the test accuracy is comparable or higher for our uniform-mode oscillator network compared to the single-layer perceptron (Table I). The test accuracy for FCNN improves drastically when we add a hidden layer, but then the number of parameters that need to be tuned also goes up.[12,38]

C. Power dissipation estimation

As mentioned in Sec. I, making oscillator-based systems scalable for data classification is a major motivation for this study. An important aspect of scalability for computing systems is that when the system is scaled, as we propose to do here by reducing the area of the oscillator by 25X from Romera *et al.*[12] the power dissipation per component (the oscillator here) also must go down and thus the power dissipation per unit area of the chip must remain roughly the same.[39]

We now show that this is indeed the case in our work. While the current density in our uniform-mode spin Hall nano oscillator is one order higher than that in the vortex-mode spin oscillator of Romera *et al.*,[12] the area of cross-section is different in the two cases. In Romera *et al.*,[12] the current flows perpendicular to the plane. So the

TABLE I. Comparison of classification accuracy numbers between uniform-mode SHNO network, simulated in this paper, and a standard single-layer perceptron/Fully Connected Neural Network without a hidden layer, trained in software.

Type of learning	WBC: Train accuracy (%)	WBC: Test accuracy (%)	Iris: Train accuracy (%)	Iris: Test accuracy (%)	MNIST: Train accuracy (%)	MNIST: Test accuracy (%)
Our SHNO network	96	92	91	94	85	83
Single-layer perceptron	96	98	97	85	97	82

© Author(s) 2021

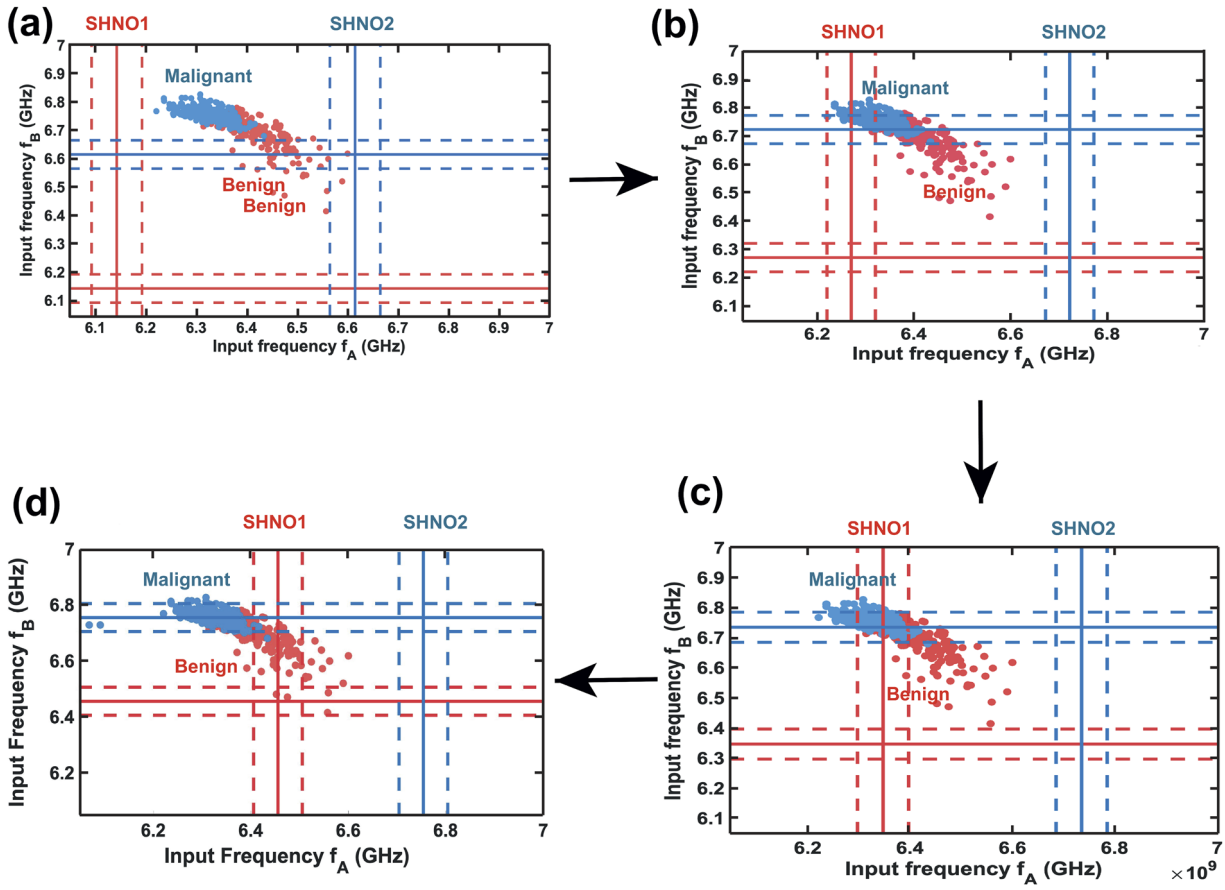


FIG. 8. (a) shows the initial stage of the training for WBC data set when the oscillators are not synchronized to the external RF magnetic fields for inputs in the given data set. A couple of intermediate stages of the training are shown in (b) and (c). (d) shows the synchronization pattern once training is achieved (after 150 iterations) and matches with Fig. 6 (a). The input frequencies f_A and f_B for the different input samples belonging to the WBC data set remain unchanged during the training process and already shown in Fig. 5 (a). During training, the natural frequencies of the two spin Hall nano oscillators (SHNOs) are tuned gradually such that SHNO1 couples to the first external RF magnetic field with input frequency f_A for samples belonging to the 'benign' category (shown in red) and SHNO2 couples to second external RF magnetic field with input frequency f_B for samples belonging to the 'malignant' category (shown in blue).

area of cross-section for the current flow is equal to the lateral area of the oscillator, which is proportional to the square of the diameter of the tunnel junction.

But the area of cross-section is different for our uniform-mode SHNO since the current flows in-plane here. Here, the area of cross-section for current flow is roughly equal to the product of the diameter of the SHNO and the thickness of the heavy metal layer in it (10 nm). The thickness of the heavy metal layer is much lower than the lateral diameter of the oscillator. Also the diameter of the oscillator in our case is 5 times lower than that in Romera *et al.* As a result, the area of cross-section here is much lower than Romera *et al.* This leads to lower current, despite the same current density. For current density of 7.5×10^{11} A/m^2 in our SHNO here, we calculate the current to be 0.56 mA. Resistance of the current path = resistivity of Pt x (length of the path/area of cross-section). Considering the length of the path roughly equal to the diameter of the oscillator, resistance of the current path = resistivity of Pt/thickness of the heavy metal (Pt) layer = 10 Ω. Thus in our case, power dissipation per oscillator is calculated to be 3 μW, which is three orders lower than that reported for the vortex oscillator in Romera *et al.*[12] Since the area of the oscillator has gone down from Romera *et al.*[12] to our proposed system here by 25 X, power dissipation per unit area remains roughly the same.

While calculating the power consumed by our spin Hall oscillators, we did not include the read power consumed in detecting the synchronization pattern of the oscillators. But the read power is dominated by the magnitude of the read current, which in spin Hall oscillators is typically several orders lower than the write current.[23,24] This is because the write current has to be strong enough to trigger magnetic oscillation through the transfer of spin-orbit torque to the magnetization. On the contrary, the read current merely detects the direction of the magnetization with respect to a fixed direction (direction of magnetization in the fixed layer of the tunnel junction) through the Tunneling MagnetoResistance (TMR) effect. Designing

© Author(s) 2021

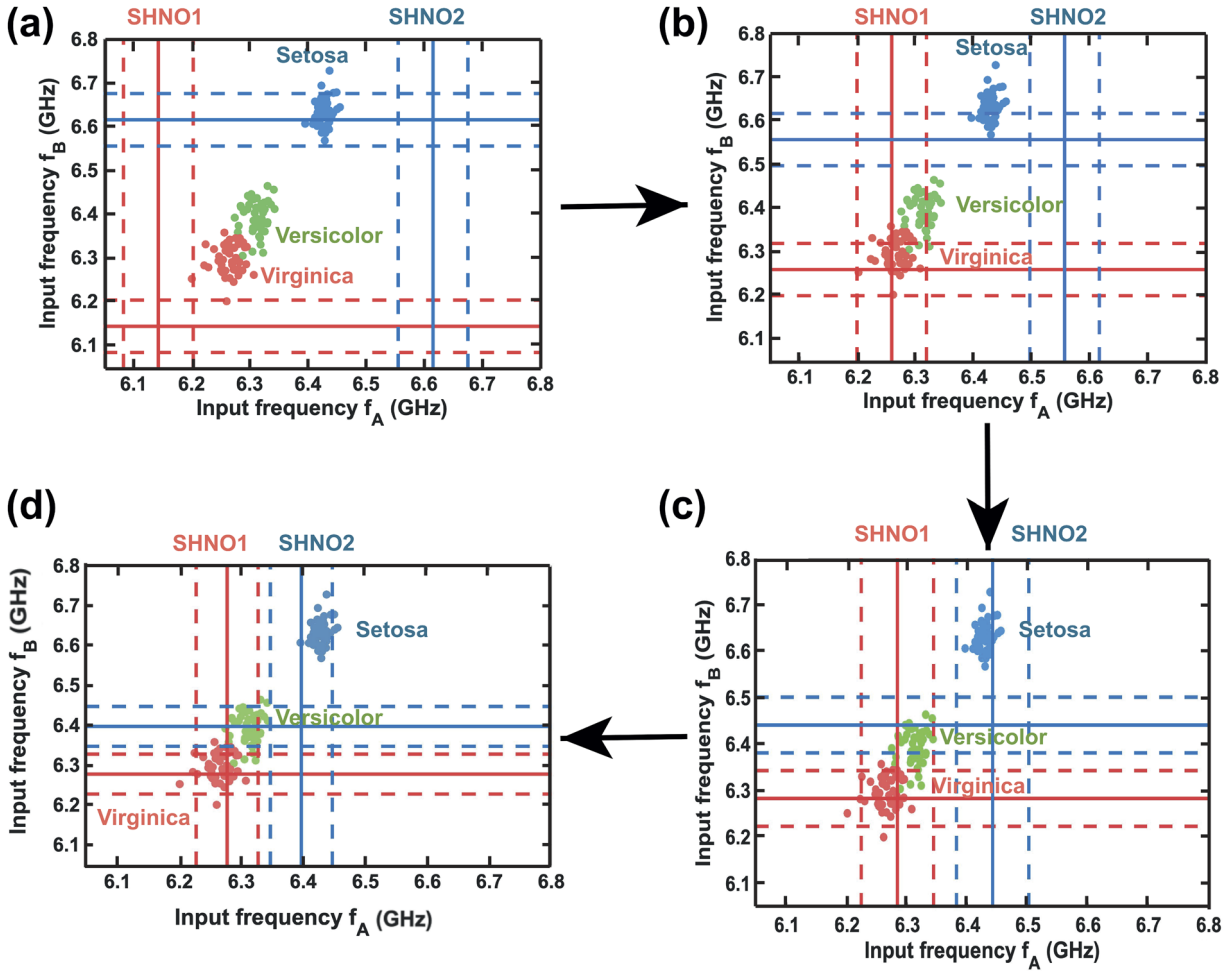


FIG. 9. (a) shows the initial stage of the training for Iris data set. A couple of intermediate stages of the training are shown in (b) and (c). (d) shows the synchronization pattern once training is achieved (after 150 iterations) and matches with Fig. 6 (b). The input frequencies f_A and f_B remain unchanged during the training process and already shown in Fig. 5 (b). During training, the natural frequencies of the two SHNOs are tuned gradually such that SHNO2 couples to the first external RF magnetic field with input frequency f_A for samples belonging to the 'setosa' category (shown in blue) and SHNO1 couples to the same external RF magnetic field with input frequency f_A for samples belonging to the 'virginica' category (shown in red). For 'versicolor,' SHNO1 synchronizes with f_A and SHNO2 synchronizes with f_B at the same time (shown in green).

the read-out circuit for detection of synchronization of the oscillators and calculation of the read power consumed in the circuit will be a part of our future study.

For the natural frequency range of our spin Hall oscillator (6.2–7.2 GHz), ignoring the read power, the energy per oscillation turns out to be 0.41–0.48 fJ since the write power/power to generate the oscillation is ≈ 3 μW. Our energy per oscillation turns out to be 1–2 orders lower than the energy per spike reported for the implementation of spiking neural networks through spintronic systems.[42]

IV. CONCLUSION

Thus, in this paper, we have shown that a system of very small SHNOs (75 nm in diameter), where the moments precess in a nearly uniform mode, can be used to classify inputs in different popular machine learning data sets. Through a combination of device-level micromagnetic study and system-level simulations, we show reasonably high classification accuracy with our SHNO-based system. Our study is important for scalable spintronic neuromorphic systems for this kind of data classification tasks.

While coupling between the oscillators and the external RF magnetic fields, the frequencies of which depend upon the input in reduced dimensions, has been considered in our simulations, the coupling among the oscillators themselves has been ignored here, as mentioned earlier. Including coupling among the oscillators in the simulation during the training process, as in Romera *et al.*,[12] Vassileva *et al.*[40] and Vodenicarevic *et al.*,[41] for the different data sets used here, will be a part of our future study.

© Author(s) 2021

SUPPLEMENTARY MATERIAL

See supplementary material for more details or additional data on the data sets used, the input-dimension reduction technique, the learning rule followed, and micromagnetic simulation of the spin oscillators.

ACKNOWLEDGMENTS

We thank the Department of Science and Technology, India, for INSPIRE Faculty Award and the Science and Engineering Research Board, India, for Early Career Research Award. These awards helped fund this research.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]J. Grollier, D. Querlioz, K. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, Nature Electronics **3**, 360 (2020).
[2]A. Sengupta, Y. Shim, and K. Roy, IEEE Transactions on Biomedical Circuits and Systems **10**, 1152 (2016).
[3]U. Saxena, D. Kaushik, M. Bansal, U. Sahu, and D. Bhowmik, IEEE Transactions on Magnetics **54**, 1 (2018).
[4]Y. Huang, W. Kang, X. Zhang, Y. Zhou, and W. Zhao, Nanotechnology **28**, 08LT02 (2017).
[5]T. Bhattacharya, S. Li, Y. Huang, W. Kang, W. Zhao, and M. Suri, IEEE Access **7**, 5034 (2019).
[6]O. Akinola, X. Hu, C. H. Bennett, M. Marinella *et al.*, J. Phys. D: Appl. Phys. **52**, 49LT01 (2019).
[7]A. Sengupta and K. Roy, Appl. Phys. Rev. **4**, 041105 (2017).
[8]D. Bhowmik, U. Saxena, A. Dankar, A. Verma, D. Kaushik, S. Chatterjee, and U. Singh, J. Magn. Magn. Mater. **489**, 165434 (2019).
[9]N. Hassan, X. Hu, L. Jiang-Wei, W. Brigner, G.-S. Akinola, M. Pasquale, C. H. Bennett, J. A. C. Incorvia, and J. S. Friedman, J. Appl. Phys. **124**, 152127 (2018).
[10]U. Sahu, A. Pandey, K. Goyal, and D. Bhowmik, AIP Adv. **9**, 125339 (2019).
[11]D. Kaushik, U. Singh, U. Sahu, I. Sreedevi, and D. Bhowmik, AIP Adv. **10**, 025111 (2020).
[12]M. Romera, P. Talatchian, S. Tsunegi, F. Abreu Araujo, V. Cros, P. Bortolotti, J. Trastoy, K. Yakushiji, A. Fukushima, H. Kubota *et al.*, Nature **563**, 230 (2018).
[13]M. Gajek, J. Nowak, J. Sun, P. Trouilloud, E. O'sullivan, D. Abraham, M. Gaidis, G. Hu, S. Brown, Y. Zhu *et al.*, Appl. Phys. Lett. **100**, 132408 (2012).
[14]M. Zahedinejad, A. A. Awad, S. Muralidhar, R. Khymyn, H. Fulara, H. Mazraati, M. Dvornik, and J. Åkerman, Nature Nanotechnol. **15**, 47 (2020).
[15]K. L. Metlov and K. Y. Guslienko, J. Magn. Magn. Mater. **343**, 55 (2013).
[16]A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, AIP Adv. **4**, 107133 (2014).
[17]R. A. Fisher, Annals of Eugenics **7**, 179 (1936).
[18]W. N. Street, W. H. Wolberg, and O. L. Mangasarian, Proc. SPIE **1905**, 861 (1993).
[19]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Proceedings of the IEEE **86**, 2278 (1998).
[20]T. Chen, R. K. Dumas, A. Eklund, P. K. Muduli, A. Houshang, A. A. Awad *et al.*, Proceedings of the IEEE **104**, 1919 (2016).
[21]A. Giordano, M. Carpentieri, A. Laudani, G. Gubbiotti, B. Azzerboni, and G. Finocchio, Appl. Phys. Lett. **105**, 042412 (2014).
[22]Z. Duan, A. Smit, L. Yang, B. Youngblood, J. Lindner, V. E. Demidov, S. O. Demokritov, and I. N. Krivorotov, Nature Communications **5**, 5616 (2014).
[23]L. Liu, C. F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, Science **336**, 555 (2012).
[24]D. Bhowmik, L. You, and S. Salahuddin, Nature Nanotechnology **9**, 59 (2014).
[25]R. H. Liu, W. L. Lim, and S. Urazhdin, Phys. Rev. Lett. **114**, 137201 (2015).
[26]B. Divinskiy, V. E. Demidov, A. Kozhanov, A. B. Rinkevich, S. O. Demokritov, and S. Urazhdin, Appl. Phys. Lett. **111**, 032405 (2017).
[27]T. Taniguchi, H. Arai, H. Kubota, and H. Imamura, IEEE Transactions on Magnetics **50**, 1 (2014).
[28]S. Yakata, H. Kubota, Y. Suzuki, K. Yakushiji, A. Fukushima, S. Yuasa, and K. Ando, J. Appl. Phys. **105**, 07D131 (2009).
[29]H. Kubota, S. Ishibashi, T. Saruya, T. Nozaki, A. Fukushima, K. Yakushiji, K. Ando, Y. Suzuki, and S. Yuasa, J. Appl. Phys. **111**, 07C723 (2012).
[30]A. Giordano, M. Carpentieri, A. Laudani, G. Gubbiotti, B. Azzerboni, and G. Finocchio, Appl. Phys. Lett. **105**, 042412 (2014).
[31]L. Q. Liu, T. Moriyama, D. C. Ralph, and R. A. Buhrman, Phys. Rev. Lett. **106**, 036601 (2011).
[32]L. Q. Liu, O. J. Lee, T. J. Gudmundsen, D. C. Ralph, and R. A. Buhrman, Phys. Rev. Lett. **109**, 096602 (2012).
[33]A. Berger, E. R. J. Edwards, H. T. Nembach, O. Karis, M. Weiler, and T. J. Silva, Phys. Rev. B **98**, 024402 (2015).
[34]D. Qu, S. Y. Huang, B. F. Miao, S. X. Huang, and C. L. Chien, Phys. Rev. B **89**, 140407 (2014).
[35]S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. D. Beach, Nature Materials **12**, 611 (2013).
[36]J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, Advances in Neural Information Processing Systems **17**, 513 (2005).
[37]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2016).
[38]Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[39]C. Hu, *MOS Transistor in Modern Semiconductor Devices for Integrated Circuits* (Pearson, 2010).
[40]E. Vassilieva, G. Pinto, J. de Barros, and P. Suppes, IEEE Transactions on Neural Networks **22**, 84 (2011).
[41]D. Vodenicarevic, N. Locatelli, J. Grollier, and D. Querlioz, J. Appl. Phys. **124**, 152117 (2018).
[42]A. Sengupta, A. Banerjee, and K. Roy, Phys. Rev. Appl. **6**, 064003 (2016).

© Author(s) 2021

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-10832-0

1174: FUTURISTIC TRENDS AND INNOVATIONS IN MULTIMEDIA SYSTEMS USING BIG DATA, IOT AND CLOUD TECHNOLOGIES (FTIMS)

Low bandwidth data hiding for multimedia systems based on bit redundancy

Neeraj Kumar[1] · Rajeev Kumar[2] · Aruna Malik[3] · Samayveer Singh[3]

Received: 12 May 2020 / Revised: 23 December 2020 / Accepted: 10 March 2021
Published online: 09 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

In recent years, low bandwidth data hiding schemes for multimedia systems are being seen as a promising new technology for multimedia information protection and rights management. More specifically, Absolute moment preserving block truncation coding (AMBTC) based data hiding schemes have gain a wide attraction among the researches due to their low complexity and high compression ratio. One such scheme proposed by Ou et al. has been a breakthrough in the field as the scheme provides high capacity with only slight degradation in image quality. The Ou et al.'s scheme basically divides the AMBTC blocks into smooth and complex categories based on a user-defined threshold and embeds the secret data only into the smooth blocks as data embedding into complex blocks result in severe degradation of image quality. To address this problem of non-embeddability in complex blocks, this paper proposes a high capacity data hiding scheme which efficiently utilizes the complex image blocks for embedding the secret data without any degradation in the image quality unlike the traditional AMBTC based schemes. The proposed scheme basically utilizes the bit redundancy of image blocks for creating a room for secret data bits inside bitmaps of AMBTC codes. Thus, the scheme significantly increases the embedding capacity without any further degradation in image quality. To validate the performance of proposed scheme, experimental results are compared with existing techniques which show significant improvement in embedding capacity while maintaining the stego-image quality.

Keywords Bit redundancy · Secret data · Embedding capacity · Data hiding · AMBTC

✉ Samayveer Singh
samayveersingh@gmail.com

Extended author information available on the last page of the article

1 Introduction

Digital transformation of information has made human life very easier and comfortable. With the help of technology, any type of information can be converted into digital form and can be easily stored, transmitted, modified and received. That's why, people are communicating by sharing/transmitting digital multimedia contents, such as audio, video, image over the public platforms [5]. However, the security of these platforms has been a concern [9, 31]. In highly sensitive fields, like defence, medical, information security has the utmost priority. Therefore, in such fields, instead of transmitting information directly, information is secretly embedded inside a cover media and then it is transmitted over the public platforms. The technique of embedding the secret data in other media is called steganography technique [23]. Therefore, even if transmitting medium is not secure, information can be covertly sent over the public platform. There have been discussed a number of steganography techniques [1–4, 6–8, 10, 12, 16–22, 24–28, 32, 33] to provide the solution for secure data embedding with an aim of high embedding capacity along with good quality stego-image. The steganography techniques have been categorized into three categories (namely spatial domain [1, 14–17, 23, 29], transform domain [32] and compression domain [2, 10, 18, 22, 26]) based on their style of working. Among them compression domain-based steganography techniques are quite popular in the application areas where communication bandwidth is a concern. In the compression domain, all of the popular data compression techniques such as Lempel–Ziv–Welch (LZW), Joint Photographic Experts Group (JPEG), Vector quantization, AMBTC have been extensively researched for efficient use in steganographic system. Among them, AMBTC based steganography schemes have attracted a lot of eye gaze of relevant researchers due to their low complexity and high compression ratio. Block Truncation coding (BTC) (from which AMBTC has been derived) is lossy compression technique which was first introduced by Delp and Mitchell [7] in 1979 and they named the technique as Moment Preserving Block Truncation Coding (MPBTC). In 1984, Lema and Mitchell disclosed a modified version of MPBTC technique for improving [7] primarily on complexity aspect. The modified version is popularly known as Absolute Moment Preserving Block Truncation Coding (AMBTC) technique [25]. The AMBTC enhances image quality and computational cost of BTC technique. The details regarding encoding process of standard Block Truncation Coding and AMBTC encoding process are presented in sub-sections 1.1 & 1.2, respectively. The AMBTC technique caught the lot of eye balls of prominent researchers due to its low computational complexity aspect. Hence, multiple variations of BTC/AMBTC techniques [4, 8, 24, 33] have been introduced to further improve their performance. Many of the improvements have mostly focussed on analysing bitmap to improve BTC/AMBTC based coding efficiency and compression rate.

1.1 Standard BTC technique

In BTC compression technique, an image is non-overlappingly partitioned into number of uniform size blocks of size$n \times n$. An average value ($\bar{x}$) and standard deviation (σ) is of each block is computed using Eqs. (1) and (2), respectively, as follows:

$$\bar{x} = \frac{1}{n*n} \sum_{i=1}^{n*n} (x_i) \tag{1}$$

$$\sigma = \sqrt{\frac{1}{n*n}\sum_{i=1}^{n*n}\left(x_i - \bar{x}\right)^2} \tag{2}$$

A bitmap B(of size$n \times n$) is created by comparing each pixel value with the average value ($\bar{x}$) of the block. If the pixel value is greater than or equal to $\bar{x}$, then replace pixel value by 1 and otherwise replace by 0. The calculated average ($\bar{x}$) and standard deviation (σ) arethen used to compute the two quantization values a and b which represent the original block along with the bitmap B, using Eqs. (3) and (4), respectively, as follows:

$$\mathrm{a} = \bar{x} + \sigma\sqrt{\frac{s}{v}} \tag{3}$$

$$\mathrm{b} = \bar{x} - \sigma\sqrt{\frac{s}{v}} \tag{4}$$

where s and $v$represent the number of 0's and 1's in the bitmap, respectively. Thus, BTC code data for each block is obtained in form of trio a, b, B.At receiver end, the image block can be reconstructed by replacing 1 with a and 0 with b in bitmap.

1.2 Standard AMBTC technique

In AMBTC compression technique, the image is partitioned in the same way as in BTC. Then, two mean quantization values, those are high and low mean quantization value are calculated by separating the pixel values of the block into two sections based on pixel comparison with average value $\bar{x}$ of the block. Pixels greater than or equal to $\bar{x}$ are replaced by 1 and otherwise replace by 0. Then, a high mean quantization value x_h is calculated by taking average of pixel values corresponding to 1 as disclosed in Eqs. 5 and similarly a low mean quantization value x_l is calculated by taking average of pixel values corresponding to 0 as disclosed in Eqs. 6.Thus, the values a and b are computed using Eqs. (5) and (6) as follows:

$$a = \frac{1}{p}\sum_{x_i \le \bar{x}} x_i \tag{5}$$

$$b = \frac{1}{q}\sum_{x_i \ge \bar{x}} x_i \tag{6}$$

where p and q represent the number of '0's and '1's in the bitmap, respectively. In other words, the values of a is calculated by taking average of pixels corresponding to 0 in the bitmap and in which b is calculated by taking average of pixels corresponding to 1 in the bitmap, respectively. Thus, both BTC and AMBTC compression techniques provide two 8-bit quantization values (low and high quantization values) $aandb$ and an associated 16-bit bitmap B for an image block of a predetermined size (e.g. 4×4). Thus, the complete image block of 4×4 pixels in compressed form can be represented by only 32-bit sized trio a, b, B.

Another dimension in the BTC/AMBTC compression techniques have been added by the security researchers, this dimension is popularly known as covert communication. The covert communication has also attracted a lot of researchers which has led to significant improvement in the performance of data hiding techniques. Various data hiding techniques based on BTC/AMBTC are disclosed in research papers [3, 6, 12, 19–21, 27, 28, 30]. From the analysis/survey of various research papers discussing the BTC/AMBTC based data hiding techniques [3, 6, 12, 19–21, 27, 28, 30], it is quite clear that most of these techniques have focussed on data hiding on smooth blocks. However, they have almost kind of ignored the complex blocks which may also contain some smooth pixels that can also be compressed as well as can be utilized for data embedding. Thus, to overcome the limitation of the existing scheme, the proposed scheme tries to make an attempt of efficiently embedding the secret data into the complex blocks in addition to smooth blocks so that data embedding can be increased. The main contribution of the proposed scheme can be summarized as follows:

- The proposed scheme divides the image into smooth & complex blocks like Ou et al. [30]. Since the procedure adopted by Ou et al. [30] for embedding the secret data into smooth blocks is one of the simplest and efficient method, the proposed scheme also uses the same procedure for embedding the secret data into smooth blocks.
- In case of complex blocks, the bitmap is divided into segments of size 2×2 bits and then the embeddability of the bitmap is decided based on the bit redundancy among the segments. The bit redundancy helps in creating a room for secret data by minimally representing the redundant sections of the bitmap using some well-defied rules described in the later sections of the paper.
- Therefore, the proposed scheme is able to efficiently utilize both the complex and smooth blocks of the AMBTC codes image for embedding the secret data which in turn significantly increases the embedding capacity without any further degradation in image quality.
- The rest of the paper is organized as follows. Section 2 discusses the related work. The proposed scheme is introduced in Section 3. Section 4 presents the experimental results and discussion. Finally, the paper is concluded in section 5.

2 Related works

In this section, BTC/AMBTC based data hiding methods are briefly reviewed. As far as we know, one of the very first BTC based data hiding method was proposed by Chuang et al. [6] in 2006 for grey scale images. Chuang et al.'s method uses BTC coded trios i.e. low quantization value(a), high quantization value(b) and bitmap (B)of each 4×4 image block to carry out secret data. The method first classifies the trios into smooth and complex classes using a user-defined threshold and then embeds the secret data into the bitmap of smooth blocks, while no data is hidden into the complex blocks. Thus, Chuang et al.'s method provides decent embedding capacity with a slight degradation in image quality. Hong et al. [12] introduced a new data hiding scheme which embeds secret data bits into the trios by interchanging the position of low and high quantization values i.e., a, b while simultaneously toggling their corresponding bits in the bit plane. In case the secret data bit is zero, order of quantization values remains unchanged and if secret data bit is one, order of high and low quantization values changed and their corresponding bits in the bitmap are also toggled. Thus, Hong et al.'s method does not degrade image quality during secret data embedding as change

in order of quantization values does not have any impact on the values of trios. However, this scheme has very less embedding capacity. Further, this scheme faces the problem when the quantization values of the block are equal. Chen et al. [3] removes the equalization problem of quantization values of [12] by proposing an improved embedding method. Chen et al.'s method proposes to use the entire bitmap for data hiding in case of equal quantization values.

In 2014, Ou et al. [30] extended the work of [3, 12] by proposing a new data hiding scheme aimed to significantly increase the embedding capacity. Ou et al.'s method re-calculates the quantization values after replacing the bits of bitmap of smooth blocks by secret data sequences so that the new quantization values can be obtained according the modified bitmap which results into higher quality stego-image. Additionally, Ou et al.'s scheme suggests a way of one-bit embedding in complex block by altering the order of high and low quantization values without impacting the image quality. Thus, the scheme improves image quality by 2 dB to 3 dB by mainly focussing on smooth blocks without fully utilizing the bitmap of the complex block. To utilize bitmaps of complex blocks for secret data embedding, Malik et al. [27, 28] introduced the concept of 32-bit bitplane. Malik et al.'s schemes enhanced the AMBTC technique to produce 32-bit bitmaps and four quantization values so that the difference between the quantization values of complex planes can be reduced. This decrease difference helped in hiding the secret data in the bitmaps without having any significant impact on image quality. Thus, the schemes provide good quality marked images with significant increase in embedding capacity. However, the image compression is reduced by half.

To maintain the image compression ratio and also utilize the complex block's bitmaps, Kumar et al. [21] proposed a new data hiding method using adaptive quantization and dynamic bitmap based AMBTC. Kumar et al.'s method discloses adaptive number of quantization values and adaptive bitmaps ranging from one to four values and 16 bits to 32 bits bitmap, respectively. Because of this adaptiveness the scheme improves image quality but with little less compression ratio. In another work [20], Kumar et al. used hamming distance and PVD (pixel value differencing) method for data hiding in complex blocks of AMBTC coded image. For this, Kumar et al.'s method [20] categories the image blocks into three categories namely smooth block, less complex block, and high complex block classes. Kumar et al.'s method hides secret data into smooth blocks, using bit replacement approach like [30], in less complex blocks by using hamming distance and in high complex blocks using PVD technique. Thus, Kumar et al.'s method is able to utilize complex blocks to some extent for embedding the secret data which in turn helped in increasing the embedding capacity with slight degradation in image quality.

Ou et al.'s scheme [30] is further enhanced by Huang et al. [13] by proposing a novel scheme. The scheme mainly utilizes the variation among pixel elements for data embedding. In case of complex blocks, data is hidden by modifying the difference between quantization values which in turn increases the data hiding capacity by $log_2(T)$per complex block, where T is a user defined threshold. Hong et al. [11] extended the work of Huang et al. method [13] by using pixel pair matching technique. The pixel pair matching based scheme uses reference tables filled with symmetrical patterns to modify the quantization levels of smooth blocks to achieve the smallest distortion. In case of complex blocks, one additional bit of secret data is embedded using a lossless embedding method. Additionally, division switching mechanism is also used to modify a block from smooth to complex the minimal distortion solution cannot be found. Thus, the scheme increases the stego-image quality.

Thus, it is seen from the aforementioned discussion and survey paper analysis that a lot of work have been done for hiding the secret data in smooth blocks, however, there is still a

dearth of techniques which can efficiently embed the secret data into the bitmaps of complex blocks. For this, a new data hiding method is being proposed which can efficiently embed the secret data in both smooth and complex blocks. The proposed scheme embeds the secret data in bitmap of the smooth block by simple bit replacement strategy like [30]. In case of complex blocks, the bitmap is partitioned into four sections and secret data is hidden based on some predetermined bit pattern/redundancy. Thus, the proposed scheme increases the embedding capacity without any further degradation in image quality. In the next section, the proposed scheme is explained in detail.

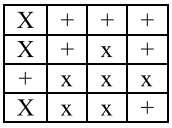
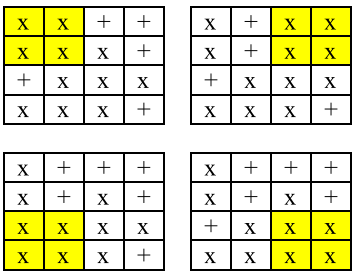
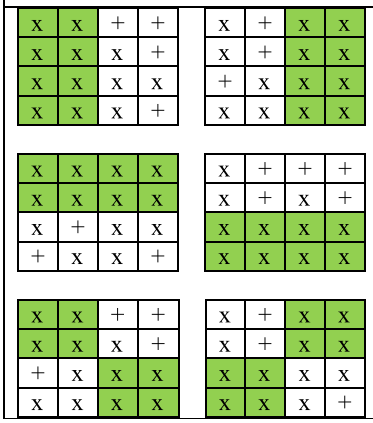
3 Proposed scheme

This section discusses the proposed scheme which embeds secret data in AMBTC coded image. As per the scheme, a cover image is initially divided into the number of non-overlapping image blocks of predetermined size, preferably 4×4 pixels. Then, each image block is coded into AMBTC trio i.e. a, b, B using standard AMBTC based encoding process (which is defined section 1.2). The obtained trios are classified into any of the two categories namely smooth and complex, based on a user-defined threshold. For this, an absolute difference between low and high quantization values i.e., *a and b*, is calculated. If the difference is less than or equal to the threshold, then the corresponding trio (which now onwards may be interchangeably called block) is a smooth block/trio otherwise it is a complex block.

In case of smooth block, secret data will be embedded in similar way to the smooth blocks of Ou et al. [30]. However, in case of complex block, bitmap is first partitioned into four sections each of size 2×2 pixels/bits which means each section will have 4 bits where a bit may be either '1' or '0'. These sections are then categorized into two categories namely compressible and non-compressible based on bit redundancy inside the section. More specifically, if all the bits in the sections are either '0' or '1', then the section belongs to compressible category otherwise to non-compressible category. If a bitmap includes two or more compressible sections with same bit value, then the bitmap is used for embedding the secret data by first representing the compressed sections using an encoded bitmap(B') and ordering the quantization values *a and b* in ascending order only inside the trio. The encoded bitmap includes i) representative bits for compressible sections ii) the bits of non- compressible sections and iii) filler bits taken from the secret data so that the encoded bitmap can be made of same size as the original bitmap to mitigate the chances of susceptibility. Thus, number of filler (hidden) bits in the encoded bitmap are dependent on the number of compressible sections in the original bitmap which can be clearly observed from Table 1.

The Table 1 illustrates all the possible bit pattern scenarios inside a bitmap and possible actions need to be taken by the hider as per the available scenario. In the Table 1, compressible sections are shown using green coloured boxes and the non-compressible sections are denoted by white as well as yellow coloured boxes. Further, the bits ('0' and '1') are denoted by x and + to clearly show the redundancy inside the bitmap. Basically, the Table 1 shows possible bit pattern scenarios in a bitmap, embeddability or non-embeddability based on scenarios, order of quantization values in trios, count of number of representative bits for compressible sections, count of number of bits for non-compressible sections and count of number of secret data bits to be embedded in bitmap and total number of bits in the encoded bitmap (which is fixed as 16 bits) using Column 1, 2, 3, 4, 5,6 and 7, respectively as follows.

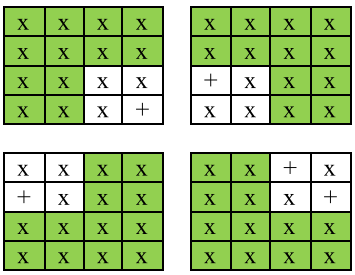
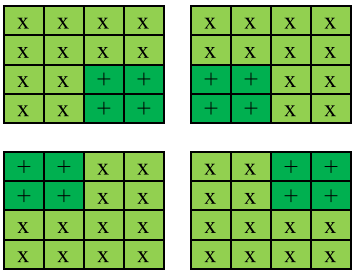
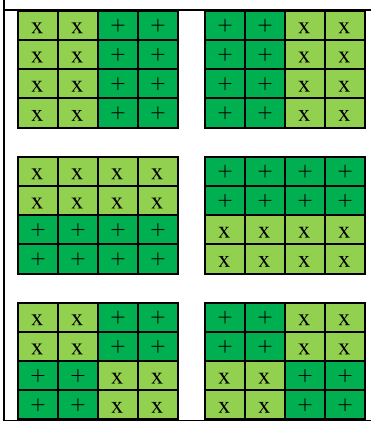
Table 1 Representation of possible bit pattern scenarios inside the bitmap and required actions for embedding the secret data

Possible bit pattern scenarios in Bitmap	Embeddability	Trio order	Number of representative bits for compressible sections	Number of bits for non-compressible sections bitmap	Filler (hidden) bits	Total bits in bitmap
	No	$a, b, B^{'}$	0	16	0	16
	No	$a, b, B^{'}$	0	16	0	16
	Yes	$b, a, B^{'}$	6	8	2	16

The construction of pattern of representative bits for different combination of compressible sections is illustrated in Tables 2 and 3. More specifically, the Table 2 represents the construction for the bitmaps having two compressible sections and the Table 3 represents the construction for the bitmaps having three compressible sections. First four columns of Tables 2 and 3 show the configuration of representative bit patterns for the embeddable bitmap as per the locations of the compressible sections and the last column shows the possible scenario inside the embeddable bitmap.

From the Table 1, it is quite clear that no secret data will be embedded inside the bitmap in case the bitmap does not have more than one compressible section. However, in case of two or more compressible sections, the secret data bits are embedded by representing the compressible section using lesser number of representative bits as per the Tables 2 and 3 so that a room for filler bits (which are from secret data) is created. Additionally, the ordering of quantization values inside the trio is changed to descending order so that receiver can clearly identify the embeddable blocks at the receiving end. Thus, the order of quantization values which is either ascending or descending order discloses whether the bitmap has the secret information or not. In this way, the proposed scheme is able to embed significant amount of the secret data in both

Table 1 continued

	Yes	$b, a, B^{'}$	5	4	7	16
	Yes	$b, a, B^{'}$	5	0	11	16
	Yes	$b, a, B^{'}$	6	0	10	16

smooth and complex blocks. The detailed embedding algorithm for the proposed data hiding scheme is outlined as follows:

3.1 Embedding algorithm

Input- I: original image of size $N \times N$ pixels, *thr*: user-defined threshold, S: secret data stream.
Output- Encoded AMBTC compressed trios.

BEGIN.

Step1: Input original image and apply standard AMBTC encoding method (which is defined in Section 1.2) to get AMBTC trios.
Step2: Read the first trio and calculate absolute difference between the quantization values.

Table 2 Representative bit pattern construction for bitmaps having two compressible sections

1st bit	2nd bit	3rd,4th and 5th bit	6th bit position	Possible bit pattern scenarios in compressible Bitmap
0 (at least two compressible sections)	0 (other two are non-compressible sections)	000	x (or +) as per values in compressible sections	x x + + x x x + x x x x x x x +
0 (at least two compressible sections)	0 (other two are non-compressible sections)	001	x (or +) as per values in compressible sections	x + x x x + x x + x x x x x x x
0 (at least two compressible sections)	0 (other two are non-compressible sections)	010	x (or +) as per values in compressible sections	x x x x x x x x x + x x + x x +
0 (at least two compressible sections)	0 (other two are non-compressible sections)	011	x (or +) as per values in compressible sections	x + + + x + x + x x x x x x x x
0 (at least two compressible sections)	0 (other two are non-compressible sections)	100	x (or +) as per values in compressible sections	x x + + x x x + + x x x x x x x
0 (at least two compressible sections)	0 (other two are non-compressible sections)	101	x (or +) as per values in compressible sections	x + x x x + x x x x x x x x x +
0 (at least two compressible sections)	1 (other two are also compressible sections)	000	x (or +) as per values in compressible sections	x x + + x x + + x x + + x x + +
0 (at least two compressible sections)	1 (other two are also compressible sections)	001	x (or +) as per values in compressible sections	+ + x x + + x x + + x x + + x x
0 (at least two compressible sections)	1 (other two are also compressible sections)	010	x (or +) as per values in compressible sections	x x x x x x x x + + + + + + + +
0 (at least two compressible sections)	1 (other two are also compressible sections)	011	x (or +) as per values in compressible sections	+ + + + + + + + x x x x x x x x
0 (at least two compressible sections)	1 (other two are also compressible sections)	100	x (or +) as per values in compressible sections	x x + + x x + + + + x x + + x x
0 (at least two compressible sections)	1 (other two are also compressible sections)	101	x (or +) as per values in compressible sections	+ + x x + + x x x x + + x x + +

Step3: If difference is less than a nuser-defined threshold (*thr*) then the block is a smooth block. In case of smooth block, the secret data embedding is done on the lines of Ou et al. scheme.

Step4: Else difference is greater than or equals to user-defined threshold (*thr*) which means the block is a complex block. In case of complex block go to **Step4.1**.

Step4.1: Divide bitmap of the complex block into four sections of size 2×2 such as S^1, S^2, S^3 *and* S^4, where S^n represents the n^{th}section and n belongs to 1, 2, 3, *or* 4.

Step4.2: Count the number of compressible sections in the bitmap by first initializing two variables as follows: $CNT = 0, H = \left[\frac{NULLNULL}{NULLNULL}\right]$ (i.e. 2×2 null matrix) and then run through the following pseudocode.

```
FOR n = 1 to 4
C^n = Xor_{l=1}^{4}(S_l^n)   //where S_l^n represents the l^th bit value of n^th section
    IF C^n == 1
      H=H+S^n
      CNT=CNT+1
    END
END
```

Step4.3: If $CNT \leq 1$, then no secret data is embedded in the bitmap (B_j) which means no change in the original AMBTC trio.

Step4.4: Elseif ($CNT = 2$ && ($\sum H = 0 \parallel \sum H = 8$)), then

$\dot{B}_j^{1} : \dot{B}_j^{6}$ *are replaced by* 6 *bit representative values as per the Table−II*

$\dot{B}_j^{7:10} = S^z, if\ c^z = 0\ (i.e.non-compressible)$,

$\dot{B}_j^{11:14} = S^{z+1}, if\ c^{z+1} = 0\ \ (i.e.non-compressible)$,

$\dot{B}_j^{15} : \dot{B}_j^{16}$ *are*replaced by two secret data bits

Interchange the order of quantization values

Step4.5: Elseif ($CNT = 3$ && ($\sum H = 0 \parallel \sum H = 12$)), then

$\dot{B}_j^{1} : \dot{B}_j^{5}$ *are replaced by* 5 *bit representative values as per the Table*−III

$\dot{B}_j^{6} : \dot{B}_j^{10} = S^z, if\ c^z = 0, i.e.non-compressible$

- $\dot{B}_j^{11} : \dot{B}_j^{16}$ *are* replaced by seven secret data bits
- Interchange the order of quantization values

Step4.5: Elseif ($CNT = 4$ && ($\sum H = 8$)), then

$\dot{B}_j^{1} : \dot{B}_j^{6}$ *are replaced by* 6 *bit representative values as per the Table−II*

$\dot{B}_j^{7} : \dot{B}_j^{16}$ *are* replaced by ten secret data bits

Interchange the order of quantization values

Step4.6: Elseif ($CNT = 4$ && ($\sum H = 12$)), then

$\dot{B}_j^{1} : \dot{B}_j^{5}$ *are replaced by* 5 *bit representative values as per the Table*−III

- $\dot{B}_j^{6} : \dot{B}_j^{16}$ *are* replaced by eleven secret data bits

Table 3 Representative bit pattern construction for bitmaps having three compressible sections

1st bit	2nd bit	3rd and 4th bit	5th bit	Possible bit pattern scenarios in compressible
1 (at least three compressible sections)	0 (other one is non compressible section)	00	x (or +) as per values in compressible sections	x x x x x x x x x x x x x x x +
1 (at least three compressible sections)	0 (other one is non compressible section)	01	x (or +) as per values in compressible sections	x x x x x x x x + x x x x x x x
1 (at least three compressible sections)	0 (other one is non compressible section)	10	x (or +) as per values in compressible sections	x x x x + x x x x x x x x x x x
1 (at least three compressible sections)	0 (other one is non compressible section)	11	x (or +) as per values in compressible sections	x x + x x x x + x x x x x x x x
1 (at least three compressible sections)	1 (other one is also compressible section)	00	x (or +) as per values in compressible sections	x x x x x x x x x x + + x x + +
1 (at least three compressible sections)	1 (other one is also compressible section)	01	x (or +) as per values in compressible sections	x x x x x x x x + + x x + + x x
1 (at least three compressible sections)	1 (other one is also compressible section)	10	x (or +) as per values in compressible sections	+ + x x + + x x x x x x x x x x
1 (at least three compressible sections)	1 (other one is also compressible section)	11	x (or +) as per values in compressible sections	x x + + x x + + x x x x x x x x

- Interchange the order of quantization values

Thus, a new encoded AMBTC trio is obtained for the complex block.

Step5: Read & calculate the absolute difference between quantization values of next trio and repeat step 3 to 4 till all trios are not processed.

Thus, the image I with hidden secret data is obtained in the form of encoded AMBTC trios which can be transmitted to receiver so that confidential message (secret data) can be safely delivered to the receiver.

At the receiving end, first of all embedded secret data is extracted and then the blocks are recovered to minimize the changes. The extraction & recovery phase is a reverse procedure of embedding phase which has been described in the detail using sections 3.1 and 3.2. Since the extraction & recovery phase is an obvious one, the details regarding the same has been omitted. In the next sub-section, Illustrative Example of proposed embedding phase are discussed in detail.

3.2 Illustrative example of proposed embedding phase

Consider an image I of size 4×8 pixels which is divided into 4×4 size blocks as shown in Fig. 1a. The first block of the image has pixels values i.e., 140, 139, 220, 141, 211, 210,214, 212, 211, 213, 212, 209, 203, 216, 209, 216 and the second block has pixels values i.e., 210, 213, 212, 209, 211, 210, 214, 212, 147, 143, 145, 141, 203, 216, 140, 216.The blocks are processes using standard AMBTC method to get the blocks in form of trio i.e. $\{a, b, B\}$. Thus processing, the obtained trio for the first block is $a = 140$, $b = 212$, $B = 0010111111111111$and for the second block is $a = 143$, $b = 211$, $B = 1111111100001101$ which can also be seen from Fig. 1b.

For the first block, absolute difference value is $|212 - 140| = 72 > 50$ (thr) and for the second block, absolute difference value is $|211 - 143| = 68 > 50$ (thr)which means both the blocks are complex blocks. (Since the embedding procedure in smooth block is same as [30], the description of embedding is omitted). In Fig. 1b, green coloured boxes inside the bitmaps disclose compressible sections, means the section having same bit values. In this example, the first image block has three compressible sections while the second image block has two compressible sections. As the bitmap of each image block has more than one compressible section with same bit value i.e., 1, means both the blocks belongs to embeddable category. Now, a new encoded bitmap is formed for each image block based on number of compressible sections to embed the secret data. The new bitmap includes representative bit values for compressible sections, bits of non-compressible sections, and secret data bits in concatenated manner. For the first image block, configured representative bit values are 10,101 as per the Table 3 are appended by bits of non-compressed sections i.e., 1101 and secret data bits say 1,000,101 to form the encoded bitmap. In case of second image block, configured representative bit values are 000101 as per the Table 2 are appended by bits of non-compressed sections i.e., 00001101 and secret data bits say 01 to form the encoded bitmap. The Encoded AMBTC trios after embedding the secret data can be visualized from Fig. 1c. In the next section, the experimental results and their comparison are presented.

4 Experimental results and discussions

In this section, the proposed data hiding scheme is evaluated on two primarily used parameters, namely, data hiding capacity (or embedding capacity) and image quality which is defined by peak signal to noise ratio (PSNR). For this, six greyscale test images, namely, Lena, Baboon, Plane, Peppers, Boats, Barbara each of size 512×512pixels as shown in Fig. 2 are used. The implementation is done using MATLAB running on Intel (R) Core (TM) i5 processor 3.20-

140	139	211	210
220	141	214	212
211	213	203	216
212	209	209	216

210	213	211	210
212	209	214	212
147	143	203	216
145	141	140	216

(a)

Quantization values $a = 140$ $b = 212$
Bitmap

0	0	1	1
1	0	1	1
1	1	1	1
1	1	1	1

Quantization values $a = 143$ $b = 211$
Bitmap

1	1	1	1
1	1	1	1
0	0	1	1
0	0	0	1

(b)

Interchanged Quantization values $a = 212$ $b = 140$
Configured Representative bit values for three compressible sections with same bit value = 10101
Bits of non-compressed sections = 1101
Seven Secret Data Bits =1000101
Encoded Bitmap

1	0	1	0
1	1	1	0
1	1	0	0
0	1	0	1

Interchanged Quantization values $a = 211$ $b = 143$
Configured Representative bit values for two compressible sections with same bit value = 000101
Bits of non-compressed sections = 00001101
Two Secret Data Bits =01
Encoded Bitmap

0	0	0	1
0	1	0	0
0	0	1	1
0	1	0	1

(c)

Fig. 1 Image partitioning and trios representation (**a**): Partitioned image (**b**): Representation of trios (**c**): Encoded AMBTC trios

GHz with 4-GB RAM hardware platform. The secret data stream used in our experiment is generated using random function.

The experimental results of total number of embedded secret data bits (at user-defined threshold 16) in all the test images are mentioned in Table 4. Additionally, the experimental results of the proposed scheme are graphically represented in Fig. 3a. The presented graph discloses the results by drawing line for PSNR versus embedding capacity analysis for all the test images. From the Fig. 3a, it is clearly visible that Lena & Peppers images (which represent most smooth images) have the highest embedding capacity as well as PSNR whereas the Baboon image (which represents most complex image) has the lowest embedding capacity as well as PSNR. The main reason behind this performance, is the number of smooth blocks as the proposed scheme also embeds large amount of secret data inside the smooth blocks like the existing schemes [13, 20, 30]. Further, the proposed scheme also makes use of complex blocks for embedding additional amount of secret data which is not visible from Fig. 3a. For this, a new graph has been presented which only indicates the amount of secret data hidden inside complex blocks for the comprehensive and in-depth evaluation. It is clearly visible from Fig. 3b that the proposed scheme significantly increases the number of secret data bits hidden in highly complex images, such as Baboon. The main reason behind this performance is the efficient utilization and number of complex blocks for embedding the secret data. Since the Baboon image has greater number of complex blocks than Lena/Peppers image, the achieved increment in the embedding capacity is also greater.

For comparative analysis, the working style and the performance in terms of experimental results of the proposed scheme are compared with some of the popular existing techniques such as Hong et al. [19], Kumar et al. [20], Ou et al. [30], Huang et al. [13], Hong et al. [11] and presented in Table 5 and Fig. 4, respectively. The Table 5 basically presents which of the following aspects such as quantization values and their order in trios, bitmaps of smooth and complex blocks have been utilized for secret data embedding by existing schemes as well as

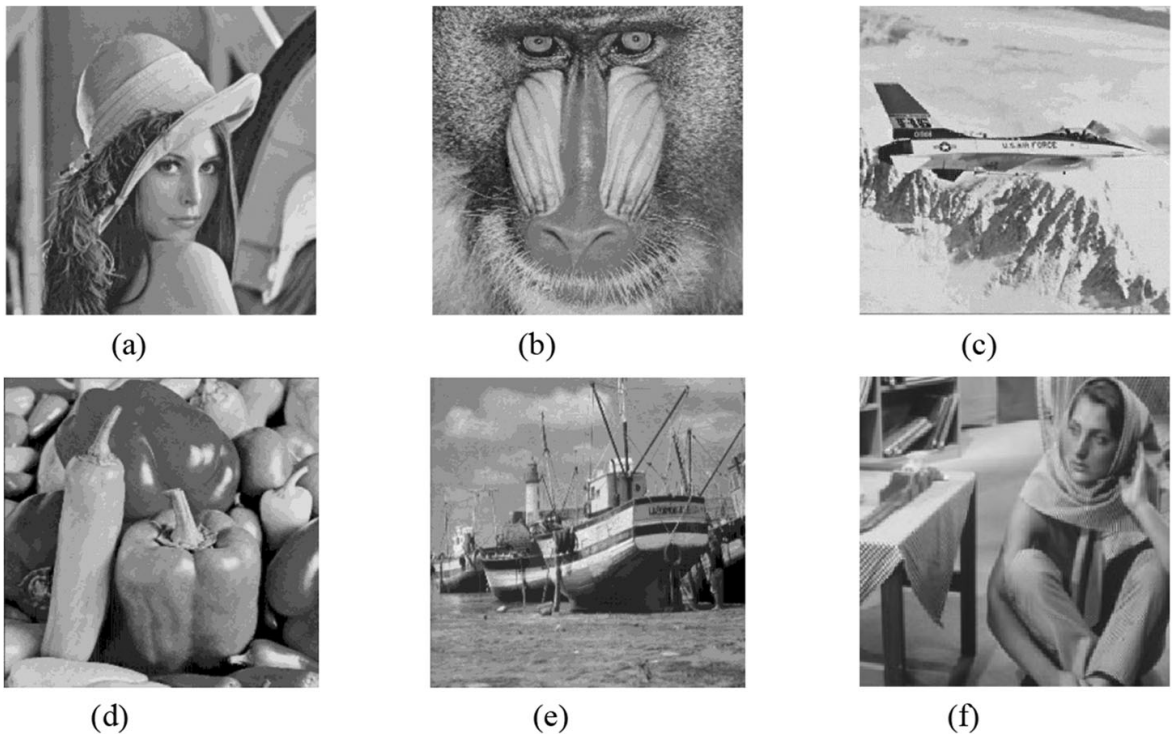


Fig. 2 Cover Images

Table 4 Embedding capacity of the proposed scheme for all the test images

Image	Lena	Baboon	Plane	Peppers	Boats	Barb
Embedding Capacity	229,827	135,414	222,881	235,277	203,652	170,203

the proposed scheme. From the Table 5, it is visible that Huang et al. [13], Hong et al. [11] and Kumar et al. [20] schemes embeds the secret data by modifying the quantization values but other schemes including the proposed scheme do not modify the quantization values for embedding. Further, all the schemes except Kumar et al. [20] embed 1-bit secret data by altering the order of quantization values. The bitmap of smooth blocks is used by all the schemes except Hong et al. [19]. However, the bitmap of complex blocks is used only by Kumar et al. [20] (in existing schemes) and the proposed scheme. The utilization of complex block's bitmaps is one of the main advantages/novelties of the proposed scheme as it helps in increasing the embedding capacity. Further, the proposed scheme does not modify the quantization values for embedding the secret data which helps in maintain the marked image quality. Furthermore, the scheme maintains the same compression ratio as the aforementioned existing schemes.

For comparative analysis of experimental results, the graphs presented in Fig. 4 are provided and their analysis is presented as follows. From the Fig. 4, it is evident that the Hong et al. [19] has the least embedding capacity but has the highest PSNR value as it neither modifies the quantization values nor the bitmaps. The Hong et al.'s scheme just alters the order of quantization values for embedding the secret data which does not reflect into the modification in any of the aspect of the compressed image. So, Hong et al.'s scheme has PSNR equal to the standard AMBTC coded image. As far as the proposed scheme is concerned, the proposed technique has higher embedding capacity than Hong et al. [19], Ou et al. [30] but has lesser embedding capacity than Huang et al. [13], Hong et al. [11]. Since Ou et al. [30] only embed the data into smooth block, while the proposed scheme also makes use of complex blocks, the resultant embedding capacity of Ou et al. [30] are comparatively on lesser side than the proposed scheme.

The Huang et al. [13] and Hong et al. [11] schemes modify the quantization values based on difference between quantization values to embed additional amount of secret data therefore

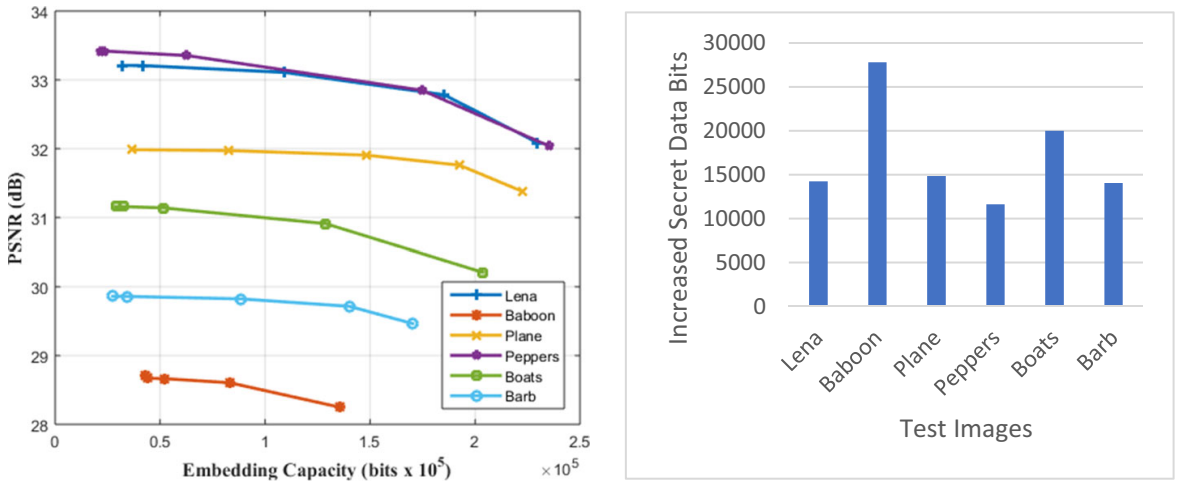


Fig. 3 Experimental results for all the test images of the proposed scheme (**a**) Representing PSNR versus Embedding Capacity results (**b**) Representing the number of bits hidden in complex blocks

Table 5 Comparison of working styles

Schemes	Quantization values *a*, *b*	Order of Quantization values	Bitmap of Smooth Block	Bitmap of Complex Block
Hong et al. [19]	No	Yes	No	No
Kumar et al. [20]	Yes	No	Yes	Yes
Ou et al. [30]	No	Yes	Yes	No
Huang et al. [13]	Yes	Yes	Yes	No
Hong et al. [11]	Yes	Yes	Yes	No
Proposed Technique	No	Yes	Yes	Yes

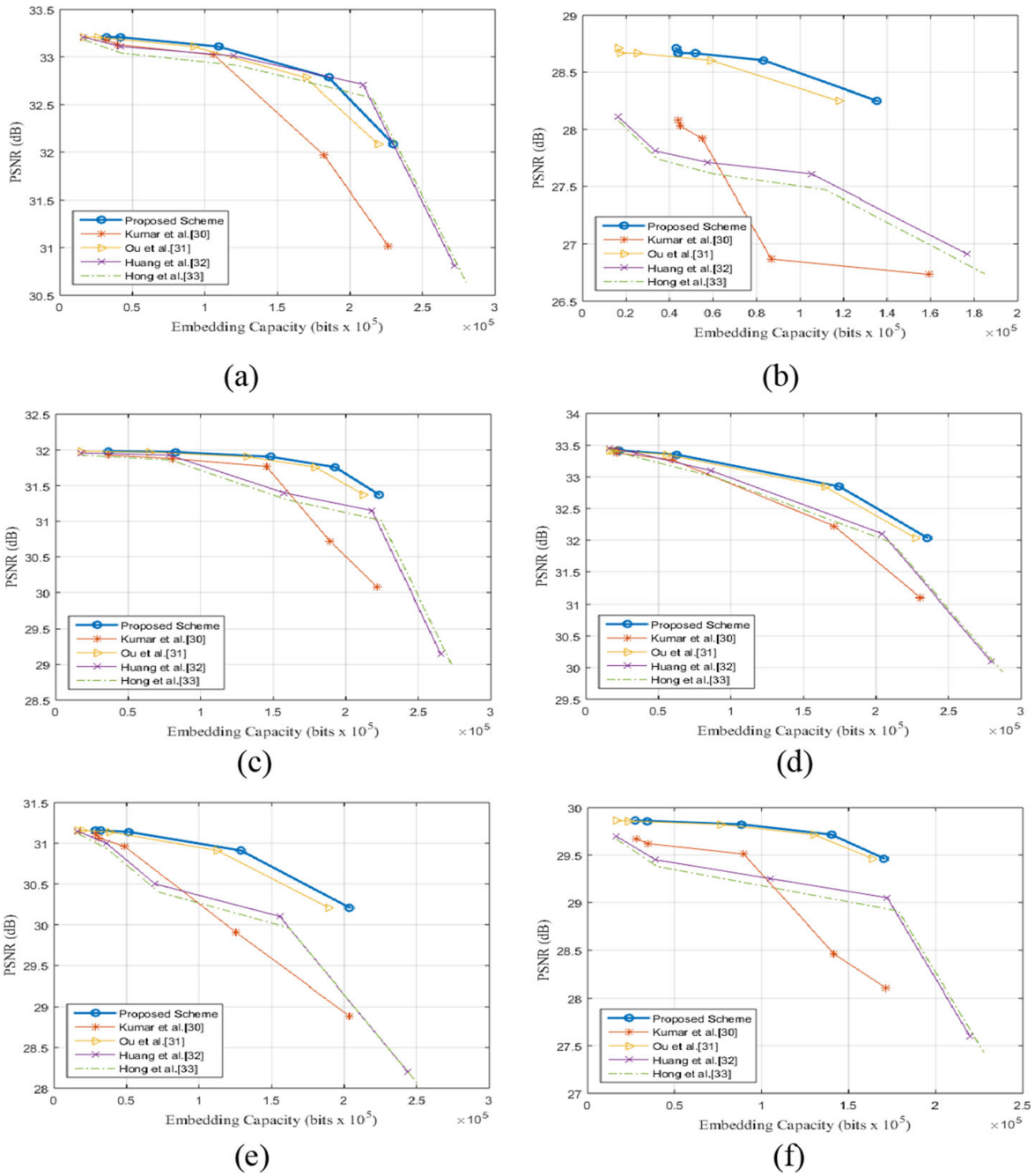


Fig. 4 Shows PSNR vs Embedding Capacity results of test images by existing techniques and proposed technique (**a**) Lena (**b**) Baboon (**c**) Plane (**d**) Peppers (**e**) Boats (**f**) Barbara

these techniques have higher embedding capacity than the proposed scheme. However, the quality of the resultant marked image in case of Huang et al. [13] and Hong et al. [11] is poorer due to the large modification into the quantization values. It is also visible from Fig. 4 that the proposed scheme and Ou et al. [30] have the highest PSNR values for all test images. This is mainly because of the recalculation of quantization values after secret data embedding. From the Fig. 4, it can also be seen that the proposed scheme has the highest PSNR value in lower range of embedding capacity and is able to take significant lead in the maximum embedding capacity in comparison to Ou et al. [30] which has the most similar performance in terms of PSNR. Thus, it can be concluded that the proposed scheme is able to significantly increase the embedding capacity while maintaining the quality of the marked image.

5 Conclusion

In this paper, a novel high capacity data hiding scheme for AMBTC encoded images has been introduced. The proposed scheme embeds the secret data into complex blocks in addition to the smooth blocks of the AMBTC coded image. The proposed scheme exploits bit redundancy to optimally represent the image blocks which in turn helps in additional embedding of secret message in bitmap of complex blocks. The concepts of bitmap division and bit redundancy have immensely helped in outperforming the existing and related schemes. To validate the point, experimental results of the proposed scheme have been compared with popular existing and related AMBTC based data hiding schemes which again confirm the superiority aspect. In future work, the redundancy exploitation aspect can be extended to colour images.

References

1. Chand S, Kumar R, Singh S (2018) A reversible high capacity data hiding scheme using combinatorial strategy. Int J Multimedia Intell Secur 3(2):146
2. Chen LS, Chang W, Lee S, Lin J (2010) Data Hiding Based on Side Match Vector Quantization and Modulus Function, 2010 Sixth international conference on intelligent information hiding and multimedia signal processing, Darmstadt, pp. 66–69.
3. Chen J, Hong W, Chen TS, Shiu CW (2010) Steganography for BTC compressed images using no distortion technique. Imaging Sci J 58(4):177–185
4. Cheng SC, Tsai WH (1994) Image compression by moment-preserving edge detection. Pattern Recogn 27(11):1439–1449
5. Christodoulou L, Abdul-Hameed O, Kondoz AM, Calic J (2016) Adaptive subframe allocation for next generation multimedia delivery over hybrid LTE unicast broadcast. in IEEE Trans Broadcast 62(3):540–551
6. Chuang JC, Chang CC (2006) Using a simple and fast image compression algorithm to hide secret information. Int J Comput Appl 28(4):329–333
7. Delp EJ, Mitchell OR (1979) Image compression using block truncation coding. IEEE Trans Commun 27(9):1335–1342
8. Dhara BC, Chanda B (2004) Block truncation coding using pattern fitting. Pattern Recogn 37:2131–2139
9. Fraga-Lamas P, Fernández-Caramés T, Suárez-Albela M, Castedo L, González-López M (2016) A review on internet of things for defense and public safety. Sensors. 16:1644. https://doi.org/10.3390/s16101644
10. Gregory KW (1991) The JPEG still picture compression standard. Commun ACM 34(4):30–44
11. Hong W (2018) Efficient Data Hiding Based on Block Truncation Coding Using Pixel Pair Matching Technique. Symmetry 10:36. https://doi.org/10.3390/sym10020036
12. Hong W, Chen TS, Shiu CW (2008) Lossless steganography for AMBTC-compressed images. Int Congr Image Sig Process\ 2:13–17
13. Huang YH, Chang CC, Chen YH (2017) Hybrid secret hiding schemes based on absolute moment block truncation coding. Multimed Tools Appl 76:6159–6174

14. Kumar R, Chand S (2016) A reversible high capacity data hiding scheme using pixel value adjusting feature. Multimed Tools Appl 75(1):241–259
15. Kumar R, Chand S (2017) A novel high capacity reversible data hiding scheme based on pixel intensity segmentation. Multimed Tools Appl 76(1):979–996
16. Kumar R, Chand S, Singh S (2018) An improved histogram-shifting-imitated reversible data hiding based on HVS characteristics. Multimed Tools Appl 77(11):13445–13457
17. Kumar R, Chand S, Singh S (2018) A reversible data hiding scheme using pixel location. Int Arab J Inf Technol 15(4):763–768
18. Kumar R, Chand S, Singh S (2019) An optimal high capacity reversible data hiding scheme using move to front coding for LZW codes. Multimed Tools Appl 78(16):22977–23001
19. Kumar R, Jung K-H (2019) A systematic survey on block truncation coding based data hiding techniques. Multimed Tools Appl 78:32239–32259. https://doi.org/10.1007/s11042-019-07997-0
20. Kumar R, Jung K-H (2019) Enhanced AMBTC based data hiding method using hamming distance and pixel value differencing. J Inf Secur Appl 47:94–103. https://doi.org/10.1016/j.jisa.2019.04.007
21. Kumar R, Kumar N, Jung K (2019) A New Data Hiding Method Using Adaptive Quantization & Dynamic Bit Plane Based AMBTC, 2019 6th international conference on signal processing and integrated networks (SPIN), Noida, India, pp. 854–858
22. Kumar R, Malik A, Singh S, Kumar B, Chand S (2016) Reversible data hiding scheme for LZW codes using even-odd embedding strategy. In: international conference on computing, communication and automation (ICCCA2016) (pp. 1399–1403)
23. Kumar R, Saini KK, Chand S (2013) A new steganography technique using snake scan ordering strategy. Int J Image, Graphics Sig Process 6:25–32
24. Kumar R, Singh S, Jung K (2019) Human Visual System Based Enhanced AMBTC for Color Image Compression Using Interpolation, 2019 6th international conference on signal processing and integrated networks (SPIN), Noida, India, DOI https://doi.org/10.1109/SPIN.2019.8711635.
25. Lema MD, Mitchell OR (1984) Absolute moment block truncation coding and its application to color image. IEEE Trans Commun 32:1148–1157
26. Lin CC, Liu XL, Tai WL, Yuan SM (2013) A novel reversible data hiding scheme based on AMBTC compression technique. Multimed Tools Appl 74(11):3823–3842
27. Malik A, Sikka G, Verma HK (2017) A high payload data hiding scheme based on modified AMBTC technique. Multimed Tools Appl 76(12):14151–14167
28. Malik A, Sikka G, Verma HK (2018a) A high capacity data hiding scheme using modified AMBTC compression technique. Int Arab J Inf Technol (IAJIT) 16(1):148–155
29. Malik A, Singh S, Kumar R (2018) Recovery based high capacity reversible data hiding scheme using even-odd embedding. Multimed Tools Appl 77(12):15803–15827
30. Ou D, Sun W (2014) High payload image steganography with minimum distortion based on absolute moment block truncation coding. Multimed Tools Appl 74:9117–9139
31. Poletti C, Michieli M. (2018) Smart cities, social media platforms and security: online content regulation as a site of controversy and conflict. City Territory Archit 5. https://doi.org/10.1186/s40410-018-0096-2.
32. Wang Y, Moulin P (2003) Steganalysis of block-DCT image steganography, vol 2003. IEEE Workshop on Statistical Signal Processing, St. Louis, MO, USA, pp 339–342
33. Yang C-K, Tsai W-H (1995) Improving block truncation coding by line and edge information and adaptive bit plane selection for gray-scale image compression. Pattern Recogn Lett 16:67–75. https://doi.org/10.1016/0167-8655(94)00068-E

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

Affiliations

Neeraj Kumar[1] · Rajeev Kumar[2] · Aruna Malik[3] · Samayveer Singh[3]

Neeraj Kumar
neeraj.mohiwal@gmail.com

Rajeev Kumar
rajivgarg@outlook.com

Aruna Malik
arunacsrke@gmail.com

[1] Department of Electronics and Communication Engineering, Nagarjuna College of Engineering and Technology, Bengaluru, India

[2] Department of Department of Computer Science & Engineering, Delhi Technological University, Delhi, India

[3] Department of Computer Science & Engineering, NIT Jalandhar, Jalandhar, India

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)

Machine Learned Classifiers for Trustworthiness Assessment of Web Information Contents

Priyanka Meel
Biometric Research Laboratory
Department of Information technology
Delhi Technological University
New Delhi, India-110042
priyankameel86@gmail.com

Dinesh Kumar Vishwakarma
Biometric Research Laboratory
Department of Information technology
Delhi Technological University
New Delhi, India-110042
dvishwakarma@gmail.com

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397228

***Abstract*—Social networking, information sharing, knowledge imparting, discussions on current happenings etc. are always a part of human society. With the fast pace of life and advancement in technology; people rely more on online information, as a result of this web platforms have become a dominant place for social interactions. This has given rise to unverified and unauthenticated news that has extremely negative effects. Fake news, rumor, misinformation, disinformation, satire, hoax, clickbait, propaganda are all different flavors of the same malice of information pollution. The research community is constantly trying to figure out a viable technical solution to this problem in different ways. In this work, we designed a framework based on five independent supervised machine-learned classifiers Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Naïve Bayes and Random Forest for trustworthiness assessment of web information contents. The classifiers are being trained and tested on two different datasets: Fake News Detection (Jruvika/FND) and Real or Fake News that contains full news articles in the form of headline and body. Experiments and result analysis verify that the highest accuracy attained by the projected method is 96.61% on the Fake News Detection dataset using the SVM classifier. The work is also compared with other contemporary techniques.**

Keywords— Machine Learning, Infodemic, Supervised Model, Support Vector Machine

I. INTRODUCTION

Humans have always interacted with their kind to survive, gain knowledge and pass on information about the happenings around them. So social networking existed right from the beginning of human history. With the advancement in technology and the ease of getting news and information handy and on the go, one may have to compromise with the authenticity of the information. A large section of society is using social networking sites such as Twitter, Facebook, Instagram, WhatsApp, YouTube, etc. to obtain news rather than using authenticated news channels and sources [1]. At the same time, unauthenticated sources through these social media platforms, websites and Twitter handle distribute fabricated news to the audience. In such a scenario it becomes difficult to understand on which source one should rely on to obtain unfabricated and genuine news. The misinformation or rumors spread across especially during emergencies can have a devastating effect on individuals and society. Spurious news in such a scenario would not only give rise to panic among the individuals but in some cases, it may also target a particular community.

Important events that leave a mark on the society such as elections, war, stock prices, business deals, politics, the health status of celebrities, cryptocurrency, religious events etc. noticeably give birth to fake news. The 2016 general elections in the USA were driven by fake news. After demonetization in India, the new 2000-rupee notes were falsely advertised to have a chip installed in it. The current coronavirus pandemic has given vertical rise to a surge of fake news being named as "**infodemic**" by WHO officials [2]. The key reason behind this is that dependable news sources are recurrently swallowed up by unconfirmed online information. All these events prove that fake news is a huge threat to our society and a problem that should be given utmost attention.

The younger generation is thought to be more tech-savvy than their parents, but they too seem to lack the ability to tell whether a piece of particular news is fabricated or genuine. The research done by Common Sense Media puts forward the fact that 44% of them have confirmed that they cannot differentiate between fake and genuine news. While it was also found that at least one online news story is being shared every day by 31% of kids aged between 10 to 18 that they found out later was inaccurate or fake [3].

The intention of sharing spurious news can vary. Few people share fake news to gain profit, while others share it to defame the opponent, some target a particular community and some try to mislead the readers and so on. During the election season, the primary motive is to mislead readers and to defame the opponents. But there are groups of people who might be sharing it to gain some monetary benefits. This was observed during the 2016 US presidential election that teenagers in the Macedonian town of Veles had gained at least $60,000 in 6 months by spreading fake news [4], [5]. So, the motive would largely vary from person to person. Researchers working in different fields are trying to provide a viable solution to this problem of misinformation and fake news that will eventually help in enhancing the reliability of online information. In this work, we tried to design an efficient framework using machine-learned classifiers to check the credibility status of online news articles. The main highlights of the work are as follows:

ISBN: 978-1-7281-8529-3/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

• A framework is designed for the trustworthiness assessment of web information contents using machine-learned classifiers.

• Five machine-learned classifiers Support Vector Machine, Logistic Regression, Random Forest, Naïve Bayes and K-Nearest Neighbor are being independently trained and tested.

• Two standard Fake News datasets Fake News Detection (Jruvika/FND) and Real or Fake News are being used for training and testing purposes in the proportion of 80% and 20% respectively.

• Accuracy, Recall (Hit Rate/Sensitivity), Precision (Positive Predictive Value), F1-score, Misclassification rate (Error rate), Type-I error (False Positive Rate/fall out) and Type-II Error (False Negative Rate/Miss Rate) parameters are being used for performance assessment of classifiers.

• The performance comparison of machine learning methods is being done among themselves and also with other state-of-the-arts.

The work is organized systematically in five different sections. Section 1 focuses on the problem of fake news and its different aspects. The literature survey is presented in section 2, section 3 emphasis on the proposed framework; experimental settings, result analysis and state-of-the-art comparison are being pronounced in section 4. Lastly, section 5 accomplishes the work by highlighting possible areas of future research.

II. Related Works

Information that is being shared on the internet is in different data presentations of script, picture, audio and video. Artificial Intelligence Techniques are being constantly used by researchers to figure out different forgery formats of text and multimodal data. Rashlin et al. [6] analyzed the linguistic characteristics and writing style of fake news, hoax, satire and propaganda to differentiate it with the real articles in news media. The authors specifically focused on political fact-checking and described how the same event can be described by different platforms to gain their vested benefits out of it. Text-based word vector representations of Bag of Words, TF-IDF, Word2Vec along with stylometric text features are explored by Reddy et al. [7] using bagging and boosting ensemble methods to classify a new article as real or fake.

Sui et al. [8] presented an elaborative review on different aspects of misinformation including stance detection, fact-checking, abstractive summarization, rumor detection and emotion analysis. They also categorized the publicly available benchmark datasets such as LIAR, FEVER, BuzzFeedNews, BuzzFace, PHEME, RumourEval, CREDBANK, BS Detector, FakeNewsNet, Fake or Real for misinformation detection. A deep neural network is designed by Liu and Wu [9] tested on Twitter15 and Weibo16 dataset for fake early detection of fake news with an accuracy of 90% before it is retweeted 50 times. The architecture is based on three innovative components: status-sensitive crowd response feature extractor, position-aware attention mechanism and multi-region mean pooling mechanism. Gangireddy et al. [10] devised an unsupervised framework driven by a graph-based technique using inter-user behavior dynamics to classify fake and legitimate news. Their proposed architecture GTUT has experimented on PolitiFact and GossipCop dataset.

Abonizio et al. [11] extracted stylometric and psychological text features of news articles written in Spanish, American English and Brazilian Portuguese to design a language-independent framework to categorize legitimate, fake and satirical news. Four machine learning algorithms SVM, KNN, Random Forest and Extreme Gradient Boosting (XGB) are implemented in this method with 85.3% highest accuracy. A rumor that is termed as unverified information which may or may not be true is also a prominent area of research. Several aspects of rumor generation, propagation, detection, resolution and its real-life effects are being explored in various studies [12] [13] [14].

In line with the work done by other researchers to overcome the problem of fake news, we tried to design a simple and efficient framework using machine-learned classifiers by extracting explicit features from news text.

III. Proposed Model

Machine Learned classifiers are supervised methods in which fully labeled data is being used for training the classifiers using the explicit features extracted from training samples. This work deals with only the headline and body of the news article and tries to distinguish between the fake and real news based on features extracted from these two fields. Five supervised machine-learned classifiers are being independently trained and tested for two different datasets according to the architecture described in Figure 1 and Algorithm 1.

Algorithm 1: Working steps in Machine Learning framework

1: Import required python libraries
2: Read the data file
3: Perform data cleaning
4: Convert text into vector form using Count Vectorizer
5: Split the dataset into train and test samples
6: Perform tf-idf feature extraction
7: Train ML classifiers (SVM, LR, RF, NB, KNN) on extracted features
8: Test the trained models using test samples

A. Preprocessing, Vectorization and Feature Extraction

The preliminary steps of preparing the data for training the classifiers are pre-processing, vectorization and feature extraction. The datasets contained a lot of extra metadata so only the required columns of headline/title, text/body and label/type are extracted. Then the complete corpus is changed into lowercase; white spaces, punctuations and stop words are removed, stemming is applied and finally, the headline and body of the article are concatenated together to make it a single instance of the corpus. To convert the text into a vector form count vectorizer is used.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

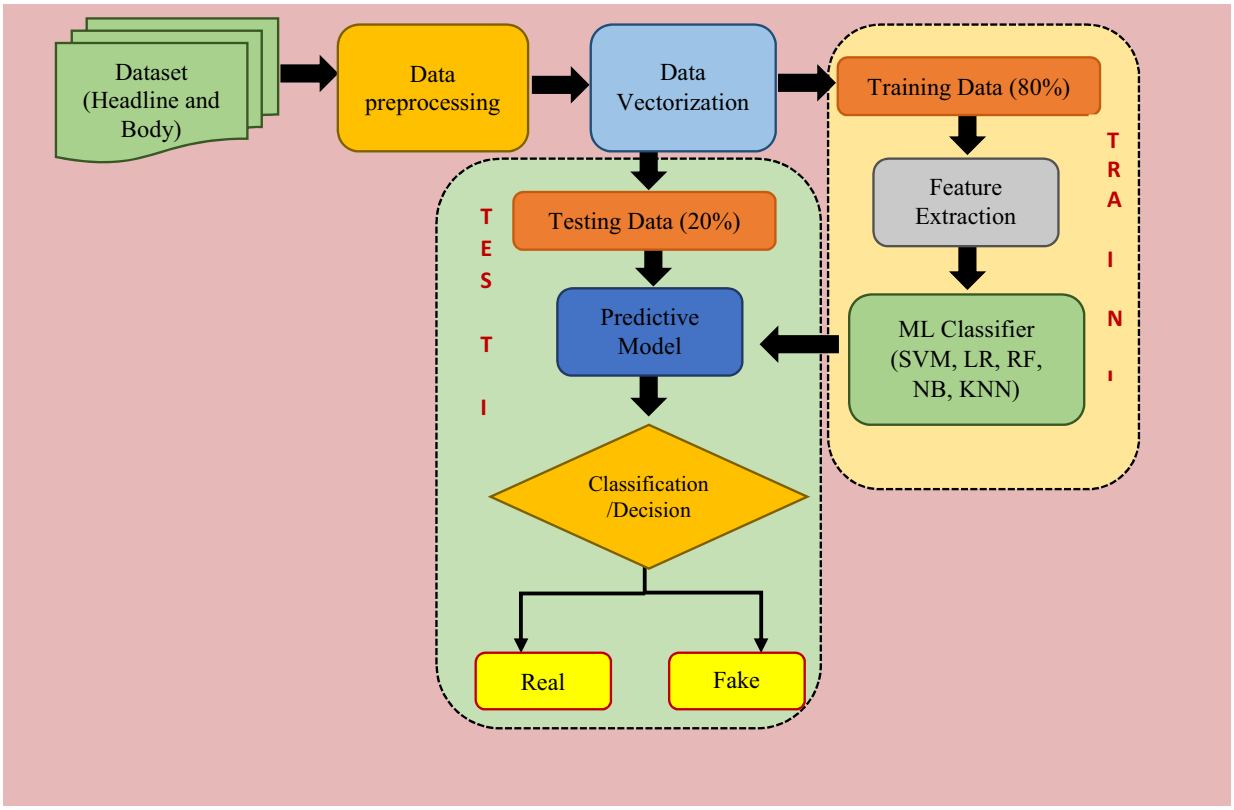


Fig. 1. Proposed Machine Learning Architecture

Feature extraction of text instances is done using tf-idf (Term Frequency-Inverse Document Frequency) which echoes the importance of each word for the given document.

B. Support Vector Machine (SVM)

Support Vector Machine follows a non-probabilistic approach as it assigns the incoming (reading) data points to a class while training, and does the same operation based on attributes learned during training in the testing phase. It is a linear classifier but can also do non-linear classification using kernel tricks (nonlinear kernel). It gives a superb performance in applications where textual data is to be classified, which makes it a really good algorithm to look at when classifying fake news. SVM model maps the training instances as points in space and classifies them in separate categories by a clear breach that is as wide as possible. Testing samples are then plotted into the identical plane and predicted to fall into a category based on the side of the breach in which they fall as illustrated in Figure 2.

C. Logistic Regression (LR)

In its basic form logistic regression is a statistical model that uses a logistic function (f(x)=1/(1+e^(-x))) to model the probability that a particular instance falls into a certain class such as pass/fail, healthy/sick, dead/alive and in our case, it is real/fake. In machine learning despite having the word "regression" in its name, logistic regression is used as a parametric classification model.

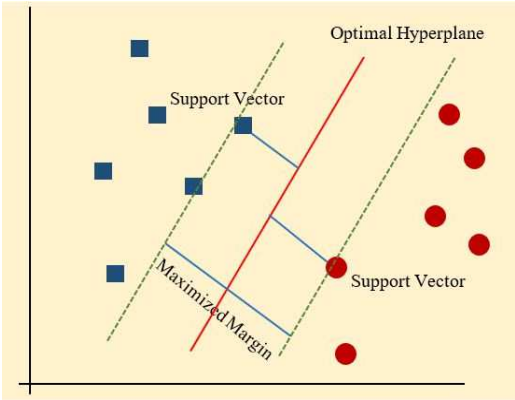


Fig. 2. Support Vector Machine

D. Random Forest (RF)

Random Forest is one of the machine learning algorithms used for classification and regression analysis. The building blocks of a Random Forest are decision trees as represented in Figure 3. A forest is made up of several trees. The central notion behind random forest is a modest but prevailing perception — the wisdom of crowds. Random forest forms decision trees on arbitrarily chosen data instances obtain the predicted output

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

from each tree and elect the optimal result using a voting mechanism. It is an ensemble method that reduces the over-fitting by averaging the result and is far better than a single decision tree. The parameters used for training Random Forest classifier are max_depth = 45, min_samples_split=7, n_estimators = 100 (denotes no. of decision trees) and random_state = 1.

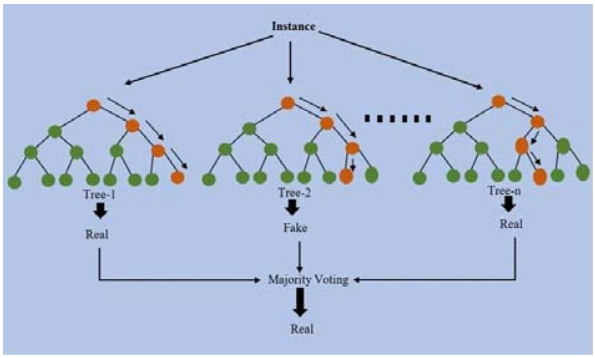


Fig. 3. Random Forest

E. Naïve Bayes (NB)

Naïve Bayes Classifiers is a complete group of algorithms based on "Bayes Theorem" sharing the common principle of mutual independence of every pair of features being classified. Mathematically Bayes theorem is characterized as:

P(A/B) =(P(B/A) P(A))/(P(B)) where
A and B are trials, P(B)≠0
P(A/B) is the likelihood of occurrence of A when B is true.
P(B/A) is the likelihood of occurrence of B when A is true.

P(A) and P(B) are the likelihoods of occurrence of A and B respectively, also known as marginal probabilities. In our fake news detection framework, we specifically use Multinomial Naïve Bayes where the value of parameter alpha=0.1.

F. K-Nearest Neighbor (KNN)

K-Nearest Neighbor also known as KNN algorithm assumes that things that have similar properties exist nearby. KNN uses the idea of similarity which can be sometimes the distance or proximity and can be calculated through various distance calculating algorithms. The majority vote of 'k' nearest data-points are selected amongst all the data-points i.e. classes of k neighbors are considered while classifying the current data point and it is being assigned to the class most common among its k nearest neighbors. We have taken n_neighbors=6, i.e., the value of k=6 so the vote of 6 closest neighbors is considered.

IV. Experiments

To validate the effectiveness of machine learning classifiers we performed experiments on windows 10 64-bit operating system and 8GB RAM with an Intel i5 processor. Programming is done with the Kaggle notebook environment in Python language using online GPU. Performance is evaluated in terms of Accuracy, Recall (Hit rate/Sensitivity), Precision (Positive Predictive Value), F1-score, Misclassification rate (Error rate), Type-I error (False Positive Rate/Fall out) and Type-II Error (False Negative Rate/Miss rate).To get the overall result validation and comparative performance analysis the model was run on two datasets hosted on the Kaggle platform (a) Fake News Detection dataset (FND) and (b) Real or Fake dataset with 80% data used for training and 20% for testing the architecture. The specifics of the datasets are described in Table I.

A. Result Analysis

The performance evaluation of proposed machine learning classifiers on the Fake News Detection (FND) dataset is presented in Table II. The experimental results show that the Support vector machine outperforms all other classifiers by providing 96.61% highest classification accuracy and least error rates. Confusion matrix and ROC curves for SVM are represented in Figure 4 and Figure 5 respectively.

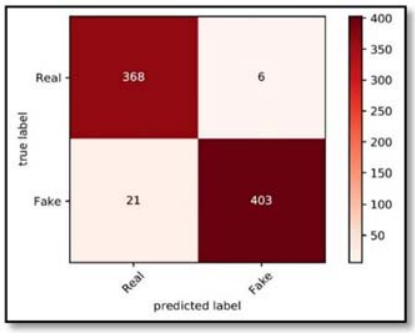


Fig. 4. Confusion Matrix For FND Dataset (SVM Classifier)

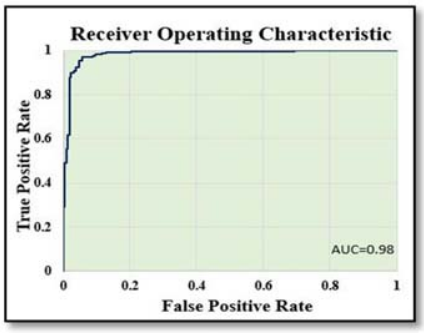


Fig. 5. ROC-AUC Curve For FND dataset (SVM Classifier)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

TABLE I. DATASET DETAILS

Sr. No.	Dataset Name	Details	# Initial Instances	# Instances after pre-processing	# Real News	# Fake News
1.	Fake News Detection (Jruvika/FND) [15]	Hosted on Kaggle platform; Contains four fields URLs, Headline, Body, Label	4009	3988	1867	2121
2.	Real or Fake [16]	Hosted on Kaggle platform, contains four fields Id, Headline, Body, Label	6335	6000	3000	3000

TABLE II. RESULT ANALYSIS ON FAKE NEWS DETECTION DATASET

Machine Learned Classifiers	Performance Parameters						
	Accuracy (%)	**Precision (%)**	**Recall/Hit Rate (%)**	**F1-score (%)**	**Misclassification Rate/Error rate (%)**	**Type-I Error/Fall out (%)**	**Type-II Error/Miss Rate (%)**
SVM	**96.61**	**94.60**	**98.39**	**96.45**	**3.39**	**4.95**	**1.61**
LR	91.22	91.30	89.84	94.71	8.78	7.55	10.16
RF	89.47	87.18	90.90	89.00	10.53	11.79	9.10
NB	95.36	92.86	97.59	95.17	4.64	6.60	2.41
KNN	92.10	88.97	94.92	91.85	7.90	10.38	5.08

TABLE III. RESULT ANALYSIS ON REAL OR FAKE DATASET

Machine Learned Classifiers	Performance Parameters						
	Accuracy (%)	Precision (%)	Recall/Hit Rate (%)	F1-score (%)	Misclassification Rate/Error rate (%)	Type-I Error/Fall out (%)	Type-II Error/Miss Rate (%)
SVM	90.42	88.68	92.67	90.63	9.58	11.83	7.33
LR	88.42	87.00	90.33	88.63	11.58	13.5	9.67
RF	79.75	78.84	81.33	80.07	20.25	21.83	18.67
NB	84.83	83.71	86.50	85.08	15.17	16.83	13.50
KNN	80.75	79.33	83.17	81.20	19.25	21.67	16.83

TABLE IV. PERFORMANCE COMPARISON ON FND AND REAL OR FAKE DATASET

Method	Description	Fake News Detection Accuracy (%)	Real or Fake Accuracy (%)
Bali et al. [17]	Random Forest (RF), K-Nearest Neighbour (KNN), Support Vector Classifier (SVC), Multi-Layer Perceptron (MLP), Naïve Bayes (NB), AdaBoost (AB) and Gradient Boosting (XGB) classifiers with GloVe word embedding, cosine similarity, n-gram count and sentiment polarity	86.20	90.02
Agarwalla et al. [18]	Support Vector Machine (SVM), Naïve Bayes (Lidstone smoothing) and Logistic Regression with Punkt statement tokenizer	83.87	88.00
O'Brian et al [19]	CNN learning with word2Vec Embedding and Adam optimizer	94.06	90.00
Reis et al. [20]	KNN, NB, RF, SVM and XGBoost with n-gram, POS tagging, LIWC, Semantic Feature	93.78	89.99
Our Method	SVM with Count Vectorizer and TF-IFD	**96.61**	**90.42**

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

Table III highlights the performance evaluation of machine learning classifiers on the Real or Fake dataset. The highest classification accuracy with this dataset is 90.42% provided by Support Vector Machine which also demonstrates the least misclassification rate, Type-I error and Type-II error in comparison with other classifiers. Figure 6 and Figure 7 represents the confusion matrix and ROC curves for Real or Fake Dataset using SVM.

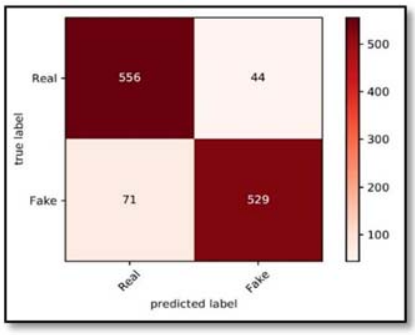


Fig. 6. Confusion Matrix for Real or Fake Dataset(SVM Cassifier)

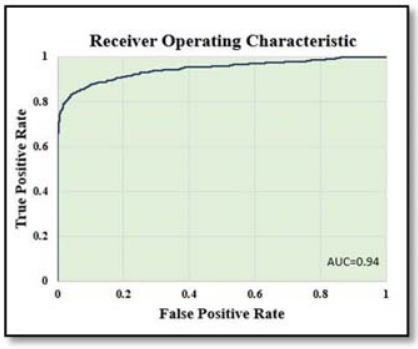


Fig.7. ROC-AUC curve for Reak or Fake Datset (SVM Dataset)

B. State-of-the-art Comparison

The outcomes of the state-of-the-art comparison of the proposed framework with other contemporary methods on identical datasets are being highlighted in Table IV. The reference methods are tested on Fake News Detection and Real or Fake dataset for identical training -testing split (80:20) with the same experimental conditions and the proposed architecture in corresponding research.

V. Conclusion and Future Works

In this work we successfully summarized, compared and analyzed the different aspects of spurious news that go around in social media. The current trend of news consumption, the scenario of fabricated news and state-of-the-art methods for the identification of fake news as well as rumors are being discussed. Five different supervised machine learning models are being trained and tested on two different text news datasets. The highest accuracy 96.61% on the Fake News Detection dataset in classifying fake news using the SVM Classifier looks very promising.

Even though a lot of work has been done by researchers specifically since 2016 to reduce the malice of information pollution from human society still a lot more is to be done. The data could be pre-processed differently; other feature extraction methods such as sentiment analysis, the polarity of the document can also be explored. Proposed work can be further extended to explore the authenticity of the URL which is also present in the dataset. The independent machine learning meta classifiers can be ensembled together through bagging, boosting or voting method. We can try to incorporate the advanced aspect of natural language processing, deep learning and transformers for extracting semantic, linguistic, stylometric hidden patterns in the text of fake writings as well for better detection in the future.

VI. References

[1] K. Sharma, F. Qian, H. Jiang and N. Ruchansky, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 10, no. 3, pp. 1-42, 2019.

[2] "United Nations Covid-19 Response," 2020. [Online]. Available: https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-covid-19.

[3] Á. Figueira and L. Oliveira, "The current state of fake news: challenges and opportunities," *Procedia Computer Science,* vol. 121, pp. 817-825, 2017.

[4] M. Freeze, M. Baumgartner, ·. P. Bruno, J. R. Gunderson, J. Olin, M. Q. Ross and J. Szafran, "Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect," *Political Behavior,* pp. 1-33, 2020.

[5] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications,* p. 112986, 2019.

[6] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017.

[7] H. Reddy, N. Raj, M. Gala and A. Basava, "Text-mining-based Fake News Detection Using Ensemble Methods," *International Journal of Automation and Computing,* vol. 17, no. 2, pp. 210-221, 2020.

[8] Q. Su, M. Wan, X. Liu and C.-R. Huang, "Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective," *Natural Language Processing Research,* vol. 1, pp. 1-13, 2020.

[9] Y. Liu and Y.-F. B. Wu, "FNED: A Deep Network for Fake News Early Detection on Social Media," *ACM Transactions on Information Systems,* vol. 38, no. 2, pp. 1-33, 2020.

[10] S. C. R. Gangireddy, D. P, C. Long and T. Chakraborty, "Unsupervised Fake News Detection: A Graph-based Approach," in *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, New Yark, USA, 2020.

[11] H. Q. Abonizio, J. I. d. Morais, G. M. Tavares and S. B. Junior, "Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features," *Future Internet,* vol. 12, no. 5, 2020.

[12] Roohani, T. Rana and P. Meel, "Rumor Propagation: A State-of-the-art Survey of Current Challenges and Opportunities," in *2nd International*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur , India, 2019.

[13] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences,* vol. 497, pp. 38-55, 2019.

[14] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys ,* vol. 51, no. 2, pp. 1-36, 2018.

[15] [Online]. Available: https://www.kaggle.com/jruvika/fake-news-detection..

[16] [Online]. Available: https://www.kaggle.com/rchitic17/real-or-fake..

[17] A. P. S. Bali, M. Fernandez, S. Choubey, M. Goel and P. K. Roy, "Comparative Performance of Machine Learning Algorithms for Fake News Detection," in *International Conference on Advances in Computing and Data Sciences , Springer* , Singapore, 2019.

[18] K. Agarwalla, S. Nandan, V. A. Nair and D. D. Hema, "Fake News Detection using Machine Learning and Natural Language Processing," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 7, no. 6, pp. 844-847, 2019.

[19] N. O'Brien, S. Latessa, G. Evangelopoulos and X. Boix, "The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors," in *32nd Conference on Neural Information Processing Systems (NIPS 2018)* , Montréal, Canada , 2018.

[20] J. C. S. Reis, A. Correia, F. Murai, A. Veloso and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intelligent Systems,* vol. 24, no. 2, pp. 76-81, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:09:08 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/350400044

Modeling and Finite Element Analysis of Universal Joint

Research · March 2021

DOI: 10.5281/zenodo.4624800

CITATIONS
0

READS
302

2 authors:

Kanishk Yadav
Delhi Technological University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Harshit Jain
Delhi Technological University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

All content following this page was uploaded by Kanishk Yadav on 26 March 2021.

The user has requested enhancement of the downloaded file.

HBRP PUBLICATION

Advancement in Mechanical Engineering and Technology
Volume 4 Issue 1
DOI: http://doi.org/10.5281/zenodo.4624800

Modeling and Finite Element Analysis of Universal Joint

Kanishk Yadav[1]*, Harshit Jain[2]
B.Tech. Student, Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Rohini, Delhi

Corresponding Author
E-mail Id:- *kanishkyadav_2k18me107@dtu.ac.in*

ABSTRACT

A universal joint, also known as a Hooke's joint, is a joint or coupling that connects rigid rods with angled axes. It's commonly found in rotary motion transmission shafts. It consists of a cross shaft linking two hinges that are close together and oriented at 90 degrees to each other. The objective of this project is to design and understand the working of the universal joint. Also, to determine the stresses, strains and element displacement in the existing design of the universal joint. If the existing design shows some failure, then suggest a minimum factor of safety for the joint. In this project, only the static FEA of the universal joint has been performed by the use of the Fusion 360 software. The simulation was carried out in the Static structural analysis module in the Fusion.

Keywords:- *Universal joint, Autodesk Fusion 360, modelling, static structural analysis, CAD- Computer Aided Designing, Meshing, FEA- Finite Element Analysis*

INTRODUCTION

Universal joints are a part of the power transmission system. They are commonly used when there needs to be angular deviations in the rotating shafts [1]. The main components of the universal joint as shown in Figure 2 are:

1. Two fork pins that are connected to the perpendicular arms of the centre block
2. Pin and collar assembly, which allows the fork to rotate along the axis of the centre block.
3. The forks allow the assembly to rotate along their respective shaft axes thus transmitting torque.

After a long period of use, the universal joint can develop faults such as looseness, wear, and torsional deformation of the intermediate shaft. It could compromise the universal joint's protection, which is one of the most important components of a vehicle's engine and transmission systems. [2]

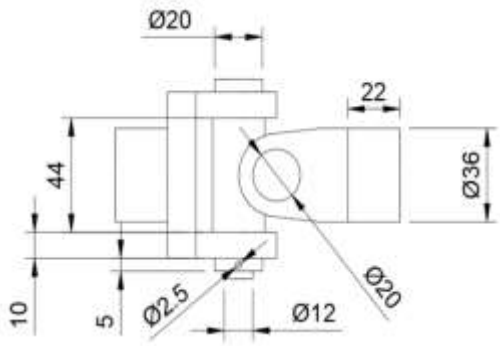


Fig.1:*-Dimensions used in designing the universal joint*
Fig.2:*-Modelling of different parts in Fusion 360*

Advancement in Mechanical Engineering and Technology
Volume 4 Issue 1
DOI: http://doi.org/10.5281/zenodo.4624800

MODELING

Fig.3:-*Universal joint assembly*

Autodesk Fusion 360 was used to build the universal joint model as shown in Figure 3. We have used dimensions as shown in Figure 1 for modelling. Fusion 360 is a collaborative product creation cloud-based CAD/CAM platform. It allows for product concept discovery and iteration, as well as collaboration within a distributed product development team. Figure 4 shows the top-view of our Universal joint assembly. Rendered image of universal joint is shown in Figure 5.

Fig.4:- *Universal joint assembly top-view*

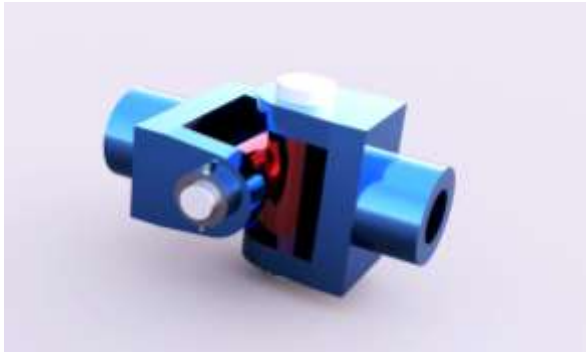


Fig.5:- *Rendered Image of Universal joint*

FINITE ELEMENT ANALYSIS

The simulation was carried out in Fusion 360 to observe the stresses and strains developed on the universal joint by static structural analysis.

The boundary conditions were given such that one end of the universal joint was kept fixed while the other end of the joint was given a torque of 6400 Nm. The material used in the analysis of universal joints was structural steel. The result of Stress analysis, Strain analysis and Displacement analysis is shown in Figure 6, 7 and 8 respectively.

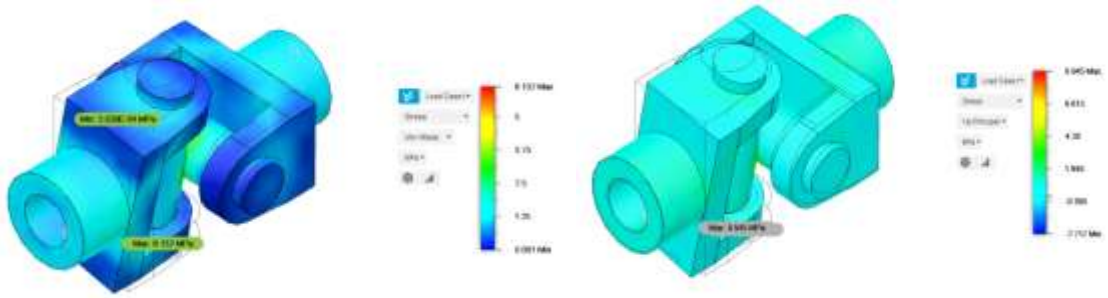


Fig.6:- *Stress Analysis*

Advancement in Mechanical Engineering and Technology
Volume 4 Issue 1
DOI: http://doi.org/10.5281/zenodo.4624800

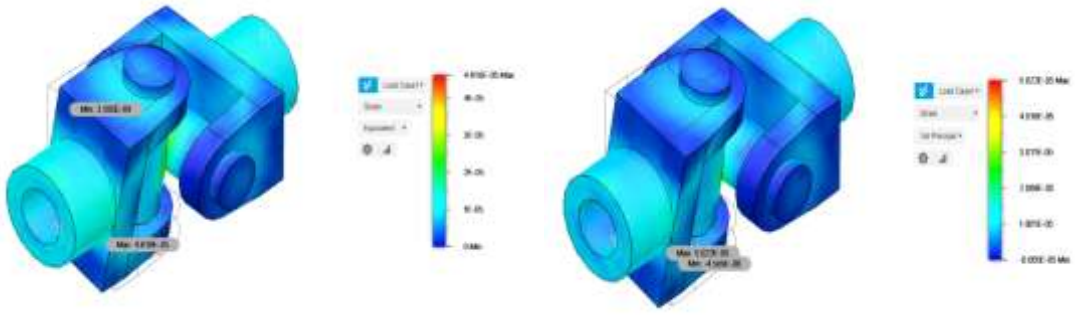


Fig.7:- *Strain Analysis*

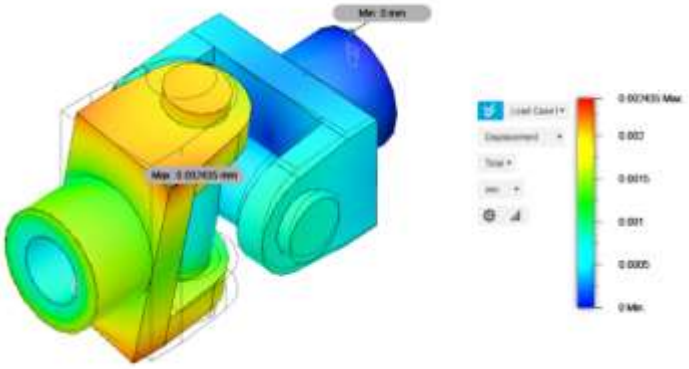


Fig.8:- *Displacement Analysis*

RESULTS AND DISCUSSION

S. No.	PARAMETERS	VALUES
1	Maximum Stress (Von-mises)	6.132 MPa
2	Maximum Principal stress (Maximum Tensile stress)	8.945 MPa
3	Minimum Principal Stress (Maximum compressive stress)	2.717 MPa
4	Maximum shear stress	1.911 MPa
5	Maximum Equivalent strain	4.61 * 10^(-5)
6.	Maximum Principal strain	5.022*10^(-5)
7.	Maximum deformation	0.002435 mm
8.	Recommended factor of safety	15

LITERATURE REVIEW

[3] Mills, control devices, gadgets, instrumentation, medical and optical gadgets, arms radio sewing machines, material hardware, and computer drives all use universal joints.

When the FEA findings were compared to the current results, they were found to be quite similar. As a result, FEA is a feasible tool for evaluating the stresses and failure points of a joint.

[4] The stresses which generally develop in the universal joint because of power transmission include bending stress in yoke, shear stresses in yoke, crushing stresses between pin and hole of yoke, shearing stress in hole of yoke and bending stress in spider assembly.

[5] The stresses were obtained by applying similar boundary conditions as in our paper, that is keeping one end constrained and applying torque on the other end. Since the stresses obtained were well within the yield

DOI: http://doi.org/10.5281/zenodo.4624800

stresses of the material, hence weight optimisation approach was also applied to the transmission yoke

[6] Java, an object-oriented and high-level programming language, is used to measure the different stresses in flange coupling components. This offers an alternative approach for precisely computing stresses in a short period of time.

.

[7] A comparison is made between the existing design and a new design by making small modifications to the existing design. In the new design the stresses obtained (i.e. Von mises stress and shear stress) are significantly reduced and are evenly distributed.

CONCLUSION

1. The modelling of the universal joint was carried out in Fusion 360 to better understand its design, working and the failure points.
2. From the simulation, it was observed that considerable compressive stresses were induced at the point of contact between the fork pin and the fork which causes premature wearing of the fork pin, resulting in failure of the universal joint.
3. High amounts of stresses were also observed at the intersection of the centre block, which could also result in failure of the universal joint.
4. The factor of safety obtained by the FEA software for the joint was 15.
5. The Von-mises Stress and maximum shear stress were found to be 6.13 MPa and 8.94 MPa respectively.
6. The part was found to be safe under given load conditions

REFERENCES

1. Vesali, F., Rezvani, M. A., & Kashfi, M. (2012). Dynamics of universal joints, its failures and some propositions for practically improving its performance and life expectancy. *Journal of mechanical science and technology*, *26*(8), 2439-2449.
2. An, K., & Wang, W. (2017). Transmission performance and fault analysis of a vehicle universal joint. *Advances in Mechanical Engineering*, *9*(5), 1687814017707478.
3. Arora, R. Modeling and Failure Analysis of Universal Joint using ANSYS.
4. Mahajan, S., Khamankar, S.,D. Finite Element Analysis of Universal Joint. *Journal of Emerging Technologies and Innovative Research*. 2349-5162.
5. Kolekar, D., Kalje, A., Kulkarni, S. Design, Development and Structural Analysis of Universal joint. *International Journal of Advanced Engineering Research and Studies*. 2249–8974
6. Vardhan, D. H., HariPrasad, M., Rafi, D. M., & Ravuri, M. Java for Mechanical Design Computation: Dimensions of the Various Flange Couplings.
7. Shah, P., & Patil, A. (2016). Modeling and FEA of universal coupling of an automobile truck. *International Journal for Innovative Research in Science and Technology*, *3*(2), 131-6.

Cite as

Kanishk Yadav, & Harshit Jain. (2021). Modeling and Finite Element Analysis of Universal Joint. Advancement in Mechanical Engineering and Technology, 4(1), 1–4. http://doi.org/10.5281/zenodo.4624800

View publication stats

Environmental Earth Sciences (2021) 80:295
https://doi.org/10.1007/s12665-021-09599-2

ORIGINAL ARTICLE

Modeling gully erosion susceptibility in Phuentsholing, Bhutan using deep learning and basic machine learning algorithms

Sunil Saha[1] · Raju Sarkar[2] · Gautam Thapa[3] · Jagabandhu Roy[1]

Received: 27 August 2020 / Accepted: 22 March 2021 / Published online: 2 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

The present study attempts to demarcate the areas susceptible to gully erosion in Phuentsholing, Bhutan, using Deep Learning CNN (convolution neural network) and artificial neuron network (ANN), Support Vector Machine (SVM) and maximum entropy, three basic machine learning techniques in the GIS setting. Application of deep learning technique is new in the field of gully erosion. Considering the 240 gully pixels and seventeen gully erosion conditioning factors (GECFs), the gully erosion susceptibility maps (GESMs) were prepared. Out of the 240 gully pixels, 70% were used as training datasets and 30% were used as validation datasets for modeling and judging the GESMs. The GECFs were selected based on the previous literatures and multi-collinearity test. The importance of the GECFs was assessed by the chi-square attribute evaluation (CSEA) and random forest (RF) methods. Finally, applying the receiver operating characteristics' area under curve (AUC-ROC), RMSE, MAE and R-index, the robustness of the GESMs was evaluated and compared. The GESMs were classified using natural break classification method into very high, high, moderate, low and very low susceptible classes. Nearly, 20% of the study area has very high susceptibility to gully erosion. As per the results of CSEA and RF methods, sand concentration, land use\cover and altitudes have the largest contribution in making the area very susceptible to gully erosion. Results of the validation techniques recognized the entire selected model as accurate and robust. Among the selected models, the capability of CNN model (AUC = 0.910, MAE = 0.029, RMSE = 0.171 for training data and AUC = 0.929, MAE = 0.089, RMSE = 0.299 for testing data) in predicting the gully erosion susceptibility is higher than other models. The produced GESMs will be helpful to the researchers as well as decision makers in establishing gully erosion management strategies.

Keywords Deep learning · Gully erosion susceptibility · CNN · Chi-square attribute evaluation · Phuentsholing

Introduction

Over the last few decades, the most critical concerns for the preservation of the Earth's biodiversity are the conservation of natural resources and environmental management. Land resource is gradually degrading due to natural as well as unscientific anthropogenic activities. For managing the land and water resources in proper way, an evaluation of soil erosion taking into account multiple environmental factors is necessary (Debanshi and Pal 2020; Saha et al. 2019). The sustainable soil management has turn into prerequisite for handling the human needs, especially the growing global population and the resulting pressure on soil (Vogel et al. 2018). The movement of weathered material from the surface by relocating agents to another location is defined as soil erosion (Parlak 2007). Naturally, when force of surface runoff exceeds the resistance power of soil particle, then soil erosion is occurred (May and Place 2005). Gully erosion is

✉ Raju Sarkar
rajusarkar@dce.ac.in

Sunil Saha
sunilgeo.88@gmail.com

Gautam Thapa
gautamama@gmail.com

Jagabandhu Roy
jagabandhuroy1991@gmail.com

[1] Department of Geography, University of Gour Banga, Malda, West Bengal, India

[2] Department of Civil Engineering, Delhi Technological University, Bawana Road, Delhi, India

[3] Consulting Engineer, Royal Govt. of Bhutan Projects, Phuentsholing, Bhutan

an intensified process of soil erosion in which the top soil has been removed by cumulative runoff of water to a particular depth (Desta and Adunga 2012) forming a narrow channel on the valley side and floors, and an epitome with steep side walls, a lower depth ratio, subjugated by surface runoff, soil pipes, detachment and stress cracking (Knighton 1998; Bull and Kirkby 1997; Gómez-Gutiérrez et al. 2015). Gully growth is also an intensifying phenomenon in semi-arid areas due to intense rainstorm events alternating with the long dry season and regional climate instability with increased anthropogenic action, i.e., overgrazing and inaccurate land management (Conoscenti et al. 2013; Rahmati et al. 2016). However, soil erosion is a hazard with many other impacts, like degradation of land, pollution of surface water, and declining agricultural productivity. As one of the most harmful forms of erosion, gully erosion is the most prevalent cause of soil loss in various parts of the world (Roy and Saha 2021; Conoscenti et al. 2014; Capra et al. 2005; Rahmati et al. 2017a, b; Mararakanye et al. 2017). It is a complex mechanism that involves lithology, water, weather, topography and vegetation (Guerra et al. 2017). Gully erosion has enormous impacts on desertification, flooding, surface water contamination and livelihood vulnerability (Swarnkar et al. 2018). Huge amount of world's agricultural land was significantly impacted by soil erosion over the past 40 years, and this is continuing at the rate of 10 million hectares each year (Sun et al. 2014). This natural cycle can nevertheless be forecast via a holistic approach to the system. Erosion pins (Castillo et al. 2007), geometric estimate of eroded material (Keay-Bright and Boardman 2009), attribute-based assessments (Hancock and Evans 2010), volumetric measurement by sediment trap buckets (Sougnez et al. 2011) and modeling pre- and post-depositional surfaces (Diaz et al. 2014) methods for estimating the volume of sediment generated from gully erosion have been proposed. While these methods may require comprehensive data and frequent field surveys, the spatial likelihood of gulling cannot be predicted. The first and utmost step to manage insecurity caused by gully erosion is to determine triggering factors and classify areas might be occurred (Saha et al. 2020). For preparing the maps of gully erosion susceptibility, various mathematical models, bivariate and multivariate, along with GIS and remote sensing, were effectively investigated and proposed. The effectiveness of all machine learning and statistical methods relies on the consistency of the model. Discrimination efficiency and reliability are the two standards for judging the model accuracy (Pearce and Ferrier 2000). Discrimination is the model's ability to discriminate between locations where natural disasters are reported and places where natural phenomena are considered missing. Reliability defines the relationship between the expected possibility of repeated events (gully erosion) and the observed ratios (Pearce and Ferrier 2000). Discrimination is low and reliability is high for a good model (Pearce and Ferrier 2000).

For carrying out gully erosion susceptibility mapping (GESM), researchers around the world employed various mathematical, machine learning and data-mining approaches over the last decade (Gayen et al. 2019; Cama et al. 2020; Zakerinejad and Märker 2014; Conoscenti et al. 2018). The most popular predictive methods are logistic regression (Cama et al. 2020; Zakerinejad and Märker 2014), correlation matrix and logistic regression (Debanshi and Pal 2020) analytical hierarchy process (Hembram et al. 2018), evidential belief function (Gayen and Saha 2017), information value (Hembram et al. 2018), frequency ratio (Rahmati et al. 2016; Zabihi et al. 2018), entropy index (Conoscenti et al. 2014; Roy et al. 2018; Azareh et al. 2019), and weights of evidence (Gayen and Saha 2017; Rahmati et al. 2016). Nevertheless, gully erosion is a product of dynamic mechanisms that are still hard to predict with great precision. Different machine learning like artificial neural network (ANN) (Pourghasemi et al. 2017; Garosi et al. 2019), classification and regression trees (Gómez-Gutiérrez et al. 2015), multivariate adaptive regression splines (Gómez-Gutiérrez et al. 2015; Conoscenti et al. 2018), support vector machines (SVM) (Pourghasemi et al. 2017), decision tree (Geissen et al. 2007), stochastic gradient boost tree (Angileri et al. 2016), multilayer perceptron neural network (Roy and Saha 2021), random subspace, dagging (Roy and Saha 2021) boosted regression tree and random forest (Saha et al. 2020), Naïve Bayes (Garosi et al. 2019), generalized additive model (Garosi et al. 2019), best-first decision tree (Hosseinalizadeh et al. 2019), bagging best-first decision tree (Hosseinalizadeh et al. 2019), random-subspace best-first decision tree (Hosseinalizadeh et al. 2019), rotation-forest best-first decision tree (Hosseinalizadeh et al. 2019) were executed in the field of natural hazard modeling.

Generally, GESMs considering machine learning are able to offer greater predictive accuracy than traditional computational techniques. Such approaches served to identify the areas are of susceptible to gully erosion. Now, different deep learning techniques have been considered for the natural hazard assessment in earth science (Lin and Weng 2018; Zuo et al. 2019). They have pointed out that the accuracy of the deep learning techniques are greater than the machine learning algorithms and statistical methods. In this analysis, we have used convolution neural network (CNN) and the benchmark machine learning techniques, i.e., ANN, SVM and maximum entropy. The application of the CNN method for gully susceptibility mapping is new. Here, the output of the CNN was compared with the three selected bench mark methods. Such methods can be helpful in predicting the spatial gully mapping or to forecast areas of gullies that are important to the planners, as well as to analyze the effects

of changing environmental on the frequency and position of gullies (Conoscenti et al. 2014). The main aim of this research is to delineate area susceptible to gully erosion using deep learning approach and compare this with three benchmark machine learning methods.

Description of the study area

Bhutan is a mountainous country which lies in eastern Himalaya bounded by India in East, West and South and Tibetan plateau in North. The entire Bhutan is divided into 20 districts which are locally known as Dzongkhags also known as district. The present study area lies in Phuentsholing (26°50′ N–26°59′ N latitude and 89°15′ E–89°28′ E longitude), the business hub of Bhutan, under Chukha Dzongkhag and elevation ranges from 180 to 2500 m. It is located along the Phuentsholing–Thimphu National Highway at a chainage of 5.5 km from Phuentsholing City covering an area of 7.31 km^2. (Fig. 1). This Highway, also known as Asian Highway, is one of the most important trade routes connecting Bhutan with neighboring country India.

The present study area is bounded in between Rinchending checkpost (first official checkpost to enter Bhutan) on the East and Reldri Higher Secondary School in the West. Between the two unstable hill slopes facing towards East and West, a small seasonal stream flows along the gorge carrying sediment load generated from the gulley erosion of the study area (N 26° 50′57″–E 89°23′22.7″). Gully erosion in the study region is attributed mainly to heavy precipitation during monsoon (June–September) and weak bedrock geology. The annual daily maximum rainfall of Phuentsholing is 285.4 mm. Figure 2 shows the daily average rainfall of Phuentsholing (2008–2018). The temperature within

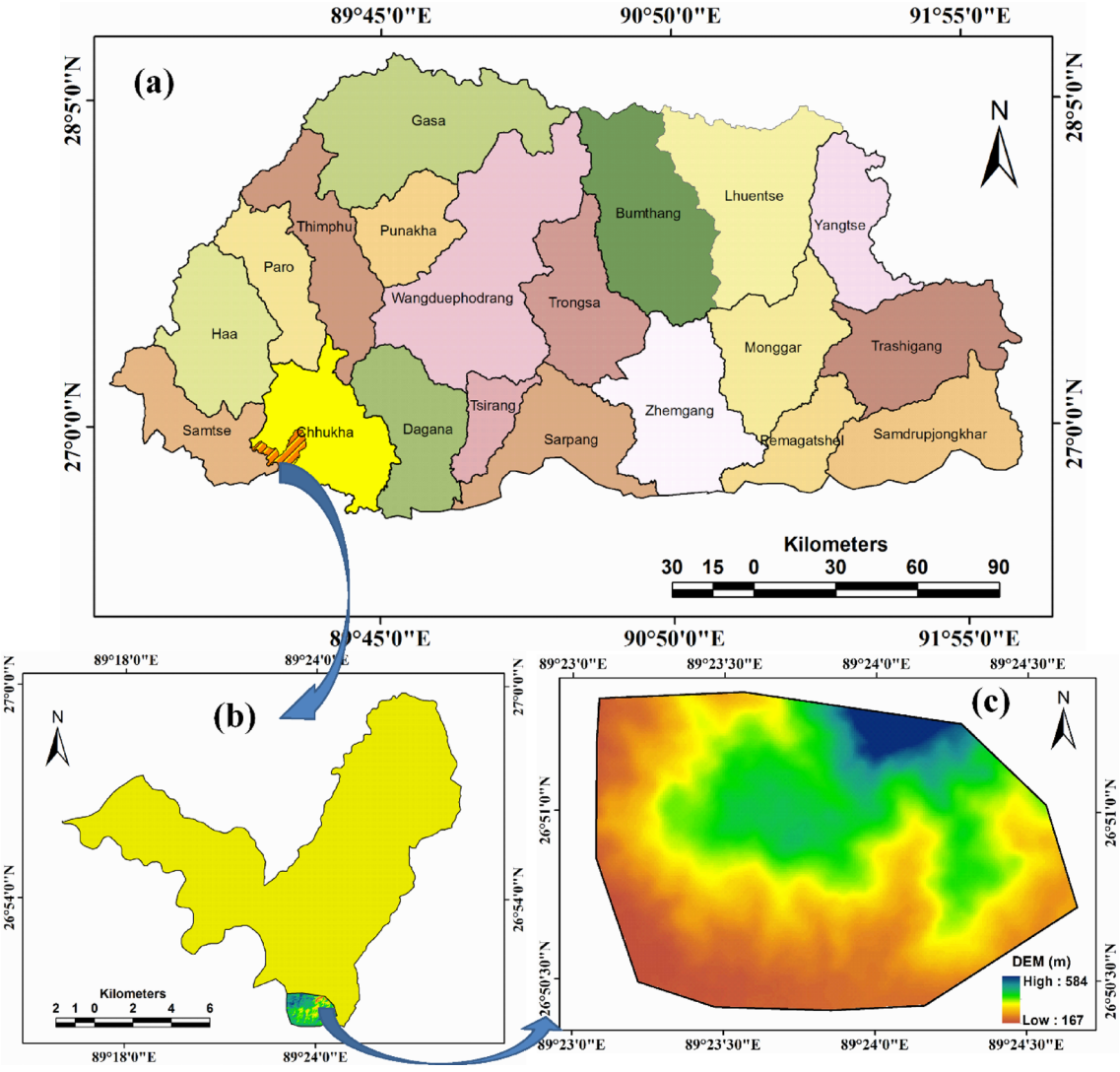


Fig. 1 Location of the study area: **a** Bhutan, **b** Phuentsholing region and **c** Study area showing the elevation

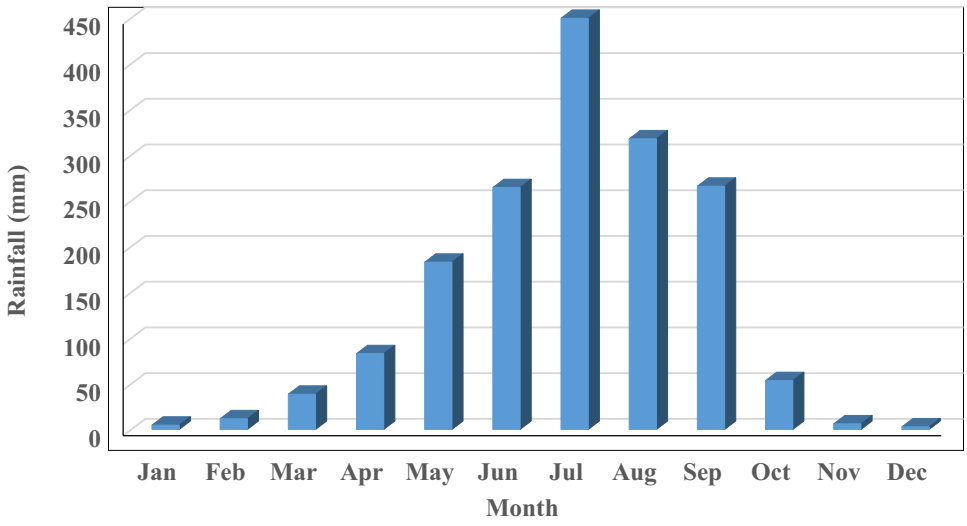


Fig. 2 Monthly average rainfall of Phuentsholing (2008–2018)

the area varies due to large variability in elevation. In the foothill plain, the temperature ranges from 15 to 35 °C, the temperature drops gradually traveling towards the interior part of Bhutan. Figure 3 shows the daily annual temperature (1996–2011) of Phuentsholing area. The areas where gully erosion is prominent are of phyllite, slate, and schist that contain high amounts of clay and these rocks are highly fractured and weathered due to the presence of several shear zones and folds (Dikshit et al. 2019; Sarkar and Dorji 2019). In this area, the main cause of gullying is the heavy rainfall and surface runoff. The gullies are concentrated on the sloppy area (15°–30°) of this mountain region which is lithologically covered by gravelly sand (Fig. 4). The width of the largest gully of the study area is 22 m and depth is 5.4 m. On the other hand, width and depth of the smallest gully are 3 and 1 m. The present area of study for gully erosion lies in the Baxa Group's Phuentsholing formations consisting of slate (dark gray to black), phyllite with thin limestone interbeds, and dolomite (cream colored) with thin beds of fine–medium-grain quartzite (dark-brown) (Gariano et al. 2019). The Phuentsholing Formation comprises dark gray to black, finely laminated slate and phyllite. The surface bedrock of the study area consists of gravely sand of alluvial deposit. Several generations of alluvial and fluvial fan deposits of Plio-Pleistocene to Holocene age lie depositionally on top of the rocks of the study area.

Materials and methodology

The study involved the different steps for carrying out the work (Fig. 5). The steps are: (i) Data collection and gully erosion inventory map construction, (ii) Gully erosion conditioning factors (GECFs) preparation in GIS platform, (iii) Selection of GECFs using multi-collinearity assessment, (iv) Application of three bench mark methods: ANN, ME, and SVM and one deep learning approach, namely convolution neural network (CNN) for gully erosion susceptibility maps

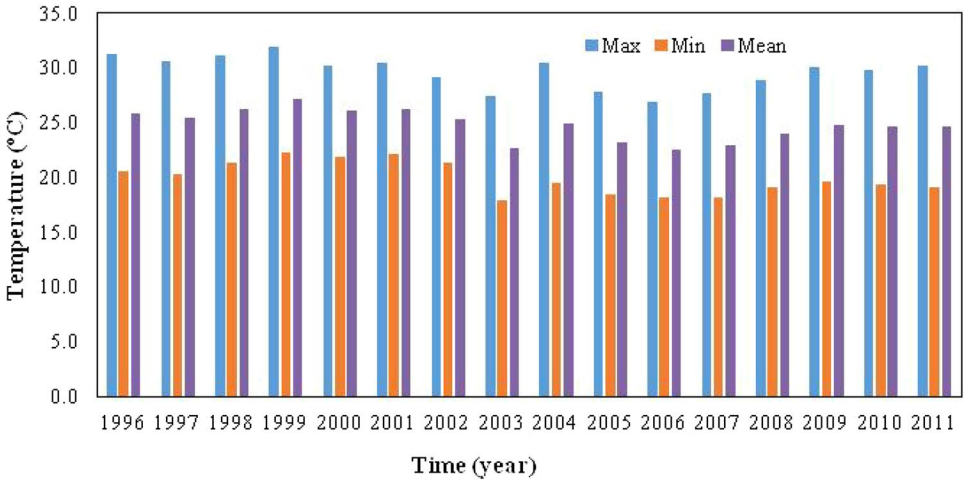


Fig. 3 Annual Temperature (1996–2011)

Fig. 4 Field photos of gully erosion

(GESMs) construction, (v) Gully erosion conditioning factors (GECFs) significance analysis using random forest (RF) model and Chi-square attribute evaluation (CSAE) techniques, (vi) Models performance analysis using Receiver Operating Characteristic (ROC) curve, statistical methods like mean absolute error (MAE), mean root square error (RMSE) and also relative gully erosion density (R-Index) methods.

Data sources

The high-resolution PALSAR DEM (phased array-type L-band synthetic aperture radar digital elevation model) with 12.5 m × 12.5 m resolution of dated 2011 was downloaded from Alaska satellite facility. The geological data of the study area have been obtained from the field study undertaken through the National Adaptation Programme of Action (NAPA-II), Royal Government of Bhutan. The rainfall data were also collected from National Centre of Hydrology and Meteorology, Bhutan. The Landsat 8OLI/TIRS data were downloaded from United State Geological Survey (USGS). Soil textures such as sand, silt, and clay have been collected from field survey and these were tested using sieve method in soil laboratory. The altitude, slope, aspect, plan curvature (PC), convergence index (CI), topographic position index (TPI), topographic ruggedness index (TRI), topographic wetness index (TWI), valley depth, relative slope position (RSP), and drainage were extracted from PALSAR DEM.

Preparing gully erosion inventory map (GEIM)

To prepare the inventory map of gully, some important data are necessary which may be spatial and non-spatial data layers. Generally, inventory datasets can be defined as the actual situation of datasets in the field (Garosi et al. 2019). These data can be shown as the point, line and polygon layers. Inventory map is essential for hazard modeling. GEIM helps to compare the predicted situation with actual situation. Several researchers used different ways like field survey with GPS, Google Earth engine, high-resolution and satellite ortho-images (Gayen et al.2019; Debanshi and Pal 2020; Saha et al. 2020).

In the current study, through the field investigation with handheld global positioning system (GPS) and Google earth engine, gully erosion areas were marked. First, mapped gully polygons were converted into raster at a resolution of 12.5 m × 12.5 m. Then, a total of 240 gully pixels were chosen randomly from the total gully pixels and 70% were used to train the models, and 30% were used to evaluate

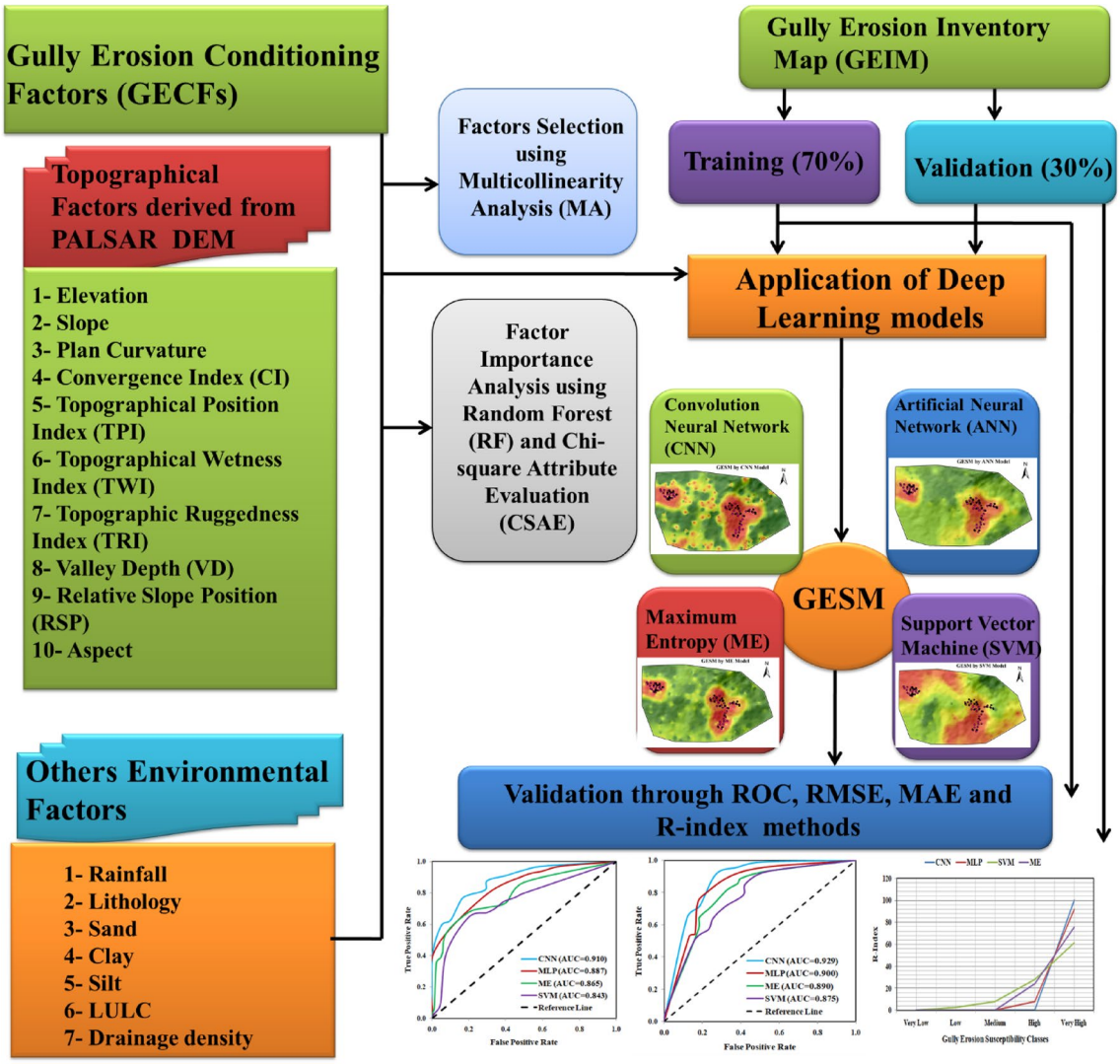


Fig. 5 Methodological flow chart

the models of susceptibility to gully erosion. The center of each cell was converted into points for training the models. The ratio of gully occurrence (denotes as 1) and non-gully occurrence (denotes as 0) should be equal to 1 for running the models (Schicker and Moon 2012). Similarly, in this study for running the models, equal numbers of non-gully pixels were selected randomly.

Gully erosion conditioning factors (GECFs)

The selection of explanatory variables that often affect model predictive efficiency is a significant step for preparing the GESM (Rahmati et al. 2017b). In this study, the GECFs were chosen after reviewing the previous literatures (Pourghasemi et al. 2017; Roy et al. 2019; Hembram et al. 2018; Hembram and Saha 2020; Saha et al. 2020). In the small area, the climatic factors are assumed to be fairly similar (Ferro et al. 1991). The climatic effect in making the areas susceptible to gully erosion is indirectly represented by the topographic attributes, such as elevation, aspect and slope (Zabihi et al. 2019). Topographic explanatory variables primarily regulate the initiation of the gully (Gómez-Gutiérrez et al. 2015). A PALSAR DEM with a 12.5 m/12.5 m cell size was used to prepare the topographical attributes. Elevation, slope aspect, slope degree and plan curvatures are primary topographical attributes; and TWI, TPI, TRI, CI, valley depth and RSP are secondary topographical attributes extracted from the PALSAR DEM. In Table 1, methods for preparing the GECFs are mentioned.

Elevation was described as one of the significant GECFs in the susceptibility assessment of gully erosion (Hembram and Saha 2020; Gayen et al. 2019). It can actually occur in various elevations, depending on the gully development mechanisms and other environmental factors. Typically,

Table 1 Methods used for preparing the GECFs

Factors	Methods	References
Altitude	12.5 m × 12.5 m digital elevation model	Li et al. (2005)
Slope	$\text{Tan}\theta = \frac{N \times i}{636.6}$; N = No. of Contour Cutting; i = Contour Interval	Wentworth (1930)
Aspect	$\text{Aspect} = 57.29 \times \alpha\text{tan2}(\left[\frac{dz}{dy}\right] - \left[\frac{dz}{dx}\right])$, where $dz/dx = ((c+2f+i) - (a+2d+g))/8$; $dz/dy = ((g+2h+i) - (a+2b+c))/8$ Here, a–i indicate the cell value of 3*3 window	Burrough et al. (1998)
Clay	Sieve method	Bieganowskiet al. (2011)
Sand		
Silt		
Convergence Index (CI)	$CI = \left(\frac{1}{8}\sum_{i=1}^{8}\theta_i\right) - 90°$, where θ denotes average angle between the aspect of adjacent cells and the direction to the central cell	Kiss (2004)
Relative Slope Position	$RSP = \frac{Zs - Zy}{Zs - Zy}$, where Zs is the elevation, Zx is the elevation of the valley and Zy is the elevation of the ridge	Conrad et al. (2015)
Topographic ruggedness index (TRI)	$\text{TRI} = \sqrt{\lvert X\rvert(\max^2 - \min^2)}$, where x is the elevation of each neighbor cell to a specific cell, and max and min are the largest and smallest elevations among the nine neighboring pixels	Riley et al. (1999)
Topographic position index (TPI)	$\text{TPI} = Z_0 - \overline{Z}$, $\overline{Z} = \frac{1}{n_R}\sum_{i \in R} Z_i$, Z_0 = the central point and $\overline{Z}$ = average elevation	Weiss (2001)
Topographic wetness index (TWI)	$\text{TWI} = \text{In}(A_S / \tan\beta)$, where 'As' is the specific catchment area in meter and slope in degrees	Moore et al. (1991)
Valley Depth (VD)	VD = Zs − Zy, where Zs is the elevation and Zx is the elevation of the valley	Rodriguez et al. (2002)
Plan Curvature (PC)	$Plan\ Curvature = \frac{q^2 \times r - 2 \times p \times q \times s + p^2 \times t}{\left(\sqrt{1+p^2+q^2}\right)^3}$. The values of H, G, p, q, r, s and t are obtained as follows: $H = \frac{df}{dy} = \frac{Z1+Z2+Z3-Z4-Z8-Z9}{6 \times P}$, $G = \frac{df}{dx} = \frac{Z3+Z6+Z9-Z1-Z4-Z7}{6 \times P}$, $p = \frac{Z3+Z6+Z9-Z1-Z4-Z7}{6 \times \Delta_S}$, $q = \frac{Z1+Z2+Z3-Z7-Z8-Z9}{6 \times \Delta_S}$, $r = \frac{Z1+Z3+Z4+Z6+Z7+Z9-2\times(Z2+Z5+Z8)}{3 \times \Delta_S^2}$, $s = \frac{-Z1+Z3+Z7-Z9}{4 \times \Delta_S^2}$, $t = \frac{Z1+Z2+Z3+Z7+8Z+Z9-2\times(Z4+Z5+Z6)}{3 \times \Delta_S^2}$, where Z1–Z9 are altitude values in 3 × 3 cellular networks and Δ_S denotes the cell size	Shary et al. (2002)
LULC	Maximum likelihood Classification	Myung et al. (2003)
Drainage Density (Dd)	$\sum D/A$, D = total length of drainage and A = Area	
Lithology	Digitization process	Anderson et al. (2004)

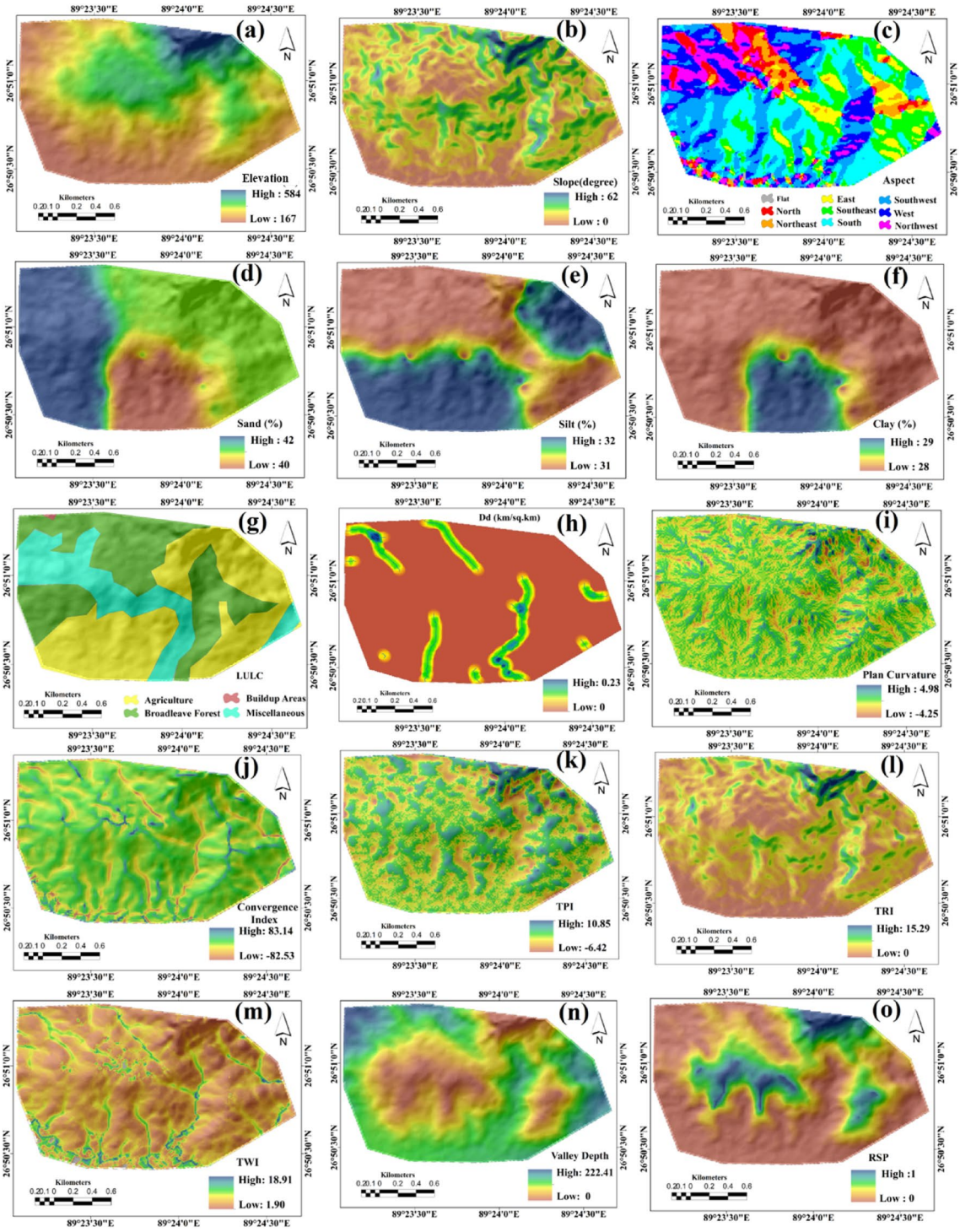


Fig. 6 Gully conditioning factors: **a** Elevation, **b** slope, **c** aspect, **d** percentage of sand, **e** percentage of silt, **f** percentage of clay, **g** land use/land cover (LULC), **h** drainage density, **i** plan curvature, **j** convergence index, **k** topographical positioning index (TPI), **l** terrain ruggedness index (TRI), **m** topographical wetness index (TWI), **n** valley depth and **o** relative slope position (RSP)

compared to the low elevated flat zones, with the existence of other conditioning factors, highland tracts have comparatively greater capacity for gullying. The altitude of this research region from sea level varies from 167 to 584 m (Fig. 6a). The slope determines the surface runoff and erosion power of drainage density which in turn influence the processes of gully erosion (Lucà et al. 2011). Gullies are subjected to initiation in gently sloping areas owing to

concentration of overland flow (Agnesi et al. 2011). The research area's average slope ranges from 0 to 62 degrees (Fig. 6b). The compass direction of the geographic slope is indicated by slope aspects and these are classified as flat, North, South-East, West, North-East, North-West South-East and South-West. The slope aspect is regarded as an important component in the study of various denudational processes (Maharaj 1993; Nagarajan et al. 2000) as well as the monitoring of natural hazards (Baeza and Corominas 2001; Umar et al. 2014). The aspect of slope (Fig. 6c) influences the weathering mechanisms, vegetation and geomorphic processes and indirectly influences the gully erosion (Gayen et al. 2019; Geroy et al. 2011). The soil texture groups are found in the three categories namely sand, silt and clay types. The percentage of sand is 40–42 and the silt and clay percentages are 31–32 and 28–29, respectively (Fig. 6d–f). Bedrock lithology is consisting of mainly gravelly sand formed due to uneven weathering and deposition of landslide material and fluvial deposit. Plan curvatures may be used to explain the impact of local morphology on overland flow, runoff and as in turn effect on gully erosion (Shary et al. 2002). This element is measured as the slope perpendicular to the slope gradient and impacts the convergence and divergence of the water flow over the slope surface. The value of plan curvature varies from − 4.25 to 4.98 (Fig. 6i).

The land use classification was done using maximum likelihood classification from Landsat 8OLI/TIRS imagery. LULC has great influence on the occurrence of the gully erosion (Hembram et al. 2018; Cama et al. 2020). The bare lands are severely prone to erosion than of others, i.e., agricultural field, forest cover, and built-up areas. The study consists of different types of land use, namely agricultural land, built-up area, board leave evergreen forest and miscellaneous types (Fig. 6g). The drainage of the study area is extracted from PALSAR DEM using the Arc hydro tool in GIS environment. The drainage density is prepared with km per sq.km grid in GIS and value of drainage density ranges from 0 to 0.23 km/per sq.km (Fig. 6h).

A terrain parameter describing the relief system as a set of divergent regions (ridges) and converging areas (channels) is called convergence index (CI). The pixel is characterized as divergent when the CI value is positive, whilst the converging pixels have a negative CI value (Kiss 2004). CI value ranges from − 82.53 to 82.14 (Fig. 6j) in the study area. The TPI shows the discrepancy between the height at the central point within a given distance and the average elevation around it (Gallant and Wilson 2000). Negative TPI values indicate that the central point is located below its normal position, while positive values imply a spot above the normal height. Analyzing the TPI the topographical undulation of an area can be identified. In this analysis, the value of topographic position index ranges from − 6.42 to 10.85 (Fig. 6k).

Riley et al. (1999) developed the terrain ruggedness index (TRI) as a metric for conveying the overall difference in elevation between adjacent cells of a digital elevation network. It computes the variation in elevation values from the middle cell and the cells that surround it immediately. So each of the elevation difference values is squared to get all of them positive and to average the squares. The topographic roughness index is then calculated by taking the square root of this mean. The terrain ruggedness index value ranges from 0 to 15.29 (Fig. 6l). The TWI quantifies topographic regulation of hydrological processes. The slope and the upstream contributing area per unit width orthogonal to the direction of flow consist of all features. The index was built for Catena hillside (Beven and Kirkby 1979). The topographic wetness index value varies from 1.90 to 18.91 (Fig. 6m). The depth of the valley is measured as the differential between the height and an interpolated ridge level (Conrad et al. 2015). The value of valley depth ranges from 0 to 222.41 (Fig. 6n). The relative slope position is defined as the position of one-point relative to the slope of the ridge and valley, with a value of 0 at the bottom of the valley and 1 at the top of the ridge. The relative slope position (RSP) value ranges from 0 to 1 (Fig. 6o).

Models

Convolution neural network (CNN)

Deep learning algorithms (DLA) are focused on human brain layout, and are grounded in a theoretical context of an artificial neutral network (ANN). The well-known DLA is CNN developed by LeCun et al. (1998). Categorization and estimation using CNN have been widely implemented in numerous fields, including the earth sciences (Lin and Weng 2018; Zuo et al. 2019). Uses of multiple layers, pooling, local links and mutual weights from a conventional neural network are differentiated by CNN. The fundamental principle that underlies CNN is the images that are input data, or can be viewed as information. It will reduce the number of parameters substantially, contributing to speedier processing. Convolutionary layers (CLs), pooling layers (PLs) and ReLU (rectified linear units) are the three layers of CNN framework (Fig. 7). Convolutions are learnt by the CLs and best categorize the data. In the present work, 15 convolutional kernels were used with 15 fully connected units. PLs control over-fitting, enable for reliable conversion, and increase computational efficiency by growing the amount of transfer frameworks (Nair and Hinton 2010). Maximum pooling method was considered for modeling the gully erosion susceptibility. Utilizing activation function ReLU enhances the nonlinear properties of the network. Researchers developed and applied various data-type, image,

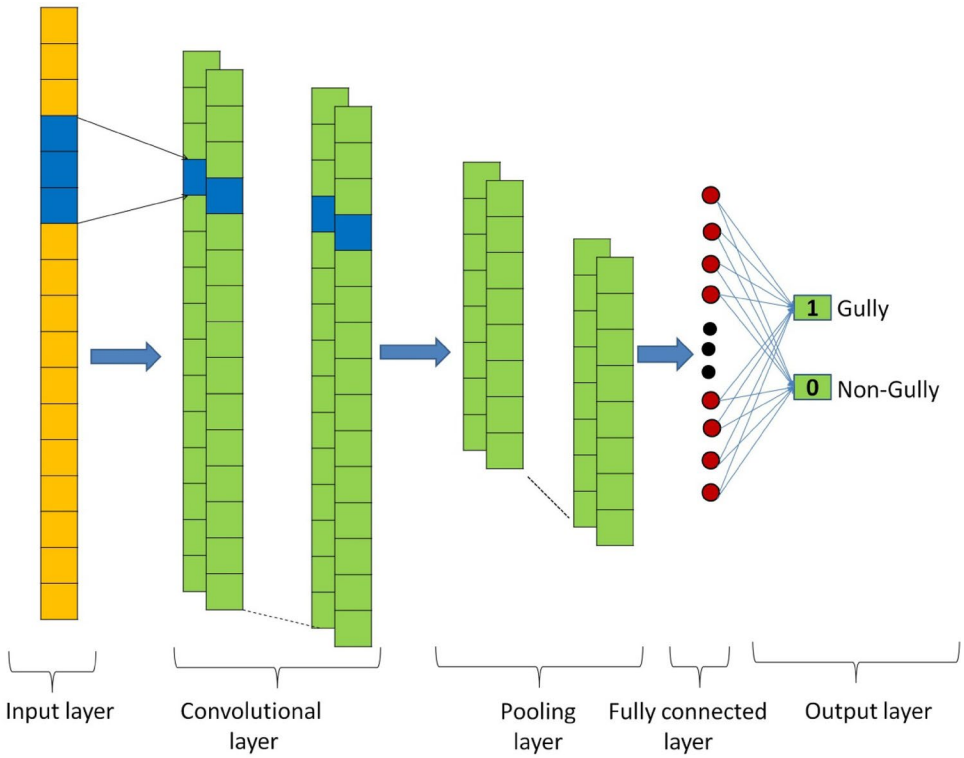


Fig. 7 Theoretical structure of convolution of Artificial Neural Network (CNN)

and objective-based research systems, including GoogleNet, ZFNet and VGGNet (Szegedy et al. 2015; Zeiler and Fergus 2014; Simonyan and Zisserman 2014). Several articles presented detailed explanations of these styles of levels, core criteria for learning and how the CNN handles training data (LeCun et al. 2015). In the present study, the framework of 2D-CNN is regarded because it was primarily used in comprehensive earth science work (Wang et al. 2019). The technique is used for susceptibility mapping of landslides. The images are broad in size since no data are marked and research is continuous. A learning rate of 0.004, Adam optimizer and RMSE loss function were used for training the CNN model. During the training process, the parameters are automatically learnt in the convolutional layers. Finally, the CNN converts the pixels into final probability results. With a loss function, feed-forward propagation parameters were calculated.

Artificial neural network (ANN)

Based on the human brain system, the artificial neural network (ANN) model was built (Cherkassky et al. 2006). This consists of three layers: input, hidden, and output (Fig. 8). The functionality of the model is controlled by highly integrated nodes. Input layer shows template meaning for one or more concealed nodes. Hidden nodes learn, classify or use weighted relations for the sequence processing. The network output is situated on the opposite side of the connection and check how well the network responds to the information it receives (Yesilnacar and Topal 2005; Prasad et al. 2012). The sigmoid stimulation role of secret neurons polarizes and settles down the network. Arora and Mathur (2001), Prasad et al. (2012) have proposed more comprehensive description. Using a back-propagation method in terms of gradient descent and momentum, ANN was performed. To avoid over-fitting, the number of hidden neurons in the secret layer and the optimum learning rate were set using a trial-and-error process. An ANN was fitted, in our study, with five secret layers, 500 epochs and 20 thresholds of validity.

Maximum entropy (ME)

The word "entropy" to explain the amount of knowledge concealed in a dataset that the researchers depend on to learn an elusive pattern was first agreed by Shannon (1948). Basically, the interaction within the data creates a property called entropy. The entropy in a dataset provides the more unpredictable outcomes, the more ready-to-explore and extract information afterwards. Jaynes (1957a, b) provided ME as the most satisfactory function of probability distribution (PDF) to be selected for an undefined PDF. The appropriate information shall be collected from the thematic geo-environmental layers. Feature types depend on the format of the present study data, whether categorical or continuous. At the sample presence locations, features enforce a set of rules on the results.

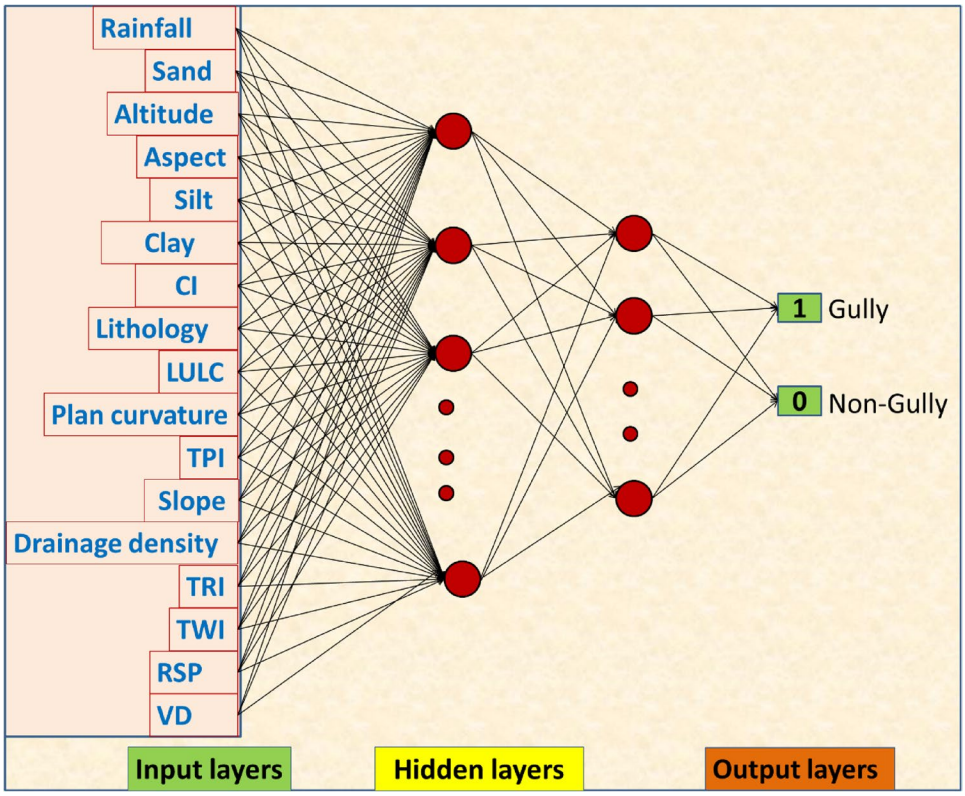


Fig. 8 Structure of Artificial Neural Network (ANN)

The correct PDF, without any subjective rule, satisfies all the criteria (Kornejady et al. 2017). To support the model improvement, it is necessary to pick some training sites, background sites: nevertheless, the model has no option but to recognize certain pixels as absence sites when computing area under curve (AUC). Regarding the template algorithm, presence-only attribute pushes ME to seek a generative approach in which it directly tackles a biased question by the Bays' rule (Rahmati et al. 2016). According to Dudík et al. (2004), ME is almost same as maximized Gibbs distribution. Since we do not need ME to be essentially the same and appropriate for the training set, the log-loss from the formula can be extracted in a way that penalizes the log-probability and circumvents an over-fitting problem. This approach should increase the accuracy and generalization capacity of the final estimated PDF at the same time. Phillips et al. (2006) and Phillips and Dudík (2008) provided more detailed regarding this method.

Support vector machine (SVM)

Boser, Guyon and Vapnik first suggested SVM as a controlled machine learning system with an intellectual design in 1992 (Jakkula 2006). Applying a learning algorithm based on the theory of predictive learning and optimization, SVM helps the program to know how to perform categorization and regression functions, improve precision and thereby escape over-fitting disadvantages (Guo et al. 2005; Huang et al. 2002). Delegating samples to two or more classes is the primary idea behind a classification scheme. Contrarily, regression functions can be incorporated to estimate values for the continuous dependent variable. When faced with an extremely unorganized set, SVM does not consider a continuous decision surface in 2D space, but it uses a kernel trick to turn the data into higher-dimensional space and then define the data by hyperplane (Kecman 2005; Statnikov et al. 2013). A kernel in a function space is a product of a dot. In the present study, one of the most common used Gaussian Kernel functions was used. Yao et al. (2008) and Marjanović et al. (2011) provided detailed explanation of the SVM model in their works.

Accuracy assessment

ROC curve

In ROC curve, sensitivity and 1 – specificity of all possible cut-off values are plotted (Akgün and Türk 2011;

Roy and Saha 2021; Gayen et al. 2019). The sensitivity consisted of all the cells detected as susceptible in gullies and the specificity consisted of all the cells without gullies detected as not susceptible. The validation data set was compared with the susceptibility maps for gully erosion and analyzed using several metrics, including: area under curve of ROC (AUCROC) (Lee and Park 2013; Pradhan et al. 2014; Bui et al. 2018; Akgün and Türk 2011; Roy and Saha 2021; Gayen et al. 2019). Even though there is no consistent criterion for choosing a cut-off value to estimate natural disasters, researchers have suggested the likelihood (P) if it is higher than 0.5 is no stable (Ahmadlou et al. 2018). AUCROC quantify the predictability of the model. If the AUC value is below 0.5, the model is not suitable and represents a random prediction, and if the value is near to 1 the model is highly suitable (Walter 2002).

Statistical methods

On the basis of the variations between the values expected by a model and the values actually observed, the root mean square error (RMSE) was determined (Eq. 1).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left[\overline{X} - X\right]^2}{N}} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n}\left|\overline{X} - X\right| \tag{2}$$

where N is sample size, $\overline{X}$ means predicted values and X means observed values. According to RMSE, mean absolute error (MAE) is described as the sum of the deviations between the predicted values of a variable and the real observed values (Eq. 2).

R-index method

The GESMs' accuracy was measured using a relative gully density index (R-index). The R-index is given as follows, by (Baeza 2001) (Eq. 3):

$$R = ((ni/Ni)/\sum(ni/Ni)) \times 100 \tag{3}$$

where ni is the percentage of the region of each susceptibility class susceptible to gullies, and Ni is the gullies number in each susceptibility class. If the R-index value increases in the different susceptibility groups, the effect would be more accurate and verified (Meena et al. 2019).

Assessing the influence of GECFs

Chi-square method

Optimization algorithms, most widely utilized in artificial intelligence, pick a limited range of features in the learning dataset to reduce the expense and time of modeling activities and achieve effective results across modeling processes (Vafaie and Imam 1994). Different factor selecting algorithms are available (Karegowda 2010). The Chi-square attribute evaluation (CSEA) methodology was utilized in this analysis. The CSEA is calculated from the Eq. 4.

$$x^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} \tag{4}$$

where E is expected values and O is observed values. The higher the chi-square value for a given conditioning variable, the more relevant it is for the study of gully susceptibility.

Random forest (RF) model

The random forest (RF) is an extension of the CART model and was first suggested by Breiman (2001). An efficient model for classification and regression problems is RF (Gayen et al. 2019). This RF is a major update of the bootstrap aggregation and belongs to the ensemble model group (Pourghasemi and Rahmati 2018). More information about the RF can be found in the Breiman (2001), Liaw and Wiener (2002), and Segal (2004). Through the mean decrease accuracy, significance of the selected factor was assessed.

In our analysis, "random forest" package in R programming language was used for running the RF model. RF was used to examine the influence of factors in mapping the gully erosion. As a consequence, the adjust RF feature of the "Random Forest" package was used to optimize the number of trees in RF and the quantity of variables attempted to obtain the lowest OOB error for each break.

Results

The output of the current work is discussed on the below sections.

Assessment of multi-collinearity

Multi-collinearity must be checked for hazard modeling to exclude strongly associated GECFs from the modeling phase and to prevent any errors in modeling outcomes (Gayen et al. 2019; Roy et al. 2019). The multi-collinearity problem actually occurs when two are very strongly linked (Roy and Saha 2019). Two measures were used to specify the

Table 2 Result of Multi-collinearity test of selected GECFs

Factors	Tolerance	VIF
Altitude	0.25	4.00
Slope	0.23	4.35
Clay	0.29	3.45
Convergence Index (CI)	0.65	1.54
Relative Slope Position	0.22	4.55
Sand	0.25	4.00
Silt	0.60	1.67
Topographic ruggedness index (TRI)	0.26	3.85
Topographic position index (TPI)	0.37	2.70
Topographic wetness index (TWI)	0.39	2.56
Valley Depth (VD)	0.23	4.35
Plan Curvature (PC)	0.42	2.38
Aspect	0.65	1.54
LULC	0.78	1.28
Drainage Density (Dd)	0.25	4.00
Lithology	0.56	1.79
Rainfall	0.37	2.70

multi-collinearity problem including tolerance and inflation factor variance (VIF). O'Brien (2007) mentioned a tolerance of 0.20 or lower and a VIF of 5 or higher indicated a multi-collinear problem. In this analysis, the minimum tolerance is 0.22 and the highest VIF is 4.35 (Table 2). Therefore, all the explanatory variables can be used for the modeling processes.

Assessment of gully erosion models

Gully erosion susceptibility maps (GESMs) were created in the current study by the CNN, ANN, ME and SVM models. Conoscenti et al. (2013) elaborated that GESMs not only identify the permanent or ephemeral gullies or gully erosion but also identify the new gullies location in future. Models of susceptibility to gully erosion using machine learning frameworks were established to predict gully erosion in the study region using training data sets. As the training datasets, gully and non-gully pixels were considered. For running the models, pixels (70%) from both groups were randomly set as training datasets. As 0 and 1 codes, respectively, the gully and non-gully were characterized. The relational weights of the models were added to determine the gully erosion susceptibility indices for all pixels until all the four models were successfully run in the training phase. Of the total area, the gully erosion susceptibility indices of all pixels were determined and a particular gully erosion susceptibility index was assigned to each pixel. Using a mathematical approach, susceptibility indices for gully erosion are reclassified. The natural break technique of Jenk was used as a computational method for reclassifying gully erosion susceptibility indices for this study. GESMs were classified into five groups, namely very low, low, medium, high, and very high levels of susceptibility to erosion in the GIS environment (Hong et al. 2016; Pradhan and Lee 2010). In fact, the classification of natural breaks determines the boundary of increasing susceptibility degree or class depending on the

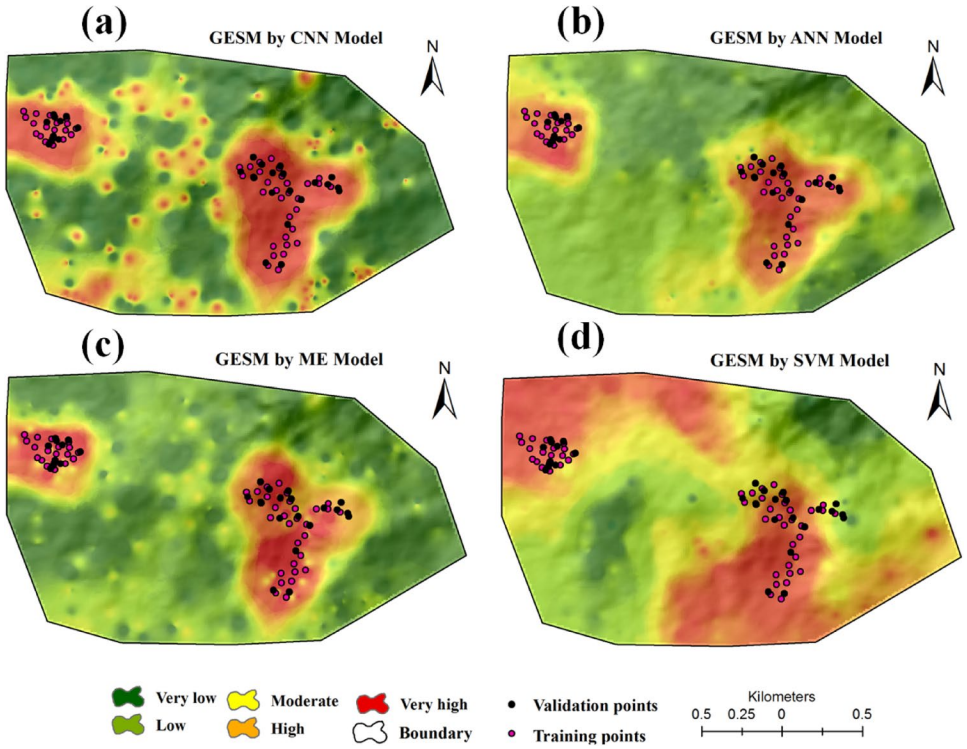


Fig. 9 Gully erosion susceptibility models: 1. CNN, 2. ANN, 3. ME and 4. SVM

intrinsic existence of the GESMs if a major value change occurs. The percentages of susceptibility areas are most vital for spatial gully erosion in existing and future. In case of convolution neural network (CNN) model, the very high susceptibility area is 20.82%, followed by very low (26.91%), low (21.80%), medium (16.77%) and high (13.70%) susceptibility classes (Fig. 9a). Based on artificial neural network (ANN) model, 24.89, 39.99, 12.71, 6.60 and 15.81% of the study area are covered by the very low, low, moderate, high and very high susceptibility classes, respectively (Fig. 9b). By contrast, study area's 29.68% in SVM model is covered by the very high susceptibility class which is significantly greater than the other models. The ME model shows that the 32.04 and 11.72% of the study area was occupied by the very low and very high levels of susceptibility (Fig. 9c).

Robustness of models

One deep learning model (CNN) and three benchmark machine learning techniques (ANN, SVM and ME) were used in this study to build GESMs, and training and test data sets were considered to verify the efficiency of these models. The testing and training datasets are familiar with succession curve rate (SCR) and prediction curve rate (PCR). The AUC for CNN, ANN, SVM and ME models was produced using ROC curve. Based on the SCR, the CNN, ANN, SVM and ME models achieved the AUC values of 0.910, 0.887, 0.712 and 0.865, respectively (Fig. 10a). On the other hand, the AUC values of 0.929 (CNN), 0.900 (ANN), 0.890 (ME) and 0.745 (SVM) were attained using prediction curve rate (PCR) (Fig. 10b).

Statistical measures using RMSE and MAE approaches are another vital accuracy assessment perspective for GESMs. To evaluate the accuracy of CNN, ANN, ME and

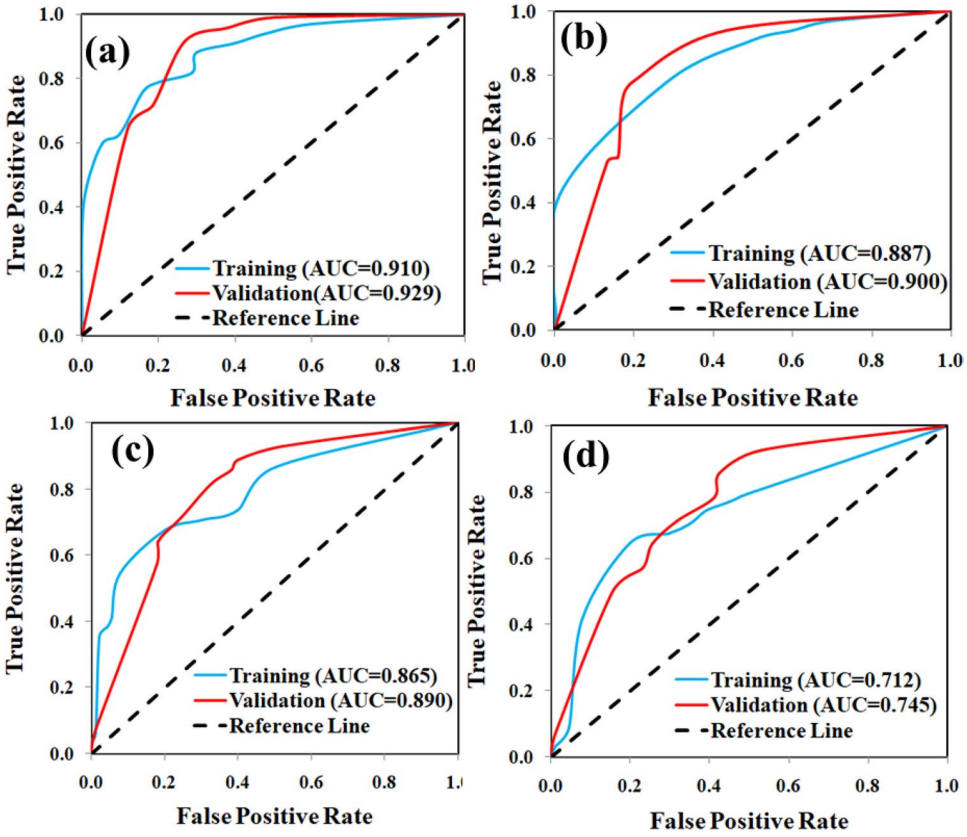


Fig. 10 ROC curves prepared for validation the models using training datasets and validation datasets: **a** CNN, **b** ANN, **c** ME and **d** SVM

Table 3 Result of validation methods used in this study

Statistical methods	Training datasets				Validation datasets			
	CNN	ANN	ME	SVM	CNN	ANN	ME	SVM
AUC	0.910	0.887	0.865	0.712	0.929	0.900	0.890	0.745
MAE	0.029	0.089	0.108	0.189	0.089	0.109	0.109	0.207
RMSE	0.171	0.298	0.328	0.435	0.299	0.329	0.339	0.455

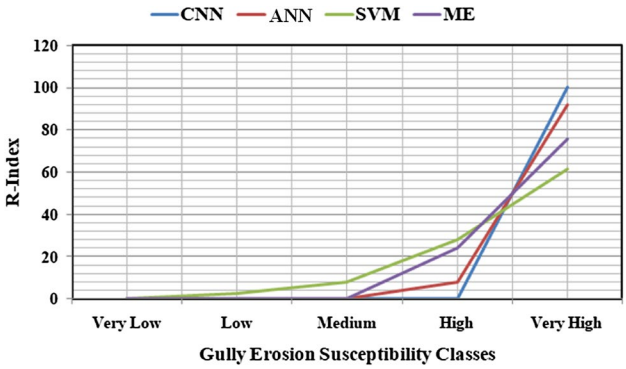


Fig. 11 Gully erosion density (R-index) under different susceptibility classes

SVM models, RMSE and MAE values were calculated using testing and training datasets. According to testing and training datasets, the CNN model achieved the lowest MAE value and the SVM achieved the highest value (Table 3). Similar results were found in case of RMSE (Table 3). But the values of RMSE and MAE for all the models are less than 0.5. Therefore, all models are correct and appropriate for GESMs assessment.

R-index is important validation technique for GESMs. According to the R-index, the very high susceptibility class will have the highest R-index considering the best suitable model. In the very high susceptibility class of CNN model, the highest R-index value is found (Fig. 11, Table 4). The CNN model's R-index values are higher in very high sensitivity level than the other ones, and lower than the other classes. Thus, it can be inferred from the study of different precision measurements that all the chosen models have the capabilities to model the susceptibility to gully erosion. Among the selected models, the CNN is the strongest model. Capability of the SVM model is lower than the other two models because this model has lowest AUC as well as R-Index value and the highest RMSE and MAE values.

Discussions

As a dynamic geomorphic process, gully erosion inflicts substantial damage to economy, infrastructure and natural resources (Azareh et al. 2019; Debanshi and Pal 2020; Roy and Saha 2021). High-precision modeling and prediction of gully erosion still seems to be difficult. A novel hybrid deep learning based on CNN, ANN, SVM and ME was implemented and tested for this purpose in this study. The benefit of the proposed approach is that profound learning is implemented to estimate the susceptibility index of gully erosion with a calculation of clarity and simplicity. However, deep learning has the ability to combine both quantitative and qualitative factors, and the significance of options can be measured in percentage. Therefore, for this study, deep learning and data mining techniques, including

Table 4 Values of R-index of different sub-categories of gully erosion susceptibility models

Models	Susceptibility classes	No of pixels	% of pixels	No of gully pixels	R-index
CNN	Very low	22,994	26.91	0	0
	Low	18,630	21.80	0	0
	Medium	14,329	16.77	0	0
	High	11,705	13.70	0	0
	Very high	17,793	20.82	5248	100
ANN	Very low	21,271	24.89	0	0
	Low	34,170	39.99	0	0
	Medium	10,862	12.71	0	0
	High	5638	6.60	186	8
	Very high	13,510	15.81	5062	92
SVM	Very low	6383	7.47	2	0
	Low	20,200	23.64	113	2
	Medium	15,456	18.09	295	8
	High	18,050	21.12	1184	28
	Very high	25,362	29.68	3654	61
ME	Very low	27,381	32.04	0	0
	Low	27,777	32.51	0	0
	Medium	12,267	14.36	1	0
	High	8014	9.38	1071	24
	Very high	10,012	11.72	4176	76

CNN, ANN, ME and SVM are proposed. The major benefit of using the deep learning and data mining algorithms in the implemented gully erosion models is that they do not even need statistical assumptions regarding the distribution of the used data. For the preparation of possible groundwater maps, Panahi et al. (2020) used a MLA and DLA, specifically SVM and CNN It has been suggested in the said research that CNN has higher prediction capacity than SVM. Chen et al. (2017) used three well-known models of machine learning, namely ME, SVM, and ANN and their ensembles and ANN has obtained the highest AUC. Pourghasami et al. (2017) used MLA to map vulnerability to gully erosion in the watershed of Aghemam, Iran, namely the ANN, SVM), ME, and their ensembles. The ANN–SVM ensemble demonstrated the highest accuracy and predictive strength. In the current study, CNN—the deep learning method, and three bench mark methods such as ANN, ME, and SVM were effectively used and compared. CNN model has given the better accuracy than machine learning techniques used in the previous gully erosion studies (Chen et al. 2017; Roy et al. 2019; Gayen et al. 2019). For precise mapping of the gully erosion susceptibility, therefore, CNN model can be used which gives a better accuracy than the benchmark models used in this study.

In the current study, a total of 17 GECFs were chosen using the multi-collinearity assessment to prepare the GESM. The multi-collinearity test results indicated that all of the selected variables are appropriate for GESMs. The evaluation of the influence of geo-environmental variables to the gully erosion models is of considerable significance and was explored extensively in the literatures (Amiri et al. 2019; Panahi 2020). Generally, the relative significance of GECFs depends on the model and the measurement criteria used, and so a factor has a lower involvement to one model that may contribute significantly to other models (Tien Bui et al. 2016). But in this study, both the methods used for measuring the contribution of the GECFs have shown the same result. Both the methods RF and CSAE showed sand concentration as the main determining factor.

The comparison and accuracy appraisal of the deep learning and standard machine learning models is one of the key goals of this research. CNN had the highest goodness-of-fit and predictive power followed by the models ANN, ME and SVM according to the AUC, RMSE, MAE and R-index. Because of various sampling approaches for training and testing data partitioning, different spatial sizes, specific predictors and different model validation techniques, it is difficult to equate our findings explicitly with other experiments, although the deep learning methodology has provided stronger results than traditional methods and machine learning methods (Conoscenti et al. 2014; Chen et al. 2017; Pourghasemi et al. 2017; Hembram and Saha 2020; Roy and Saha 2019; Bosino et al. 2020; Cama et al. 2020; Roy and Saha 2021). By the natural break classification method in the present study, GESM was categorized into five classes, including very low, low, moderate, high and very high levels of susceptibility. The natural break classification scheme is perhaps the most effective to distinguish the susceptibility zones of gully erosion (Saha et al. 2020). The CNN model showed that the 20.8% of the study area is very highly susceptible (Fig. 12). In this analysis, the most vulnerable areas of gully erosion are located in the area of the high slopes, and having high concentration of sand. Above all, the most important matter of this region is the gully erosion-prone environment. Natural habitats, farm property, social housing and road infrastructure can be ruined by gully erosion and the economic development of the country may also disappear.

Previous research found that gully erosion studies could be best supported by machine learning algorithms

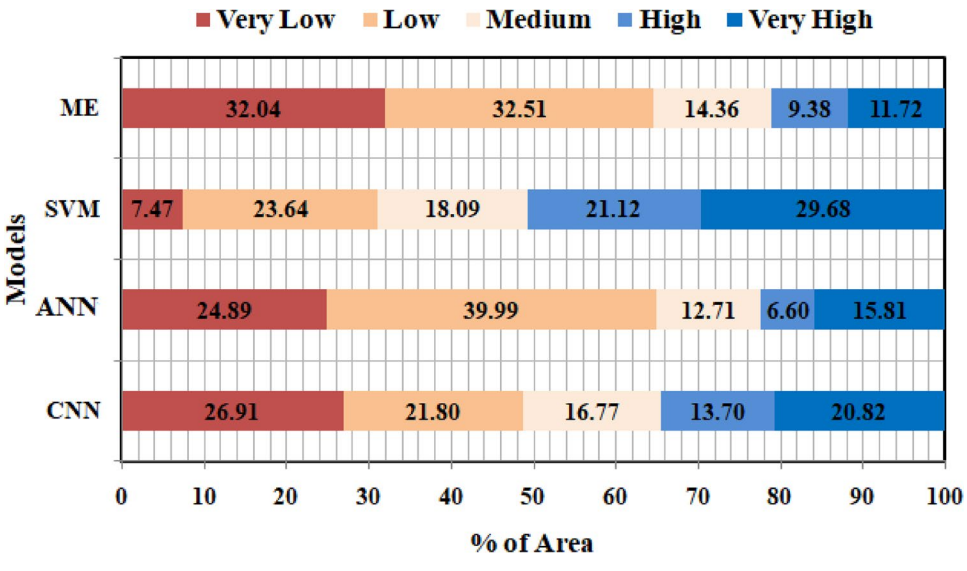


Fig. 12 Area coverage of different gully erosion susceptibility classes

than conventional statistical approaches (Zabihi et al. 2019; Gayen et al. 2019; Pourghasemi et al. 2018; Roy and Saha 2021; Saha et al. 2020). Nevertheless, our observations can differ since they rely on the position and range of gully conditioning variables. The deep learning techniques, therefore, are sufficiently powerful and important methods for mapping the susceptibility to gully erosion in today's environment. The models can also be used to forecast gully erosion occurrences in other regions of Bhutan, in particular in areas with similar conditions of lithological, pedological and land use. In addition, the method may be generalized to predict susceptibilities of other landforms correctly mapped and predictor variables are available (e.g., susceptibilities to soil erosion in various environmental conditions). These GESMs will, however, assist decision makers to devise plans for minimizing gully erosion and environmental management. It is impossible to avoid or overcome the question of over-fitting. Several explanations are specified as follows: First in the case of this analysis, the points of non-gully are chosen to be the same as the points of gully erosion, whereas the real area of non-gully erosion is greater than the area of gully erosion. Second, the sampling ratio of 70:30 without conducting a sampling ratio accuracy test is specified by the training and research datasets. A significant benefit of the algorithms implemented in this analysis for deep learning and machine learning is that they simplify the task of researching several datasets to obtain valuable information. These algorithms can account for specific assumptions that can be used, along with the automatic analysis of massive databases, to help plan decisions. In future experiments to improve GESM precision, the following should be carried out: (1) accurate review of the gully-to-non-gully area ratio, (2) assessment of the sampling ratio of the training and validation dataset, and (3) recognition of the variations between the various feature selection methods.

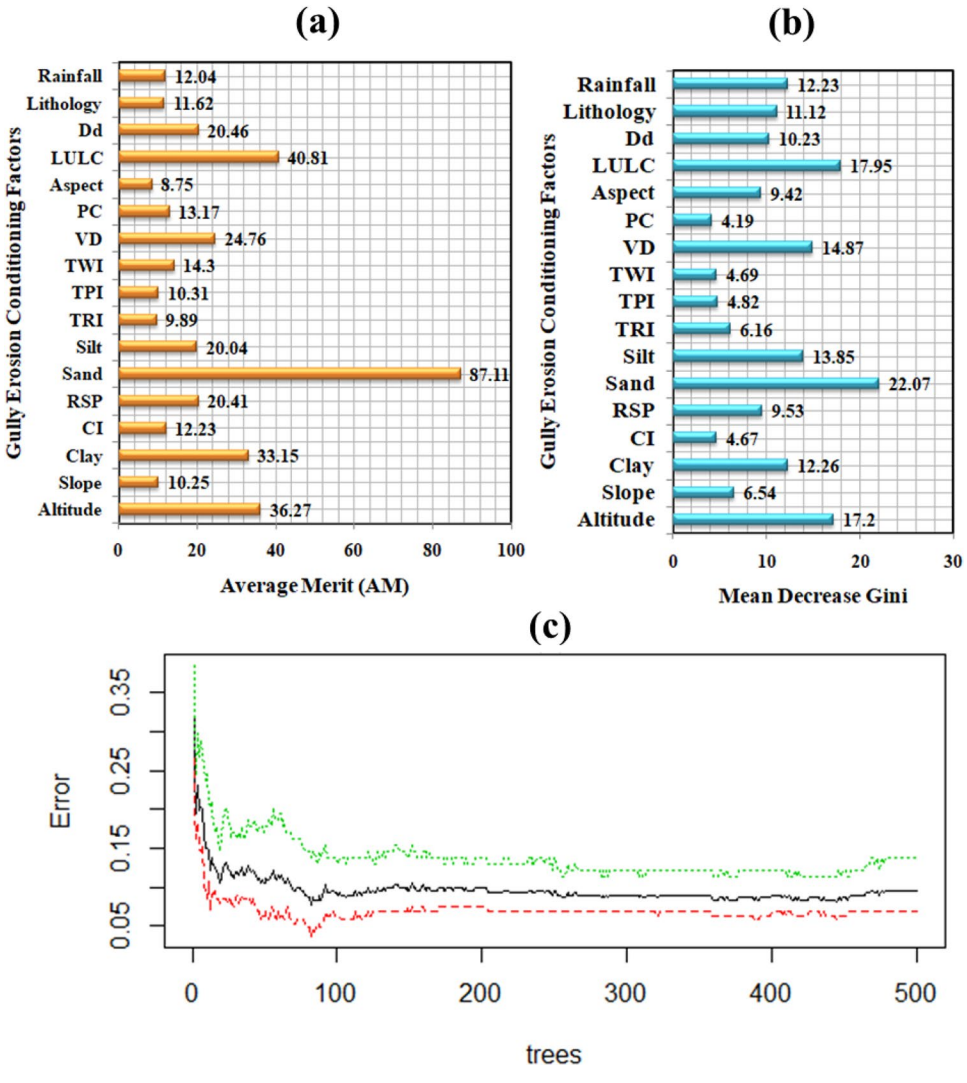


Fig. 13 Relative influence of different gully conditioning factors **a** Chi-square attribute evaluation, **b** RF and **c** error graph of RF model

Factors importance analysis

The significance of GECFs was assessed using two machine learning methods, namely Chi-square attribute evaluation (CSAE) and Random forest (RF) approaches. Chi-square attribute evaluation (CSAE) is an important technique that evaluates the significance of GECFs. In the current study, significance of the GECFs was also evaluated by the RF model in R programming software using the "Random Forest" package. The 500 number of tree and 4 number of variable were tried at each split which was built by RF model (Fig. 13c). The OOB error is 3.62% (Fig. 13c). The mean decrease Gini coefficient was utilized for assessing the importance of GECFs. The result of CSAE and RF models revealed that sand concentration (AM = 87.11 and RF = 22.07) has the highest contribution in making the area susceptible to gully erosion followed by LULC, altitude, clay, TRI, drainage density, slope, valley depth, silt, rainfall, lithology, aspect, TWI, topographic position index, relative slope position, convergence index and plan curvature (Fig. 13). Previous studies of Pourghasemi et al. (2017), Gayen et al. (2019), Roy and Saha (2019), etc. on gully erosion susceptibility assessment depicted slope, elevation and rainfall as the major gully conditioning factors. The first and utmost step to manage danger and hazards caused by gully erosion is to determine triggering factors.

Conclusions

In the present work, deep learning algorithm CNN was used to classify the area into different gully erosion susceptibility zones. Predicting gully susceptibility area is useful to take necessary steps for protective measures and reducing future potential damage. Gully erosion inventory map (GEIM) and 17 GECFs were constructed in GIS environment to produce the spatial gully erosion susceptibility map. GEIM was categorized into two groups such as the training and testing datasets. A critical field of concern in the watersheds vulnerable to this devastating phenomenon is the sense of the selected study area. Owing to the possession of GECFs with a footprint of man-made interruptions, this area is the most prone to gully erosion in the Phuentsholing region of Bhutan. Among the selected factors sand concentration, LULC, clay and altitude are the most important and have higher contribution in making the area prone to gully erosion. As far as the model results are concerned, CNN has provided better result in compared to the other models as they. It may be due to the algorithm used by the models while CNN benefited from learning algorithms for back-and-forth propagation that allowed the model to provide better result. Nearly, 20% of the study area has very high gully erosion susceptibility. In conclusion, the findings indicate that the deep learning and machine learning techniques integrated with remote sensing and GIS can be used for the accurate prediction of area susceptible to gully erosion. Considering the high precision and efficiency of the methods included in this study in recognizing the areas prone to gully erosion, it could be used in countries of similar climatic and geographical scenarios. This work will be helpful to the decision makers in soil and water management, and infrastructure planning of Phuentsholing region of Himalayan Kingdom, Bhutan.

Author contributions Conceptualization and methodology, writing—original draft preparation SS, RS and JR; software, JR; formal analysis, SS, RS, GT and JR; writing—review and editing, RS and GT; All authors have read and agreed to the published version of the manuscript.

Funding No funding received for this study.

References

Ahmadlou M, Karimi M, Alizadeh S, Shirzadi A, Parvinnejhad D, Shahabi H, Panahi M (2018) Flood susceptibility assessment using integration of adaptive network- based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms (BA). Geocarto Int. https://doi.org/10.1080/10106049.2018.14742

Akgün A, Türk N (2011) Mapping erosion susceptibility by a multivariate statistical method: a case study from the Ayvalık region, NW Turkey. Comput Geosci 37:1515–1524. https://doi.org/10.1016/j.cageo.2010.09.006

Amiri M, Pourghasemi HR, Ghanbarian GA, Afzali SF (2019) Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma 340:55–69. https://doi.org/10.1016/j.geoderma.2018.12.042

Anderson CG, Maxwell DC (2004) Starting a Digitization Center. Elsevier, Amsterdam, The Netherlands (**ISBN 978-1843340737**)

Angileri SE, Conoscenti C, Hochschild V, Märker M, Rotigliano E, Agnesi V (2016) Water erosion susceptibility mapping by applying stochastic gradient treeboost to the imeraMeridionale River basin (Sicily, Italy). Geomorphology 262:61–76. https://doi.org/10.1016/j.geomorph.2016.03.018

Azareh A, Rahmati O, Rafiei-Sardooi E, Sankey JB, Lee S, Shahabi H et al (2019) Modelling gully-erosion susceptibility in a semi-arid region, Iran: investigation of applicability of certainty factor and maximum entropy models. Sci Total Environ 655:684–696. https://doi.org/10.1016/j.scitotenv.2018.11.235

Baeza C, Corominas J (2001) Assessment of shallow landslide susceptibility by means of multivariate statistical techniques. Earth Surf Process Landf 26:1251–1263. https://doi.org/10.1002/esp.263

Beven KJ, Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology. Hydrol Sci Bull 24(1):43–69. https://doi.org/10.1080/02626667909491834

Bieganowski A, Ryżak M (2011) Soil texture: measurement methods. In: Gliński J, Horabik J, Lipiec J (eds) Encyclopedia of agrophysics. Encyclopedia of Earth Sciences Series. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3585-1_157

Bosino A, Giordani P, Quénéhervé G, Maerker M (2020) Assessment of calanchi and rill–interrill erosion susceptibilities using terrain analysis and geostochastics: a case study in the Oltrepo Pavese, Northern Apennines, Italy. Earth Surface Process Landforms 45(12):3025–3041

Breiman L (2001) Random forests. Mach Learn 45:5–32

Bui DT, Shahabi H, Shirzadi A, Chapi K, PradhanB CW, Khosravi K, Panahi M, Bin Ahmad B, Saro L (2018) Land subsidence susceptibility mapping in South Korea using machine learning algorithms. Sensors 18(8):2464. https://doi.org/10.3390/s18082464

Bull LJ, Kirkby MJ (1997) Gully processes and modelling. Prog Phys Geogr 21(3):354–374

Burrough PA, McDonell RA (1998) Principles of geographical information systems. Oxford University Press, New York, NY, USA, p 190

Cama M, Schillaci C, Kropáček J, Hochschild V, Bosino A, Märker M (2020) A probabilistic assessment of soil erosion susceptibility in a head catchment of the Jemma Basin, Ethiopian Highlands. Geosciences 10(7):248

Capra A, Mazzara LM, Scicolone B (2005) Application of the EGEM model to predict ephemeral gully erosion in Sicily, Italy. CATENA 59(2):133–146

Castillo VM, Mosch WM, Garcia CC, Barbera GG, Cano JAN, Lopez-Bermudez F (2007) Effectiveness and geomorphological impacts of check dams for soil erosion control in a semiarid Mediterranean catchment: El Carcavo (Murcia, Spain). CATENA 70:416–427. https://doi.org/10.1016/j.catena.2006.11.009

Chen W, Pourghasemi HR, Kornejady A, Zhang N (2017) Landslide spatial modeling: introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. Geoderma 305:314–327. https://doi.org/10.1016/j.geoderma.2017.06.020

Cherkassky V, Krasnopolsky V, Solomatine DP, Valdes J (2006) Computational intelligence in earth sciences and environmental applications: issues and challenges. Neural Netw 19:113

Conoscenti C, Agnesi V, Angileri S, Cappadonia C, Rotigliano E, Märker M (2013) A GIS-based approach for gully erosion susceptibility modelling: a test in Sicily, Italy. Environ Earth Sci 70:1179–1195. https://doi.org/10.1007/s12665-012-2205-y

Conoscenti C, Angileri S, Cappadonia C, Rotigliano E, Agnesi V, Märker M (2014) Gully erosion susceptibility assessment by means of GIS-based logistic regression: a case of Sicily (Italy). Geomorphology 204:399–411. https://doi.org/10.1016/j.geomorph.2013.08.021

Conoscenti C, Agnesi V, Cama M, Caraballo-Arias NA, Rotigliano E (2018) Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity. Land Degrad Dev 29(3):724–736. https://doi.org/10.1002/ldr.2772

Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J (2015) System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci. Model Dev 8:1991–2007 (**8(7), 2271–2312**)

Debanshi S, Pal S (2020) Assessing gully erosion susceptibility in Mayurakshi river basin of eastern India. Environ Dev Sustain 22(2):883–914. https://doi.org/10.1007/s10668-018-0224-x

Desta L, Adunga B (2012) A field guide on gully prevention and control Nile Basin initiative, eastern Nile subsidiary action program (ENSAP), Eastern Nile, Technical Regional Office (ENTRO), Eastern Nile Watershed Management Project, Addis Ababa

Diaz V, Mongil J, Navarro J (2014) Topographical surveying for improved assessment of sediment retention in check dams applied to a Mediterranean badlands restoration site (Central Spain). J Soils Sediments 14:2045–2056. https://doi.org/10.1007/s11368-014-0958-5

Dikshit A, Sarkar R, Pradhan B, Acharya S, Dorji K (2019) Estimating rainfall thresholds for landslide occurrence in the Bhutan Himalayas. Water 11(8):1616. https://doi.org/10.3390/w11081616

Dudík M, Phillips SJ, Schapire RE (2004) Performance guarantees for regularized maximum entropy density estimation. In: Learning Theory, Conference on Learning Theory, COLT 2004, Banff, Canada, July 1–4, 2004, Proceedings, 472–486.

Ferro V, Giordano G, Iovino M (1991) Isoerosivity and erosion risk map for Sicily. J Hydrol Sci 36:549–564. https://doi.org/10.1080/02626669109492543

Gallant JC, Wilson JP (2000) Primary topographic attributes. In: Wilson JP, Gallant JC (eds) Terrain analysis: principles and applications. Wiley, New York, pp 51–85

Gariano S, Sarkar R, Dikshit A, Dorji K, Brunetti MT, Peruccacci S, Melillo M (2019) Automatic calculation of rainfall thresholds for landslide occurrence in Chukha Dzongkhag, Bhutan. Bull Eng Geol Environ 78:4325–4332. https://doi.org/10.1007/s10064-018-1415-2

Garosi Y, Sheklabadi M, Conoscenti C, Pourghasemi HR, Van Oost K (2019) Assessing the performance of GIS-based machine learning models with different accuracy measures for determining susceptibility to gully erosion. Sci Total Environ 664:1117–1132. https://doi.org/10.1016/j.scitotenv.2019.02.093

Gayen A, Pourghasemi HR, Saha S, Keesstra S, Bai S (2019) Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. Total Environ Sci. https://doi.org/10.1016/j.scitotenv.2019.02.436

Gayen A, Saha S (2017) Application of weights-of-evidence (WoE) and evidential belief function (EBF) models for the delineation of soil erosion vulnerable zones: a study on Pathro river basin, Jharkhand, India. Model Earth Syst Environ 3(3):1123–1139. https://doi.org/10.1007/s40808-017-0362-4

Geissen V, Kampichler C, López-de Llergo-Juárez J, Galindo-Acántara A (2007) Superficial and subterranean soil erosion in Tabasco, tropical Mexico: development of a decision tree modeling approach. Geoderma 139:277–287. https://doi.org/10.1016/j.geoderma.2007.01.002

Geroy IJ, Gribb MM, Marshall HP, Chandler DG, Benner SG, McNamara JP (2011) Aspect influences on soil water retention and storage. Hydrol Process 25:3836–3842. https://doi.org/10.1002/hyp.8281

Gómez-Gutiérrez Á, Conoscenti C, Angileri SE, Rotigliano E, Schnabel S (2015) Using topographical attributes to evaluate gully erosion proneness (susceptibility) in two mediterranean basins: advantages and limitations. Nat Hazards 79(1):291–314. https://doi.org/10.1007/s11069-015-1703-0

Guerra AJT, Fullen MA, Jorge MDCO, Bezerra JFR, Shokr MS (2017) Slope processes, mass movement and soil erosion: a review. Pedosphere 27:27–41. https://doi.org/10.1016/S1002-0160(17)60294-7

Guo Q, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol Model 182:75–90. https://doi.org/10.1016/j.ecolmodel.2004.07.012

Gutiérrez ÁG, Conoscenti C, Angileri SE, Rotigliano E, Schnabel S (2015) Using topographical attributes to evaluate gully erosion proneness (susceptibility) in two mediterranean basins: advantages and limitations. Nat Hazards 79:291–314. https://doi.org/10.1007/s11069-015-1703-0

Hancock GR, Evans KG (2010) Gully, channel and hillslope erosion—an assessment for a traditionally managed catchment. Earth Surf Process Landf 35:1468–1479. https://doi.org/10.1002/esp.2043

Hembram TK, Saha S (2020) Prioritization of sub-watersheds for soil erosion based on morphometric attributes using fuzzy AHP and compound factor in Jainti River basin, Jharkhand, Eastern India. Environ Dev Sustain 22(2):1241–1268. https://doi.org/10.1007/s10668-018-0247-3

Hembram T, Paul GC, Saha S (2018) Spatial prediction of susceptibility to gully erosion in Jainti River basin, Eastern India: a comparison of information value and logistic regression models. Model Earth Syst Environ. https://doi.org/10.1007/s40808-018-0560-8

Hong H, Naghibi SA, Pourghasemi HR, Pradhan B (2016) GIS-based landslide spatial modeling in Ganzhou City, China. Arab J Geosci 9(2):1–26. https://doi.org/10.1007/s12517-015-2094-y

Hosseinalizadeh M, Kariminejad N, Chen W, Pourghasemi HR, Alinejad M, Behbahani AM, Tiefenbacher JP (2019) Spatial modelling of gully headcuts using UAV data and four best-first decision classifier ensembles (BFTree, Bag-BFTree, RS-BFTree, and RF-BFTree). Geomorphology 329:184–193

Huang C, Davis LS, Townshend JRG (2002) An assessment of support vector machines for land cover classification. Int J Remote Sens 23:725–749. https://doi.org/10.1080/01431160110040323
Jakkula V (2006) Tutorial on Support Vector Machine (svm). 37 School of EECS, Washington State University.
Jaynes ET (1957a) Information theory and statistical mechanics. Phys Rev 106:620. https://doi.org/10.1103/PhysRev.106.620
Jaynes ET (1957b) Information theory and statistical mechanics. II Phys Rev 108:171. https://doi.org/10.1103/PhysRev.108.171
Karegowda AG, Manjunath A, Jayaram M (2010) Comparative study of attribute selection using gain ratio and correlation based feature selection. Int J Inf Technol Knowl Manag 2:271–277
Keay-Bright J, Boardman J (2009) Evidence from field-based studies of rates of soil erosion on degraded land in the central Karoo, South Africa. Geomorphology 103:455–465. https://doi.org/10.1016/j.geomorph.2008.07.011
Kecman V (2005) Support Vector Machines—An Introduction. Springer, Berlin Heidelberg
Kiss R (2004) Determination of drainage network in digital elevation model, utilities and limitations. J Hung Geomath 2:16–29
Knighton D (1998) Fluvial forms and processes: a new perspective 1998. Arnold, London
Kornejady A, Ownegh M, Bahremand A (2017) Landslide susceptibility assessment using maximum entropy model with two different data sampling methods. CATENA 152:144–162. https://doi.org/10.1016/j.catena.2017.01.010
LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
Lee S, Park I (2013) Application of decision tree model for the ground subsidence hazard mapping near abandoned underground coal mines. J Environ Manag 127:166–176. https://doi.org/10.1016/j.jenvman.2013.04.010
Li Z, Zhu Q, Gold C (2005) Digital Terrain Modeling: Principles and Methodology. CRC Press, Boca Raton, FL, USA
Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22
Lin L, Weng F (2018) Estimation of hurricane maximum wind speed using temperature anomaly derived from advanced technology microwave sounder. IEEE Geosci Remote Sens Lett 15(5):639–643. https://doi.org/10.1109/LGRS.2018.2807763
Lucà F, Conforti M, Robustelli G (2011) Comparison of GIS-based gullying susceptibility mapping using bivariate and multivariate statistics: Northern Calabria, South Italy. Geomorphology 134(3–4):297–308. https://doi.org/10.1016/j.geomorph.2011.07.006
Maharaj RJ (1993) Landslide processes and landslide susceptibility analysis from an upland watershed: a case study from St. Andrew, Jamaica West Indies. Eng Geol 34(1–2):53–79
Mararakanye N, Sumner PD (2017) Gully erosion: a comparison of contributing factors in two catchments in South Africa. Geomorphology 288:99–110
Marjanović M, Kovačević M, Bajat B, Voženílek V (2011) Landslide susceptibility assessment using SVM machine learning algorithm. Eng Geol 123:225–234. https://doi.org/10.1016/j.enggeo.2011.09.006
May L, Place C (2005) A GIS-Based Model of Soil Erosion and Transport. Freshwater Forum. 23. 48–61.http://aquaticcommons.org/id/eprint/4694
Meena SR, Ghorbanzadeh O, Blaschke T (2019) A comparative study of statistics-based landslide susceptibility models: a case study of the region affected by the gorkha earthquake in Nepal. ISPRS Int J Geo-inf 8(2):94. https://doi.org/10.3390/ijgi8020094
Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol Process 5:3–30
Myung IJ (2003) Tutorial on maximum likelihood estimation. J Math Psychol 47:90–100
Nagarajan R, Roy A, Kumar RV, Mukherjee A, Khire MV (2000) Landslide hazard susceptibility mapping based on terrain and climatic factors for tropical monsoon regions. Bull Eng Geol Environ 58(4):275–287
Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), 807–814
O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41:673–690. https://doi.org/10.1007/s11135-006-9018-6
Panahi M, Sadhasivam N, Pourghasemi HR, Rezaie F, Lee S (2020) Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.125033
Parlak M (2007) Determination of erosion risk according to CORINE methodology (a case study: Kurtboğazı Dam). Int Congress River Basin Manage 1:856
Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Model 133:225–245. https://doi.org/10.1016/S0304-3800(00)00322-7
Phillips SJ, Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x
Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026
Pourghasemi HR, Rahmati O (2018) Prediction of the landslide susceptibility: which algorithm, which precision? CATENA 162:177–192. https://doi.org/10.1016/j.catena.2017.11.022
Pourghasemi HR, Yousefi S, Kornejady A, Cerdà A (2017) Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. Sci Total Environ 609:764–775
Pradhan B, Lee S (2010) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ Model Softw 25:747–759. https://doi.org/10.1016/j.envsoft.2009.10.016
Pradhan B, Abokharima MH, Jebur MN, Tehrany MS (2014) Land subsidence susceptibility mapping at Kinta Valley (Malaysia) using the evidential belief function model in GIS. Nat Hazards 73:1019–1042. https://doi.org/10.1007/s11069-014-1128-1
Prasad R, Pandey A, Singh KP, Singh VP, Mishra RK, Singh D (2012) Retrieval of spinach crop parameters by microwave remote sensing with back propagation artificial neural networks: a comparison of different transfer functions. Adv Space Res 50:363–370. https://doi.org/10.1016/j.asr.2012.04.010
Rahmati O, Haghizadeh A, Pourghasemi HR, Noormohamadi F (2016) Gully erosion susceptibility mapping: the role of GIS-based bivariate statistical models and their comparison. Nat Hazards 82:1231–1258. https://doi.org/10.1007/s11069-016-2239-7
Rahmati O, Tahmasebipour N, Haghizadeh A, Pourghasemi HR, Feizizadeh B (2017b) Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. Geomorphology 298:118–137
Rahmati O, Tahmasebipour N, Haghizadeh A, Pourghasem HR, Feizizadeh B (2017a) Evaluating the influence of geo-environmental factors on gully erosion in a semi-arid region of Iran: an

integrated framework. Sci Total Environ 579:913–927. https://doi.org/10.1016/j.scitotenv.2016.10.176
Riley SJ, DeGloria SD, Elliot R (1999) A terrain ruggedness that quantifies topographic heterogeneity. Intermt J Sci 5(1–4):23–27
Rodriguez F, Maire E, Courjault-Rade P, Darrozes J (2002) The Black Top Hat function applied to a DEM: a tool to estimate recent incision in a mountainous watershed (Estibere Watershed, Central Pyrenees). Geophys Res Lett 29(6):9-1-9–4
Roy J, Saha S (2019) GIS-based gully erosion susceptibility evaluation using frequency ratio, cosine amplitude and logistic regression ensembled with fuzzy logic in hinglo river basin, India. Remote Sens Appl Soc Environ 15:100247. https://doi.org/10.1016/j.rsase.2019.100247
Roy J, Saha S (2021) Integration of artificial intelligence with Meta classifiers for the gully erosion susceptibility assessment in Hinglo river basin, Eastern India. Adv Space Res 67(1):316–333. https://doi.org/10.1016/j.asr.2020.10.013
Roy J, Saha S, Arabameri A, Blaschke T, Bui DT (2019) A novel ensemble approach for landslide susceptibility mapping (lsm) in darjeeling and kalimpong districts, west bengal, india. Remote Sensing 11(23):2866. https://doi.org/10.3390/rs11232866
Saha S, Gayen A, Pourghasemi HR, Tiefenbacher JP (2019) Identification of soil erosion-susceptible areas using fuzzy logic and analytical hierarchy process modeling in an agricultural watershed of Burdwan district. India. Environ Earth Sci 78(23):649. https://doi.org/10.1007/s12665-019-8658-5
Saha S, Roy J, Arabameri A, Blaschke T, Bui DT (2020) Machine learning-based gully erosion susceptibility mapping: a case study of Eastern India. Sensors 20(5):1313. https://doi.org/10.3390/s20051313
Sarkar R, Dorji K (2019) Determination of the probabilities of landslide events—a case study of Bhutan. Hydrology 6:52. https://doi.org/10.3390/hydrology6020052
Schicker R, Moon V (2012) Comparison of bivariate and multivariate statistical approaches in landslide susceptibility mapping at a regional scale. Geomorphology 161:40–57. https://doi.org/10.1016/j.geomorph.2012.03.036
Segal MR (2004) Machine Learning Benchmarks and Random Forest Regression. Center for Bioinformatics and Molecular Biostatistics UC, San Francisco http://eprints.cdlib.org/uc/item/35x3v9t4. Accessed 22 June 2020
Shannon CE (1948) Amathematical theory of communication. Bell Syst Tech J 27:379–423 (**Math. Rev. (MathSciNet: MR10, 133e)**)
Shary PA, Sharaya LS, Mitusov AV (2002) Fundamental quantitative methods of land surface analysis. Geoderma 107:1–32. https://doi.org/10.1016/S0016-7061(01)00136-7
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint rXiv:1409.1556.
Sougnez N, van Wesemael B, Vanacker V (2011) Low erosion rates measured for steep, sparsely vegetated catchments in southeast Spain. CATENA 84:1–11. https://doi.org/10.1016/j.catena.2010.08.010
Statnikov A, Aliferis CF, Hardin DP, Guyon I (2013) A Gentle Introduction to Support Vector Machines in Biomedicine: Volime 2: Case Studies and Benchmarks, vol 2. World Scientific Publishing Co Inc
Sun W, Shao Q, Liu J, Zhai J (2014) Assessing the effects of land use and topography on soil erosion on the Loess Plateau in China. CATENA 121:151–163. https://doi.org/10.1016/j.catena.2014.05.009
Swarnkar S, Malini A, Tripathi S, Sinha R (2018) Assessment of uncertainties in soil erosion and sediment yield estimates at ungauged basins: an application to the Garra River basin. India Hydrol Earth Syst Sci 22:2471–2485. https://doi.org/10.5194/hess-22-2471-2018
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9.
Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides 13:361–378. https://doi.org/10.1007/s10346-015-0557-6
Umar Z, Pradhan B, Ahmad A, Jebur MN, Tehrany MS (2014) Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in west Sumatera Province, Indonesia. CATENA 118:124–135
Vafaie H, Imam IF (1994) Feature selection methods: Genetic algorithms vs. Greedy-like search. In International Conference on Fuzzy and Intelligent Control Systems; Walt Disney World: Orlando, FL, USA, 28
Vogel HJ, Bartke S, Daedlow K, Helming K, Kogel-Knabner I, Lang B, Rabot E, Russell D, Stosel B, Weller U, Wiesmeier M, Wollschlager U (2018) A systemic approach for modeling soil functions. SOIL 4:83–92. https://doi.org/10.5194/soil-4-83-2018
Walter SD (2002) Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 21(9):1237–1256. https://doi.org/10.1002/sim.1099
Wang Y, Fang Z, Hong H (2019) Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County. China Sci Total Environ 666:975–993. https://doi.org/10.1016/j.scitotenv.2019.02.263
Weiss A. (2001). Topographic Position and Landforms Analysis. Poster presentation. ESRI User Conference, San Diego, CA.
Wentworth CKA (1930) simplified method of determining the average slope of land surfaces. Am J Sci 117:184–194
Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on Support Vector Machine: a case study on natural slopes of Hong Kong, China. Geomorphology 101:572–582. https://doi.org/10.1016/j.geomorph.2008.02.011
Yesilnacar E, Topal T (2005) Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng Geol 79:251–266. https://doi.org/10.1016/j.enggeo.2005.02.002
Zabihi M, Mirchooli F, Motevalli A, Darvishan AK, Pourghasemi HR, Zakeri MA et al (2018) Spatial modelling of gully erosion in Mazandaran Province, northern Iran. CATENA 161:1–13. https://doi.org/10.1016/j.catena.2017.10.010
Zabihi M, Pourghasemi HR, Motevalli A, Zakeri MA (2019) Gully erosion modeling using GIS-based data mining techniques in northern Iran: a comparison between boosted regression tree and multivariate adaptive regression spline. In: Pourghasemi HR, Rossi M (eds) Natural Hazards GIS-Based Spatial Modeling Using Data Mining Techniques. Springer, pp 1–29
Zakerinejad R, Märker M (2014) Prediction of Gully erosion susceptibilities using detailed terrain analysis and maximum entropy modeling: a case study in the Mazayejan Plain, Southwest Iran. Geogr Fis Din Quat 37(1):67–76
Zeiler MD, Fergus R (2014) Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014. Proceedings, Part I, 818: 833
Zuo R, Xiong Y, Wang J, Carranza EJM (2019) Deep learning and its application in geochemical mapping. Earth-Sci Rev 192:1–14. https://doi.org/10.1016/j.earscirev.2019.02.023

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

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Object tracking and counting in a zone using YOLOv4, DeepSORT and TensorFlow

Dr. Shailender Kumar
Computer Science & Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

Vishal
Computer Engineering
Delhi Technological University
Delhi, India
vishal_2k17co376@dtu.ac.in

Pranav Sharma
Computer Engineering
Delhi Technological University
Delhi, India
pranavsharma_2k17co229@dtu.ac.in

Nitin Pal
Computer Engineering
Delhi Technological University
Delhi, India
nitinpal_2k17co215@dtu.ac.in

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395971

***Abstract*— The main objective of this research work is to solve multiple object tracking problems in a given frame, wherein the proposed model intends to identify and track various objects. The problem has been solved in three stages viz. detecting, identifying, and tracking the object in a particular zone, i.e., but it is observed that something more could be done in this field, mostly the MOT-A score was not up to the mark; hence the proposed research work utilized Kalman filters for obtaining enhanced results and compared the obtained MOT-A metric with previous works, and the results were good. Object detection and recognition occur via the YOLO algorithm, which enables us to classify the objects into 80 classes. Then, Motion Prediction and feature generation occur in which an estimation model is created, and Kalman filters are used to model these states for capturing moving objects in the frame. Finally, tracking takes place with Kalman filters in the previous frame, and newly detected objects are placed in the current frame, after which an association is made for new detection. All this is done via the DeepSORT algorithm, which is essentially a Deep association metric with the SORT algorithm. Here, Kalman filters are used as they improved the accuracy of the proposed model and yielded better results. On the same lines, YOLO is used to perform object detection and recognition at the same time. It is also a detector, which by applying a single neural network, it can predict the bounding boxes and perform multi-class classification. This problem can have various applications, especially in traffic management. It can also prevent people from gathering during COVID times and raising an alert for all the authorities. Hence, this is a multidisciplinary approach wherein the work of object detection is being used in various fields like crowd assembling, Surveillance, Animal management in Zoo/Biodiversity parks as well as in the case of traffic systems as well. The real motivation behind this work was using the state-of-the-art technology to solve the modern-day problems. Manually human vigilance in large areas is a utopian task and especially when surveillance and security is big threat out there. Adding to it is the COVID Pandemic which has claimed millions of lives and yet vaccination is a still a dream yet to be realized. Hence, cluster identification problem is considered and also the ways to solve this challenge using the state-of-the-art technology. Once it has been achieved, it is initiated further to find more applications of this technology in various other domains and found out how it is applied in various use cases in case of crowd gathering at a single point and how this is a similar problem in case of animal gathering at a point and how forest rangers can solve the problem in a more efficient way.**

Keywords—Kalman Filters, Multiple Object Tracking, DeepSORT, YOLOv4, Target tracking, TensorFlow, Image Processing, Image Segmentation.

I. INTRODUCTION

Multiple object tracking has become very much popular since its inception. This is due to the broad range of applicability and ease of use. It is also state-of-the-art technology, and with minimum inputs, this one can provide for best possible results. However, it is a little bit complicated, and it may produce a larger number of false positives [1]. For the reduction of unreliable detections, earlier it has been planned to use only the SORT algorithm, but now it is intended to use the Deep SORT algorithm. The proposed model has been evaluated by using the MOT17 dataset, with the advancement of the machine learning algorithm and introduction of new possibilities to our traditional multi-object tracking problem and also to come up with several changes in the technology front. Let's start with the beginning of such models.

Let's start with what exactly is the MOT problem? Multiple object tracking(MOT) problem is looking at the picture once and then detecting what exactly is there in the object. Many objects are analyzed in a single frame, and detecting is the first and foremost step, which requires many challenges and then tracking is another task. So there is an idea to divide the problem into two parts: detection and association. The model is also able to perform surveillance and move object detection and tracking in real-time. Also, the model is able to alert the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

system in case there is a surge in the number of people present in a particular area.

First of all, a complete set of objects in the box's frame are detected. Image segmentation is also an essential part of object detection [2]. Once the detection is over, a matching sequence is set to be performed for similar detections within the frame of the bounding box, and they are tracked till they remain within it. The next question arises as to how we are doing object tracking. For that, we initially created a set of object detections. Then we created a unique ID for each of the detections. After this, we tracked the object over a period of time till it remains in the frame. Finally, we maintained the ID till the object is in the frame. Once it moves out of the frame, the object's ID was destroyed, and this is how we are tracking the objects. For tracking cars, we used the SORT algorithm. It stands for Simple Online Real-time tracking, which followed the following steps:

- It applies Kalman Filtering and Hungarian method to handle motion prediction and data association
- Object models contain object position, scale, bounding box ratio and motion prediction for the next frame
- SORT solves the data association by calculating bounding box Intersection over union (IOU) distance
- Based on the loU distance, the final assignment is done
- It is fast and simple and has high precision and accuracy

Now let's see how to perform object tracking and detection. This majorly consists of 2 parts which are firstly detecting the objects and then put a bounding box around it in each and every frame of the system and then tracking the object in the bounding box till it moves out of the scope of the box. But there was one major drawback of this approach that it cannot handle occlusion very well as SORT relies on a simple motion model and it couldn't work well as per desired results. Hence, we chose the DeepSORT algorithm for the same. In this algorithm, we have used the SORT algorithm with an extension, i.e. a Deep Association matrix is used which improves the matching procedures and reduces the number of identity switches by adding visual appearance descriptor or appearance features. As a result of this, it was able to perform better and give desired results. We could also improve the accuracy by using motion measurement and via appearance features as well.

Now, let's see the working of YOLO and DeepSORT technique:

- Object detection and recognition take place via the YOLO technique.
- Motion Prediction and feature generation take place in which an estimation model is created and Kalman filters are used to model these states.
- Finally, tracking takes place with the help of Kalman filters in the previous frame and newly detected objects are placed in the current frame, after which an association is made for new detection.

We used Kalman filters as they improved the accuracy of our model as yielded better results. On the same lines, we used YOLO as it is able to perform object detection and recognition at the same time. It is also a detector, which by applying a single neural network can predict the bounding boxes and can perform multi-class classification. Further, we discuss previous related works in this field before coming to our approach.

II. Related Work

Previous work has shown that the Neural Network Proves to be an excellent algorithm for faster detection and classification of real-life objects. Neural Networks are mathematical models inspired by the human brain to store information that has been very successful in classifying and segmentation images. In the medical industry, deep learning (DL) can play an important part. In the past, DL is used to detect tumors in the brain, breast cancer, and more recently, with the ongoing pandemic, researchers have used DL models to detect novel coronavirus (SARS-CoV-2) through X-rays and CT scans.

CNN (Convolutional Neural Network) is a Deep Neural Network that analyzes visual images and performs tasks like object classification, detection, and segmentation. Its input is an image that gives a class/classes that best describes the input image. In CNN, the input layer's information is an image, the hidden layers do the feature extraction, and the output layer gives the output to the system. Some typical applications of Convolution Neural Network are identifying vehicles, Decoding Facial Recognition, Historic and Environmental Collections, and Cars and Animals on streets [3]. The convolution neural network (CNN) model called Single Shot Multibox Detection makes the detection process significantly faster [4]. A convolutional neural network(CNN) is a type of neural network to work on image data. The CNN architecture has many layers in it. There is a layer for input and output called the input layer and output layer, respectively. Apart from these, there are many intermediate layers such as the CLs(convolutional layers), PLs(Pooling layers), and (fully connected layers). The input layer has the image as input. The image can be 2D or 3D, depending on whether the image is greyscale or color. As the name CNN indicates that the convolution layer is an essential and fundamental part of CNN.[5]

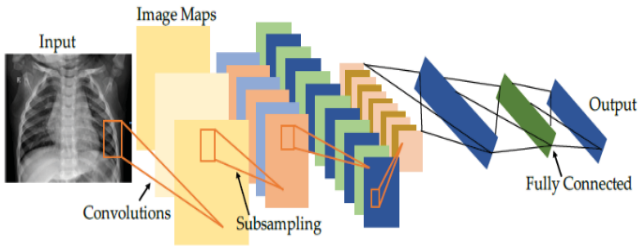


Fig. 1. Architectural overview of CNN

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

For creating an accurate neural network, the following steps are as follows:

- Image Capturing
- Image Processing
- Image Segmentation
- Feature Extraction
- Classifying the image - associating the class to which the image belongs

In recent times, the YOLO model has become famous for object detection. It is different from other techniques as it uses a single CNN that predicts the different bounding boxes around the object in an image and finds various probabilities for associating different classes to different bounding boxes. Consequently, its presentation can be streamlined from start to finish straightforwardly and therefore processing the input information rapidly. Deep Learning is one of the algorithms that has proved its importance again and again in the area of object detection, but YOLO improves inference speed for the detection of real-life objects and it makes the object detection process faster with tremendous accuracy and real-time robustness and more practical to use in real-life scenarios[6].

Moving forward to the second part of the problem i.e object tracking, which in itself is a challenging task in real life as well as during video analysis where we have to keep track of an object in different subsequent frames by creating a bounding box around it and assign a unique identity to the object across all the frames. It is easier when we have only one object in the frame at a time, but the problem becomes complex when we have multiple objects at the same instant in the frame. In the Multiple Object Tracking(MOT) problem, we try to keep track of multiple objects in the different frames in an image sequence. MOT detects all the objects in the frame and assigns a unique identity to each object and uses these identities to keep track of different objects in the frame over time [7]. The object tracking system has a flow in which they perform the task of locating various objects in an imaging sequence which can be either real-time or recorded video. Firstly an object is detected and then the object's movement is tracked in subsequent frames[8].

In this project, we implement the object Tracking System to detect multiple objects in a zone within a frame of image sequences and keep the total object's count in the frame and the total object in a particular zone specified [9]. For tracking, we use the Extension of Simple Online and Realtime Tracking (SORT) technique combining with Deep Association Metric called DeepSORT algorithm. We use DeepSORT with YOLO to detect and track multiple objects based on various attributes of the objects such as the size of the object, velocity of object, shape and physical appearance of the object etc[10]. Kalman Filters are used when there is any occlusion in the frame. Kalman Filters Works on constant velocity and linear observation and estimates the object's path when that object is not visible due to blockage like fog, smoke, flying birds etc. In object detection and tracking systems, DeepSORT is one of the essential algorithms which is very fast and accurate [11].

III. PROPOSED WORK

In This section, we discuss the proposed methodology and complete workflow of our project in detail. The projects consist of three phases, i.e., detecting the object, tracking the object and counting the object in the zone. Each of these phases is explained below.

A. Phase 1: Object Detection

This phase is concerned with the detection of objects. For the detection of objects, we used YOLOv4 [12] with pre-trained weights. YOLOv4 is an extension to YOLOv3, with improved speed and accuracy. All of the YOLO models are built and used for the detection of objects. Object detection models are trained to classify the objects in an image. When the object classes are detected, they are enclosed in the bounding boxes, and they are classified. Object detection models are generally trained and evaluated on the COCO (Common Objects in Context) dataset [13], containing a wide range of 80 object classes. The main reasons for using YOLOv4 in our project are:

- For our project, real-time object detection is essential. The advantage of real-time detection models is that they are easy to wield by all developers.
- YOLOv4 makes it a priority for real-time object detection. The authors' intended for computer vision engineers and developers to use their YOLOv4 framework in custom domains quite easily.

B. Phase 2: Tracking of Objects

For tracking objects, we used the Deep SORT algorithm. It is an extension of the Simple Real-Time Tracker (SORT) algorithm. The Deep SORT algorithm uses the Kalman filters, and it is a crucial component to estimate the tracks of existing objects in the current frame. It has states that contain eight variables; (u, u', v, v', a, a', h, h'), where:

- a – aspect ratio
- (u,v) – centers of the bounding boxes
- h – the height of the image
- other variables are respective their velocities

The DeepSORT uses the standard and straightforward form of Kalman filters that use constant velocities and linear observations.

Let’s walk through the working of DeepSORT algorithm:

- When every new frame comes, the position of each and every track is estimated on the basis of its previous locations. Only the spatial information is used for track estimation.
- Then we use the appearance feature vector to describe all the features of a given image.
- This vector obtains the appearance information of detections by extracting features and tracking images from previous frames.
- This vector is a convolution neural network trained on a large re-identification dataset.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

- It extracts features such that the features with different identities are at a considerable distance apart, and the features with the same identity are close together.
- The new detection results are associated with each coming frames' existing tracks using the appearance feature vector and the tracks' predicted position.
- Detections with low confidence value than the threshold are filtered out. New detections have to pass this threshold to become the candidates of data association.

C. *Phase 3: Zonal Counting*

A zone is made using by setting the coordinates in the frame. We used the OpenCV library for visualizing the zone into the frame. We count the number of objects with the unique ID for calculating the number of objects in the zone. Then a limit is set for the zone. Initially, the zone is of green color, and if the number of objects exceeds the limit, then the alarm is raised, and the color of the zone changes to red.

Tracking works in such a way that when it detects a particular object it assigns a specific id number to that particular object and afterward this unique id becomes the only source for tacking that particular object. By assigning the unique IDs, it makes it easier for the tracker to keep track of objects throughout the video or imaging sequence. The assigned id for the different objects can be seen in Fig. 2.

During the tracking Process, there may occur a scenario when a particular object is occluded by another object which may or may not be the point of study and can make the object in the study disappear from the view entirely or partially. By applying the Kalman filters, this problem of occlusion is resolved. One example of occlusion occurrence and continuously tracking is shown in fig. 3.

One of the critical applications of this object detection and tracking system in a particular zone is to prevent the spread of coronavirus by putting this system in the covid-19 hotspot areas and creating multiple zones, and raising the alarm when the number of people in the zone becomes greater than the threshold value. Fig. 4 and 5 show that when the number of people in the zone is less than 5 (threshold value), the zone becomes green, and when the number of people is greater than or equal to 5, then it becomes red.

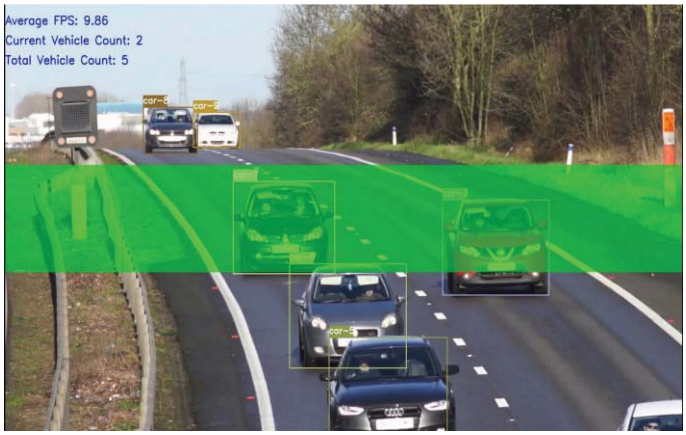


Fig. 2. Detected objects assigned with the unique ID

Fig. 3. Occlusions

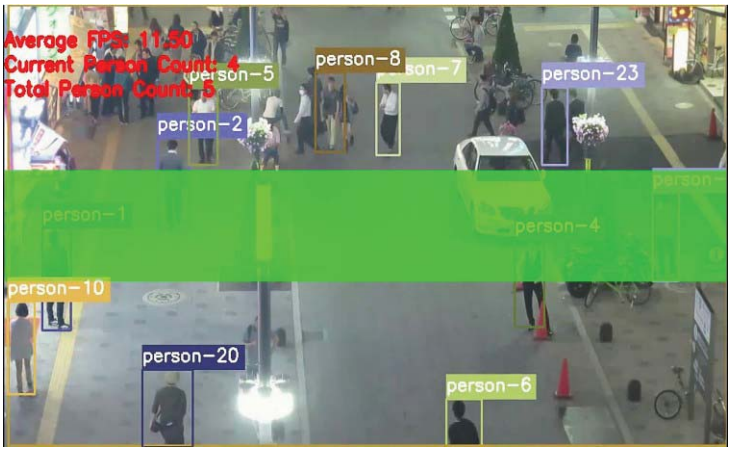


Fig. 4. The zone becomes green when the number of objects is less than 5

Fig. 5. The zone becomes red when the number of objects is greater than or equal to 5

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

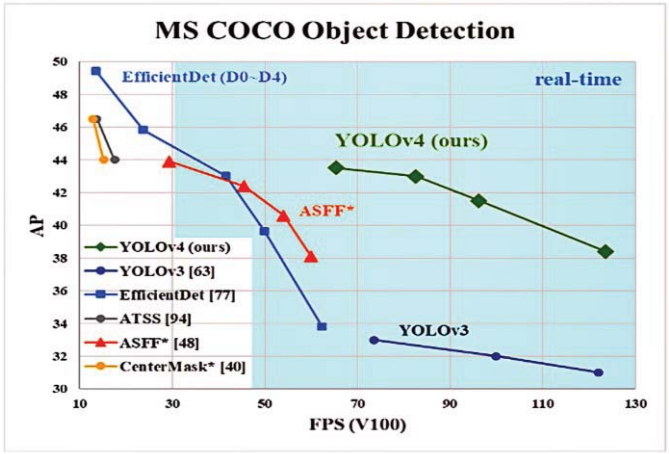


Fig. 6. YOLOv4 comparison with YOLOv3

IV. Result Analysis

We used YOLOv4 for object detection; therefore, the image classifier's accuracy is the same as that of YOLOv4. YOLOv4 shows a 10 percent and 12 percent increase in AP and FPS, respectively, compared to YOLOv3. Fig. 6, compares the YOLOv4 with YOLOv3.

YOLO detectors are best for object detection, which provides higher accuracy with higher fps. Table 1, compares YOLO models' performance with other state of are models on the real-time COCO benchmark dataset [14].

We used the MOT17 dataset [15] for evaluating our model. This is a vast dataset containing **17757 frames, 2355 tracks and 564228 bounding boxes**. The dataset is formulated by extending version of MOT16 [16] and MOT15 datasets [17], which has various image sequences including sequences with moving cameras shot from a bus, pedestrian scenes at night, people walking in large square etc.

TABLE I. Performance on COCO Benchmark Dataset

Model	mAP	FPS
CenterNet HarDNet-85	43.6	45
SSD512-HarDNet85	35.1	39
TTFNet	35.1	54.4
YOLOv3-418	31.0	34.0
YOLOv3-608	33.0	20.0
YOLOv4-512	43	83
YOLOv4-608	43.5	62

Evaluating Multi-Object Tracking causes some problems as:

- Datasets are challenging to establish.
- Multiple Object Tracking (MOT) algorithms are dependent on the detector, and the researchers do not always agree on the protocol to follow.
- Whether to evaluate the tracker with the same detector for all or to evaluate the whole system (detector plus tracker) together?
- Although some performance measures such as MOTA (Multiple Object Tracking Accuracy) are commonly used, some researchers still do not adhere to this metric.

We have used MOTA (Multiple Object Tracking Accuracy) for evaluating our model. The MOTA is the most used summary metric for MOT. It is defined as:

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDS_t}{\sum_t GT_t}$$

where,

- FP_t = number of false positives (ghost trajectories)
- IDS_t = number of identity switches at time t
- FN_t = number of false negatives (missed targets)
- GT_t = ground truth

Note: A target is considered missed if the IoU with the ground truth is less than the given threshold.

We evaluated our model with the MOT17 dataset on Jetson Xavier NX using py-motmetrics. When using public detections from the MOT17 dataset, the MOTA scores are close to state-of-the-art trackers. Tracking speed can reach up to 31 FPS (frames per second) depending on the object density. On a desktop CPU/GPU, FPS should be even higher. Even though the tracker runs much faster, it is still highly accurate. Lightweight detectors such as YOLOv4-tiny can potentially be used to obtain higher fps up to 44(fps).

TABLE II. MOTA Scores

Sequence	Avg. Density	MOTA	FPS
MOT17-03	69.8	61.03%	12.61
MOT17-04	45.3	60.17%	19.43
MOT17-13	15.5	58.67%	30.84

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

V. CONCLUSION

With this, we proceed towards the end of the paper. To summarize what we have done in this paper, we would again like to reiterate the problem which we are catering to, which is object detection and tracking in a particular area where we first detect the object via the YOLO model and identify the same. After identifying all the objects in a given frame, we track them down till they are in the frame. For better and enhanced results, we have used Kalman Filters, and an earlier SORT algorithm was used, which stands for Simple Online Realtime Tracking(SORT), but now we have used DeepSORT, which is a SORT algorithm with a Deep association metric. Lastly, we used the MOT-A score and were able to produce around 60% accuracy during detection. This model can have multipurpose uses which tracking of cars can have various applications such as counting how many people are there in the parking of an area, secondly, it can be used to check the demographic stats and see as to how many people are transmigrating from one place to another and through this data, govt. Can provide better incentives and services to such people, and the government can also demarcate the tourist spots and tighten the security. Lastly, it can be used to keep check in case of any theft in a particular area, and we can alert the various authorities involved so that the case can be solved more efficiently and smoothly. It can be used to keep a check on traffic rule violations such as lane-changing, overspeeding and rash driving. Also, it can be used to reduce traffic jams resolving time.

Because of some shortcomings like due to bad weather, unsuitable lighting conditions during night and less availability of dataset, we were not fetching great results, but still, for further scope, we can improve the accuracy of the model by training the model on a larger dataset and hence increase the accuracy of the model. Also, for future references, we can store this data for data analytics and use it to equip the administration with better results for finding a way out in the time of crises.

REFERENCES

[1] Mahalingam, T. and Subramoniam, M. (2020), "A robust single and multiple moving object detection, tracking and classification", Applied Computing and Informatics, Vol. ahead-of-print No. ahead-of-print, 2017.

[2] Guodong, C., Xia, Z., Sun, R., Wang, Z. and Sun, L. (2013), "A learning algorithm for model□based object detection", Sensor Review, Vol. 33 No. 1, pp. 25-39, 2013.

[3] S. Mane and S. Mangale, "Moving object detection and tracking using convolutional neural networks," in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018.

[4] S. Kanimozhi, G. Gayathri, and T. Mala, "Multiple Real-time object identification using Single shot Multi-Box detection," in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019.

[5] T. Rahman *et al.*, "Transfer learning with deep Convolutional Neural Network (CNN) for pneumonia detection using chest X-ray," *Appl. Sci. (Basel)*, vol. 10, no. 9, p. 3233, 2020.

[6] H.-W. Luo, C.-S. Zhang, F.-C. Pan, and X.-M. Ju, "Contextual-YOLOV3: Implement better small object detection based deep learning," in 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019.

[7] H. Karunasekera, H. Wang, and H. Zhang, "Multiple object tracking with attention to appearance, structure, motion and size," IEEE Access, vol. 7, pp. 104423–104434, 2019.

[8] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, "Learning people detectors for tracking in crowded scenes," in 2013 IEEE International Conference on Computer Vision, 2013.

[9] Y. Xu, X. Liu, L. Qin, and S. C. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," 31st AAAI Conf. Artif. Intell. AAAI 2017, pp. 4299–4305, 2017.

[10] X. Hou, Y. Wang, and L.-P. Chau, "Vehicle tracking using deep SORT with low confidence track filtering," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019.

[11] N. Wojke, A. Bewley, and D. Paulus, "Simple online and real-time tracking with a deep association metric," in 2017 IEEE International Conference on Image Processing (ICIP), 2017.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv [cs.CV], 2020.

[13] "COCO - common objects in context," *Cocodataset.org*. [Online]. Available: https://cocodataset.org/. [Accessed: 10-October-2020].

[14] "Papers with code - COCO benchmark (real-time object detection)," *Paperswithcode.com*. [Online]. Available: https://paperswithcode.com/sota/real-time-object-detection-on-coco. [Accessed: 12-October-2020].

[15] "MOT Challenge - Data," Motchallenge.net. [Online]. Available: https://motchallenge.net/data/MOT17/. [Accessed: 13-October-2020].

[16] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv [cs.CV]*, 2016.

[17] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv [cs.CV]*, 2015.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:26:26 UTC from IEEE Xplore. Restrictions apply.

Computational Methods and Function Theory
https://doi.org/10.1007/s40315-021-00386-5

On Certain Generalizations of $\mathcal{S}^*(\psi)$

Kamaljeet Gangania[1] · **S. Sivaprasad Kumar**[1]

Received: 10 November 2020 / Revised: 5 December 2020 / Accepted: 4 January 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
In this paper, we consider functions analytic in the unit disk that are subordinate to functions of the same type that are defined by certain differential subordinations. We prove several sharp majorization theorems and a product theorem.

Keywords Subordination · Majorization · Ma–Minda classes · Radius problem

Mathematics Subject Classification 30C45 · 30C50 · 30C80

1 Introduction

Let $\mathcal{A}$ denote the class of functions of the form $f(z) = z + \sum_{k=2}^{\infty} a_k z^k$ which are analytic in the open unit disk $\mathbb{D} := \{z : |z| < 1\}$. Using subordination, Ma–Minda [5] introduced the unified class of starlike and convex functions defined as follows:

$$\begin{aligned} \mathcal{S}^*(\psi) &:= \left\{ f \in \mathcal{A} : \frac{zf'(z)}{f(z)} \prec \psi(z) \right\}, \\ \mathcal{C}(\psi) &:= \left\{ f \in \mathcal{A} : 1 + \frac{zf''(z)}{f'(z)} \prec \psi(z) \right\}, \end{aligned} \tag{1}$$

where ψ is a Ma–Minda function, which is analytic and univalent with $\mathrm{Re}\,\psi(z) > 0$, $\psi'(0) > 0$, $\psi(0) = 1$ and $\psi(\mathbb{D})$ is symmetric about the real axis. Note that $\psi \in$

Communicated by Dmitri Khavinson.

K. Gangania thanks to University Grant Commission, New-Delhi, India for providing Junior Research Fellowship under UGC-Ref. No.:1051/(CSIR-UGC NET JUNE 2017).

✉ Kamaljeet Gangania
gangania.m1991@gmail.com

S. Sivaprasad Kumar
spkumar@dce.ac.in

[1] Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India

$\mathcal{P}$, the class of normalized Carathéodory functions. The class $\mathcal{S}^*(\psi)$ unifies various subclasses of starlike functions, which can be obtained by an appropriate choice of ψ. Ma–Minda discussed many properties of the class $\mathcal{S}^*(\psi)$. Now recall the following definition and a result due to MacGregor [6]:

Definition 1 [6] Let f and g be analytic in $\mathbb{D}$. A function $g(z)$ is said to be majorized by $f(z)$, denoted by $g << f$, if there exists an analytic function $\Phi(z)$ in $\mathbb{D}$ satisfying $|\Phi(z)| \leq 1$ and $g(z) = \Phi(z) f(z)$ for all $z \in \mathbb{D}$.

Theorem 1 [6] *Let g be majorized by f in $\mathbb{D}$ and $g(0) = 0$. If $f(z)$ is univalent in $\mathbb{D}$, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq 2 - \sqrt{3}$. The constant $2 - \sqrt{3}$ is sharp.*

Recently Tang and Deng [11] obtained majorization results for $\mathcal{S}^*(\psi)$ for some specific choices of ψ. Motivated by this, in Sect. 2, we devise a general approach to handle the same for $\mathcal{C}(\psi)$, which is precisely stated as: if $g \in \mathcal{A}$, $f \in \mathcal{C}(\psi)$ and g is majorized by f in $\mathbb{D}$, then we find the largest radius $r_\psi \leq 1$ such that $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_\psi$. Several other results in this direction are also obtained. In Sect. 3, we consider the radius problem posed by Obradović and Ponnusamy [9] namely: Let $g \in \mathcal{S}^*(\psi_1)$ and $h \in \mathcal{S}^*(\psi_2)$, then find the largest radius $r_0 \leq 1$ such that the function $F(z) = (g(z)h(z))/z$ belongs to a certain well-known class of starlike functions in $|z| < r_0$. As a special case, we also obtain a result of Obradović and Ponnusamy [9]. Throughout this paper we shall assume that the function ϕ has real coefficients in it's power series expansion.

2 Majorization

Let us consider the analytic function $\psi(z) := 1 + B_1 z + B_2 z^2 + \cdots$. Here $B_1 = \psi'(0)$, the coefficient of z, plays a major role in deciding the orientation of the function ψ. Thus whether ψ is positively or negatively oriented depends on whether B_1 is positive or negative. Ma–Minda only considered the case $\psi'(0) > 0$, as in the case $\psi'(0) < 0$, many postulates for the class $\mathcal{S}^*(\psi)$ are not longer valid. With this perspective, we begin with the following:

Theorem 2 *Let $\mathrm{Re}\phi(z) > 0$ and ϕ be convex in $\mathbb{D}$ with $\phi(0) = 1$. Suppose that ψ satisfies $m_r := \min_{|z|=r} |\psi(z)|$ and also the differential equation*

$$\psi(z) + \frac{z\psi'(z)}{\psi(z)} = \phi(z). \tag{2}$$

Let $g \in \mathcal{A}$ and $f \in \mathcal{C}(\phi)$. If g is majorized by f in $\mathbb{D}$, then

$$|g'(z)| \leq |f'(z)| \quad \text{in } |z| \leq r_\psi, \tag{3}$$

where r_ψ is the least positive root of the equation

$$(1 - r^2)m_r - 2r = 0. \tag{4}$$

The result is sharp for the case $m_r = \psi(-r)$.

Proof Let us define $p(z) := zf'(z)/f(z)$. Since $f \in \mathcal{C}(\phi)$, therefore we have $1 + zf''(z)/f'(z) \prec \phi(z)$, which can be equivalently written as

$$p(z) + \frac{zp'(z)}{p(z)} = 1 + \frac{zf''(z)}{f'(z)} \prec \phi(z). \tag{5}$$

Since $\mathrm{Re}\phi(z) > 0$ and ϕ is convex in $\mathbb{D}$, therefore using [8, Thm. 3.2d, p. 86] the solution ψ of the differential Eq. (2) is analytic in $\mathbb{D}$ with $\mathrm{Re}\psi(z) > 0$ and has the following integral form given by

$$\psi(z) := h(z)\left(\int_0^z \frac{h(t)}{t}dt\right)^{-1},$$

where

$$h(z) = z\exp\int_0^z \frac{\phi(t)-1}{t}dt.$$

Since $\mathrm{Re}\psi(z) > 0$ and p satisfies the subordination (5), therefore using [8, Lem. 3.2e, p. 89] we conclude that ψ is univalent and $p \prec \psi$, where ψ is the best dominant. Thus we have obtained that $f \in \mathcal{C}(\phi)$ implies $zf'(z)/f(z) \prec \psi(z)$ and ψ is the best dominant (which is important for the sharpness of the result), which is a univalent Carathéodory function. Now as $g \in \mathcal{A}$ and $f \in \mathcal{C}(\phi)$, therefore we obtain the following equality

$$\frac{f(z)}{f'(z)} = \frac{z}{\psi(\omega(z))}, \quad z \in \mathbb{D},$$

where ω is a Schwarz function. Hence, using $\min_{|z|=r}|\psi(\omega(z))| \geq \min_{|z|=r}|\psi(z)|$ and the hypothesis $\min\limits_{|z|=r}|\psi(z)| = m_r$, we obtain that

$$\left|\frac{f(z)}{f'(z)}\right| \leq \frac{r}{m_r}, \quad 0 < r < 1. \tag{6}$$

Now if g is majorized by f, then by definition, we have $g(z) = \Psi(z)f(z)$, where Ψ is analytic and satisfies $|\Psi(z)| \leq 1$ in $\mathbb{D}$ such that $g'(z) = \Psi(z)f'(z) + \Psi'(z)f(z)$. Thus using (6) together with the following Schwarz-Pick inequality

$$|\Psi'(z)| \leq \frac{1-|\Psi(z)|^2}{1-|z|^2},$$

we obtain

$$|g'(z)| \leq |f'(z)|\left(|\Psi(z)| + \frac{1-|\Psi(z)|^2}{1-r^2}\frac{r}{m_r}\right) = |f'(z)|h(\beta, r), \tag{7}$$

where $|\Psi(z)| := \beta$ and

$$h(\beta, r) = \beta + \frac{1-\beta^2}{1-r^2}\frac{r}{m_r}.$$

Thus to arrive at (3), it suffices to showing that $h(\beta, r) \leq 1$, which is equivalent to show that

$$k(\beta, r) := (1-r^2)m_r - (\beta+1)r \geq 0. \tag{8}$$

Since $\frac{\partial}{\partial\beta}k(\beta, r) = -r < 0$, (8) holds whenever

$$k(r) := \min_{\beta} k(\beta, r) = k(1, r) \geq 0.$$

Note that $k(r)$ is a continuous function of r and further $k(0) = m_0 = \psi(0) = 1 > 0$ and $k(1) < 0$. Thus there exists a point $r_\psi \in (0, 1)$ such that $k(r) \geq 0$ for all $r \in [0, r_\psi]$, where r_ψ is the least positive root of (4).

Proof of sharpness: Now let $m_r = \psi(-r)$. Choose $f(z) \in \mathcal{C}(\phi)$ such that $zf'(z)/f(z) = \psi(-z)$ and $\Psi(z) = (z+\alpha)/(1+\alpha z)$, where $-1 \leq \alpha \leq 1$. Let r_0 be the second consecutive positive root (if it exists) of the equation (4), otherwise choose $r_0 = 1$. We show that for each $r_\psi < r < r_0$ we can choose α so that $g'(r) > f'(r) > 0$, which implies that g' is not majorized by f' outside of $|z| \leq r_\psi$. First note that

$$\frac{f(r)}{f'(r)} = \frac{r}{\psi(-r)}. \tag{9}$$

Since

$$g'(r) = f'(r)\left(\frac{r+\alpha}{1+\alpha r} + \frac{1-\alpha^2}{(1+\alpha r)^2}\frac{f(r)}{f'(r)}\right) =: f'(r)h(r, \alpha)$$

and $h(r, 1) = 1$, it suffices to show that $\partial h(r, \alpha)/\partial\alpha < 0$ at $\alpha = 1$ in order to establish that $h(r, 1-\epsilon) > 1$, and hence $g'(r) > f'(r) > 0$. But at $\alpha = 1$, we have:

$$\begin{aligned}\frac{\partial h(r, \alpha)}{\partial\alpha} &= \frac{2}{(1+r)^2}\left(\frac{1-r^2}{2} - \frac{f(r)}{f'(r)}\right)\\ &= \frac{2}{(1+r)^2}\left(\frac{1-r^2}{2} - \frac{r}{\psi(-r)}\right)\\ &< 0,\end{aligned}$$

using the equations (4), (8), (9) and the fact that $k(r) < 0$ for all $r \in (r_\psi, r_0)$. □

Remark 1 The following result was proved by MacGregor [6]: Let $g \in \mathcal{A}$ and $f \in \mathcal{C}$. If g is majorized by f in $\mathbb{D}$, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq 1/3$. The result is sharp.

In our next result, we show the application to the Janowski class [3], which covers many well-known classes. Here $\mathcal{C}[D, E] := \mathcal{C}((1 + Dz)/(1 + Ez)$.

Corollary 1 *Let f belong to $\mathcal{C}[D, E]$, where $-1 \leq E < D \leq 1$ along with $1+D/E \geq 0$ and $-1 \leq E < 0$. If g is majorized by f, then*

$$|g'(z)| \leq |f'(z)| \quad in \quad |z| \leq r_0,$$

where r_0 is the smallest positive root of the equation

$$(1 - r^2)\left({}_2F_1\left(1 - \frac{D}{E}, 1, 2; \frac{-Er}{1 - Er}\right)\right)^{-1} - 2r = 0.$$

The result is sharp.

Proof In Theorem 2, put $\phi(z) = (1 + Dz)/(1 + Ez)$. Then we have $\psi(z) := 1/q(z)$, where

$$q(z) = \begin{cases} \int_0^1 \left(\frac{1+Etz}{1+Ez}\right)^{(D-E)/E} dt, & \text{if } E \neq 0; \\ \int_0^1 e^{D(t-1)z} dt, & \text{if } E = 0, \end{cases}$$

which can be further represented in terms of confluent and Gaussian hypergeometric functions, respectively as follows:

$$q(z) = \begin{cases} {}_2F_1\left(1 - \frac{D}{E}, 1, 2; \frac{Ez}{1+Ez}\right), & \text{if } E \neq 0; \\ {}_1F_1\left(1, 2; -Dz\right), & \text{if } E = 0. \end{cases}$$

Since $1 + D/E \geq 0$ and $-1 \leq E < 0$, we have

$$\min_{|z|=r} \mathrm{Re}\psi(z) = \psi(-r) = \frac{1}{q(-r)} = \left({}_2F_1\left(1 - \frac{D}{E}, 1, 2; \frac{-Er}{1 - Er}\right)\right)^{-1}.$$

Since $\mathrm{Re}\psi(z) > 0$ and $\min_{|z|=r} \mathrm{Re}\psi(z) = \psi(-r)$, we conclude that $\min_{|z|=r} |\psi(z)| = \psi(-r)$ and hence, the result follows from Theorem 2. □

Now we have the result for the class of convex functions of order α using Corollary 1:

Corollary 2 *Let f belong to $\mathcal{C}[1 - 2\alpha, -1]$, where $0 \leq \alpha < 1$. If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the smallest positive root of the equation*

$$(1 - r^2)\left({}_2F_1\left(2(1 - \alpha), 1, 2; \frac{r}{1 + r}\right)\right)^{-1} - 2r = 0.$$

The result is sharp.

Corollary 3 *Let f belong to $\mathcal{C}[D, 0]$. If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the smallest positive root of the equation*

$$(1 - r^2)\frac{Dre^{-Dr}}{e^{-Dr} - 1} + 2r = 0.$$

The result is sharp.

Proof From the proof of Corollary 1, we obtain that $\psi(z) = Dze^{Dz}/(e^{Dz} - 1)$, when $\phi(z) = 1 + Dz$. Now with a little computation, we find that the function $l(z) = ze^z/(e^z - 1)$ is convex univalent in $\mathbb{D}$. Therefore, the function $\psi(z) = l(Dz)$ is also convex in $\mathbb{D}$ for each fixed $0 < D \leq 1$. Since ψ is also symmetric about the real axis, we conclude that $\min_{|z|=r} |\psi(z)| = \psi(-r)$. □

Theorem 3 *Let ϕ be convex in $\mathbb{D}$, with $\mathrm{Re}\phi(z) > 0$, $\phi(0) = 1$ and suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)} + z\left(\frac{zf'(z)}{f(z)}\right)' \prec \phi(z). \tag{10}$$

If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1 - r^2) \min_{|z|=r} \mathrm{Re}\psi(z) - 2r = 0,$$

where

$$\psi(z) := \frac{1}{z}\int_0^z \phi(t)dt.$$

The result is sharp for the case $\min_{|z|=r} \mathrm{Re}\psi(z) = \psi(\pm r)$.

Proof Let $p(z) = zf'(z)/f(z)$. Then the subordination (10) can equivalently be written as:

$$p(z) + zp'(z) \prec \phi(z).$$

A simple calculation show that the analytic function $\psi(z) := (1/z)\int_0^z \phi(t)dt$ satisfies

$$\psi(z) + z\psi'(z) = \phi(z).$$

Now from the Hallenbeck and Ruscheweyh result [8, Thm. 3.1b, p. 71], we have $p \prec \psi$, where ψ is the best dominant and also convex. Further, since $\mathrm{Re}\phi(z) > 0$, using the integral operator [8, Thm. 4.2a, p. 202] preserving functions with positive real part, we see that ψ is a Caratheódory function. Thus we have

$$\frac{f(z)}{zf'(z)} \prec \frac{1}{\psi(z)} \quad \text{which implies} \quad \left|\frac{f(z)}{f'(z)}\right| \leq \frac{r}{\min_{|z|=r} |\psi(z)|} = \frac{r}{\min_{|z|=r} \mathrm{Re}\psi(z)}.$$

Now proceeding as in Theorem 2 the result follows. □

Corollary 4 *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)} + z\left(\frac{zf'(z)}{f(z)}\right)' \prec \frac{1+z}{1-z}.$$

If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1-r^2)(2\log(1+r) - r) - 2r^2 = 0.$$

The result is sharp.

Corollary 5 *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)} + z\left(\frac{zf'(z)}{f(z)}\right)' \prec e^z.$$

If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1-r^2)(1-e^{-r}) - 2r^2 = 0.$$

The result is sharp.

Theorem 4 *Let ϕ be convex in $\mathbb{D}$, with $\mathrm{Re}\phi(z) > 0$, $\phi(0) = 1$ and suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)}\left(\frac{zf'(z)}{f(z)} + 2z\left(\frac{zf'(z)}{f(z)}\right)'\right) \prec \phi(z), \quad \alpha \in [0,1). \tag{11}$$

If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1-r^2)\min_{|z|=r} |\sqrt{\psi(z)}| - 2r = 0,$$

where

$$\psi(z) := \frac{1}{z}\int_0^z \phi(t)dt.$$

The result is sharp when $\min_{|z|=r} |\sqrt{\psi(z)}| = \sqrt{\psi(\pm r)}$.

Proof Let $p(z) = zf'(z)/f(z)$. Then the subordination (11) can be equivalently written as:

$$p^2(z) + 2zp(z)p'(z) \prec \phi(z),$$

which, using the change of variable $P(z) = p^2(z)$, becomes

$$P(z) + zP'(z) \prec \phi(z).$$

Now proceeding as in Theorem 3, we see that $p(z) \prec \sqrt{\psi(z)}$ and $\sqrt{\psi(z)}$ is the best dominant. Further, since $\mathrm{Re}\phi(z) > 0$, using [8, Thm. 4.2a, p. 202], we see that ψ is a Carathéodory function. Therefore,

$$|\arg\sqrt{\psi(z)}| = \frac{1}{2}|\arg\psi(z)| \leq \frac{\pi}{4},$$

which implies $\mathrm{Re}\sqrt{\psi(z)} > 0$. Thus we have

$$\frac{f(z)}{zf'(z)} \prec \frac{1}{\sqrt{\psi(z)}} \quad \text{which implies} \quad \left|\frac{f(z)}{f'(z)}\right| \leq \frac{r}{\min_{|z|=r}|\sqrt{\psi(z)}|}.$$

Now proceeding as in Theorem 2 the result follows. □

Corollary 6 *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)}\left(\frac{zf'(z)}{f(z)} + 2z\left(\frac{zf'(z)}{f(z)}\right)'\right) \prec \frac{1+(2\alpha-1)z}{1+z}.$$

If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1-r^2)\min_{|z|=r}\mathrm{Re}\sqrt{\psi(z)} - 2r = 0,$$

where

$$\psi(z) := \frac{1}{z}\int_0^z \frac{1+(2\alpha-1)t}{1+t}dt.$$

Corollary 7 *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)}\left(\frac{zf'(z)}{f(z)} + 2z\left(\frac{zf'(z)}{f(z)}\right)'\right) \prec 1+\alpha z, \quad (\alpha \in (0,1]).$$

If g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1-r^2)\sqrt{1-\beta r} - 2r = 0, \quad \text{where} \quad \beta = \frac{\alpha}{2}.$$

The result is sharp.

We state the following result without proof as it follows from Theorem 2.

Theorem 5 *Let $\psi \in \mathcal{P}$ be a univalent function such that*

$$m_r := \min_{|z|=r} |\psi(z)| = \begin{cases} \psi(-r), & \text{if } \psi'(0) > 0; \\ \psi(r), & \text{if } \psi'(0) < 0. \end{cases}$$

Let $g \in \mathcal{A}$ and $f \in \mathcal{S}^*(\psi)$. If g is majorized by f in $\mathbb{D}$, then

$$|g'(z)| \leq |f'(z)| \quad \text{in } |z| \leq r_\psi,$$

where r_ψ is the least positive root of the equation

$$(1-r^2)m_r - 2r = 0.$$

The result is sharp.

Example 1 Let us consider the analytic functions $\psi_1(z) = \sqrt{1-z}$ and $\psi_2(z) = \sqrt{1+z}$. Note that $\psi_1'(0) < 0$, $\psi_2'(0) > 0$ and for $|z| = r$,

$$m_{r_1} = \min_{|z|=r} |\psi_1(z)| = \psi_1(r) = \sqrt{1-r} = \psi_2(-r) = \min_{|z|=r} |\psi_2(z)| = m_{r_2}.$$

Now from Theorem 5, we obtain the following result:
If $g \in \mathcal{A}$, $f \in \mathcal{S}^*(\psi_i)$, where $i = 1, 2$ and g is majorized by f, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_0$, where r_0 is the least positive root of the equation

$$(1-r^2)\sqrt{1-r} - 2r = 0.$$

Interestingly, the desired radius in both the cases remain the same as $\psi_1(\mathbb{D}) = \psi_2(\mathbb{D})$, though ψ_1 and ψ_2 are oppositely oriented.

Remark 2 Taking $\alpha = 0$ or $\eta = 1$ in Corollary 8, case (ii) and (iii), respectively, we obtain the result proved by T. H. MacGregor [6], namely: *Let $g \in \mathcal{A}$ and $f \in \mathcal{S}^*$. If $g << f$ in $\mathbb{D}$, then $|g'(z)| \leq |f'(z)|$* in $|z| \leq 2 - \sqrt{3}$. The result is sharp.

Next we obtain the following majorization results for some known classes as well those introduced and studied in [1,2,7,10].

Corollary 8 *Let $g \in \mathcal{A}$ and $f \in \mathcal{S}^*(\psi)$. If $g << f$ in $\mathbb{D}$, then $|g'(z)| \leq |f'(z)|$ in $|z| \leq r_\psi$, where r_ψ is the least positive root of the equation $P(r) = 0$ and the result follows for each one of the following cases:*

(i) $P(r) = (1-r^2)\frac{1-Dr}{1-Er} - 2r$ *when* $\psi(z) = \frac{1+Dz}{1+Ez}$, *where* $-1 \le E < D \le 1$.
(ii) $P(r) = (1-r)(1-(1-2\alpha)r) - 2r$ *when* $\psi(z) = \frac{1+(1-2\alpha)z}{1-z}$, *where* $0 \le \alpha < 1$.
(iii) $P(r) = (1-r^2)\left(\frac{1-r}{1+r}\right)^{\eta} - 2r$ *when* $\psi(z) = \left(\frac{1+z}{1-z}\right)^{\eta}$, *where* $0 < \eta \le 1$.
(iv) $P(r) = (1-r^2)\left(\sqrt{2} - (\sqrt{2}-1)\sqrt{\frac{1+r}{1-2(\sqrt{2}-1)r}}\right) - 2r$ *when* $\psi(z) = \sqrt{2} - (\sqrt{2}-1)\sqrt{\frac{1-z}{1+2(\sqrt{2}-1)z}}$.
(v) $P(r) = (1-r^2)(b(1-r))^{1/a} - 2r$ *when* $\psi(z) = (b(1+z))^{1/a}$, *where* $a \ge 1$ *and* $b \ge 1/2$.
(vi) $P(r) = (1-r^2) - 2re^r$ *when* $\psi(z) = e^z$.
(vii) $P(r) = (1-r^2)(\sqrt{1+r^2} - r) - 2r$ *when* $\psi(z) = z + \sqrt{1+z^2}$.
(viii) $P(r) = (1-r^2) - r(1+e^r)$ *when* $\psi(z) = \frac{2}{1+e^{-z}}$.
(ix) $P(r) = (1-r^2)(1-\sin r) - 2r$ *when* $\psi(z) = 1 + \sin z$.

The results are sharp.

Remark 3 In Corollary 8, case (ix), we obtained the radius $r_\psi \approx 0.312478$ which improves the majorization radius $r_s \approx 0.309757$ obtained in [12].

Let $\psi(z) = 1 + z/(1-\alpha z^2)$, $0 \le \alpha < 1$, introduced and studied by Kargar et al. [4]. Clearly $\psi \in \mathcal{P}$ only when $\alpha = 0$ and hence Theorem 5 holds when $\psi(z) = 1+z$. Moreover, for some $r > 0$, the quantity $z/\psi(z)$ does not exist for all $|z| = r$. In view of the same, the result proved by Tang and Deng [11], needs correction and the corrected version is stated in the following corollaries:

Corollary 9 *Let* $g \in \mathcal{A}$ *and* $f \in \mathcal{S}^*(1+\beta z)$, $0 < \beta \le 1$. *If* $g << f$ *in* $\mathbb{D}$, *then*

$$|g'(z)| \le |f'(z)| \quad in \quad |z| \le r_\beta,$$

where r_β *is the least positive root of the equation*

$$(1-r^2)(1-\beta r) - 2r = 0.$$

The result is sharp.

Now we obtain the result related to $\mathcal{BS}(\alpha)$, the class of Booth lemniscate starlike functions, when $\alpha \ne 0$.

Corollary 10 *Let* $0 < \alpha < 1$ *and* r_α *be the unique root of the equation*

$$\alpha r^2 + r - 1 = 0. \tag{12}$$

Let $g \in \mathcal{A}$ *and* $g << f$ *in* $\mathbb{D}$, *where* $f \in \mathcal{BS}(\alpha)$. *Then*

$$|g'(z)| \le |f'(z)| \quad in \quad |z| \le r_{B(\alpha)} := \min\{r_\alpha, r_0\},$$

where r_0 is the least positive root of the equation

$$(1-r^2)\left(1-\frac{r}{1-\alpha r^2}\right)-2r=0.$$

The result is sharp.

Proof Observe that $\operatorname{Re}\left(1+z/(1-\alpha z^2)\right)>0$ for $|z|<r_\alpha$, where r_α is the unique root of (12). Thus the inequality in (6) holds for $|z|=r<r_\alpha$ and the result follows at once. □

3 Product of Starlike Functions

Assume that ψ_1 and ψ_2 belong to $\mathcal{P}$ and satisfy the following conditions for $|z|=r$ and $i=1,2$

$$\max_{|z|=r}\operatorname{Re}\psi_i(z)=\psi_i(r)\quad\text{and}\quad\min_{|z|=r}\operatorname{Re}\psi_i(z)=\psi_i(-r). \tag{13}$$

Motivated by Obradović and Ponnusamy [9], in this section, we consider a radius problem which generalizes their result and also establishs a similar result for the Uralegaddi class $\mathcal{M}(\beta):=\{f\in\mathcal{A}:\operatorname{Re}(zf'(z)/f(z))<\beta,\ \beta>1\}$.

Theorem 6 *Let $g\in\mathcal{S}^*(\psi_1)$ and $h\in\mathcal{S}^*(\psi_2)$, where ψ_i satisfy the first condition in (13). Then the function F defined by*

$$F(z)=\frac{g(z)h(z)}{z} \tag{14}$$

belongs to $\mathcal{M}(\beta)$ in the disk $|z|<r_\beta=\min\{1,r_0(\beta)\}$, where $r_0(\beta)$ is the least positive root of the equation

$$\psi_1(r)+\psi_2(r)-1-\beta=0. \tag{15}$$

The radius r_β is sharp.

Proof Let $g\in\mathcal{S}^*(\psi_1)$ and $h\in\mathcal{S}^*(\psi_2)$. Then in view of (13) and the subordination principle, it follows that

$$\operatorname{Re}\frac{zg'(z)}{g(z)}\le\psi_1(r)\quad\text{and}\quad\operatorname{Re}\frac{zh'(z)}{h(z)}\le\psi_2(r)$$

in $|z|\le r<1$. Since

$$\frac{zF'(z)}{F(z)}=\frac{zg'(z)}{g(z)}+\frac{zh'(z)}{h(z)}-1,$$

we have for $|z| = r$,

$$\operatorname{Re}\frac{zF'(z)}{F(z)} \leq \psi_1(r) + \psi_2(r) - 1 \leq \beta,$$

whenever $r \leq \min\{1, r_0(\beta)\}$, where $r_0(\beta)$ is the least positive root of the equation (15). The sharpness follows by considering the functions

$$g(z) = z\exp\int_0^z \frac{\psi_1(t)-1}{t}dt \quad \text{and} \quad h(z) = z\exp\int_0^z \frac{\psi_2(t)-1}{t}dt.$$

□

Corollary 11 *Let $g \in \mathcal{S}^*(\gamma)$ and $h \in \mathcal{S}^*(\tau)$. Then the function F defined in (14) belongs to $\mathcal{M}(\beta)$ in the disk $|z| < \min\{1, r_0(\beta)\}$, where*

$$r_0(\beta) = \frac{\beta - 1}{3 + \beta - 2(\gamma + \tau)}.$$

The proof of the following result is quite similar to the proof of Theorem 6, so is omitted here.

Theorem 7 *Let $g \in \mathcal{S}^*(\psi_1)$ and $h \in \mathcal{S}^*(\psi_2)$, where ψ_i satisfies the second condition in (13). Then the function F defined in (14) is starlike of order γ in the disk $|z| < r_\gamma$, where r_γ is the least positive root of the equation*

$$\psi_1(-r) + \psi_2(-r) - 1 - \gamma = 0.$$

The radius r_γ is sharp.

Now using Theorem 7, we obtain the following result proved by Obradović and Ponnusamy [9].

Remark 4 Let $g \in \mathcal{S}^*(\gamma)$ and $h \in \mathcal{S}^*(\tau)$. Then the function F defined in (14) is starlike of order γ_0 in the disk

$$|z| < \frac{1 - \gamma_0}{\gamma_0 + 3 - 2(\gamma + \tau)}.$$

Remark 5 Note that the identity function $z \in \mathcal{S}^*(\psi)$. Thus if we choose $g(z) = z$ (or $h(z) = z$) in (14), then the problem reduces to obtaining the $\mathcal{M}(\beta)$-radius (or $\mathcal{S}^*(\gamma)$-radius) of the class $\mathcal{S}^*(\psi_2)$ (or $\mathcal{S}^*(\psi_1)$). It is also evident that the conditions given in (13) establish the inclusion relations $\mathcal{S}^*(\psi) \subseteq \mathcal{M}(\psi(1))$ and $\mathcal{S}^*(\psi) \subseteq \mathcal{S}^*(\psi(-1))$, respectively.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Cho, N.E., Kumar, V., Kumar, S.S., Ravichandran, V.: Radius problems for starlike functions associated with the sine function. Bull. Iran. Math. Soc. **45**, 213–232 (2019)
2. Goel, P., Kumar, S.S.: Certain Class of Starlike Functions Associated with Modified Sigmoid Function. Bull. Malays. Math. Sci. Soc. (2019). https://doi.org/10.1007/s40840-019-00784-y
3. Janowski, W.: Extremal problems for a family of functions with positive real part and for some related families. Ann. Polon. Math. **23**, 159–177 (1970)
4. Kargar, R., Ebadian, A., Sokół, J.: On Booth lemniscate and starlike functions. Anal. Math. Phys. **9**(1), 143–154 (2019)
5. Ma, W. C., Minda, D.: A unified treatment of some special classes of univalent functions. In: Proceedings of the Conference on Complex Analysis (Tianjin: 157–169. Proc. Lecture Notes Anal., I Int. Press, Cambridge, MA, Conf (1992)
6. MacGregor, T.H.: Majorization by univalent functions. Duke Math. J. **34**, 95–102 (1967)
7. Mendiratta, R., Nagpal, S., Ravichandran, V.: On a subclass of strongly starlike functions associated with exponential function. Bull. Malays. Math. Sci. Soc. **38**, 365–386 (2015)
8. Miller, S.S., Mocanu, P.T.: Differential subordinations. In: Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker Inc, New York (2000)
9. Obradović, M., Ponnusamy, S.: Product of univalent functions. Math. Comput. Model. **57**, 793–799 (2013)
10. Raina, R.K., Sokół, J.: Some properties related to a certain class of starlike functions. C. R. Math. Acad. Sci. Paris. **353**, 973–978 (2015)
11. Tang, H., Deng, G.: Majorization problems for some subclasses of starlike functions. J. Math. Res. Appl. **39**, 153–159 (2019)
12. Tang, H., Srivastava, H.M., Li, Sh., Deng, G.T.: Majorization results for subclasses of starlike functions based on the sine and cosine functions. Bull. Iran. Math. Soc. (2019). https://doi.org/10.1007/s41980-019-00262-y

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

arXiv:2104.08533v1 [math.CV] 17 Apr 2021

ON OBLIQUE DOMAINS OF JANOWSKI FUNCTION

S. SIVAPRASAD KUMAR AND POOJA YADAV

Abstract. We investigate certain properties of tilted (oblique) domains, associated with the Janowski function $(1+Az)/(1+Bz)$, where $A, B \in \mathbb{C}$ with $A \neq B$ and $|B| \leq 1$. We find several bounds for these oblique domains. Further, we extend the various known argument, radius, and subordination results involving Janowski functions with complex parameters.

1. Introduction

Let $\mathcal{H}(\mathbb{D})$ denote the class of analytic functions defined on the open unit disk $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$. Assume $\mathcal{H}[a,n] := \{f \in \mathcal{H}(\mathbb{D}) : f(z) = a + a_n z^n + a_{n+1} z^{n+1} + \cdots\}$, where $n = 1, 2, \cdots$ with $a \in \mathbb{C}$ and $\mathcal{H}_1 := \mathcal{H}[1,1]$. Let $\mathcal{A}_n := \{f \in \mathcal{H}(\mathbb{D}) : f(z) = z + a_{n+1} z^{n+1} + a_{n+2} z^{n+2} + \cdots\}$ and $\mathcal{A} := \mathcal{A}_1$. A subclass of $\mathcal{A}$ consisting of all univalent functions is denoted by $\mathcal{S}$. We say, f is subordinate to g, written as $f \prec g$, if $f(z) = g(\omega(z))$ where f, g are analytic functions and $\omega(z)$ is a Schwarz function. Moreover, if g is univalent, then $f \prec g$ if and only if $f(\mathbb{D}) \subseteq g(\mathbb{D})$ and $f(0) = g(0)$. For $-\pi/2 < \lambda < \pi/2$, Wang in [24] introduced the tilted Carathéodory class by angle λ as:

$$\mathcal{P}_\lambda := \left\{ p \in \mathcal{H}_1 : e^{i\lambda} p(z) \prec \frac{1+z}{1-z} \right\}. \tag{1.1}$$

Here $\mathcal{P}_0 = \mathcal{P}$, the well known Carathéodory class. A Janowski function is a bilinear transformation, which was first investigated by Janowski in [4]. He introduced the class $\mathcal{P}(A,B)$, where $-1 \leq B < A \leq 1$, which comprises of the set of all p in $\mathcal{H}_1$ such that

$$p(z) \prec \frac{1+Az}{1+Bz}.$$

$\mathcal{S}^*(A,B)$, the class of Janowski starlike functions and $\mathcal{C}(A,B)$, the class of Janowski convex functions, consist of the set of all $f \in \mathcal{S}$ satisfying

$$\frac{zf'(z)}{f(z)} \prec \frac{1+Az}{1+Bz} \qquad \text{and} \qquad 1 + \frac{zf''(z)}{f'(z)} \prec \frac{1+Az}{1+Bz},$$

respectively. For $0 \leq \alpha < 1$, $\mathcal{S}^*(1-2\alpha, -1) = \mathcal{S}^*(\alpha)$, $\mathcal{C}(1-2\alpha,-1) = \mathcal{C}(\alpha)$, the classes of starlike and convex functions of order α respectively and $\mathcal{S}^*(1, 2\alpha - 1) = \mathcal{RS}^*(\alpha)$, the class of starlike functions of reciprocal order α, which can also be stated as:

$$\mathcal{RS}^*(\alpha) = \left\{ f \in \mathcal{S} : \operatorname{Re} \frac{f(z)}{zf'(z)} > \alpha \right\}.$$

Clearly, $\mathcal{RS}^*(\alpha) \subset \mathcal{RS}^*(0) = \mathcal{S}^*(0) = \mathcal{S}^*$. And for $\beta > 1$, $\mathcal{S}^*(1-2\beta, -1) = \mathbb{M}(\beta)$, the class of Uralegaddi functions, which is stated as:

$$\mathbb{M}(\beta) = \left\{ f \in \mathcal{S} : \operatorname{Re} \frac{zf'(z)}{f(z)} < \beta \right\}.$$

2010 *Mathematics Subject Classification.* 30C45,30C50, 30C80.

Key words and phrases. Analytic, Subordination, Janowski function, Argument.

The second author is supported by The Council of Scientific and Industrial Research(CSIR). Ref.No.:08/133(0030)/2019-EMR-I.

For $-\pi/2 < \alpha < \pi/2$ and $0 \leq \beta < 1$, $\mathcal{S}^*(e^{-i\alpha}\left(e^{-i\alpha} - 2\beta\cos\alpha\right), -1) = \mathcal{S}^*_\alpha(\beta)$, the class of $\alpha-$spirallike of order β. From the above classes the range of A in $\mathcal{P}_\lambda$, $\mathbb{M}(\beta)$ and $\mathcal{S}^*_\alpha(\beta)$ is not as given by Janowski. This motivates us to extend and study Janowski function with complex parameters. In the past many authors have studied $(1+Az)/(1+Bz)$, where $A, B \in \mathbb{C}$ with $|B| \leq 1$ and $A \neq B$. For instance see [1, 5, 6]. In the present paper we investigate the class $\mathcal{P}(A, B, \alpha)$, where $A, B \in \mathbb{C}$ with $|B| \leq 1$, $A \neq B$ and $0 < \alpha \leq 1$, which includes the set of all $p \in \mathcal{H}_1$ satisfying

$$p(z) \prec \left(\frac{1+Az}{1+Bz}\right)^\alpha.$$

The class of Janowski strongly starlike functions of order α, denoted by $\mathcal{SS}^*(A, B, \alpha)$, which is the set of all $f \in \mathcal{S}$ such that

$$\frac{zf'(z)}{f(z)} \prec \left(\frac{1+Az}{1+Bz}\right)^\alpha.$$

In particular, $\mathcal{SS}^*(1, -1, \alpha) = \mathcal{SS}^*(\alpha)$, the class of strongly starlike functions of order α, which can also be stated as:

$$\mathcal{SS}^*(\alpha) = \left\{f \in \mathcal{S} : \left|\arg\frac{zf'(z)}{f(z)}\right| < \frac{\alpha\pi}{2}\right\}$$

and note that $\mathcal{SS}^*(1, -1, 1) = \mathcal{SS}^*(1) = \mathcal{S}^*$. Apart from the above classes we also establish various subordination, radius, argument problems involving Janowski function with complex parameters. Also we point out many known results in this direction which are especial cases of our result.

2. Basic properties of the class $\mathcal{P}(A, B, \alpha)$

In the present section we discuss various geometric properties concerned with functions belonging to $\mathcal{P}(A, B, \alpha)$. We begin with the following bound estimate result.

Theorem 2.1. *Let $h \in \mathcal{H}_1$ and $A, B \in \mathbb{C}$ with $A \neq B$ and $|B| \leq 1$. Further if $|A - B| \leq |1 - A\overline{B}|$ and*

$$h(z) \prec (1-\gamma)\left(\frac{1+Az}{1+Bz}\right)^\alpha + \gamma, \tag{2.1}$$

for some $0 < \alpha \leq 1$, $\gamma \in \mathbb{C}\backslash\{1\}$, then for $|z| = r < 1$, we have

(i) $\left|\arg(h(z)-\gamma) - \alpha\tan^{-1}\frac{\mathrm{Im}(A\overline{B})r^2}{\mathrm{Re}(A\overline{B})r^2 - 1}\right| < \alpha\sin^{-1}\frac{|A-B|r^2}{|1-A\overline{B}|r^2}$

(ii) $\left(\frac{|1-A\overline{B}r^2| - |A-B|r}{1-|B|^2r^2}\right)^\alpha \leq \left|\frac{h(z)-\gamma}{1-\gamma}\right| \leq \left(\frac{|1-A\overline{B}r^2| + |A-B|r}{1-|B|^2r^2}\right)^\alpha$

(iii) $M(t_1+\pi)\cos(N(t_1+\pi)) \leq \mathrm{Re}\left(\frac{h(z)-\gamma}{1-\gamma}\right) \leq M(t_1)\cos(N(t_1))$

(iv) $M(\tau - t_2)\sin(N(\tau - t_2)) \leq \mathrm{Im}\left(\frac{h(z)-\gamma}{1-\gamma}\right) \leq M(t_2)\sin(N(t_2))$,

where $M(t) = \left(\sqrt{(u(t))^2 + (v(t))^2}\right)^\alpha$ and $N(t) = \alpha\tan^{-1}\left(\frac{v(t)}{u(t)}\right)$, with $u(t) = \frac{1 - \mathrm{Re}(A\overline{B})r^2 + |A-B|r\cos t}{1-|B|^2r^2}$ and $v(t) = \frac{|A-B|r\sin t - \mathrm{Im}(A\overline{B})r^2}{1-|B|^2r^2}$. Further t_1 and t_2 are the roots of

$$\frac{u(t)u'(t) + v(t)v'(t)}{u(t)v'(t) - v(t)u'(t)} = \tan\left(\alpha\tan^{-1}\frac{v(t)}{u(t)}\right) \tag{2.2}$$

and

$$\frac{v(t)u'(t) - u(t)v'(t)}{u(t)u'(t) + v(t)v'(t)} = \tan\left(\alpha\tan^{-1}\frac{v(t)}{u(t)}\right) \tag{2.3}$$

respectively.

Proof. Let $H(z) := ((h(z)-\gamma)/(1-\gamma))^{\frac{1}{\alpha}}$. Since $h(z) \neq 0$ and $h(0) = 1$, we have $H \in \mathcal{H}_1$ and

$$H(z) \prec \frac{1+Az}{1+Bz}. \tag{2.4}$$

As $A, B \in \mathbb{C}$, clearly $H(z)$ is contained in the oblique circle, shown in the Figure 1, whose radius and center are given by $R := |A-B|r/(1-|B|^2r^2)$ and $C := 1-A\overline{B}r^2/(1-|B|^2r^2)$, respectively with angles $\tau(r) := \tan^{-1}(\mathrm{Im}(A\overline{B})r^2/(\mathrm{Re}(A\overline{B})r^2-1))$ and $\zeta(r) := \sin^{-1}(|A-B|r^2/(|1-A\overline{B}|r^2))$. By taking argument estimate of (2.4) and using the Figure 1 with the fact that circle is symmetric about the line passing through origin and center, we obtain (i). Let $|z| = r < 1$ and $t \in [0, 2\pi)$, we have $(h(re^{it})-\gamma)/(1-\gamma) \in \Omega := ((1+Az)/(1+Bz))^{\alpha}$, which implies $\partial\Omega : (C+Re^{it})^{\alpha} := M(t)e^{iN(t)}$. To find modulus, real and imaginary parts estimate, we need the critical points. By a simple computation, we obtain $M'(t) = 0$ at $t = \tau(1) = \tau$ and $\tau+\pi$, $(M(t)\cos N(t))' = 0$ at the roots t_1 and $t_1+\pi$ of the equation (2.2) and $(M(t)\sin N(t))' = 0$ at the roots t_2 and $\tau - t_2$ of the equation (2.3), all these values eventually yield (ii), (iii) and (iv) respectively. ∎

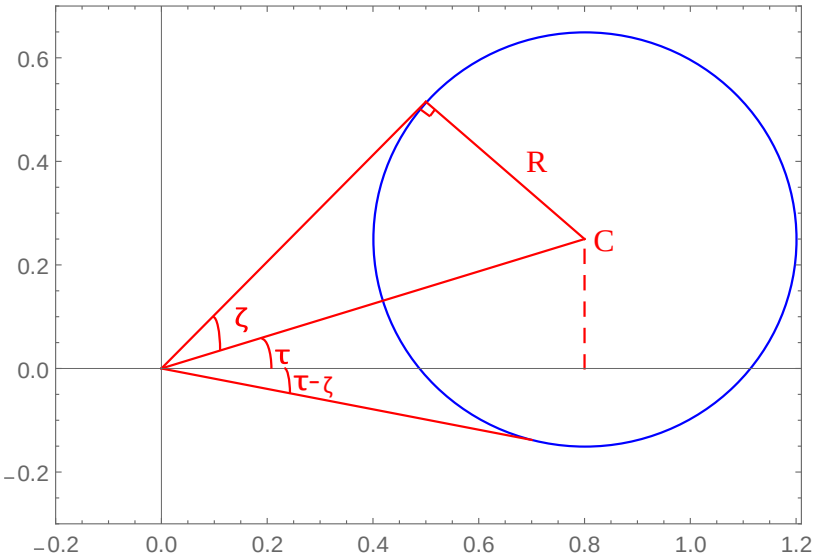


FIGURE 1. The image of $\partial\mathbb{D}$ under $(1+Az)/(1+Bz)$.

Remark 2.2. 1. When $\gamma = 0$, we can obtain from Theorem 2.1 various bound estimates for functions in $\mathcal{P}(A, B, \alpha)$.
2. By taking $\alpha = 1/2$, $\gamma = 0$, $A = 1$ and $B = 0$ in Theorem 2.1, we have $h(z) \prec \sqrt{1+z}$. Therefore the estimates are $|\arg h(z)| \leq \sin^{-1} r/2$ and $\sqrt{1-r} \leq |h(z)| \leq \sqrt{1+r}$. If $r = 1$, we have $t_1 = 0$ and $t_2 = 2\pi/3$, which implies $0 < \mathrm{Re}\, h(z) < \sqrt{2}$ and $-0.5 < \mathrm{Im}\, h(z) < 0.5$.
3. Note that if $w(z) = (1+Az)/(1+Bz)$, then $w(0) = 1 \in w(\mathbb{D})$. Thus the image domain $w(\mathbb{D})$ will always intersect real axis even if it is an oblique domain, non-symmetric with respect to real axis.

For $\alpha = 1$ and $\gamma = 0$, Theorem 2.1, reduces to the following sharp bounds:

Corollary 2.3. *Let $h \in \mathcal{H}_1$ and $A, B \in \mathbb{C}$ with $A \neq B$, $|A| \leq 1$ and $|B| \leq 1$. Further if*

$$h(z) \prec \frac{1+Az}{1+Bz},$$

then for $|z| = r < 1$, we have

(i) $\left|\arg h(z) - \tan^{-1}\dfrac{\mathrm{Im}(A\overline{B})r^2}{\mathrm{Re}(A\overline{B})r^2-1}\right| < \sin^{-1}\dfrac{|A-B|r}{|1-A\overline{B}r^2|}$

(ii) $\dfrac{|1-A\overline{B}r^2|-|A-B|r}{1-|B|^2r^2} \le |h(z)| \le \dfrac{|1-A\overline{B}r^2|+|A-B|r}{1-|B|^2r^2}$

(iii) $\dfrac{1-\mathrm{Re}(A\overline{B})r^2-|A-B|r}{1-|B|^2r^2} \le \mathrm{Re}\, h(z) \le \dfrac{1-\mathrm{Re}(A\overline{B})r^2+|A-B|r}{1-|B|^2r^2}$

(iv) $\dfrac{\mathrm{Im}(A\overline{B})r^2+|A-B|r}{|B|^2r^2-1} \le \mathrm{Im}\, h(z) \le \dfrac{\mathrm{Im}(A\overline{B})r^2-|A-B|r}{|B|^2r^2-1}.$

Note that if $A = ae^{im\pi}$ and $B = be^{in\pi}$, where $a \ge 0, 0 \le b \le 1, -1 \le m, n \le 1$ and $A \ne B$ then $(1+Az)/(1+Bz)$ maps the unit disk onto

$$H(\mathbb{D}) := \left\{ w \in \mathbb{C} : \left| w - \frac{1-A\overline{B}}{1-|B|^2} \right| < \frac{|A-B|}{1-|B|^2} \right\}. \tag{2.5}$$

Clearly, $H(\mathbb{D})$ represents a disk when $b < 1$ and a half plane when $b = 1$. We see that $w = 0$ is an exterior or interior or boundary point of $H(\mathbb{D})$ is decided by the value of a as $a < 1$ or $a > 1$ or $a = 1$ respectively. Therefore to investigate argument related problems of a Janowski function, we take $0 \le a \le 1$ or $|A-B| \le |1-A\overline{B}|$ so that $w = 0$ is not an interior point of $H(\mathbb{D})$. Following are the radius and center of the disk (2.5):

$$R := \frac{|A-B|}{1-|B|^2} = \frac{\sqrt{a^2+b^2-2ab\cos((n-m)\pi)}}{1-b^2},$$
$$C := \frac{1-A\overline{B}}{1-|B|^2} = \frac{1-ab\cos((n-m)\pi)+iab\sin((n-m)\pi)}{1-b^2}.$$

We observe that whenever the difference of m and n is same then the corresponding R and C also remain same. Therefore without lose of generality, we can fix $n = 1$. Accordingly we confine our study of Janowski function by considering

$$\frac{1+Az}{1-bz}, \tag{2.6}$$

where $A+b \ne 0$, $A \in \mathbb{C}$, $0 \le b \le 1$ and the class $\tilde{P}(A,b) = \{p \in \mathcal{H}_1 : p(z) \prec (1+Az)/(1-bz)\}$. The corresponding class of Janowski strongly starlike functions of order α is denoted by $\widetilde{SS}^*(A,b,\alpha) = \{f \in \mathcal{S} : zf'(z)/f(z) \prec ((1+Az)/(1-bz))^\alpha\}$. As a consequence of Theorem 2.1, the following result yields an equivalence relation between half plane Janowski sector whose boundary passes through origin and its argument bounds.

Theorem 2.4. *For $0 < \alpha \le 1$ and $-1 < m < 1$, then the function*

$$h(z) = \left(\frac{1+e^{im\pi}z}{1-z}\right)^\alpha, \tag{2.7}$$

is analytic, univalent and convex in $\mathbb{D}$ with

$$h(\mathbb{D}) = \left\{ w \in \mathbb{C} : -\alpha(1-m)\frac{\pi}{2} \le \arg w \le \alpha(1+m)\frac{\pi}{2} \right\}. \tag{2.8}$$

Proof. By using Theorem 2.1, the function $h(z)$ given in (2.7) satisfy

$$\left| \arg h(z) - \alpha \tan^{-1}\left(\tan\frac{m\pi}{2}\right) \right| < \frac{\alpha\pi}{2},$$

which gives the desired result. ∎

Remark 2.5. 1. From the domain (2.8) we have:

$$(h(\mathbb{D}))^{1/\alpha} = \{w \in \mathbb{C} : \operatorname{Re} e^{-im\pi/2} w > 0\}. \tag{2.9}$$

2. For $0 < \alpha_1 \leq 1$ and $0 < \alpha_2 \leq 1$, if $m = (\alpha_1 - \alpha_2)/(\alpha_1 + \alpha_2)$ and $\alpha = (\alpha_1 + \alpha_2)/2$ in (2.7) then Theorem 2.4 reduces to [9, Lemma 3].

The following result generalizes Theorem 2.4.

Theorem 2.6. *Let $h \in \mathcal{H}_1$ and $0 \leq b \leq 1$ with $b + e^{im\pi} \neq 0$, where $-1 \leq m \leq 1$. Also, if*

$$h(z) \prec \frac{1 + e^{im\pi} z}{1 - bz}, \tag{2.10}$$

then

$$\operatorname{Re} e^{-i\lambda} h(z) > 0, \tag{2.11}$$

where $\lambda = \tan^{-1}\left(\dfrac{b \sin(m\pi)}{b\cos(m\pi) + 1}\right)$.

Proof. To obtain (2.11), it suffices to show that $|\arg(e^{-i\lambda} w)| < \pi/2$ or $|\arg w - \lambda| < \pi/2$, where $w = h(z)$. By using Theorem 2.1 with $\alpha = 1$, the function $h(z)$ given in (2.10) satisfy

$$\left| \arg h(z) - \tan^{-1} \frac{b \sin(m\pi)}{b\cos(m\pi) + 1} \right| < \sin^{-1} \frac{|e^{im\pi} + b|}{|1 + be^{im\pi}|} = \frac{\pi}{2},$$

which leads to the desired result. ∎

Remark 2.7. 1. When $b = 1$ then Theorem 2.6 provides sufficient condition for functions to be in the class $\mathcal{P}_{-\lambda}$.
2. Let $|A| \leq 1$ and $|B| \leq 1$ with $A \neq B$. Assume $a(\alpha) := \arg((1 + Az)/(1 + Bz))^{\alpha}$. Then from Theorem 2.1, we observe that $\max a(\alpha_1) \leq \max a(\alpha_2)$ and $\min a(\alpha_1) \geq \min a(\alpha_2)$, whenever $0 < \alpha_1 \leq \alpha_2 \leq 1$. Therefore we have

$$\left(\frac{1 + Az}{1 + Bz}\right)^{\alpha_1} \prec \left(\frac{1 + Az}{1 + Bz}\right)^{\alpha_2} \quad (0 < \alpha_1 \leq \alpha_2 \leq 1).$$

Theorem 2.8. *Let $|A_j| \leq 1$ and $|B_j| \leq 1$ $(j = 1, 2)$ with $A_1 \neq B_1$ and $A_2 \neq B_2$. Let C_j and R_j be the centres and radii of $(1 + A_j z)/(1 + B_j z) = \phi_j(z)$ $(j = 1, 2)$, respectively. Then (i) $\mathcal{P}(A_1, B_1) \subseteq \mathcal{P}(A_2, B_2)$ if and only if the line segment C_1C_2 lies entirely in the domain $\phi_1(\mathbb{D})$, whenever $|C_1 - C_2| \leq R_1$.*
(ii) $\mathcal{P}(A_1, B_1) \subseteq \mathcal{P}(A_2, B_2)$ if and only if the line segment C_1C_2 does not lie entirely in the domain $\phi_1(\mathbb{D})$, whenever $|C_1 - C_2| \geq R_1$.

The proof is skipped here, as the result is evident from the Figure 2.

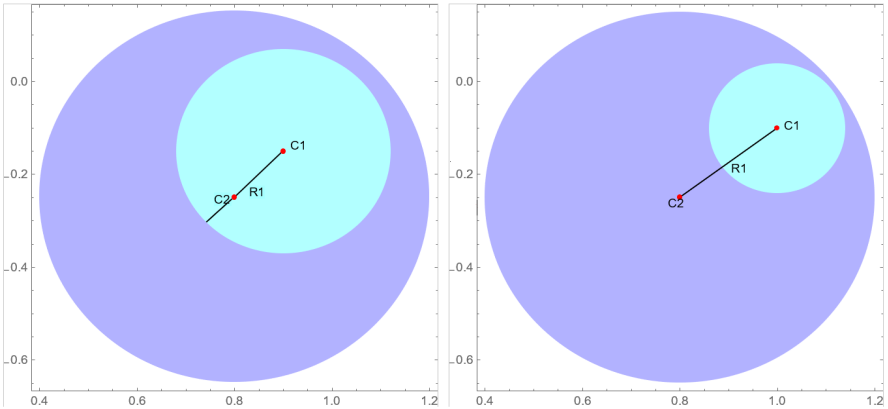


FIGURE 2. The image of $\mathbb{D}$ under $\phi_i(z)$ $(i = 1, 2)$.

We note that if $\mathcal{P}(A_1, B_1) \subseteq \mathcal{P}(A_2, B_2)$, then $(1 + A_1 z)/(1 + B_1 z) \prec (1 + A_2 z)/(1 + B_2 z)$, thus from Theorem 2.8, we obtain the following result:

Corollary 2.9. *If $\mathcal{P}(A_1, B_1) \subseteq \mathcal{P}(A_2, B_2)$, then for $0 < \alpha \leq 1$, we also have*

$$\left(\frac{1 + A_1 z}{1 + B_1 z}\right)^{\alpha} \prec \left(\frac{1 + A_2 z}{1 + B_2 z}\right)^{\alpha}.$$

Note that the observations made in [20] are generalized in part 2 of Remark 2.7 and Corollary 2.9.

3. ARGUMENT RELATED RESULTS

In this section, we generalized various subordination results using Theorem 2.4 and the following versions of the Jack's Lemma:

Lemma 1. [12, p. 234-235] *Let $h \in \mathcal{H}[1, n]$. If there exists a point $z_0 \in \mathbb{D}$, such that*

$$|\arg p(z)| < |\arg p(z_0)| = \frac{\pi\beta}{2}, \quad (|z| < |z_0|),$$

for some $\beta > 0$, then we have

$$\frac{z_0 p'(z_0)}{p(z_0)} = \frac{2ik \arg p(z_0)}{\pi},$$

for some $k \geq n(a + a^{-1})/2 > n$, where $p(z_0)^{1/\beta} = \pm ia$, and $a > 0$.

Lemma 2. [14] *Let $h(z)$ be analytic in $\mathbb{D}$, with $h(0) = 1$ and $h(z) \neq 0$. If there exist two points $z_1, z_2 \in \mathbb{D}$, such that*

$$-\frac{\alpha_1 \pi}{2} = \arg h(z_1) < \arg h(z) < \arg h(z_2) = \frac{\alpha_2 \pi}{2}$$

for $\alpha_1, \alpha_2 \in (-1, 1)$ and $|z| < |z_1| = |z_2|$, then we have

$$\frac{z_1 h'(z_1)}{h(z_1)} = -i\frac{\alpha_1 + \alpha_2}{2}k \quad \text{and} \quad \frac{z_2 h'(z_2)}{h(z_2)} = i\frac{\alpha_1 + \alpha_2}{2}k,$$

where

$$k \geq \frac{1 - |a|}{1 + |a|} \quad \text{and} \quad a = i \tan \frac{\pi}{4}\left(\frac{\alpha_2 - \alpha_1}{\alpha_2 + \alpha_1}\right).$$

The first result of this section is the generalization of [15, Theorem 1.6], which produces various corollaries and also generalizes Pommerenke's result [18]: Let $f \in \mathcal{A}$, $g \in \mathcal{C}$ and $0 < \alpha \leq 1$, then

$$\left|\arg\left(\frac{f'(z)}{g'(z)}\right)\right| < \frac{\alpha\pi}{2} \qquad \Longrightarrow \qquad \left|\arg\left(\frac{f(z)}{g(z)}\right)\right| < \frac{\beta(\alpha)\pi}{2}.$$

Theorem 3.1. *Let $f, g \in \mathcal{A}$ and $0 < \alpha \leq 1$. For some $m \in [-1, 1)$ and $\beta \in (0, 1)$, let $a = i \tan \frac{m\pi}{4}$ and $|g(z)/(zg'(z))| > \beta$. If*

$$\frac{f'(z)}{g'(z)} \prec \left(\frac{1 + e^{i\mu\pi}z}{1 - z}\right)^{\alpha(\mu_1+\mu_2)/2}, \tag{3.1}$$

then

$$\frac{f(z)}{g(z)} \prec \left(\frac{1 + e^{im\pi}z}{1 - z}\right)^{\alpha}, \tag{3.2}$$

where $\mu_j = 1 + (-1)^j m + \dfrac{2}{\alpha\pi} \tan^{-1} \dfrac{\alpha\beta(1 - |a|) \cos\left(\arg \frac{g(z)}{zg'(z)}\right)}{1 + |a| + (-1)^{j+1}\alpha\beta(1 - |a|) \sin\left(\arg \frac{g(z)}{zg'(z)}\right)}$ $(j = 1, 2)$ *and* $\mu = \dfrac{\mu_2 - \mu_1}{\mu_1 + \mu_2}$.

Proof. Let $p(z) := f(z)/g(z)$. Then in view of Theorem 2.4, to prove (3.2), it is sufficient to show $-\alpha(1 - m)\pi/2 \leq \arg p(z) \leq \alpha(1 + m)\pi/2$. On the contrary, if there exists two points $z_1, z_2 \in \mathbb{D}$ such that

$$-\alpha(1 - m)\frac{\pi}{2} = \arg p(z_1) < \arg p(z) < \arg p(z_2) = \alpha(1 + m)\frac{\pi}{2}$$

for $|z| < |z_1| = |z_2|$, then by Lemma 2, we have

$$\frac{z_1 p'(z_1)}{p(z_1)} = -i\alpha k \quad \text{and} \quad \frac{z_2 p'(z_2)}{p(z_2)} = i\alpha k.$$

Since $p(z) = f(z)/g(z)$, thus we have

$$\frac{f'(z)}{g'(z)} = p(z)\left(1 + \frac{zp'(z)}{p(z)}\frac{g(z)}{zg'(z)}\right)$$

and

$$\arg\left(\frac{f'(z)}{g'(z)}\right) = \arg p(z) + \arg\left(1 + \frac{zp'(z)}{p(z)}\frac{g(z)}{zg'(z)}\right).$$

For $z = z_1$, we have

$$\begin{aligned}\arg\left(\frac{f'(z_1)}{g'(z_1)}\right) &\leq -\alpha(1 - m)\frac{\pi}{2} + \arg\left(1 - i\alpha k\beta\left(\cos\left(\arg\frac{g(z_1)}{z_1 g'(z_1)}\right) + i\sin\left(\arg\frac{g(z_1)}{z_1 g'(z_1)}\right)\right)\right) \\ &\leq -\alpha(1 - m)\frac{\pi}{2} + \tan^{-1}\left(\frac{\alpha\beta(|a|-1)\cos\left(\arg\dfrac{g(z_1)}{z_1 g'(z_1)}\right)}{1 + |a| + \alpha\beta(1 - |a|)\sin\left(\arg\dfrac{g(z_1)}{z_1 g'(z_1)}\right)}\right),\end{aligned}$$

which contradicts (3.1). Similarly for $z = z_2$, we have

$$\arg\left(\frac{f'(z_2)}{g'(z_2)}\right) \geq \alpha(1+m)\frac{\pi}{2} + \arg\left(1 + i\alpha m\beta\left(\cos\left(\arg\frac{g(z_2)}{z_2 g'(z_2)}\right) + i\sin\left(\arg\frac{g(z_2)}{z_2 g'(z_2)}\right)\right)\right)$$
$$\geq \alpha(1+m)\frac{\pi}{2} + \tan^{-1}\left(\frac{\alpha\beta(1-|a|)\cos\left(\arg\frac{g(z_2)}{z_2 g'(z_2)}\right)}{1+|a|-\alpha\beta(1-|a|)\sin\left(\arg\frac{g(z_2)}{z_2 g'(z_2)}\right)}\right),$$

which again contradicts (3.1), that completes the proof. ∎

If we choose $g(z) = z$ in Theorem 3.1, it reduces to the following corollary:

Corollary 3.2. *Let $f \in \mathcal{A}$ and $0 < \alpha \leq 1$. For some $m \in [-1, 1)$, let $a = i\tan\frac{m\pi}{4}$. If*

$$f'(z) \prec \left(\frac{1+e^{i\mu\pi}z}{1-z}\right)^{\alpha(\mu_1+\mu_2)/2},$$

then

$$\frac{f(z)}{z} \prec \left(\frac{1+e^{im\pi}z}{1-z}\right)^{\alpha}, \tag{3.3}$$

where $\mu_j = 1 + (-1)^j m + \frac{2}{\alpha\pi}\tan^{-1}\frac{\alpha(1-|a|)}{1+|a|}$ $(j = 1, 2)$ and $\mu = \frac{\mu_2 - \mu_1}{\mu_1 + \mu_2}$.
Further, $f \in \mathcal{S}^*(\beta)$, where $\beta = \alpha\pi + \tan^{-1}\frac{\alpha(1-|a|)}{1+|a|}$.

Proof. The subordination (3.3) holds, by using Theorem 3.1. Now just to prove $f \in \mathcal{S}^*(\beta)$. Since

$$\arg\frac{zf'(z)}{f(z)} = \arg\frac{z}{f(z)} + \arg f'(z),$$

using Theorem 2.4, we obtain

$$-\alpha(1+m+\mu_1)\frac{\pi}{2} < \arg\frac{zf'(z)}{f(z)} < \alpha(1-m+\mu_2)\frac{\pi}{2},$$
$$-\alpha\pi - \tan^{-1}\frac{\alpha(1-|a|)}{1+|a|} < \arg\frac{zf'(z)}{f(z)} < \alpha\pi + \tan^{-1}\frac{\alpha(1-|a|)}{1+|a|},$$

which implies $f \in \mathcal{S}^*(\beta)$. ∎

Remark 3.3. If $m = 0$, then Corollary 3.2 reduces to [15, Theorem 1.7].

The next corollary is a generalization of [13, Theorem 1].

Corollary 3.4. *Let $f \in \mathcal{A}$, $g \in \mathcal{C}$ and $g \in \mathcal{RS}^*(\beta)$, where $0 \leq \beta < 1$. Suppose $0 < \alpha \leq 1$ and for some $m \in [-1, 1)$, let $a = i\tan\frac{m\pi}{4}$. If*

$$\frac{f'(z)}{g'(z)} \prec \left(\frac{1+e^{i\mu\pi}z}{1-z}\right)^{(\alpha(\mu_1+\mu_2))/2},$$

then

$$\frac{f(z)}{g(z)} \prec \left(\frac{1+e^{im\pi}z}{1-z}\right)^{\alpha}.$$

where $\mu_j = 1 - m + \frac{2}{\alpha\pi}\tan^{-1}\frac{\alpha\beta(1-|a|)}{1+|a|+\alpha(1-|a|)}$ $(j = 1, 2)$ and $\mu = \frac{\mu_2 - \mu_1}{\mu_1 + \mu_2}$.

Proof. Since $g \in \mathcal{C}$, from Marx-Strohhäcker's theorem, we have $\mathrm{Re}(zg'(z)/g(z)) > 1/2$, which implies that $|g(z)/(zg'(z)) - 1| < 1$. Thus we obtain

$$\left|\mathrm{Im}\left(\frac{g(z)}{zg'(z)}\right)\right| < 1 \quad \text{and} \quad \mathrm{Re}\left(\frac{g(z)}{zg'(z)}\right) > \beta. \tag{3.4}$$

Now using (3.4) and the methodology of Theorem 3.1, the result follows at once. ∎

Remark 3.5. When we take $m = 0$ in Corollary 3.4, then the result reduces to [13, Theorem 1].

Theorem 3.6. *Let $p \in \mathcal{H}_1$ and $m \in [-1, 1)$. Also, for a fixed $\gamma \in [0, 1]$ and $\alpha > 0$, let $\beta > \beta_0 (\geq 0)$, where β_0 is the solution of the equation*

$$\alpha\beta(1 - m) + \frac{2\gamma}{\pi}\tan^{-1}\eta = 0$$

and for a suitable fixed $\eta \geq 0$, let $\lambda(z) : \mathbb{D} \to \mathbb{C}$ be a function satisfying

$$\frac{\beta\,\mathrm{Re}\,\lambda(z)}{1 + \beta|\mathrm{Im}\,\lambda(z)|} \geq \eta. \tag{3.5}$$

If

$$(p(z))^{\alpha}\left(1 + \lambda(z)\frac{zp'(z)}{p(z)}\right)^{\gamma} \prec \left(\frac{1 + e^{i\mu\pi}z}{1 - z}\right)^{\delta}, \tag{3.6}$$

then

$$p(z) \prec \left(\frac{1 + e^{im\pi}z}{1 - z}\right)^{\beta}. \tag{3.7}$$

where $\mu_j = \alpha\beta(1 + (-1)^j m) + \dfrac{2\gamma}{\pi}\tan^{-1}\eta$ $(j = 1, 2)$, $\delta = \dfrac{\mu_1 + \mu_2}{2}$ and $\mu = \dfrac{\mu_2 - \mu_1}{\mu_1 + \mu_2}$.

Proof. According to Theorem 2.4, to prove (3.7), it is sufficient to show that $-\beta(1 - m)\frac{\pi}{2} < \arg p(z) < \beta(1 + m)\frac{\pi}{2}$. On the contrary if there exists two points $z_1, z_2 \in \mathbb{D}$ such that

$$-\beta(1 - m)\frac{\pi}{2} = \arg p(z_1) < \arg p(z) < \arg p(z_2) = \beta(1 + m)\frac{\pi}{2}$$

for $|z| < |z_1| = |z_2|$, then from Lemma 2, we have

$$\frac{z_1 p'(z_1)}{p(z_1)} = -ik\beta \quad \text{and} \quad \frac{z_2 p'(z_2)}{p(z_2)} = ik\beta$$

where, $k \geq \frac{1-|a|}{1+|a|}$ and $a = i\tan\frac{m\pi}{4}$. For $z = z_1$, using (3.5) we have

$$\begin{aligned}\arg\left((p(z_1))^{\alpha}\left(1 + \lambda(z_1)\frac{z_1 p'(z_1)}{p(z_1)}\right)^{\gamma}\right) &\leq -\alpha\beta(1 - m)\frac{\pi}{2} - \gamma\tan^{-1}\frac{\beta m\,\mathrm{Re}\,\lambda(z_1)}{1 + \beta m|\mathrm{Im}\,\lambda(z_1)|} \\ &\leq -\left(\alpha\beta(1 - m)\frac{\pi}{2} + \gamma\tan^{-1}\eta\right),\end{aligned}$$

which contradicts (3.6). Similarly, for $z = z_2$, we have

$$\begin{aligned}\arg\left((p(z_2))^{\alpha}\left(1 + \lambda(z_2)\frac{z_2 p'(z_2)}{p(z_2)}\right)^{\gamma}\right) &\geq \alpha\beta(1 + m)\frac{\pi}{2} + k\tan^{-1}\frac{\beta\gamma\,\mathrm{Re}\,\lambda(z_2)}{1 + \beta\gamma|\mathrm{Im}\,\lambda(z_2)|} \\ &\geq \alpha\beta(1 + m)\frac{\pi}{2} + \gamma\tan^{-1}\eta,\end{aligned}$$

which also contradicts (3.6). Hence the result follows. ∎

Remark 3.7. If we take $m = 0$ in Theorem 3.6 then the result reduces to [22, Theorem 1].

4. Radius results

This section aims to find the largest radius R of a property $\mathfrak{P}$ such that every function of a set $\mathcal{M}$ has the property $\mathfrak{P}$ in the disk $\mathbb{D}_r$, where $r \leq R$. First result of this section generalizes [1, Theorem 2.1-2.2].

Theorem 4.1. *Let $\gamma, \delta \in \mathbb{C}\backslash\{1\}$ and $A, B, C, D \in \mathbb{C}$ with $A \neq B$, $|B| \leq 1$, $D \neq C$ and $|D| \leq 1$. If*

$$f(z) \prec (1-\delta)\left(\frac{1+Cz}{1+Dz}\right)^{\beta} + \delta,$$

then

$$f(z) \prec (1-\gamma)\left(\frac{1+Az}{1+Bz}\right)^{\alpha} + \gamma,$$

in $|z| \leq R$, where

$$R = \min\left\{\frac{\alpha|(A-B)(1-\gamma)^{\frac{1}{\alpha}}|}{|(C-D)\beta(1-\gamma)^{\frac{1}{\alpha}-1}(\delta-1)|+\beta|(1-\gamma)^{\frac{1}{\alpha}-1}(AD(\gamma-1)+B(C(1-\delta)+D(\delta-\gamma)))|}, 1\right\}.$$

Proof. Let P and Q be functions defined as

$$P(z) = (1-\gamma)\left(\frac{1+Az}{1+Bz}\right)^{\alpha} + \gamma \quad \text{and} \quad Q(z) = (1-\delta)\left(\frac{1+Cz}{1+Dz}\right)^{\beta} + \delta.$$

Further, define the function M as

$$M(z) = P^{-1}(Q(z)) = \frac{\left((1-\delta)(1+Cz)^{\beta} + (\delta-\gamma)(1+Dz)^{\beta}\right)^{\frac{1}{\alpha}} - (1+Dz)^{\frac{\beta}{\alpha}}(1-\gamma)^{\frac{1}{\alpha}}}{A(1+Dz)^{\frac{\beta}{\alpha}}(1-\gamma)^{\frac{1}{\alpha}} - B\left((1-\delta)(1+Cz)^{\beta} + (\delta-\gamma)(1+Dz)^{\beta}\right)^{\frac{1}{\alpha}}}.$$

Replacing z by Rz so that $|z| = R \leq 1$, we obtain

$$|M(Rz)| \leq \frac{|(C-D)\beta(1-\gamma)^{\frac{1}{\alpha}-1}(\delta-1)|R}{\alpha|(A-B)(1-\gamma)^{\frac{1}{\alpha}} - \beta|(1-\gamma)^{\frac{1}{\alpha}-1}(AD(\gamma-1)+B(C-C\delta+D(\delta-\gamma)))|R} \leq 1$$

for

$$R \leq \frac{\alpha|(A-B)(1-\gamma)^{\frac{1}{\alpha}}|}{|(C-D)\beta(1-\gamma)^{\frac{1}{\alpha}-1}(\delta-1)|+\beta|(1-\gamma)^{\frac{1}{\alpha}-1}(AD(\gamma-1)+B(C(1-\delta)+D(\delta-\gamma)))|}.$$

Thus $f(z) \prec Q(z)$ for $|z| \leq \min\{R, 1\}$. Hence the result follows. ∎

Taking $\gamma = \delta = 0$ in the Theorem 4.1, we obtain the following corollary.

Corollary 4.2. *Let $A, B, C, D \in \mathbb{C}$ with $A \neq B$, $|B| \leq 1$, $D \neq C$ and $|D| \leq 1$. Then $\mathcal{S}^*(C, D, \beta) \subseteq \mathcal{S}^*(A, B, \alpha)$ if and only if*

$$\beta(|C-D|+|AD-BC|) \leq \alpha|A-B|.$$

In particular,

(1) for $\beta_1 \in (0, 1]$, we have $\mathcal{SS}^*(\beta_1) \subseteq \mathcal{S}^*(A, B, \alpha)$ if and only if

$$\beta(2+|A+B|) \leq \alpha|A-B|.$$

(2) for $\alpha_1 \in (0,1]$, we have $\mathcal{S}^*(C,D,\beta) \subseteq \mathcal{SS}^*(\alpha_1)$ if and only if

$$\beta(|C+D|+|C-D|) \leq 2\alpha.$$

Remark 4.3. Let $\alpha_1, \alpha_2 \in [0,1)$, $A = 1-2\alpha_1$, $C = 1-2\alpha_2$, $B = D = -1$ and $\alpha = \beta = 1$, then the corollary 4.2 reduces to the well known fact that $\mathcal{S}^*(\alpha_2) \subseteq \mathcal{S}^*(\alpha_1)$ if and only if $\alpha_2 \leq \alpha_1$.

Corollary 4.4. *Let $A, B \in \mathbb{C}$ with $A \neq B$, $|B| \leq 1$ and $\beta_2 > 0$. If $f \in \mathcal{S}^*(A,B,\alpha)$, then*

(1) $f \in \mathbb{M}(\beta_2)$ in $|z| \leq R_{\mathbb{M}}(\beta_2)$ for $\beta > 1$, where

$$R_{\mathbb{M}}(\beta_2) = \min\left\{\frac{a|A-B|}{2(\beta_2+1)+|A-B(2\beta-1)|}, 1\right\}. \tag{4.1}$$

(2) $f \in \mathcal{RS}^*(\beta_2)$ in $|z| \leq R_{\mathcal{RS}^*}(\beta_2)$ for $0 \leq \beta < 1$, where

$$R_{\mathcal{RS}^*}(\beta_2) = \min\left\{\frac{a|A-B|}{2\beta_2+|A-B(2\beta-1)|}, 1\right\}. \tag{4.2}$$

Theorem 4.5. *Let $f(z) \in \mathcal{A}$ with $f'(z) \neq 0$ in $\mathbb{D}$. Also let $0 < A \leq 1$, $-1 \leq B < 0$, $\alpha = 0.38344486\cdots$ and $\beta \geq 0.61655\cdots$ which satisfy the conditions $\tan^{-1}\alpha = (1-2\alpha)/2$ and $\tan^{-1}\alpha \leq (\beta-\alpha)\pi/2$, respectively. If*

$$f'(z) \prec \left(\frac{1+Az}{1+Bz}\right)^{\beta}, \tag{4.3}$$

then $f(z)$ is starlike in $|z| < r_0$, where

$$r_0 = \frac{-(A-B)+\sqrt{(A-B)^2+4AB\left(\sin(\alpha+\frac{2}{\pi}\tan^{-1}\alpha)\frac{\pi}{2\beta}\right)^2}}{2AB\sin\left((\alpha+\frac{2}{\pi}\tan^{-1}\alpha)\frac{\pi}{2\beta}\right)}, \tag{4.4}$$

is the smallest positive root of the equation

$$\left(\alpha+\frac{2}{\pi}\tan^{-1}\alpha\right)\frac{\pi}{2} = \beta\sin^{-1}\left(\frac{(A-B)x}{1-ABx^2}\right). \tag{4.5}$$

Proof. By Theorem 2.1, the subordination (4.3) yields that

$$|\arg f'(z)| \leq \beta\sin^{-1}\frac{(A-B)|z|}{1-AB|z|^2}. \tag{4.6}$$

Let $p(z) = f(z)/z$, we suppose that there exists a point z_0, $|z_0| = r_0 < 1$, such that $|\arg p(z)| < \alpha\pi/2$ for $|z| < r_0$ and $|\arg p(z_0)| = \alpha\pi/2$ then by Lemma 1, we have

$$\frac{z_0 p'(z_0)}{p(z_0)} = \frac{2ik\arg p(z_0)}{\pi}, \tag{4.7}$$

where $k \geq m(a+a^{-1})/2 \geq 1$, $p(z_0) = \pm ia$ and $a > 0$. Also if $\arg p(z_0) = \alpha\pi/2$ then using (4.5), we have

$$\begin{aligned}
\arg f'(z_0) &= \arg(p(z_0)+zp'(z_0)) \\
&= \arg p(z_0) + \arg\left(1+\frac{z_0 p'(z_0)}{p(z_0)}\right) \\
&= \frac{\alpha\pi}{2} + \arg(1+i\alpha k) \\
&\geq \frac{\alpha\pi}{2} + \tan^{-1}\alpha \\
&= \beta\sin^{-1}\left(\frac{(A-B)r_0}{1-ABr_0^2}\right),
\end{aligned}$$

which contradicts (4.6). Similar contrary conclusion will arrive for the case when $\arg p(z_0) = -\alpha\pi/2$. Therefore we have

$$|\arg p(z_0)| = \left|\arg\left(\frac{f(z)}{z}\right)\right| < \frac{\alpha\pi}{2}. \tag{4.8}$$

Now using (4.5), (4.6) and (4.8), we have

$$\begin{aligned}\left|\arg\left(\frac{zf'(z)}{f(z)}\right)\right| &\leq |\arg f'(z)| + \left|\arg\left(\frac{z}{f(z)}\right)\right| \\ &< \left(2\alpha + \frac{2}{\pi}\tan^{-1}\alpha\right)\frac{\pi}{2}.\end{aligned}$$

For f to be starlike, we must have $2\alpha + (2\tan^{-1}\alpha)/\pi = 1$, which gives $\alpha = 0.38344486\cdots$ and since

$$\left(\alpha + \frac{2}{\pi}\tan^{-1}\alpha\right)\frac{\pi}{2} = \beta\sin^{-1}\left(\frac{(A-B)r_0}{1-ABr_0^2}\right),$$

we have $f(z)$ is starlike in $|z| < r_0$, where r_0 is given in (4.4), that completes the proof. ∎

5. Subordination results

In this section we discuss subordination related results for the Janowski function. The book [11] provides numerous results pertaining to differential subordination and some of the results which we need in context of our study are listed below:

Lemma 3. [11, Theorem 3.1d,p.76] *Let h be analytic and starlike univalent in $\mathbb{D}$ with $h(0) = 0$. If g is analytic in $\mathbb{D}$ and $zg'(z) \prec h(z)$, then*

$$g(z) \prec g(0) + \int_0^z \frac{h(t)}{t}dt.$$

Lemma 4. [11, Theorem 3.4h,p.132] *Let $g(z)$ be univalent in $\mathbb{D}$, Φ and Θ be analytic in a domain Ω containing $g(\mathbb{D})$ such that $\Phi(w) \neq 0$, when $w \in g(\mathbb{D})$. Now letting $G(z) = zg'(z)\cdot\Phi(g(z))$, $h(z) = \Theta(g(z)) + G(z)$ and either h or g is convex. Further, if*

$$\operatorname{Re}\frac{zh'(z)}{G(z)} = \operatorname{Re}\left(\frac{\Theta'(g(z))}{\Phi(g(z))} + \frac{zG'(z)}{G(z)}\right) > 0.$$

as well as p is analytic in $\mathbb{D}$, with $p(0) = g(0)$, $p(\mathbb{D}) \subset \Omega$ and

$$\Theta(p(z)) + zp'(z)\cdot\Phi(p(z)) \prec \Theta(g(z)) + zg'(z)\cdot\Phi(g(z)) := h(z)$$

then $p \prec g$, and g is the best dominant.

Antonino and Romaguera in [2] gave the following result:

Lemma 5. [2] *Let $p(z)$ be analytic in $\mathbb{D}, p(0) = c$, let $q(z)$ be univalent and convex in $\mathbb{D}$ and let $g(z)$ be analytic in $\mathbb{D}$ such that $zg'(z)/g(z)$ is analytic and different from zero in $\mathbb{D}$. Let β and γ be complex constants such that*

$$\operatorname{Re}\left(\frac{\xi g'(\xi)}{g(\xi)}(\beta q(z) + \gamma)\right) > 0 \quad (z, \xi \in \mathbb{D}).$$

then

$$p(z) + \frac{g(z)}{zg'(z)}\left(\frac{zp'(z)}{\beta p(z) + \gamma}\right) \prec q(z) \implies p(z) \prec q(z).$$

Several authors studied various combinations of $zf'(z)/f(z)$ and $1+zf''(z)/f'(z)$. One such combination is their quotient, which was first investigated by Silverman in [21]. Silverman considered the following class for $\beta \in (0,1]$,

$$\mathcal{G}_\beta = \left\{ f \in \mathcal{A} : \frac{1+zf''(z)/f'(z)}{zf'(z)/f(z)} - 1 \prec \beta z \right\}.$$

Theorem 5.1. *Let* $|A| \leq 1$ *and* $0 \leq b \leq 1$ *with* $A+b \neq 0$ *and* $\beta(1+b)^{\alpha-1}(1+|A|)^{\alpha+1} \leq \alpha|A+b|$. *If* $f \in \mathcal{G}_\beta$, *then* $f \in \widetilde{\mathcal{SS}}^*(A,b,\alpha)$.

Proof. For $f \in \mathcal{G}_\beta$, we define the function $p(z) = zf'(z)/f(z)$. By standard calculations we obtain that

$$\frac{1+zf''(z)/f'(z)}{zf'(z)/f(z)} - 1 = \frac{zp'(z)}{p^2(z)}.$$

For $f \in \widetilde{\mathcal{SS}}^*(A,b,\alpha)$, it is enough to show $p(z) \prec ((1+Az)/(1-bz))^\alpha := q(z)$. For if, there exist points $z_0 \in \mathbb{D}$, $\varsigma_0 \in \partial\mathbb{D}\backslash\{-1/A, 1/b\}$ and $m \geq 1$ such that $p(|z|< |z_0|) \subset q(\mathbb{D})$ with $p(z_0) = q(\varsigma_0)$ and $z_0p'(z_0) = m\varsigma_0 q'(\varsigma_0)$. If

$$\frac{\alpha|A+b|}{(1+|A|)^{\alpha+1}(1+b)^{\alpha-1}} \geq \beta \implies \left| \frac{\alpha\varsigma_0(A+b)}{(1+A\varsigma_0)^{\alpha+1}(1-b\varsigma_0)^{\alpha-1}} \right| \geq \beta.$$

Thus,

$$\left| \frac{m\varsigma_0 q'(\varsigma_0)}{q^2(\varsigma_0)} \right| \geq \beta \quad \text{for all } m \geq 1.$$

or, equivalently,

$$\left| \frac{z_0p'(z_0)}{p^2(z_0)} \right| \geq \beta,$$

which contradicts $f \in \mathcal{G}_\beta$. Thus $p(z) \prec q(z)$ and that completes proof. ∎

Remark 5.2. If $\alpha = 1$ and $-1 \leq B < A \leq 1$ in Theorem 5.1 then the result reduces to [23, Theorem 3.2,p.9].

Theorem 5.3. *If* $f \in \mathcal{G}_\beta$, *then* f *is starlike of reciprocal order* α.

Proof. For $f \in \mathcal{G}_\beta$, we define the function $p(z)$ by $f(z)/(zf'(z)) = \alpha + (1-\alpha)p(z)$. Clearly $p \in \mathcal{H}_1$ and $\alpha + (1-\alpha)p(z) \neq 0$ for $z \in \mathbb{D}$. Now by a simple computation we obtain that

$$1 - \frac{1+zf''(z)/f'(z)}{zf'(z)/f(z)} = (1-\alpha)zp'(z).$$

Since $f \in \mathcal{G}_\beta$, therefore we have

$$(1-\alpha)zp'(z) \prec \beta z. \tag{5.1}$$

Now using Lemma 3, (5.1) leads to

$$p(z) \prec 1 + \frac{\beta}{1-\alpha}z,$$

which shows that

$$|p(z) - 1| < \frac{\beta}{1-\alpha}.$$

Now using Theorem 2.1, we obtain that

$$\left| \operatorname{Arg}\left(\frac{f(z)}{zf'(z)} - \alpha \right) \right| = |\operatorname{Arg} p(z)| < \sin^{-1}\frac{\beta}{1-\alpha} = \frac{\delta\pi}{2}.$$

Since $\alpha \in [0,1)$, and $\beta \in (0,1]$. Therefore $\delta \in (0,1]$, which completes our result. ∎

In [19], authors have considered Janowski function with complex parameters and therefore the corresponding Janowski disk is non symmetric with respect to real axis. However their findings were based on the Janowski disk that are symmetric with respect to real axis, which is possible only if the parameters are real in Janowski function. Therefore by eliminating this limitation, we obtain the following result as an extension of [19, Lemma 2.17].

Theorem 5.4. *Let $\alpha \in (0,1]$, $l,m \in [-1,1]$ and $a,b,c,d \in [0,1]$ such that $ae^{il\pi}+b \neq 0$ and $ce^{im\pi}+d \neq 0$. Let $Q \in \mathcal{H}[1,n]$ satisfy*

$$Q(z) \prec \frac{1+ae^{il\pi}z}{1-bz} \tag{5.2}$$

and

$$Q(z)p^{\alpha}(z) \prec \frac{1+ce^{im\pi}z}{1-dz}, \tag{5.3}$$

for $p \in \mathcal{H}_1$. If

$$\mu := \sin^{-1}\sqrt{\frac{c^2+d^2+2cd\cos(m\pi)}{1+c^2d^2+2cd\cos(m\pi)}} + \sin^{-1}\sqrt{\frac{a^2+b^2+2ab\cos(l\pi)}{1+a^2b^2+2ab\cos(l\pi)}} \leq \frac{\alpha\pi}{2}, \tag{5.4}$$

then

$$\mathrm{Re}(e^{-i\gamma\pi/2}p(z)) > 0,$$

where $\gamma = (\tan^{-1}A - \tan^{-1}B)/\mu$ with $A = \dfrac{cd\sin(m\pi)}{cd\cos(m\pi)+1}$ and $B = \dfrac{ab\sin(l\pi)}{ab\cos(l\pi)+1}$.

Proof. From Theorem 2.1, (5.2) yields

$$\left|\arg Q(z) - \tan^{-1}\left(\frac{ab\sin(l\pi)}{ab\cos(l\pi)+1}\right)\right| < \sin^{-1}\sqrt{\frac{a^2+b^2+2ab\cos(l\pi)}{1+a^2b^2+2ab\cos(l\pi)}}. \tag{5.5}$$

Similarly, from (5.3), we obtain

$$\left|\arg Q(z) + \alpha\arg p(z) - \tan^{-1}\left(\frac{cd\sin(m\pi)}{cd\cos(m\pi)+1}\right)\right| < \sin^{-1}\sqrt{\frac{c^2+d^2+2cd\cos(m\pi)}{1+c^2d^2+2cd\cos(m\pi)}}. \tag{5.6}$$

After some computations using (5.5) and (5.6), we obtain

$$-\frac{\pi}{2}\left(\frac{2}{\alpha\pi}(\mu-\gamma\mu)\right) \leq \arg p(z) \leq \frac{\pi}{2}\left(\frac{2}{\alpha\pi}(\mu+\gamma\mu)\right) \tag{5.7}$$

which eventually yields

$$p(z) \prec \frac{1+e^{i\gamma\pi}z}{1-z},$$

that completes the proof. ∎

Theorem 5.5. *Let $p(z) \in \mathcal{H}_1$, $A \in \mathbb{C}$ and $0 \leq b \leq 1$ with $|A| \leq 1$, $A+b \neq 0$ and $\mathrm{Re}(1+Ab) \geq |A+b|$. Further let α,γ are two real parameters lying in $[0,1]$ and μ,δ,ρ and η are complex parameters such that $\mathrm{Re}(\mu/\eta) > 0$, $\mathrm{Re}\,\delta > 0$ and $\mathrm{Re}\,\rho \geq 0$. If*

$$\mu(p(z))^{\alpha}(\delta+\rho p(z)) + \eta z p'(z)(p(z))^{\alpha-1} \prec h(z),$$

then $p \in \widetilde{\mathcal{S}\mathcal{S}}^{*}(A,b,\gamma)$, where

$$h(z) = \left(\frac{1+Az}{1-bz}\right)^{\alpha\gamma}\left(\mu\delta + \mu\rho\left(\frac{1+Az}{1-bz}\right)^{\gamma} + \eta\gamma\frac{(A+b)z}{(1+Az)(1-bz)}\right).$$

Proof. Let us choose $g(z) = ((1+Az)/(1-bz))^{\gamma}$, $\Phi(w) = \eta w^{\alpha-1}$ and $\Theta(w) = \mu w^{\alpha}(\delta + \rho w)$ then clearly $g \in \mathcal{P}$, univalent and convex in $\mathbb{D}$. Φ, Θ are analytic in a domain Ω containing $g(\mathbb{D})$, with $\Phi(w) \neq 0$ when $w \in g(\mathbb{D})$. If $G(z) = zg'(z)\Phi(g(z))$ then

$$\operatorname{Re}\frac{zG'(z)}{G(z)} = \operatorname{Re}\left(\frac{\alpha+1}{1-bz} - \frac{\alpha-1}{1+Az} - 1\right) \geq \frac{\alpha+1}{1+b} - \frac{\alpha-1}{1+|A|} - 1 \geq 0$$

and

$$\operatorname{Re}\frac{zh'(z)}{G(z)} = \operatorname{Re}\left(\frac{\mu\alpha\delta}{\eta} + \frac{\mu(\alpha+1)\rho g(z)}{\eta} + \frac{zG'(z)}{G(z)}\right) > 0.$$

Thus by Lemma 4, we obtain that $p \prec g$ and g is the best dominant. ∎

By taking $\mu = 1, \delta = 1-\lambda, \rho = \lambda, \eta = \lambda$ and $p(z) = zf'(z)/f(z)$ in Theorem 5.5, where $\lambda \in [0,1]$, we obtain the following more modified and simplified form of results in [5].

Corollary 5.6. *Let $f(z) \in \mathcal{A}$, such that $f(z)/z \neq 0$ in $\mathbb{D}$, $A \in \mathbb{C}$ and $0 \leq b \leq 1$ with $|A| \leq 1$, $A + b \neq 0$ and $\operatorname{Re}(1+Ab) \geq |A+b|$. Suppose also that the real parameters λ, α and γ are such that they lies in $[0,1]$. If*

$$\left(\frac{zf'(z)}{f(z)}\right)^{\alpha}\left(1 + \lambda\frac{zf''(z)}{f'(z)}\right) \prec h(z),$$

then $f \in \widetilde{\mathcal{S}}^{*}(A,b,\gamma)$, where

$$h(z) = \left(\frac{1+Az}{1-bz}\right)^{\alpha\gamma-1}\left((1-\lambda)\frac{1+Az}{1-bz} + \frac{\lambda(1+Az)^{1+\gamma}(1-bz)^{1-\gamma} + \lambda\gamma(A+b)z}{(1-bz)^2}\right).$$

By taking $\delta = 1$ and $\rho = 0$ in Theorem 5.5, we obtain the following result.

Corollary 5.7. *Let $p(z) \in \mathcal{H}_1$ and $0 \leq b \leq 1$ with $b + e^{im\pi} \neq 0$, where $-1 \leq m \leq 1$. Also let $\alpha > -1$ and if*

$$\mu(p(z))^{\alpha} + \eta z p'(z)(p(z))^{\alpha-1} \prec \left(\frac{1+e^{im\pi}z}{1-bz}\right)^{\alpha\gamma}\left(\mu + \eta\gamma\frac{(b+e^{im\pi})z}{(1+e^{im\pi}z)(1-bz)}\right) := h(z),$$

for some $\mu, \eta \in \mathbb{C}$ such that $\operatorname{Re}(\mu/\eta) \geq 0$, then

$$\operatorname{Re} e^{-i\beta}(p(z))^{1/\gamma} > 0,$$

where $\beta = \tan^{-1}\dfrac{b\sin(m\pi)}{b\cos(m\pi)+1}$. And the inequality is sharp for the function $p(z)$ defined by

$$p(z) = \left(\frac{1+e^{im\pi}z}{1-bz}\right)^{\gamma}.$$

Proof. By Theorem 5.5 we have $p(z) \prec ((1+e^{im\pi}z)/(1-bz))^{\gamma} := g(z)$ and g is the best dominant. Further by Theorem 2.6 we obtain the desired conclusion. ∎

By taking $\mu = 1-\lambda, \alpha = 1$ and $\eta = \lambda$ in Corollary 5.7, we obtain the following result.

Corollary 5.8. *Let $p(z) \in \mathcal{H}_1$ and $0 \leq b \leq 1$ with $b + e^{im\pi} \neq 0$, where $-1 \leq m \leq 1$. Now if*

$$(1-\lambda)p(z) + \lambda z p'(z) \prec \left(\frac{1+e^{im\pi}z}{1-bz}\right)^{\gamma}\left(1 - \lambda + \lambda\gamma\frac{(b+e^{im\pi})z}{(1+e^{im\pi}z)(1-bz)}\right) := h(z),$$

for some $0 < \lambda \leq 1$ and $0 < \gamma \leq 1$, then

$$\operatorname{Re} e^{-i\beta}(p(z))^{1/\gamma} > 0,$$

where $\beta = \tan^{-1}\dfrac{b\sin(m\pi)}{b\cos(m\pi)+1}$. *Further the inequality is sharp for the function* $p(z)$ *given by*

$$p(z) = \left(\frac{1+e^{im\pi}z}{1-bz}\right)^{\gamma}.$$

Remark 5.9. When $m = 0$ and $b = 1$ then Corollary 5.8 reduces to the [17, Theorem 1].

In case when $m = 0$, $\gamma = 1$ and $\lambda = 0.5$, Corollary 5.8 yields:

Corollary 5.10. *Let* $p(z) \in \mathcal{H}_1$ *and* $0 \le b \le 1$. *Suppose*

$$p(z) + zp'(z) \prec \frac{1+2z-bz^2}{(1-bz)^2},$$

then

$$p(z) \prec \frac{1+z}{1-bz}.$$

Theorem 5.11. *Let* $\alpha \in (0,1], \beta, \gamma \in \mathbb{C}$ *with* $\beta \neq 0$, $\mathrm{Re}(\beta+\gamma) > 0$ *and* $A, B \in \mathbb{C}$ *with* $A \neq B$, $|B| \le 1$ *and* $|A-B| \le 1 - \mathrm{Re}(A\overline{B})$. *Let* $\lambda(z)$ *be Carathéodory and if* $p(z)$ *is analytic such that*

$$p(z) + \lambda(z)\frac{zp'(z)}{\beta p(z)+\gamma} \prec \left(\frac{1+Az}{1+Bz}\right)^{\alpha},$$

then

$$p(z) \prec q(z) := H(z)\left(\int_0^z \frac{H(t)g'(t)dt}{g(t)}\right)^{-1} \prec \left(\frac{1+Az}{1+Bz}\right)^{\alpha},$$

where

$$g(z) = z\exp\int_0^z \frac{(\lambda(t))^{-1}-1}{t}dt \quad \textit{and} \quad H(z) = g(z)\exp\int_0^z \frac{g'(t)\left(\left(\frac{1+At}{1+Bt}\right)^{\alpha}-1\right)dt}{g(t)}.$$

The function $q(z)$ *is convex and is the best (1,n)-dominant.*

Proof. Let $f \in \mathcal{A}_n$ with $f(z)/z \neq 0$ in $\mathbb{D}$ and

$$p(z) = \frac{F'(z)g(z)}{F(z)g'(z)}, \tag{5.8}$$

where

$$F(z) = \left(\frac{\beta+\gamma}{g^{\gamma}(z)}\int_0^z f^{\beta}(t)g^{\gamma-1}(t)g'(t)dt\right)^{\frac{1}{\beta}}. \tag{5.9}$$

Since $\mathrm{Re}(\beta+\gamma) > 0$, by using an argument similar to that of [11, Lemma 1.2c,p.11], we can show that the function $S(z)$ defined by

$$S(z) = \frac{z^{\beta+\gamma-1}}{z^{\beta+\gamma}g^{\gamma-1}g'(z)f^{\beta}(z)}\int_0^z \left(\frac{f(t)}{t}\right)^{\beta} t^{\beta+\gamma-1}\left(\frac{g(t)}{t}\right)^{\gamma-1} g'(t)dt, \tag{5.10}$$

is analytic in $\mathbb{D}$ and $S \in \mathcal{H}[\frac{1}{\beta+\gamma}, n]$. From (5.9) and (5.10) we obtain

$$\frac{F(z)}{z} = \left((\beta+\gamma)S(z)\frac{zg'(z)}{g(z)}\right)^{\frac{1}{\beta}}\left(\frac{f(z)}{z}\right).$$

Since both the expressions on the right are analytic and nonzero, therefore we conclude that $F \in \mathcal{A}_n$ and $F(z)/z \neq 0$ for $z \in \mathbb{D}$. Now by a simple computation using 5.8 and 5.9, we obtain

$$p(z) + \lambda(z)\frac{zp'(z)}{\beta p(z)+\gamma} = \frac{f'(z)g(z)}{f(z)g'(z)} \prec \left(\frac{1+Az}{1+Bz}\right)^{\alpha}.$$

Since $zg'(z)/g(z)$ and $((1+Az)/(1+Bz))^{\alpha}$ are Carathéodory, also $\mathrm{Re}(\beta+\gamma)>0$, thus we have

$$\mathrm{Re}\left(\frac{zg'(z)}{g(z)}\left(\beta\left(\frac{1+Az}{1+Bz}\right)^{\alpha}+\gamma\right)\right)>0.$$

Therefore the result holds by Lemma 5. ∎

Theorem 5.11 is a generalisation of [7, Theorem 2] and also we have the following result.

Corollary 5.12. *Let $\alpha\in(0,1], \beta,\gamma\in\mathbb{C}$ with $\mathrm{Re}(\beta+\gamma)>0$ and $A,B\in\mathbb{C}$ with $A\neq B, |B|\leq 1$ and $|A-B|\leq 1-\mathrm{Re}(A\overline{B})$. Suppose that $g(z)\in\mathcal{S}^*$ and $f\in\mathcal{A}$ satisfies*

$$\frac{f'(z)g(z)}{f(z)g'(z)}\prec\left(\frac{1+Az}{1+Bz}\right)^{\alpha}$$

then the function $F(z)$ defined in (5.9) is in $\mathcal{A}$, $F(z)/z\neq 0$, for $z\in\mathbb{D}$ and

$$\frac{F'(z)g(z)}{F(z)g'(z)}\prec\left(\frac{1+Az}{1+Bz}\right)^{\alpha}.$$

As stated in (2.6), we can consider $B\in[-1,0]$ and by taking $\lambda(z)\equiv 1$, $\beta=1$ and $\gamma=0$ in Theorem 5.11, and letting $p(z)=zf'(z)/f(z)$, for $f\in\mathcal{A}$, we obtain the following subordination implication:

Corollary 5.13. *If $\alpha\in(0,1]$ and $A\in\mathbb{C}$, $b\in[0,1]$ such that $A+b\neq 0$ and $|A+b|\leq 1+\mathrm{Re}(Ab)$. Also, if $f(z)\in\mathcal{A}$ satisfies*

$$1+\frac{zf''(z)}{f'(z)}\prec\left(\frac{1+Az}{1-bz}\right)^{\alpha}$$

then

$$\frac{zf'(z)}{f(z)}\prec\frac{zK'(z)}{K(z)},$$

where

$$K(z)=\begin{cases}\int_0^z\exp\left(\int_0^w\frac{(\frac{1+At}{1-bt})^{\alpha}-1}{t}dt\right)dw, & A\neq 0,\ b\neq 0\\ \int_0^z\exp(\alpha bt[{}_3F_2[1,1,1+\alpha,2,2,bt]])dt, & A=0,\ b\neq 0,\\ \int_0^z\exp(\alpha At[{}_3F_2[1,1,1-\alpha,2,2,-At]])dt, & A\neq 0,\ b=0.\end{cases}$$

Remark 5.14. When $\alpha=1$, Corollary 5.13 reduces to [7, Corollary 3.2], for which

$$K(z)=\begin{cases}((1-bz)^{-A/b}-1)/A, & A\neq 0,\ b\neq 0,\\ -\log(1-bz)/b, & A=0,\ b\neq 0,\\ (e^{Az}-1)/A, & A\neq 0,\ b=0.\end{cases}$$

Moreover, for $A=1-2\beta$ $(0\leq\beta<1)$ and $b=1$, with $\alpha=1$ in Corollary 5.13, we obtain that, if $f(z)\in\mathcal{C}(\beta)$ then $f(z)\in\mathcal{S}^*(\gamma)$, where

$$\gamma=\gamma(\beta)=\begin{cases}\frac{1-2\beta}{2(2^{1-2\beta}-1)}, & \beta\neq\frac{1}{2},\\ 1/(2\log 2), & \beta=\frac{1}{2}.\end{cases}$$

Note that this relationship was proved by MacGregor in [10].

Acknowledgement

The work of the second author is supported by The Council of Scientific and Industrial Research(CSIR). Ref.No.: 08/133(0030)/2019-EMR-I.

References

[1] R. M. Ali, M. H. Mohd and L. S. Keong, Radii of starlikeness, parabolic starlikeness and strong starlikeness for Janowski starlike functions with complex parameters, Tamsui Oxf. J. Inf. Math. Sci. **27** (2011), no. 3, 253–267.

[2] J. A. Antonino and S. Romaguera, Strong differential subordination to Briot-Bouquet differential equations, J. Differential Equations **114** (1994), no. 1, 101–105.

[3] S. D. Bernardi, Convex and starlike univalent functions, Trans. Amer. Math. Soc. **135** (1969), 429–446.

[4] W. Janowski, Some extremal problems for certain families of analytic functions. I, Ann. Polon. Math. **28** (1973), 297–326.

[5] K. Kuroki, S. Owa and H. M. Srivastava, Some subordination criteria for analytic functions, Bull. Soc. Sci. Lett. Łódź Sér. Rech. Déform. **52** (2007), 27–36.

[6] K. Kuroki and S. Owa, Some subordination criteria concerning the sălăgean operator, J. Ineq. Pure Appl. Math. **10** (2009), no. 2, 11 pp.

[7] K. Kuroki, H. M. Srivastava and S. Owa, Some applications of the principle of differential subordination, Electron. J. Math. Anal. Appl. **1** (2013), 40–46.

[8] R. J. Libera, Some classes of regular univalent functions, Proc. Amer. Math. Soc. **16** (1965), 755–758.

[9] J. L. Liu and R. Srivastava, The order of starlikeness of some classes of strongly starlike functions, Quaest. Math. **41** (2018), no. 5, 707–718.

[10] T. H. MacGregor, A subordination for convex functions of order α, J. London Math. Soc. **9** (1974/75), 530–536.

[11] S. S. Miller and P. T. Mocanu, *Differential subordinations*, Monographs and Textbooks in Pure and Applied Mathematics, 225, Marcel Dekker, Inc., New York, 2000.

[12] M. Nunokawa, On the order of strongly starlikeness of strongly convex functions, Proc. Japan Acad. Ser. A Math. Sci. **69** (1993), no. 7, 234–237.

[13] M. Nunokawa, S. Owa, J. Nishiwaki, K. Kuroki and T. Hayami, Differential subordination and argumental property, Comput. Math. Appl. **56** (2008), no. 10, 2733–2736.

[14] M. Nunokawa, S. Owa, H. Saitoh, N. E. Cho and N. Takahashi, Some properties of analytic functions at extremal points for arguments, (preprint, 2002).

[15] M. Nunokawa and J. Sokół, On the order of strong starlikeness and the radii of starlikeness for of some close-to-convex functions, Anal. Math. Phys. **9** (2019), no. 4, 2367–2378.

[16] M. Nunokawa, J. Sokół and N. E. Cho, Sufficient conditions for univlence and starlikeness, Bull. Iranian Math. Soc. **42** (2016), no. 4, 933–939.

[17] S. Owa and M. Obradović, An application of differential subordinations and some criteria for univalency, Bull. Austral. Math. Soc. **41** (1990), no. 3, 487–494.

[18] C. Pommerenke, On close-to-convex analytic functions, Trans. Amer. Math. Soc. **114** (1965), 176–186.

[19] S. Ponnusamy and V. Singh, Criteria for strongly starlike functions, Complex Variables Theory Appl. **34** (1997), no. 3, 267–291.

[20] R. K. Raina, P. Sharma and J. Sokół, A class of strongly close-to-convex functions, Bol. Soc. Paran. Mat. **38** (2020), no. 6, 9–24.

[21] H. Silverman, Convex and starlike criteria, Int. J. Math. Math. Sci. **22** (1999), no. 1, 75–79.

[22] S. Singh and S. Gupta, Non-autonomous differential subordinations related to a sector, J. Ineq. Pure Appl. Math. (electronic only) **4** (2003), no. 4.

[23] J. Sokół and L. Trojnar-Spelina, On a sufficient condition for strongly starlikeness, J. Inequal. Appl. **2013**, 2013:383, 11 pp.

[24] L. M. Wang, The tilted Carathéodory class and its applications, J. Korean Math. Soc. **49** (2012), no. 4, 671–686.

Department of Applied Mathematics, Delhi Technological University, Delhi–110042, India
Email address: spkumar@dce.ac.in

Department of Applied Mathematics, Delhi Technological University, Delhi–110042, India
Email address: poojayv100@gmail.com

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Optimal DG Allocation Using Particle Swarm Optimization

Prem Prakash
Assistant Professor
Electrical Engineering Department
Delhi Technological University, Delhi
Email: ppyadav1974@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395798

***Abstract*— This paper has proposed a PSO based optimization approach to optimize the sizing and position of distributed generation (DG). This research article discusses about the DG units, which are considered for placement is capable to inject active as well as reactive power. The basic objective of optimum installation and estimation of capacity of DG units is to reduce system loss of active power, enhance bus voltage. It is already known that, the distribution network has more power loss due to its higher R/X ratio. Henceforth, it is clear that, when DG is placed in a location other than optimal location, it leads to more system losses and voltage instability. Therefore, it is most important to optimize the position and size of DG unit in the network of distribution to achieve the designed objectives. The highlighted technique is utilized on 85-bus network. Programs for load flow of radial distribution system are prepared in MATLAB environment. The Forward and Backward Sweep (FBS) based method is applied to implement the PSO based optimization technique. PSO based optimization is known as a very popular optimization technique among all AI based techniques as it require very less computational time but at the same time, this technique provides very accurate results. Therefore, the proposed research work motivates to apply this technique for optimal capacity and siting of DG in distribution network**

Keywords—PSO, DG, load flow, distribution network, reduction of power loss of network, enhancement bus voltage

I. INTRODUCTION

The main function of distributed generators are basically the power generating units, which generates power at small scale are connected to the load center. They are capable of producing clean and green power as they are energized by renewable sources of energy like wind and solar. Therefore, the emission of greenhouse gases is reduced significantly, which in turn reduces global warming. Further, DGs are capable to diminish system loss of real power component with enhancement in bus voltage if their sizing and siting are performed in a strategic manner [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. The contribution of DG with respect to power generation is increasing day-by-day due to its enumerable advantages such as environmental, technical and economical [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Power flow equations are applied for position and capacity of DG in distribution network optimally. For this purpose, mathematical optimization techniques are used. These methods are capable of providing accurate results simultaneously associated with some inaccuracy because some assumptions are made to developing the formulations. The authors of [6], [12], [15], have investigated an analytical based technique for optimum position and estimate capacity of DG in distribution network to reduce power loss and enhance bus voltage in their research. Metaheuristic optimization-based approach such as GA, PSO, ant lion optimization (ALO), evolutionary programming-based method, immune algorithm and flower pollination based algorithm, supervised firefly algorithm etc. [1], [4], [5], [8], [10], [19] based optimization techniques have been applied for DG in distribution network so as the power can be minimized by considering multi-objective and single objective functions.

Most of the author considered constant power load at peak load level in their research for integration of DG optimally [12], [14], [15]. Jamil *et al*. [14] proposed an analytical approach for optimal integration of SPV based DG at multiple locations in the network to assess the performance of the system. Daud *et al*. [1] highlighted the comparison among improved gravitational search method (IGSA), gravitational search method (GSA) and PSO based algorithm for installation and sizing of DG unit in distribution network. Bayat and Bagheri [7] proposed a novel artificial intelligent based method for optimum installation of DG in the system and concluded that the proposed method has a faster convergence rate. Babu and Singh [13] proposed a general algebraic modeling system (GAMS) based novel software for load flow solutions and optimal installation and estimate the capacity of DG unit for reduce network active power loss.

To satisfy the consistently expanding need in the deregulated and rebuilt the power system. The Govt. arrangements and expanded accessibility of little limit age advancements are supporting the expanded improvement and organization of conveyed age. Natural concerns have persuaded the utilization of inexhaustible disseminated age. The objectives of DG placement in distribution network are summarize in order to achieve power loss minimization, improvement of reliability, maximization of loadability, minimization of cost, improvement of voltage profile etc.

It is very essential to integrate the DG unit (s) in network of distribution for reduction of loss and enhancement of bus voltage by BFS method of load flow to obtain the bus voltage, branch current and branch power flow.

The installation of distribution generators in network is worth note and crucial because it is capable of improvement in system voltage profile, reduce system power loss, enhance system reliability, reduce emission of toxic gases in environment. This is possible only when the placement and sizing of DG unit is performed in optimal manner. If their sizing and siting is not performed optimally then they lead to increase system loss and voltage instability. To achieve these objectives an efficient optimization technique should be applied. There are two types of optimization techniques are available one is classical optimization technique in which the expressions for DG capacity are derived by applying mathematical methods of optimization. These methods are capable to provide accurate results or close to accurate results and require less computational time. In these methods

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:38:25 UTC from IEEE Xplore. Restrictions apply.

the obtained results are suffer from accuracy if assumptions are made during developing the expressions. One more drawback of these methods is that the optimization is not obtained as global optimization. The other methods of optimization are metaheuristic based techniques. These methods are generally nature inspired techniques. And these methods are working on the principle of biological properties. These methods require more time for simulation and achieved results may or may not be accurate exactly. The optimization achieved by these methods is globally optimized.

Therefore, these methods are universally accepted in order to achieve optimum results in engineering and other field of life. In this research article artificial intelligent based PSO approach is applied. Because it requires less computational time than other nature inspired optimization technique such as GA, ACO etc. and many more, also provide better results almost similar to other technique with less computational time. In present paper the DG unit which is to be placed is capable to insert real power only. The real example of such type of DG is solar photovoltaic (SPV) based DG

II. FORMULATION OF PROBLEM

In this research article the main focus is put forward that in the form of objective function. The designed function comprises three functions therefore the designed function is considered as multi-objective function. This multi-objective function is formed by giving relative weightage to the respective functions. Hence, these weights are given as per functions required outcome of the objective function. In order to avoid violation of constraints the suitable value of penalty factor is added.

$$F = w_1 f_1 + w_2 f_2 + w_3 f_3 \tag{1}$$

Where, w_1, w_2, and w_3 are weight factor and are decided by priority of function and also follow following property

$$|w_1| + |w_2| + |w_2| = 1 \tag{2}$$

In this research article the value of respective weight factors are as $w_1 = 0.50$, $w_2 = 0.25$ and $w_3 = 0.25$

A. Objective Function

Prime objective of this study is to design the objective function to achieve the desired output. The basic issue is to decide the best areas and rating of the DGs which will streamline different tasks identified with framework execution. As expressed before for multi objective streamlining issue, objective function is not to be limited to framework power loss. The considered function is streamlining issue, 3 goals f_1, f_2, and f_3 are studied. These goals are reduction of network power loss, enhancement in voltage profile and stability as far as possible separately

B. System Real Power Loss

The function f_1 is part of objective function and is defined as follows

$$f_1 = \frac{\sum_{i=1}^{n-1} \left(P_{Li}\right)_{after\ DG\ Placement}}{\sum_{i=1}^{n-1} \left(P_{Li}\right)_{before\ DG\ Placement}} \tag{3}$$

Here, there are n buses and (n-1) branches exist in the network P_{Li} is network power loss with and without DG installation

C. Voltage Profile Deviation

The second function of objective function is f_2, and it is defined as follows

$$f_2 = \frac{\sum_{i=1}^{n} \left|V_i - V_{iref}\right|_{after\ DG\ Placement}}{\sum_{i=1}^{n} \left|V_i - V_{iref}\right|_{before\ DG\ Placement}} \tag{4}$$

Here, V_i is bus voltage at i^{th} bus and other voltage is considered as reference voltage after DG placement and before DG placement

D. Voltage Stability Index

The third function (f_3) is voltage stability index of objective function (1) is defined as follows

$$VSI(k) = |V_i|^4 - 4\left(P_k X_{ik} - Q_k R_{ik}\right)^2 - 4\left(P_k R_k + Q_k X_k\right)^2 |V_i|^2 \tag{5}$$

Where, f_3 is reciprocal of $VSI(k)$ is defined as follows

$$f_3 = \frac{1}{VSI(k)_{after\ DG\ \ Placement}} \tag{6}$$

Therefore, final fitness function is as follows which is to be minimized

$$F = \min\left(w_1 f_1 + w_2 f_2 + w_2 f_3\right) \tag{7}$$

E. Constraints

The following constraints are discussed herewith in order to obtain optimum simulation results

a) Real and imaginary power balancing constraint

Real and imaginary power are expressed as follows to maintain equation of power balance

$$\sum_{j=1}^{n} P_{DG(j)} + P_{sw} = \sum_{j=1}^{n} P_{D(j)} + \sum_{j=1}^{n-1} P_{branch(j)} \tag{8}$$

$$\sum_{j=1}^{n} Q_{DG(j)} + Q_{sw} = \sum_{j=1}^{n} Q_{D(j)} + \sum_{j=1}^{n-1} Q_{branch(j)} \tag{9}$$

b) Constraint for Voltage Limit

To guarantee the bus voltage should remain in prescribed limit of voltage as it is decided by authorities:

$$V_{\min} \le |V_i| \le V_{\max} \tag{10}$$

Where,

V_i = Voltage of i^{th} bus

V_{min} = lowest voltage value at bus

V_{max} = highest voltage value at bus

In this study deviation of ± 6 % of nominal voltage value is permissible

c) DG Size

It is worth note that size of DG unit affects the system performance significantly if size is than optimum size network loss increases and DG size less than optimum size then also system increases. Therefore, it is very essential that the optimum size of DG should be installed.

$$P_{DG}^{\min} \le P_{DG(i)} \le P_{DG}^{\max} \tag{12}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:38:25 UTC from IEEE Xplore. Restrictions apply.

Furthermore, installation of DG in network changes the power flow and the power flow in conventional network is unidirectional and it becomes bidirectional when DGs are connected in the network.

$$Q_{DG}^{\min} \le Q_{DG(i)} \le Q_{DG}^{\max} \quad (13)$$

III. PARTICLE SWARM OPTIMIZATION

F. Methodology of PSO

Eberhart and Kennedy have first presented particle swarm enhancement (PSO) in 1995 as another heuristic technique. The first goal of their examination was to numerically recreate the social conduct of winged flocking of birds. The examination of birds advanced, it is found that by certain alterations, could likewise fill in as an amazing streamlining agent. The main form of PSO was proposed to deal with just nonlinear nonstop enhancement issues. PSO is a computational insight based strategy that isn't greatly affected by the size and non linearity.

x_i^k – Particle location

v_i^k – Particle speed

w – Inertia weight

p_i^k – Location of particle as individual best

p_g^k – Swarm location as global best

c_1, c_2 – PSO parameters

r_1, r_2 – Parameters in random form

Location of individual particles is simplified below

$$s_i^{k+1} = s_i^k + v_i^{k+1}$$

The speed is determined as

$$v_i^{k+1} = w \times v_i^k + c_1 \times rand\left(x_i^k - pb\right) + c_2 \times rand\left(x_i^k - gb\right)$$

G. Algorithm for Basic PSO

The following steps are involved in order to implement the PSO based optimization technique

1. Initialize the particle positions $x_i^0 \in D$ in IR^n and particle speeds $0 \le v_0^{ix} \le v_0^{\max}$ for $i = 1 \ldots n$ randomly
2. Set constant $k_{\max}, c_1, c_2$
3. Set $k = 1$
4. Assess the value of considered fitness function f_i^k by means of x_i^k when objective function is determined
5. If $F_i^k \le F_i^{best}$ then $F_i^{best} \le F_i^k, q_i^k = s_i^k$ and if $F_i^k \le F_g^{best}$ then $F_g^{best} \le F_i^k, q_g^k = s_i^k$
6. Revise particle velocities v_i^k and locations of particles s_i^k for $i = 1, \ldots\ldots\ldots n$
7. Initialize counter $k = k + 1$
8. If stopping criteria is met
9. Print results
10. Stop

IV. BASIC CHARACTERISTICS OF PSO

There are some basic descriptions for PSO based method which makes very effective are as given

- Algorithm parameters are small in number
- No derivatives are used
- Very effective for global search
- Easy to implement than other intelligent techniques
- Fitness function in PSO is the objective function thus direct optimization of problem is achieved

TABLE I. PSO, PARAMETERS CHOSEN FOR PROPOSED TECHNIQUE

S. No.	Parameter (s)	Number Chosen for Particular Parameter
1.	Initial population size	50
2.	Maximum number of iterations	20
3.	Accelerating coefficient 1	0.1
4.	Accelerating coefficient 2	0.1
5.	Inertial weight	0.9
6.	Final weight	0.4

V. RESULT AND DISCUSSION

The projected PSO method is applied to optimize the objective function given in (1) with given constraints as DG size, and voltage at different buses etc. In present research article 85-bus distribution network is taken to apply the proposed methodology for DG unit which injects the active power. In real time, the solar photovoltaic (SPV) based DG is taken into account to test the proposed method.

Further, in present paper the constant power load modeling type load profile is considered to implement proposed algorithm. The programming of load flow by BFS based method is performed in MATLAB environment. The code is formulated to achieve the solution of DG unit(s) in the form of optimum installation and sizing of DG in distribution network. The network parameters for the calculation depicted in Table II.

TABLE II. PARAMETERS CHOSEN FOR DG

S. No.	Network Parameter (s)	Particular
1.	Minimum range of size of DG	10 kVA
2.	Maximum range of size of DG	2 MVA
3.	Number of DGs	3
4.	Load model	Constant power load model

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:38:25 UTC from IEEE Xplore. Restrictions apply.

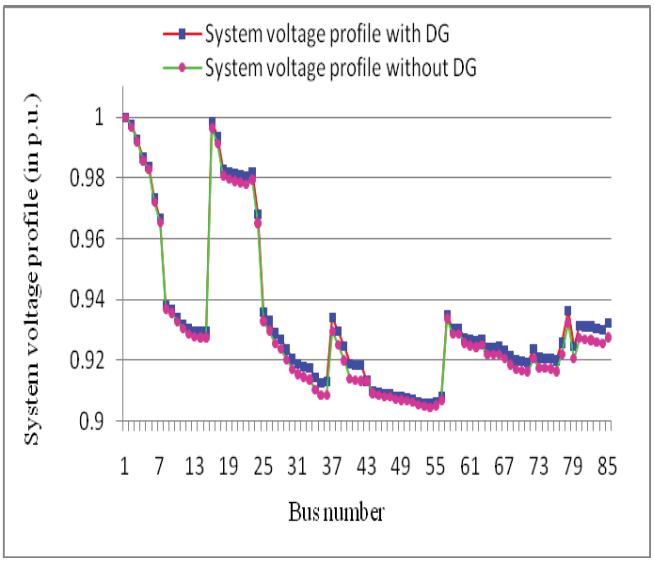


Fig.1 Bus voltage profile before and after DG placement

Fig. 1 is given for bus voltage of considered network system before and after DG installation in the network. From this figure it is evident that the bus voltage is improved significantly when DG unit (s) is/are placed at bus number 36, 69, and 55 in the system respectively

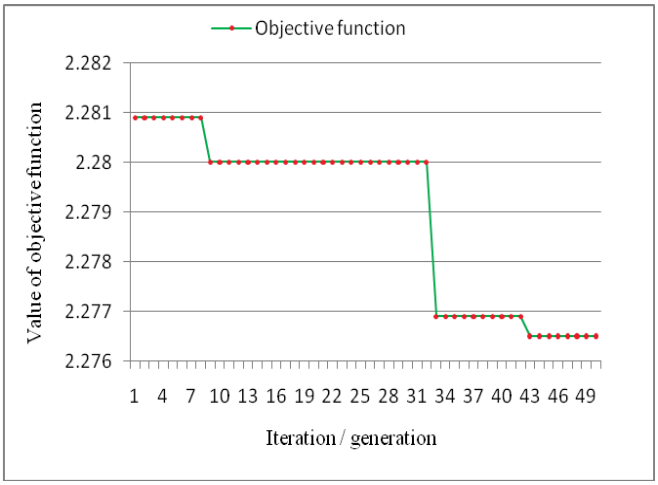


Fig.2 Variation of objective function

In Fig. 2 the variant of considered fitness function is given; in this figure numbers of iterations are considered as 50. Moreover, from this figure it is observed that the optimum value of objective is 2.2765

Moreover, the functions f_1, f_2, and f_3 are considered, after simulation their value is found as 0.2616, 0.9610, and 1.3226 respectively.

Additionally, from obtained simulation results it is observed that bus number 36, 69 and 55 are serve as candidate bus for DG unit placement and their corresponding size are 0.6665 MW, 0.7614 MW and 0.5518 MW. Moreover, the highlighted approach is very much capable to lessen the network power loss to 61.3243 kW from 234.4670 kW which is almost 73.84% reduction of system power loss is observed by placement of DG units at above mentioned locations.

Fig. 3 represents the variation of voltage stability index (VSI) at different buses, it also indicates the value of VSI at various buses which is used to identify the responsive bus for DG unit (s) installation.

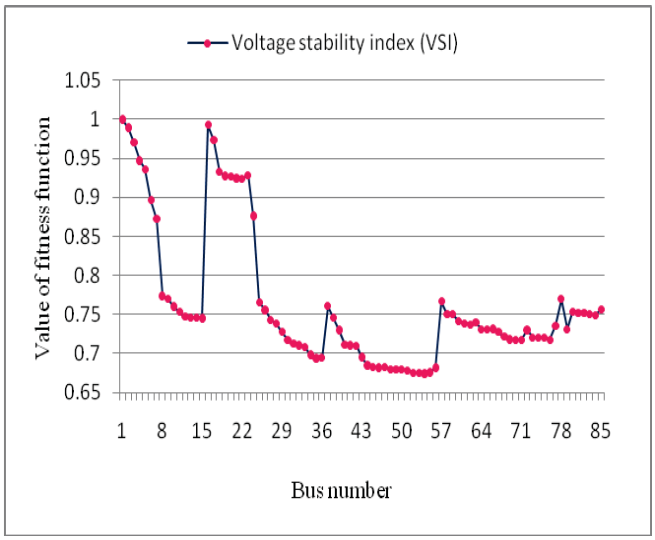


Fig.4 Value of fitness function at different buses

TABLE III. OPTIMIZATION RESULTS USING PSO BASED OPTIMIZATION TECHNIQUE

S. No.	Size (MVA)	Location (bus)	Objective Functions
1.	0.1677	36	f_1 = 0.2616
2.	0.7334	69	f_2 = 0.9234
3.	0.4961	55	f_3 = 1.3226

CONCLUSIONS

In present paper PSO based approach is investigated. It is the nature inspired heuristic optimization technique for optimum installation and optimum capacity of SPV based DG in distribution system. The highlighted method is tested on IEEE 85-bus distribution system. The voltage sensitivity index (VSI) is considered in order to find sensitive bus in form of candidate bus to place the DG unit(s) in distribution system. The sensitive buses are selected to limit the search space. If search space is more the simulation time is large. There are several AI based optimization like GA, BFO, ABC and ACO. These methods have drawback as they require more computation time. The simulated results obtained by PSO based technique provide that the proposed method is very much competent to minimize the network active power loss. Additionally, the proposed is also competent to enhance the system bus voltage profile. The PSO based method is capable to minimize the computational burden and simulation time of PSO is very much less than other method like GA etc. Moreover, PSO based. The further work can be extended by taking hybrid AI based techniques which are enable to solve designed function and reduce drawbacks of AI based technique.

ACKNOWLEDGMENT

The author is associated with Electrical Engineering Department of DTU, Main Bawana Road, Shahbad Daulatpur, Delhi-110042.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:38:25 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] S. G. Daud, A. F. A. Kadir, C. K. Gan, A. Mohamed, and T. Khatib, "A comparison of heuristic optimization techniques for optimal placement and sizing of photovoltaic based distributed generation in a distribution system," *Sol. Energy*, vol. 140, pp. 219–226, 2016.

[2] S. K. Injeti, "A Pareto optimal approach for allocation of distributed generators in radial distribution systems using improved differential search algorithm," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 908–927, 2018.

[3] S. Xie, W. Lin, Y. Zheng, S. Kong, D. Zhou, Z. Hu, and Y. Li, "Multi-objective active distribution networks expansion planning by scenario-based stochastic programming considering uncertain and random weight of network," *Appl. Energy*, vol. 219, no. February, pp. 207–225, 2018.

[4] D. B. Prakash and C. Lakshminarayana, "Multiple DG Placements in Distribution System for Power Loss Reduction Using PSO Algorithm," *Procedia Technol.*, vol. 25, pp. 785–792, 2016..

[5] A. Bayat and A. Bagheri, "Optimal active and reactive power allocation in distribution networks using a novel heuristic approach," *Appl. Energy*, vol. 233–234, pp. 71–85, 2019.

[6] R. Ishak, A. Mohamed, A. N. Abdalla, and M. Z. Che Wanik, "Optimal placement and sizing of distributed generators based on a novel MPSI index," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 389–398, 2014.

[7] D. Rama Prabha, T. Jayabarathi, D. R. Prabha, and T. Jayabarathi, "Optimal placement and sizing of multiple distributed generating units in distribution networks by invasive weed optimization algorithm," *Ain Shams Eng. J.*, vol. 7, no. 2, pp. 1–12, 2016.

[8] M. M. Othman, W. El-khattam, Y. G. Hegazy, and A. Y. Abdelaziz, "Optimal placement and sizing of voltage controlled distributed generators in unbalanced distribution networks using supervised firefly algorithm," *Int. J. Electr. Power Energy Syst.*, vol. 82, pp. 105–113, 2016.

[9] P. Mehta, P. Bhatt, and V. Pandya, "Optimal selection of distributed generating units and its placement for voltage stability enhancement and energy loss minimization," *Ain Shams Eng. J.*, vol. 9, no. 2, pp. 187–201, 2018.

[10] M. Junjie, W. Yulong, and L. Yang, "Size and Location of Distributed Generation in Distribution System Based on Immune Algorithm," *Syst. Eng. Procedia*, vol. 4, pp. 124–132, 2012.

[11] S. Ghosh, S. P. Ghoshal, and S. Ghosh, "Optimal sizing and placement of distributed generation in a network system," *Int. J. Electr. Power Energy Syst.*, vol. 32, no. 8, pp. 849–856, 2010.

[12] N. Acharya, P. Mahat, and N. Mithulananthan, "An analytical approach for DG allocation in primary distribution network," *Int. J. Electr. Power Energy Syst.*, vol. 28, no. 10, pp. 669–678, 2006.

[13] P. V. Babu and S. P. Singh, "Optimal Placement of DG in Distribution network for Power loss minimization using NLP & PLS Technique," *Energy Procedia*, vol. 90, pp. 441–454, 2016.

[14] M. Jamil and A. S. Anees, "Optimal sizing and location of SPV (solar photovoltaic) based MLDG (multiple location distributed generator) in distribution system for loss reduction, voltage profile improvement with economical benefits," *Energy*, vol. 103, pp. 231–239, 2016.

[15] R. Viral and D. K. Khatod, "An analytical approach for sizing and siting of DGs in balanced radial distribution networks for loss minimization," *Int. J. Electr. Power Energy Syst.*, vol. 67, pp. 191–201, 2015.

[16] M.H.Haque, "Efficient load flow method for distribution systems with radial or mesh configuration," *IEE Proc. - Gener. Transm. Distrib.*, vol. 143, no. 1, pp. 33–38, 1996.

[17] P. Prakash and D. K. Khatod, "Review of methods for optimal sizing and siting of Distributed Generation," *12th IEEE Int. Conf. , INDICON- 2015*, pp. 1–6, 2016.

[18] G. Celli and F. Pilo, "Optimal distributed generation allocation in MV distribution networks," *IEEE Proceeding*, pp. 81–86, 2001.

[19] D. K. Khatod, V. Pant, and J. Sharma, "Evolutionary programming based optimal placement of renewable distributed generators," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 683–695, 2013.

[20] A. J. G. Mena and Juan Andrés Martín García, "An efficient approach for the siting and sizing problem of distributed generation," *Int. J. Electr. Power Energy Syst.*, vol. 69, pp. 167–172, 2015.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:38:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

Partial Series Resonance Pulse Commutated Current-Fed Three-phase Current Sharing DC/DC Converter: ZCS Analysis, Design and Experimental Results

Swati Tandon, *Student Member, IEEE,* Akshay Kumar Rathore, *IEEE Fellow, and Jose Rodriguez, Life Fellow, IEEE*
Gina Cody School of Engineering and Computer Science, Concordia University, Montreal, QC, Canada
Universidad Andrés Bello Santiago, Chile
s_tando@encs.concordia.ca, akshay.rathore@concordia.ca, jose.rodriguez@unab.cl

***Abstract*—A three-inductor current-fed converter leveraging short pulse series resonance to facilitate zero-current switching (ZCS) and voltage clamping of the semiconductor devices is proposed and investigated in detail. The circuit exploits series resonance characteristics during conduction overlap of any two switching signals that naturally reduces the current in outgoing switch to zero accomplishing zero-current turn-off. It succeeds in natural switch voltage clamping eliminating the traditional snubber requirement. Owing to the short resonance period, circulating current through the converter limiting the conduction losses particularly at light load, improving the light-load efficiency. In addition, current-sharing topology offers reduced current stresses through the primary side devices and simpler gate driving requirements because of common source grounded with power supply. Detailed analysis and design are reported. Experimental results from 1 kW proof-of-concept hardware prototype are established to validate the projected claims and confirm the performance and the efficacy of proposed converter. This concept has been verified in single-phase topologies and first time studied for three-phase current-fed circuit.**

Keywords—*Current-fed, ZCS, resonance-pulse*

I. INTRODUCTION

With the progressing urbanization and industrialization, electrification via renewable sources has become a priority due to global climate change and increased carbon level. However, owing to the unregulated characteristics, sluggish dynamics, and discontinuous nature of non-conventional source like solar/fuel-cell, it is established that a power electronics interface becomes inevitable to provide seamless power transfer. Therefore, realizing high-efficiency, light-weight, high-density front-end dc-dc converters are necessary for applications like solar inverter, UPS, solar to battery charging, solar water pumping, DC microgrid, etc. [1]–[2]. Single phase converters with high voltage gain, galvanic isolation and low ripple input current are widely explored to meet standards for low voltage solar/fuel cells enabled microgrid system (380V). In addition, interleaved, modular and three-phase configurations are studied for high power applications owing to reduced losses, reduced mass and smaller form factor [3]-[4].

It is worth noting that the recent studies have focused on current-fed topologies because of their obvious merits of high voltage gain, stiff dc source current, reduced transformer KVA and inherent short circuit protection making it desirable choice for low voltage high current application as compared to its voltage-fed counterpart [5]-[6][8]. Moreover, high-density being one of the most desirable features for such applications, it calls for high frequency operation of the power converters. However, at high device switching frequency, high switching losses together with large switch turn-off voltage spike due to transformer leakage inductance, restricts usage of traditional hard-switched converter [9]-[10]. In addition, the diode reverse recovery loss is significantly high with undesirable conducted and radiated EMI compelling over-rated components [11]. To obtain acceptable efficiency and safe operation under high switching frequency operation, novel soft-switching methods are embraced with effectively utilizing wide bandgap devices (SiC, GaN) meeting space constraints and desired form factor.

Traditionally, clamping methods, such as the RCD lossy snubber and efficient active clamping were implemented in single-phase and three-phase current-fed converters to clamp the switch over-voltage spike. However, the additional auxiliary components increase the complexity, cost and volume for the same rated power [12]–[16][17]. Extended range soft-switching have been established through various approach reported in [18]– [19] where naturally clamped secondary modulation technique is proposed, analyzed and demonstrated for soft-switching and effective device voltage clamping over wide source voltage range. These converters are designed and built for bidirectional power conversion applications. For unidirectional applications with fewer elements, advanced soft-switching methods utilizing resonant circuits consisting of transformer leakage inductance and circuit parasitic were explored. Resonant converters majorly employ variable frequency modulation to achieve ZCS of the devices as reported in [20]–[21][22]. However, traditional resonant converters in most cases experience high circulating and peak currents compelling overrated components.

To further exemplify the benefits of resonant behavior with reduced leakage energy and reduced peak current , impulse-commutation approach was implemented in various single-phase and three-phase configurations like half-bridge, full-bridge and push-pull [23]-[24]. These converters focused on parallel resonant characteristics to achieve ZCS, with fixed resonant energy for wide range of load causing higher leakage energy at light load. Furthermore, large switching frequency range is witnessed for the slight variation in the input voltage. On the contrary, load adaptive series resonance-pulse assisted ZCS commutation for single-phase current-fed topologies has been explored and reported in [25] by the authors.

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

Table I: COMPARISON OF THE PROPOSED TOPOLOGY WITH OTHER THREE-PHASE TOPOLOGIES

	Active Clamping [17]	Snubber less Commutation [19]	Impulse Resonance [24]	Series Resonance Pulse [Proposed]
Semiconductor count	6 MOSFETs + 6 Diodes	9 MOSFETs	3 MOSFETs + 6 Diodes	3 MOSFETs + 6 Diodes
Switch voltage stress	$\frac{V_o}{1-D}$	$\frac{V_o}{n}$	$\frac{V_o}{n}$	$\frac{V_o}{2n} + V_{cr_pk}$
Modulation technique	Variable duty	Variable duty	Variable frequency	Variable Frequency
Soft Switching method	Active clamping	Secondary modulation	Quasi resonance	Quasi resonance
Additional Components	Clamping capacitor and 3 auxiliary devices	Load side active devices	HF capacitor and inductor	HF capacitor and inductor
Soft switching feature	ZVS of all switches	ZCS of primary switches, ZVS of load side switches	ZCS of all switches	ZCS of all switches
Circulating Current	High	Low	High	Low
Duty Cycle loss	High	-	High	Low
Circuit Complexity	Complex	Moderate	Simple	Simple
Voltage Gain range	30~60V/380V	42V/300V	42~48V/380V	42~48V/380V
Switching frequency	40kHz	100kHz	81-110kHz	85-150kHz

The first attempt to transform the concept and to implement series resonance pulse theory in three-phase current-fed topology is introduced in [26]. This paper investigates the series resonant-pulse concept that prompts the switch current to attain zero in a sinusoidal manner assisting ZCS turn-off of the switches. Variable resonant energy is realized by virtue of series resonant capacitor to limit leakage and peak current. As a result, lower resonant energy is observed at light loads, improving efficiency. With current-sharing three inductor configuration, small current ripple at the input reduces the gate driving and filter requirements making it suitable for PV/fuel-cell based applications. At the same time, using frequency control solely may not satisfy the ZCS requirement under wide load fluctuations. Therefore, a hybrid modulation technique which is a combination of frequency variation with slight change in duty ratio is implemented to extend ZCS range.

The salient features of the proposed converter is summarized as follows: 1) ripple-free input current due to interleaved current sharing inductors; 2) fewer number of components and simpler associated gate drive circuitry; 3) ZCS turn-off of the power switches; 4) ZCS turn-off of the diodes without reverse recovery issue. Steady-state analysis and operation of proposed three-phase converter is explained in Section II. Converter design procedure with a design example is given in Section III. Detailed experimental results from a 1kW laboratory hardware prototype are demonstrated to verify the expected claims in Section IV.

II. STEADY STATE OPERATION AND ANALYSIS

Schematic of the proposed series LC resonance-pulse three-phase current sharing topology is shown in Fig. 1. The leakage inductances (L_{lk1}, L_{lk2} and L_{lk3}) with additional series inductances and capacitances (C_{r1}, C_{r2} and C_{r3}) constitute the resonant tank that facilitates ZCS turn-off and natural current commutation of the semiconductor devices during the overlap period. Variable frequency modulation is implemented to control the output power transfer and regulate the load voltage. The steady-state operating waveforms of the proposed converter are depicted in Fig. 2. Switches in the primary are

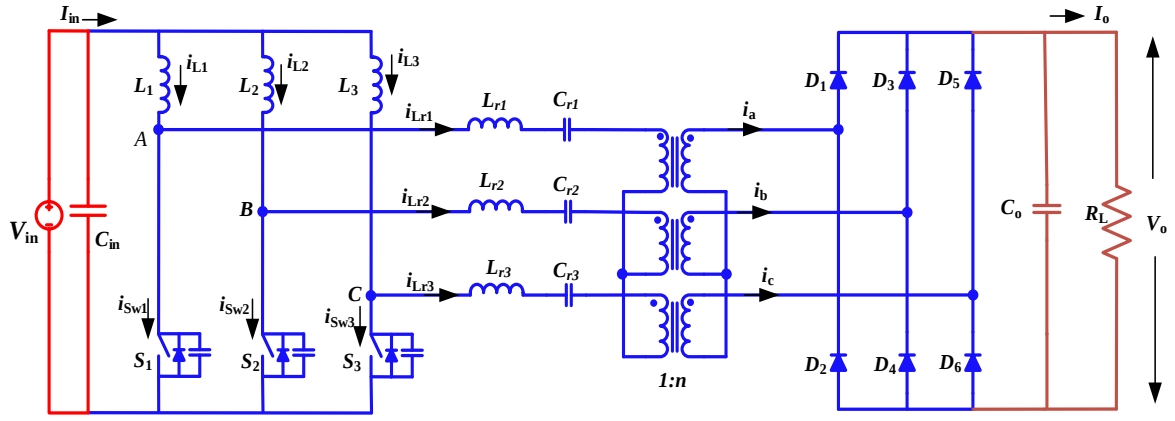


Fig. 1. Schematic for the proposed series-LC resonant-pulse current-fed three-phase current sharing topology.

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

modulated with gate signals phase shifted by 120º having a small overlap with duty ranging between $0.33 < D < 0.66$. The mathematical analysis has been carried for one third of the switching cycle (t_1-t_8), and the intervals recur in the same order for two more times during one switching cycle, with other symmetrical devices conducting. The operation of the converter during (T_s/3) is divided into 7 modes. Current and voltage equations derived help in deciding component ratings and validating converter performance in theory.

For the simplified analysis, few assumptions are made.

a) Sufficiently large value interleaved boost inductors are chosen to maintain low ripple current at the input.
b) All semiconductor devices are ideal and lossless.
c) The output capacitor is large enough to maintain the constant output voltage.
d) High value of magnetizing inductance of HF transformers is considered.

Mode 1 (t_0-t_1) *(refer Fig. 3(a))*: In this mode, switch S_3 is conducting with a constant current I_{in} through it while the other two switches are in blocking state. The boost inductor L_3 and resonant capacitors in phase A and phase B (C_{r1} and C_{r2}) get charged with constant current $I_{in}/3$ during this time. Power transfer happens through rectifier diodes. The values at the end of this interval are:

$$i_{Lr1}(t_1) = \frac{I_{in}}{3}, \quad i_{Lr2}(t_1) = \frac{I_{in}}{3}, \quad i_{Lr3}(t_1) = -\frac{2I_{in}}{3} \tag{1}$$

$$i_{S1}(t_1) = 0, \quad i_{S2}(t_1) = 0, \quad i_{S3}(t_1) = I_{in} \tag{2}$$

$$V_{cr}(t) = V_{cr}(t_o) + \frac{1}{C_r}\int_{t_o}^{t} -I_{in}\, dt \tag{3}$$

At instant $t=t_1$, $V_{cr}(t_1) = V_{cr1} = V_{cr}(t_o) - \frac{I_{in}}{C_r}T_{10}$ (4)

$$V_{ab}(t_1) = V_{cr1} - \frac{V_o}{2n} \tag{5}$$

$$T_{10} = \frac{T_s}{2} - T_{42} - T_{21} \tag{6}$$

And voltage V_{ab} clamped to zero at the end of this interval.

Mode 2 (t_1-t_2) *(refer Fig. 3(b))*: During this mode, S_3 remains conducting while S_1 is turned on at instant t_1. The parallel device capacitor discharges rapidly through the switch S_1. Consequently, incoming device current i_{S1} increases whereas the current through S_3 decreases linearly. At the same time, resonant current through phase *A* (i_{Lr1}) and phase *C* (i_{Lr3}) decreases and increases respectively. Power transfer occurs through rectifier diodes D_1, D_2, and D_3. The dynamic equations representing this mode are:

$$i_{Lr1}(t) = \frac{I_{in}}{3} - \frac{1}{2L_r}\left(\frac{V_o}{n} + 2V_{crp}\right)(t - t_1)\ ,\ i_{Lr2}(t) = \frac{I_{in}}{3} \tag{7}$$

$$i_{Lr3}(t) = \frac{1}{2L_r}\left(\frac{V_o}{n} + 2V_{crp}\right)(t - t_1) - \frac{2I_{in}}{3} \tag{8}$$

$$i_{S1}(t) = \frac{1}{2L_r}\left(\frac{V_o}{n} + 2V_{crp}\right)(t - t_1) \tag{9}$$

$$i_{S2}(t) = 0 \tag{10}$$

$$i_{S3}(t) = I_{in} - \frac{1}{2L_r}\left(\frac{V_o}{n} + 2V_{crp}\right)(t - t_1) \tag{11}$$

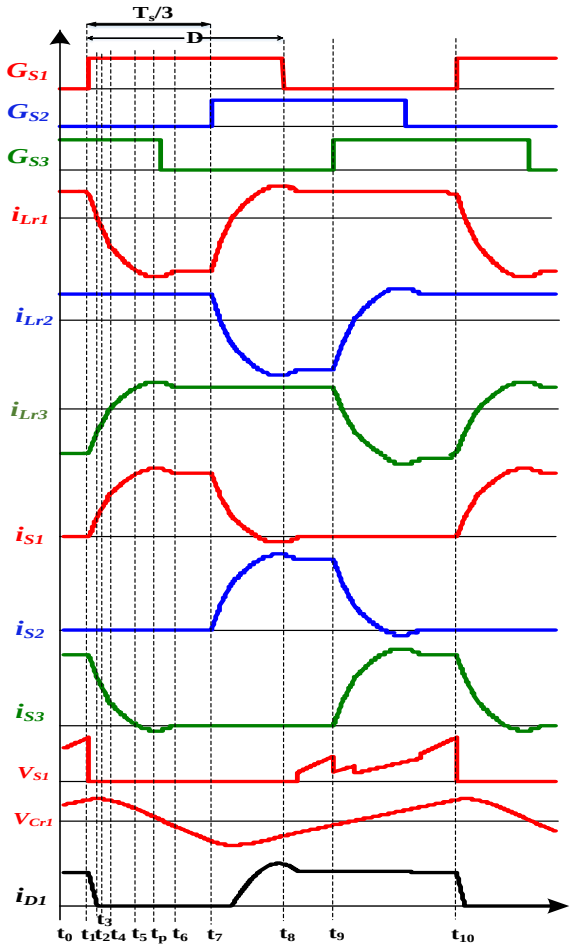


Fig. 2. Steady state operating waveforms of the proposed converter topology.

$$T_{21} = \frac{1}{\omega_r} sin^{-1}\left(\frac{I_{in}Z_o}{\frac{V_o}{n} + V_{cr}(t_2)}\right),\ V_{cr}(t_2) = V_{crp} \tag{12}$$

At the end of this mode, resonant current i_{Lr1} linearly reduces to zero and resonant capacitor C_{r1} reaches its peak value.

Mode 3 (t_2-t_3) *(refer Fig. 3(c))*: In this mode, S_3 and S_1 remains ON while diode D_1 turns off transferring current to another branch. Series resonance initiates between leakage inductance (L_{r1}+ L_{r3}) and C_{req} causing sinusoidally changing current i_{S1} and i_{S3} with opposite polarity. Resonance occurs at frequency f_r with characteristics impedance Z_r. Equations governing this mode are given below:

$$f_r = \frac{1}{2\pi\sqrt{L_{req}C_{req}}}, \quad Z_r = \sqrt{\frac{L_{req}}{C_{req}}} \tag{13}$$

$$where \quad L_{req} = L_{r1} + L_{r3} \ and \ C_{req} = \frac{C_{r1} * C_{r2}}{C_{r1} + C_{r2}} \tag{14}$$

$$i_{Lr1}(t) = -\left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r\,(t - t_2) \tag{15}$$

$$i_{Lr2}(t) = \frac{I_{in}}{3} \tag{16}$$

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

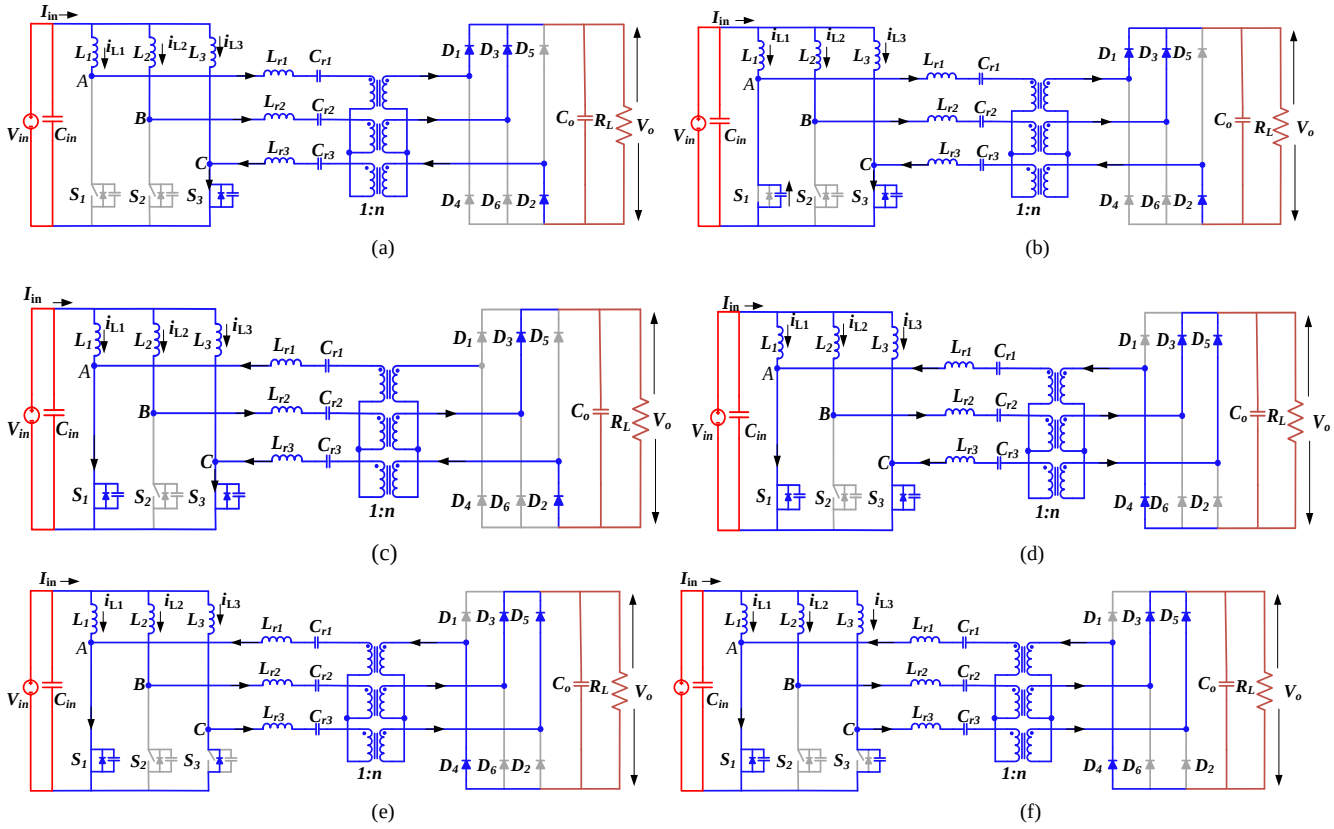


Fig. 3. Equivalent circuit for operating modes of the proposed three-phase configuration.

$$i_{Lr3}(t) = \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_2) - \frac{I_{in}}{3} \quad (17)$$

$$i_{S1}(t) = \frac{I_{in}}{3} + \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_2) \quad (18)$$

$$i_{S2}(t) = 0 \quad (19)$$

$$i_{S3}(t) = \frac{2I_{in}}{3} - \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_2) \quad (20)$$

$$T_{32} = \frac{1}{\omega_r} sin^{-1}\left(\frac{I_{in}Z_r}{6\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) \quad (21)$$

Mode 4 (t_3-t_4) *(refer Fig. 3(d))*: During this mode resonance continues with leakage energy circulating through branch *A* and *C*. During this period i_{Lr3} increases to zero. Resonant capacitor voltage V_{Cr1} starts to decrease due to negative current through L_{r1}. Dynamic equations describing this mode are:

$$i_{Lr1}(t) = -\frac{I_{in}}{6} - \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_3) \quad (22)$$

$$i_{Lr2}(t) = \frac{I_{in}}{3} \quad (23)$$

$$i_{Lr3}(t) = -\frac{I_{in}}{6} + \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_3) \quad (24)$$

$$i_{S1}(t) = \frac{I_{in}}{2} + \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_3) \quad (25)$$

$$i_{S2}(t) = 0 \quad (26)$$

$$i_{S3}(t) = \frac{I_{in}}{2} - \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_3) \quad (27)$$

$$T_{43} = \frac{1}{\omega_r} sin^{-1}\left(\frac{I_{in}Z_r}{6\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) \quad (28)$$

Mode 5 (t_4-t_5) *(refer Fig. 3(d))*: In this mode, switch current i_{S1} rises sinusoidally in positive direction and reaches I_{in} whereas i_{Lr3} reaches at $\frac{I_{in}}{3}$ level. Towards the end of this interval, switch current i_{S3} naturally reduces to zero due to resonant characteristics. Expressions describing this mode are:

$$i_{Lr1}(t) = -\frac{I_{in}}{3} - \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_4) \quad (29)$$

$$i_{Lr3}(t) = \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_4) \quad (30)$$

$$i_{S1}(t) = \frac{2I_{in}}{3} + \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right)\sin\omega_r(t - t_4) \quad (31)$$

$$i_{S2}(t) = 0 \quad (32)$$

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

$$i_{S3}(t) = \frac{I_{in}}{3} - \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right) \sin \omega_r (t - t_4) \quad (33)$$

$$T_{54} = \frac{1}{\omega_r} \sin^{-1}\left(\frac{I_{in} Z_r}{3\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) \quad (34)$$

Mode 6 (t_5-t_6) *(refer Fig. 3(e))*: In this mode, switch current i_{S3} further crosses zero in a sinusoidal fashion allowing anti-parallel diode (e.g., Intrinsic diode) to conduct. The internal diode of switch S_3 remains conducting during this interval therefore, the gating signal is forced-off during this period to achieve successful ZCS condition. At the end of this interval i_{S3} again reaches zero and the switch is blocked by a positive voltage across its terminals.

$$T_{65} = \frac{1}{\omega_o}(\pi - T_{52}) \quad (35)$$

Mode 7 (t_6-t_7) *(refer Fig. 3(f))*: This mode begins with ZCS turn-off of switch S_3 eliminating overshoot in the device voltage. The device snubber capacitor gets charged instantly and the switch voltage increases linearly owing to the resonant capacitor charge.

After this mode, the charging and discharging of the series resonant capacitors take place in two different phases with constant current flowing through them. Therefore, duration of this mode can be considered as: $T_{76} = T_{10} = \frac{T_s}{3} - T_{61}$. As a result, the analysis repeats after $\frac{T_s}{3}$ interval in the same sequence for the remaining part of the switching cycle.

III. Converter Design

This section elaborates the systematic design approach to interface low voltage sources like fuel-cell, solar panel etc. to a high voltage dc bus rated at 380V. Critical design parameters are obtained to ensure still output voltage and soft switching against fluctuations in input voltage and load changes.

A. DC voltage gain

Voltage gain (M) relation with converter parameters helps in determining the desired operating point to regulate 380V at the output for different loading condition. Fig.4 depicts the operating region of the converter for different combination of switching frequency and turn-ratio. The appropriate operating point (f_{sw}, n) also assists in keeping minimized losses for the given specifications. Assuming lossless power transfer to the load, input to output voltage ratio is obtained by analysing one of the three phases considering operation symmetry.

$$M = \frac{V_o}{V_{in}} = \frac{n\left(1 - \frac{3Xf_n r_n n(1-k^2)}{\pi}\right)}{\left(\frac{2}{3} - f_n\left(\frac{1}{2} + \frac{M}{r_n X n}\right)\right)} \quad (36)$$

Where,

n: Turn ratio of HFT, $k = \frac{M}{n r_n X}$;

f_n: Normalized frequency, $= \frac{f_{sw}}{f_r}$;

r_n: Normalized load resistance, $= \frac{R_{FL}}{n^2 Z_o}$;

X: Constant defined as, $= \left(1 + \frac{2nV_{crp}}{V_o}\right)$;

Where V_o, V_{crp}, R_{FL} represent output voltage, peak resonant capacitor charge and equivalent full-load resistor respectively.

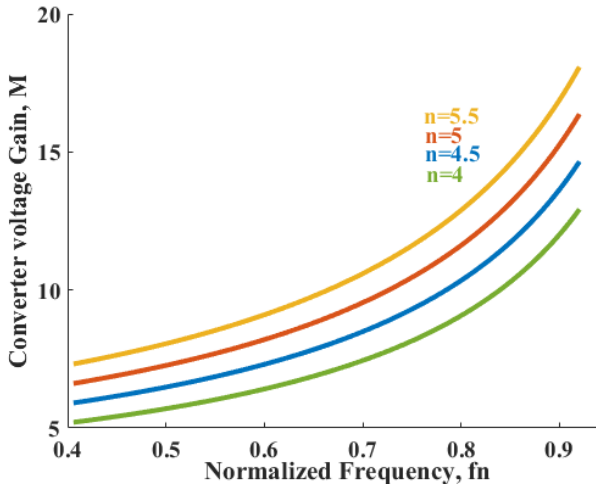


Fig. 4. Converter voltage gain characteristics curve.

As a result, constant output voltage can be assured by varying converter switching frequency for all possible extreme operating points.

B. Voltage and current stress in the converter

The peak voltage stress across the primary side MOSFETs are influenced by resonant capacitor charge and HF transformer turns ratio whereas output voltage is reflected across rectifier diodes.

$$V_{sw1-3,rated} = V_{crp} + \frac{V_o}{n} \quad (37)$$

$$V_{D1-6} = V_o \quad (38)$$

Current rating of the semiconductor devices is decided based on the rms current through these components. The approximated expressions for the rms values are given as:

$$I_{sw1-3,rms} = I_{in}\sqrt{\left(\frac{1}{3} - f_n\left(\frac{4}{9} - \frac{1}{2}\left(\frac{nXr_n}{M}\right)^2 + \frac{7}{9}\frac{M}{2nXr_n}\right)\right)} \quad (39)$$

$$I_{D1-6,avg} = I_o/3 \quad (40)$$

C. Selection of Transformer Turns Ratio

Transformer turns ratio assists in meeting voltage gain requirement while keeping lower conduction losses in the primary. HF transformer turns ratio significantly influences rms current and voltage stresses associated with the semiconductor switches given by (37) and (39). Semiconductor devices experience higher voltage stress for a lower turns ratio value. This in turn necessitates high rating devices with high ON-state resistance causing more conduction losses. Whereas transformer with high turns ratio requires larger core area and increased copper, leading to high transformer losses. Consequently, selection of turn ratio involves a trade-off between switch voltage rating and conduction losses. Considering efficiency as a prominent factor, a lower value of turns ratio is advisable to have lower component ratings with limited conduction losses.

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

D. Critical condition for ZCS

The resonant tank energy dictates the soft switching range by providing adequate energy for body-diode conduction before switch turn-off. This is possible only when the resonant current peak in all three phases exceed $I_{in}/3$ value during the switch overlap period. However, the resonant currents in all three phases are inversely related to Z_r which implies higher current stress for lower values of Z_r. Selection of Z_r implicates a compromise between ZCS operating boundary and peak current stress. Therefore, maximum value of Z_r can be estimated using the inequality given below:

$$i_{Lr3}(t) > \frac{I_{in}}{3} \quad \Rightarrow \left(\frac{V_o/n + 2V_{crp}}{Z_r}\right) > \frac{I_{in}}{3} \tag{41}$$

$$\Rightarrow Z_{rmax} < 3\left(\frac{V_o/n + 2V_{crp}}{I_{in}}\right) \tag{42}$$

As a result, a value smaller than Z_{rmax} has been chosen for this design example.

E. Duty Ratio

As the resonant period dictates the soft-switching boundary, appropriate duty ratio must be chosen to provide necessary overlap between the switches. Therefore, the resonant pulse must sustain until the semiconductor device turns-off. The desired operating boundary for the duty ratio can be given as.

$$D_{min} < D < D_{max} \tag{43}$$

The minimum and maximum duty ratios can be obtained as;

$$D_{min} = 0.33 + \frac{2I_{in}L_r f_{sw}}{3\left(\frac{V_o}{n} + 2V_{crp}\right)} + \frac{f_{sw}}{\pi f_r} sin^{-1}\left(\frac{I_{in}Z_r}{6\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) + \frac{f_{sw}}{2\pi f_r} sin^{-1}\left(\frac{I_{in}Z_r}{3\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) \tag{44}$$

$$D_{max} = 0.33 + \frac{2I_{in}L_r f_{sw}}{3\left(\frac{V_o}{n} + 2V_{crp}\right)} + \frac{f_{sw}}{2\pi f_r} sin^{-1}\left(\frac{I_{in}Z_r}{6\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) + \frac{f_{sw}}{2\pi f_r} sin^{-1}\left(\frac{I_{in}Z_r}{3\left(\frac{V_o}{n} + 2V_{crp}\right)}\right) + \frac{f_{sw}}{2f_r} \tag{45}$$

Therefore, to ensure ZCS for the entire range of source voltage and load, a duty ratio must lie within this rage.

F. Design of the series resonant circuit

Resonant tank impedance Z_r and resonant frequency f_r, are chosen minimum input and rated load conditions. Considering resonant frequency f_r as 200 kHz, and Z_r = 9.3 Ω, the computed L_r and C_r values are 3.5 μH and 160 nF respectively.

G. Boost inductor calculation

In the proposed three-phase configuration, three interleaved inductors are utilized at the input. Each inductor carries 1/3rd of total source current. Therefore, for a 1kW rated power and input voltage of 42V, a maximum current through each inductor is 8A at a peak efficiency of 97%. Hence, for a permissible current ripple of 10%, inductor value is calculated as 230μH. Design parameters calculated are listed in Table II.

TABLE II: DESIGN PARAMETERS

Parameter	Value
Resonant Inductor, $L_{r1\text{-}3}$	3.5μH
Resonant Capacitor, $C_{r1\text{-}3}$	160nF
Input Inductor, L_1 L_2 L_3	200μH
Output Capacitor, C_o	100μF
Turns ratio, n	4

TABLE III: CONVERTER SPECIFICATIONS

Parameter	Value
Input Voltage, V_{in}	42-48V
Output voltage, V_o	380 V
Maximum output power, P_o	1kW
Switching frequency, f_{sw}	80-150 kHz
Duty ratio, D	0.48-0.55

IV. EXPERIMENTAL RESULTS

This section exhibits comprehensive steady-state results obtained from laboratory prototype as shown in Fig. 5 built for converter specifications listed in Table III to verify ZCS soft turn-off and device voltage clamping. Table IV presents the component specifications used for the development of the hardware prototype. Digital simulation and experimental analysis of the proposed converter have been carried out to validate the expected claims for source voltage range of 42V-48V with load variation from full load down to 20%. The proposed dc-dc converter allows conversion from low-voltage ranging between 42 V to 48 V into 380V dc with a wide voltage-gain. The converter employs high-speed MOSFETs with inherent reverse-blocking capability; SiC Schottky diodes for output diode bridge to avail minimum reverse recovery loss; three single-phase high-frequency transformers with 1:4 turns ratio which are realized using high-permeability ferrite E-cores and litz wire. Open loop control is executed using a TMS320F28335 digital signal processor.

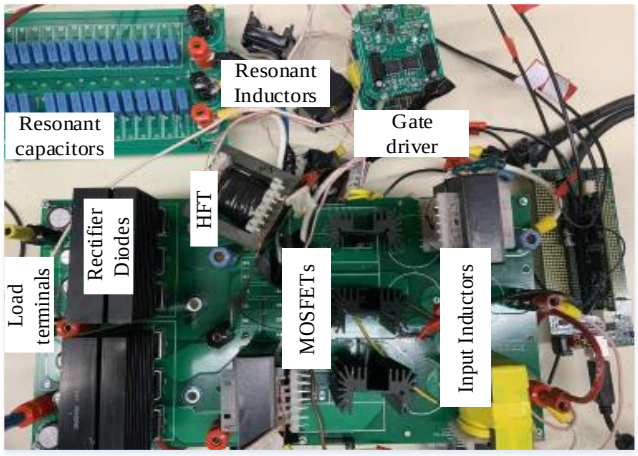


Fig. 5. Experimental prototype of the proposed converter.

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

TABLE IV
COMPONENT SPECIFICATION OF THE LABORATORY PROTOTYPE

Component	Specifications
Boost inductors L_1, L_2, L_3	55 x 28 x 21 EE ferrite Core, 300μH
Primary Switches, S_1, S_2, S_3	SCT3060ALGC11 (650V, 39A,60mΩ)
Resonant Inductor, $L_{r1\text{-}3}$	EE ferrite core, 1.48μH
Resonant Capacitors, $C_{r1\text{-}3}$,	160nF 1kV Film capacitor
3 Single phase HF Transformer	EE ferrite core, Primary turns, N_1=10, secondary turns N_2=50, L_{lk}= 1.64μH
Secondary Diodes $D_{1\text{-}6}$	STPSC20065D (650V, 20A)
Output Capacitor, C_o	80μF 450V electrolytic capacitor, 10nF film capacitor
Gate driver IC	HCNW3120
Digital Controller	TMS230F28335 (Texas Instruments DSP)

Fig. 6(a) shows the measured input current I_{in} and three boost inductors current $I_{L1,}$ $I_{L2,}$ I_{L3} with nearly 10% ripple current for V_{in}= 42V at full-load condition. It is also highlighted that interleaving legs with 120° phase shifted currents improve the quality of overall input power as the input current ripple remains reasonably low having ripple frequency 3*f_{sw}. Moreover, the interleaved legs reduce the component stress and increase the power handling capacity of the converter. Fig. 6(b) depicts the series resonant current in phase A, phase B and phase C which are shifted by 120°. Fig. 6(c) demonstrates smooth current commutation or ZCS of switch S_2 during overlap period. The gating signal for the targeted switch is removed once the current through it becomes zero, and consequently, the switch voltage (V_{ds2}) linearly ramps up to V_o/n. Successful ZCS operation was possible due to sinusoidally varying transformer current which eventually resulted in sinusoidal current through the switch ensuring body diode conduction before the switch turn-off instant. Moreover, during the overlap period the switch voltage drops to a lower value due to simultaneous charging/discharging of the series resonant capacitor in two branches. In addition, the resonant current higher than I_{in}/3 mark appears during this overlap time ($t_{overlap} \approx 1.11\mu s$), allows antiparallel body diode conduction and effective ZCS turn-off as shown in Fig. 6(c). It is noted that the peak switch voltages are limited to ≈220V and a resonant peak of 11.2A flows through each phase for nominal converter voltage of 42V. Similarly, ZCS operation of other two switches occur 120◦ and 240° after the targeted switch S_2 is turned off. Therefore, overvoltage spike in the switch voltage is prevented, avoiding any axillary clamping or snubbing requirement.

Fig. 7(a) highlights the current profiles at the input side along with 120° shifted boost inductors having current ripple of ≈ 0.5A for V_{in} = 42V at light-load condition. Fig. 7(b) depicts the phase shifted resonant current in phase A, phase B and phase C with short duration resonant pulse. Fig. 7(c) illustrates zero current commutation and voltage spike elimination of the primary switches at 200W power with V_{in} = 42V and switching frequency 90kHz. Lower switch voltages are witnessed for part load condition due to lower resonant capacitor charge and low input current. Moreover, owing to the load adaptive series resonant energy, reasonably low $I_{Lrp}/(I_{in}/3)$ ratio is observed at light load condition leading to improved system efficiency.

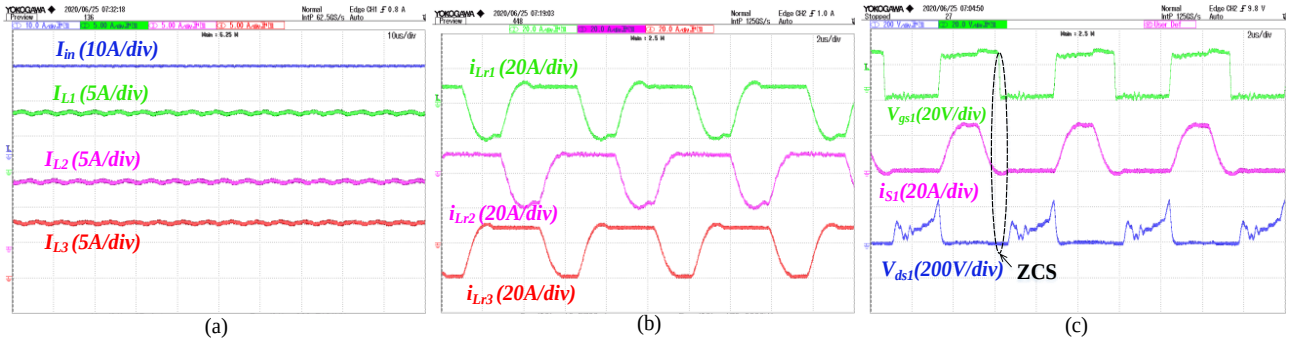
The experimentation is also performed with 48V source voltage at two extreme load scenarios i.e., full-load and light-load condition respectively in Fig. 8 and Fig. 9 to comply with applications requiring variable source voltage and load. The proposed claims hold true under these conditions as well confirming wide range ZCS and voltage clamping across semiconductor switches. However, slightly higher peaks in switch current and higher $I_{Lrp}/(I_{in}/3)$ ratio are observed for higher source voltages at different load points. Also, a swell in in resonant capacitor voltage is observed for higher source voltages leading to relatively higher switch voltage stress. Fig. 10 demonstrates the output rectifier diodes waveforms at turn off instant with zero current. Moreover, use of ultra-fast diodes prevents reverse recovery loss and voltage ringing, further eliminating the need for dissipative snubbing on the secondary side. The experimental results demonstrate wide range ZCS and device voltage clamping encompassing full-load to light-load (200W) condition with devices operating at high switching frequency. It should also be noticed that the circulating energy at rated load is limited by the proposed converter design resulting in lower losses. Bringing down the current stresses due to interleaved legs at the source along with ZCS turn-off and reduced semiconductor count, the overall efficiency has been improved making the converter well suited for low voltage high current applications.

Fig. 11 depicts practical switching frequency band of 85-150kHz for operating range of load and source, thereby allowing HF magnetics and smaller filter requirement. However, a major fluctuation in switching frequency is witnessed for load variations compared to source voltage variations, which makes the converter more suitable for sources with wide input range. It can be inferred that the change in f_{sw} is not very substantial for input voltage variations. Hence, variable frequency operation is indispensable to maintain regulated voltage, i.e. $V_o \approx$ 380V at load terminals under different scenarios.

The findings in Fig. 12 elucidate experimental efficiency trend for load variation. The recorded peak efficiency is 96.7% which is above par with the state-of-the-art dc-dc topologies. Also, the experimental efficiency under light-load is recorded to be 94.5%. The converter efficiency can be enhanced greatly by optimizing component selection, gate drive circuit and printed circuit board for the hardware prototype. Fig. 13. depicts the converter power loss distribution graph for different load condition at input voltages. From the graph it can be inferred that majority of the losses occur in semiconductor devices and boost inductors covering more than 50% of the total losses. However, device intrinsic characteristic parameters e.g. $R_{ds,on}$ and C_{oss} influence conduction and switching losses. Therefore, selecting a MOSFET with lower $R_{ds,on}$ is conducive for maintaining higher efficiency with reduced conduction losses. moreover, in order to further enhance converter efficiency, ultra-fast SiC diodes with low V_{fd} and low Q_{rr} are to be utilized for output rectifier.

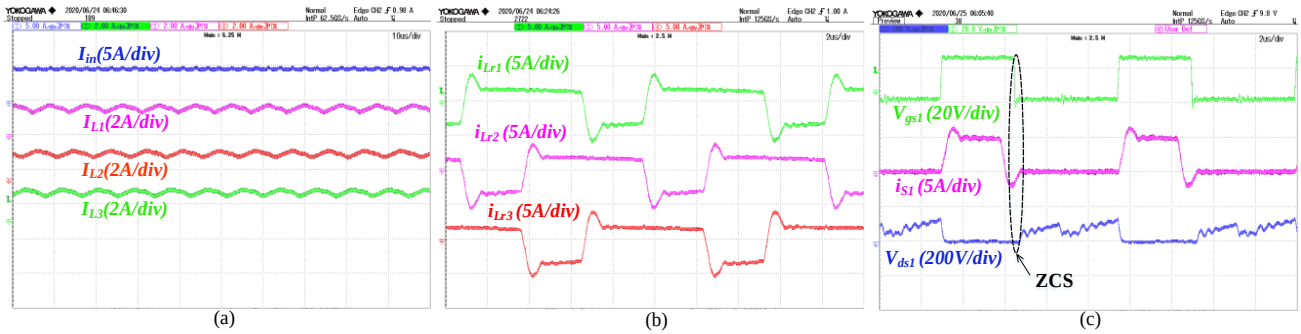
2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics



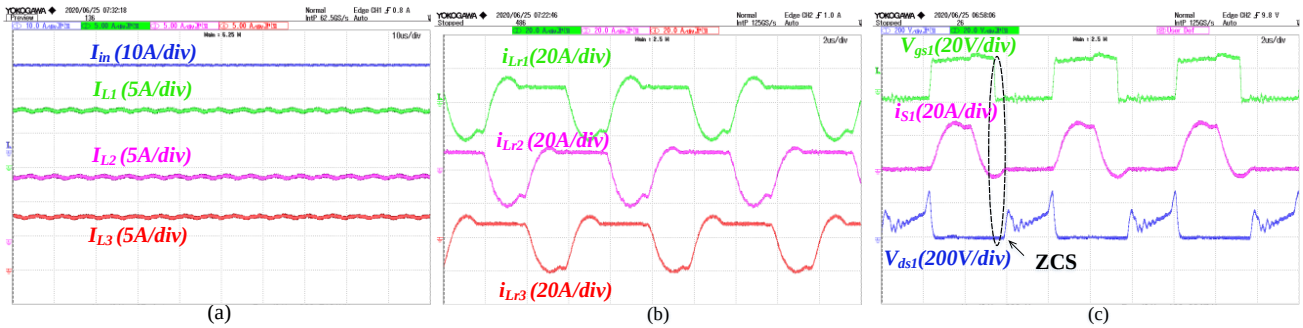
(a) (b) (c)

Fig. 6. Experimental steady-state waveforms for V_{in} = 42V at full-load (1000W) (a) Input current I_{in}, boost inductor currents I_{L1}, I_{L2}, I_{L3} (time scale: 10μs/div) (b) series resonant current through phase A, B and C i_{Lr1}, i_{Lr2}, i_{Lr3} (time scale: 2μs/div) and (c) Gate-source voltage V_{GS1}, switch current i_{Sw1}, and drain-to-source voltage V_{ds1} (time scale: 5μs/div).



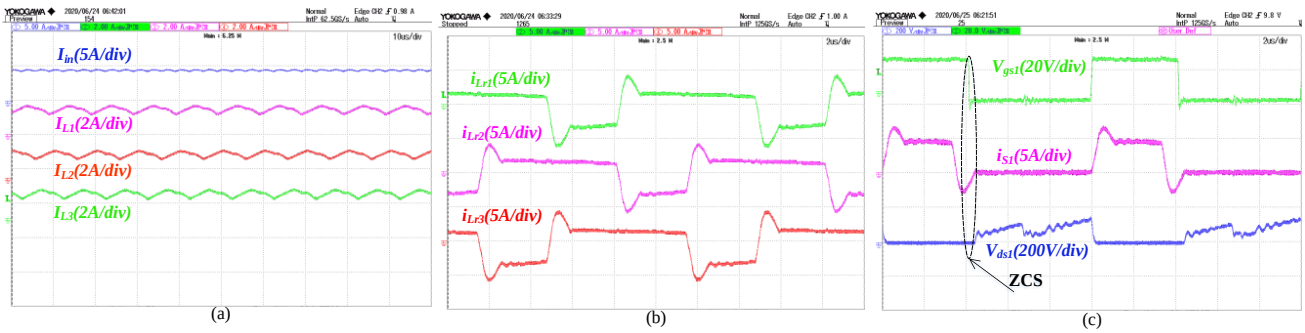
(a) (b) (c)

Fig. 7. Experimental steady-state waveforms for V_{in}= 42V at light-load (200W) (a) Input current I_{in}, boost inductor currents I_{L1}, I_{L2}, I_{L3} (time scale: 10μs/div) (b) series resonant current through phase A, B and C i_{Lr1}, i_{Lr2}, i_{Lr3} (time scale: 2μs/div) and (c) Gate-source voltage V_{GS1}, switch current i_{Sw1}, and drain-to-source voltage V_{ds1} (time scale: 5μs/div).



(a) (b) (c)

Fig. 8. Experimental steady-state waveforms for V_{in} = 48V at full-load (1000W) (a) Input current I_{in}, boost inductor currents I_{L1}, I_{L2}, I_{L3} (time scale: 10μs/div) (b) series resonant current through phase A, B and C i_{Lr1}, i_{Lr2}, i_{Lr3} (time scale: 2μs/div) and (c) Gate-source voltage V_{GS1}, switch current i_{Sw1}, and drain-to-source voltage V_{ds1} (time scale: 5μs/div).



(a) (b) (c)

Fig. 9. Experimental steady-state waveforms for V_{in} = 48V at light-load (200W) (a) Input current I_{in}, boost inductor currents I_{L1}, I_{L2}, I_{L3} (time scale: 10μs/div) (b) series resonant current through phase A, B and C i_{Lr1}, i_{Lr2}, i_{Lr3} (time scale: 2μs/div) and (c) Gate-source voltage V_{GS1}, switch current i_{Sw1}, and drain-to-source voltage V_{ds1} (time scale: 5μs/div).

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

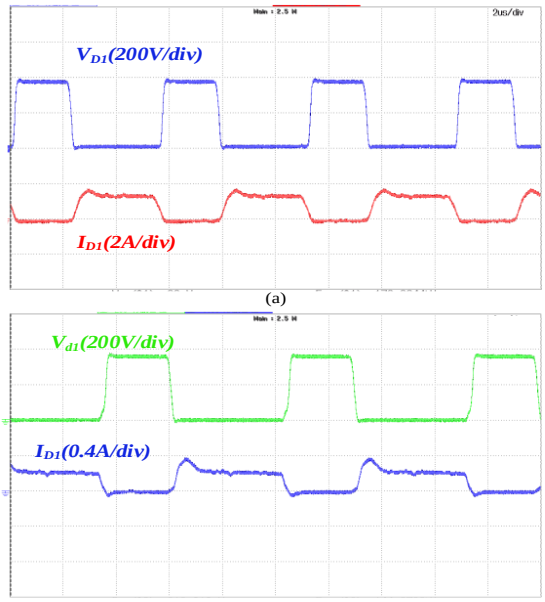


Fig. 10. Experimental results for voltage across and current through the rectifier diode for V_{in} = 42V (a) full-load and (b) light-load condition.

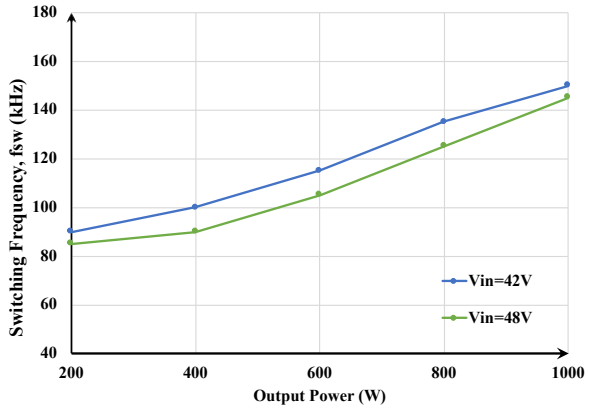


Fig. 11. Switching frequency graph with output power variation.

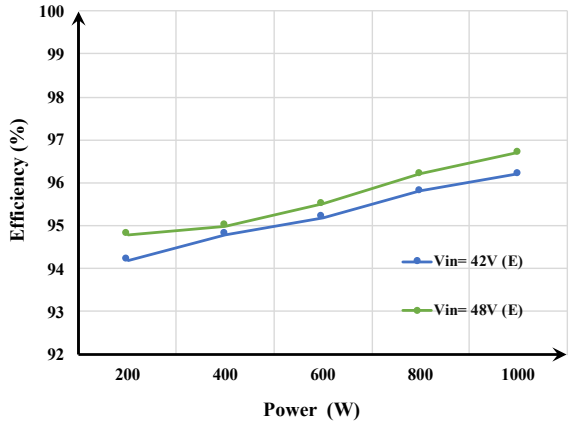


Fig. 12. Experimental efficiency trend with output power variation.

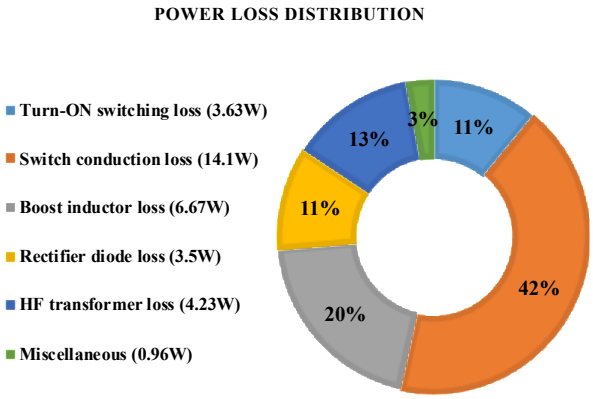


Fig. 13. Loss distribution for V_{in}=42V at rated power.

V. Conclusion

A series resonant-pulse commutated current-fed three-phase current-sharing topology has been studied and analyzed. Interleaving legs at the input facilitate low current through the devices leading to reduced conduction and switching losses. Series resonance-pulse results in ZCS turn-off of the semiconductor devices and clamps the turn-off voltage spike across the devices, thereby avoiding any snubber requirement. Moreover, load adaptive resonant energy regulates and maintains low circulating energy and low current peak under all loading conditions considerably improving the converter efficiency with reduced thermal requirement. In addition, higher switching frequency operation leads to reduced volume and light weight system. Comprehensive design of the converter with experimental verification has been provided. With merits such as high voltage gain, ripple-free source current, reduced passive/active component count and modular high-frequency isolation, the proposed converter is a promising candidate for interfacing unidirectional sources like fuel-cells, solar panel and energy storage to utility grid or DC microgrid. This topology is studied and analyzed first time with the series resonance pulse concept. The theory reported for the topology and the experimental results are new and original demonstrating the expected merits.

References

[1] J. Wang, F. Z. Peng, J. Anderson, A. Joseph and R. Buffenbarger, "Low Cost Fuel Cell Converter System for Residential Power generation," *IEEE Transactions on Power Electronics*, vol. 19, No. 5, pp. 1315-1322, 2004.

[2] M. Harfman Todorovic, L. Palma and P. Enjeti, "Design of a wide input range DC-DC converter with a robust power control scheme suitable for fuel cell power conversion," in *Proc. IEEE Applied Power Electronics Conference and Exposition*, 2004, pp.374-379.

[3] F. Blaabjerg, Z. Chen, and S. B. Kjaer, "Power electronics as efficient interface in dispersed power generation systems," *IEEE Trans. Power Electron.*, vol. 19, no. 5, pp. 1184–1194, Sep. 2004.

[4] G. Fontes, C. Turpin, S. Astier, and T. A. Meynard, "Interactions between fuel cell and power converters: Influence of current harmonic on a fuel cell stack," *IEEE Trans. Power Electron.*, vol. 22, no. 2, pp. 670–678, Mar. 2007.

[5] L. D. Salazar and P. D. Ziogas, "Design oriented analysis of two types of three-phase high frequency forward SMR topologies," in *Proc. 5th Annu. IEEE APEC*, Mar. 11–16, 1990, pp. 312–320.

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

[6] L. Changrong, A. Johnson, and J.-S. Lai, "A novel three-phase high-power soft-switched DC/DC converter for low-voltage fuel cell applications," *IEEE Trans. Ind. Appl.*, vol. 41, no. 6, pp. 1691–1697, Nov./Dec. 2005.

[7] M. Mohr and F. W. Fuchs, "Voltage Fed and Current Fed Full Bridge Converter for the Use in Three Phase Grid Connected Fuel Cell Systems," *IEEE 5th International Power Electronics and Motion Control Conference*, Shanghai, 2006, pp. 1-7.

[8] R. L. Andersen and I. Barbi, "A three-phase current-fed push-pull DC-DC converter," *IEEE Trans. Power Electron.*, vol. 24, no. 2, pp. 358–368, Feb. 2009.

[9] A. K. Rathore, A. K. S. Bhat, R. Oruganti, "A comparison of soft-switched DC–DC converters for fuel-cell to utility-interface application", *Proc. IEEE Power Convers. Conf.*, pp. 588-594, 2007.

[10] A. K. Rathore and U. R. Prasanna, "Comparison of soft-switching voltage-fed and current-fed bi-directional isolated dc/dc converters for fuel cell vehicles*," in Proc. IEEE Int. Symp. Ind. Electron*., 2012, pp. 3212–3219.

[11] J. Zhang, J. S. Lai, R.-Y.Kim, and W.Yu,"High-power density design of a soft-switching high-power bidirectional DC–DC converter," *IEEE Trans. Power Electron.*, vol. 22, no. 4, pp. 1145–1153, Jul. 2007.

[12] R. Watson and F. C. Lee, "A soft-switched, full-bridge boost converter employing an active-clamp circuit," *in Proc. IEEE Power Electron. Spec. Conf.*, 1996, pp. 1948–1954.

[13] V. Yakushev, V. Meleshin, and S. Fraidlin, "Full-bridge isolated current fed converter with active clamp," *in Proc. 14th IEEE Appl. Power Electron. Conf. Expo*., 1999, pp. 560–566.

[14] J.-T. Kim, B.-K. Lee, T.-W. Lee, S.-J. Jang, S.-S. Kim, and C.-Y. Won, "An active clamping current-fed half-bridge converter for fuel-cell generation systems," *in Proc. IEEE Power Electron. Spec. Conf*., 2004, pp. 4709– 4714.

[15] S.-K. Han, H.-K. Yoon, G.-W. Moon, M.-J. Youn, Y.-H. Kim, and K.-H. Lee, "A new active clamping zero-voltage switching PWM current fed half-bridge converter," *IEEE Trans. Power Electron.*, vol. 20, no. 6, pp. 1271–1279, Nov. 2005.

[16] H. Cha, J. Choi, and P. N. Enjeti, "A three-phase current-fed DC/DC converter with active clamp for low-DC renewable energy sources," *IEEE Trans. Power Electron.*, vol. 23, no. 6, pp. 2784–2793, Nov. 2008.

[17] Y. Song et al., "A current-fed three-phase half-bridge dc-dc converter with active clamping," 2009 IEEE Energy Conversion Congress and Exposition, San Jose, CA, USA, 2009, pp. 1362-1366, doi: 10.1109/ECCE.2009.5316124.

[18] P. Xuewei, A. K. Rathore, "Novel Bidirectional snubberless naturally commutated soft-switching current-fed full bridge isolated DC/DC Converter for fuel cell vehicles", *IEEE Trans. Ind. Electro.*, vol. 61, no. 5, pp. 2307-2315, May 2014.

[19] P. Xuewei and A. K. Rathore, "Naturally clamped soft-switching current- fed three-phase bidirectional DC/DC converter," *IEEE Trans. Ind. Elec- tron.*, vol. 62, no. 5, pp. 3316–3324, May 2015.

[20] S. V. G. Oliveira and I. Barbi, "A Three-phase step-up DC–DC converter with a three-phase high-frequency transformer for DC renewable power source applications," *IEEE Trans. Ind. Electron.*, vol. 58, no. 8, pp. 3567–3580, Aug. 2011.

[21] Z. Wang and H. Li, "Soft switching three-phase current-fed bidirectional DC-DC converter with high efficiency over a wide input voltage range," *IEEE Trans. Power Electron.*, vol. 27, no. 2, pp. 669–684, Feb. 2012.

[22] K. Modepalli, R. Suryadevara and L. Parsa, "High-frequency isolated DC-DC converter for offshore wind energy systems," 2016 IEEE Energy Conversion Congress and Exposition (ECCE), Milwaukee, WI, 2016, pp. 1-6, doi: 10.1109/ECCE.2016.7855160.

[23] K. R. Sree, A. K. Rathore, "Impulse commutated zero current switching current-fed three-phase DC/DC converter", *IEEE Trans. Ind. Appl.*, vol. 52, no. 2, pp. 1855-1864, Mar. 2016.

[24] K. R. Sree and A. K. Rathore, "Analysis and design of impulse-commutated zero-current-switching single-inductor current-fed three-phase push–pull converter*," IEEE Trans. On Ind. Appl., vol. 53,s no. 2, Mar./Apr. 2017.*

[25] S. Tandon and A. K. Rathore, "Novel Series LC Resonance-Pulse-Based ZCS Current-Fed Full-Bridge DC–DC Converter: Analysis, Design, and Experimental Results," *in IEEE Transactions on Power Electronics*, vol. 36, no. 2, pp. 1844-1855, Feb. 2021.

[26] S. Tandon, A. K. Rathore and J. Rodriguez, "Partial Series Resonance-Pulse Assisted Zero-Current-Switching Current-Fed Three-phase Current Sharing DC-DC Converter," *2020 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, Jaipur, India, 2020, pp. 1-6.

APPENDIX

A. *Voltage Gain Formulation*

The average current expression in one of the three identical phases considering operation symmetry is computed using steady state waveforms shown in Fig. 2 and relevant resonant current expressions reported in Section II. Considering, the sum of the average resonant inductor current in three phases equal to the primary referred average load current.

$$3I_{Lr1,avg} = nI_o \Rightarrow \frac{I_{Lr1,avg}}{n} = \frac{V_o}{3R_{FL}}$$

The average resonant inductor current in the can be computed by utilizing time interval information from Section II. as,

$$I_{Lr1,avg} = \frac{1}{T_s}\int_0^{T_s} i_{Lr1}(t)\ dt$$

$$\frac{I_{Lr1,avg}}{n} = \frac{f_s}{n}\left[\int_{t_4}^{t_6} i_{Lr1}(t)\ dt + \int_{t_6}^{t_9} I_{in}\ dt + \int_{t_9}^{t_{10}} i_{Lr1}(t)\ dt\right]$$

Expression for $I_{\mathrm{Lr1,avg}}$ and assuming lossless power transfer, the input and output relationship in terms of converter parameters like V_{o}, V_{in} and load R_{fl} has been formulated with necessary assumptions.

B. *Converter Loss calculation*

TABLE A.1: LOSS FORMULAS TO COMPUTE THE CONVERTER LOSSES

Description	Equation
Switch conduction loss	$I_{sw,rms}{}^2 * R_{Ds,on}$
Switch turn-on loss	$\frac{1}{2} C_{oss} V_{sw}{}^2 f_{sw}$
Overall rectifier diode loss	$V_{Df} * I_{D,avg} + I_{D,rms}{}^2 * R_D + V_d\ Q_{rr}\ f_{sw}$
HF Transformer loss	$\left(I_{Lr,rms}{}^2 * R_{winding} + P_{C,limit}\ V_e\right)$
Boost inductor loss	$\left(I_{L1}{}^2 * R_{L,DC} + P_{C,limit}\ V_e\right)$
Gate drive loss	$C_{iss} V_g^2 f_{sw}$
Capacitor loss	$I_{cout,rms}{}^2 * R_{ESR}$

where

$I_{sw,s}$ = Switch rms current, A, $R_{DS,n}$ = Switch drain-to-source on-state resistance, Ω, V_{sw}, = Switch average voltage during turn-off, V

f_{sw} = Switching frequency, Hz, C_{oss} = Switch output capacitance, F

$I_{D,s}$ = Diode rms current, A, R_D = Diode turn-on resistance, Ω

$I_{D,g}$ = Diode average current, A, V_{Df} = Diode forward voltage, V

$R_{winding}$ = transformer dc resistance, Ω, $I_{L,s}$ = inductor rms current, A

$R_{L,dc}$ = Inductor DC resistance, Ω, $I_{Cout,ms}$ = Capacitor ripple rms current, A, R_{ESR} = Capacitor equivalent series resistance, Ω

$P_{c,mit}$ = core loss limit, mW/cm^3, V_e = Effective core volume, cm^3

C_{iss} = Switch input capacitance, F, f_{sw} = Switching frequency, Hz

V_g = Switch driving voltage, V, V_d = Diode blocking voltage, V

Q_{rr}=Diode reverse recovery charge.

Acknowledgement

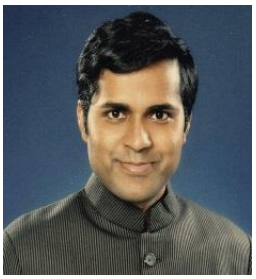
"J. Rodriguez acknowledges the support of ANID through projects FB0008, ACT192013 and 1210208".

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JESTIE.2021.3074876, IEEE Journal of Emerging and Selected Topics in Industrial Electronics

Swati Tandon (Student Member, IEEE) received her master's degree in power system from the Delhi Technological University, Delhi, India in 2013. She is currently working towards her Ph.D. degree at the department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada. Her primary research interests include soft-switching techniques, development of high frequency resonant based current-fed DC-DC converter topologies for microgrid or distributed energy systems.

From 2013 to 2017, she has worked as an Assistant Manager at Design and development center, Electrical & Automation, Larsen & Toubro limited, Mumbai, India.



Akshay Kumar Rathore (IEEE Fellow) received his M.Tech. degree in electrical machines and drives from the Indian Institute of Technology (BHU) Varanasi, India, in 2003, and the Ph.D degree in power electronics from the University of Victoria, BC, Canada in 2008.

He had two subsequent postdoctoral research appointments with the University of Wuppertal, Wuppertal, Germany, and the University of Illinois at Chicago, Chicago, IL, USA. From November 2010 to February 2016, he was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada. He has published about 270 research articles in international journals and conferences, including 90 IEEE TRANSACTIONS. His research is mainly focused on current-fed converters and multilevel inverters.

He is serving as the IEEE Industry Applications Society (IAS) Awards Department Chair, an IAS Prominent Lecturer and Industrial Electronics Society (IES) Member-at-Large. Dr. Rathore received the Gold Medal during his M.Tech. degree for securing the highest academic standing among all electrical engineering specializations. He is a recipient of the 2013 IEEE IAS Andrew W. Smith Outstanding Young Member Achievement Award, the 2014 Isao Takahashi Power Electronics Award, the 2017 IEEE IES J David Irwin Early Career Award, 2020 IEEE IAS Outstanding Area Chair Award and the 2020 IEEE Bimal Bose Award for Industrial Electronics Applications in Energy Systems. He is also the Paper Review Chair of the IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS FOR RENEWABLE AND SUSTAINABLE ENERGY CONVERSION SYSTEMS. He is an Associate Editor of IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS and an editor of IEEE Transactions on Sustainable Energy. He was a Distinguished Lecturer from 2017-18 and an IAS Executive Board Member-at-Large from 2017-19 of the IAS Society. He chaired the IEEE IAS Industrial Automation and Control Committee from 2018-19 and the IEEE IES Technical Committee on Transportation Electrification from 2016-17. He served as the Editor-in-Chief of the IEEE Industrial Electronics Technology News (ITeN) from 2016-18 and received recognition award in 2019.

Jose Rodriguez (M´81-SM´94-F´10-LF´20) received the Engineer degree in electrical engineering from the Universidad Tecnica Federico Santa Maria, in Valparaiso, Chile, in 1977 and the Dr.-Ing. degree in electrical engineering from the University of Erlangen, Erlangen, Germany, in 1985. He has been with the Department of Electronics Engineering, Universidad Tecnica Federico Santa Maria, since 1977, where he was full Professor and President. Since 2015 he was the President and since 2019 he is full professor at Universidad Andres Bello in Santiago, Chile. He has coauthored two books, several book chapters and more than 400 journal and conference papers. His main research interests include multilevel inverters, new converter topologies, control of power converters, and adjustable-speed drives. He has received a number of best paper awards from journals of the IEEE. Dr. Rodriguez is member of the Chilean Academy of Engineering. In 2014 he received the National Award of Applied Sciences and Technology from the government of Chile. In 2015 he received the Eugene Mittelmann Award from the Industrial Electronics Society of the IEEE. In years 2014 to 2020 he has been included in the list of Highly Cited Researchers published by Web of Science.

2687-9735 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:54:07 UTC from IEEE Xplore. Restrictions apply.

JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 14, 2020

PERFORMANCE EVALUATION OF VERMIFILTRATION TECHNOLOGY FOR THE TREATMENT OF WASTEWATER USING DETOXIFYING EARTHWORMS: A CASE STUDY

Prof. S. K Singh[1], Twinkle Bhatia[2]

[1,2] Department of Environmental Engineering, Delhi Technological University, Delhi (110042), India

Received: 14 March 2020 Revised and Accepted: 8 July 2020

ABSTRACT: Wastewater is defined as the contaminated sewage discharging from various sources such as residential, commercial, agricultural as well as industrial sites. It is produced by anthropogenic activities which is done at source through various processes. The wastewater from residences and organizations carries wastes such as urine and human feces, food waste and wastes coming from washing etc. are categorized as sanitary sewage or domestic sewage. The sewage containing liquid and solids coming from the domestic sites can be treated with the help of various conventional processes such as Trickling Filter, Activated Sludge Process, Rotating Biological Contractors, etc. However, each of the conventional process have some or the other disadvantages which slows down their efficiencies and thereby treated water is difficult to acquire from them. These disadvantages include more treating time, uneconomical, high operation and maintenance cost which may cause fly nuisance etc. Therefore, to treat such contaminated domestic wastewater, an economical, odor-free, eco-friendly with low operation and maintenance cost, technology is required. A novel technology called as Vermifiltration technology is one such eco-friendly technology which fulfills all the above criteria to treat the domestic sewage. This technology uses earthworms which act as a 'biofilter'. In a vermifilter, removal ratios for BOD_5, TSS, COD, and Fecal coliforms is about 90%, 85%, 98% and 99% respectively. The paper majorly discusses about the Vermifiltration process, its performance evaluation through sampling and data analysis of treated sewage of educational institute, its advantages and disadvantages, its scope in future along with some research work done using this novel technology to treat the wastewater.

KEYWORDS: domestic wastewater, Vermifiltration technology, treatment processes, water quality parameters

I. INTRODUCTION

Wastewater is any contaminated water which originates predominantly from Domestic, Commercial, and Industrial sites along with Surface water, Storm water as well as ground water. There are a number of treatment processes to remove the contamination and toxicity in wastewater. There are various sources of wastewater generation which includes residential, commercial, agricultural, and industrial sites. Toxic wastewater may cause health risks such as Cholera, Amoebiasis, Giardia (also known as beaver fever) etc. There are various technologies that have been used in past to treat the wastewater but it has been found that none of the conventional technologies have qualities like odor-free, sludge-free, eco-friendly or is economical with low operation and maintenance cost altogether. One of the novel technologies which is being used nowadays which tackles all such problems is the Vermifiltration technology. It is similar to trickling filter as it is also an attached growth system. Vermifiltration technology do not require any external energy, except for pumping and due to this, it can be useful for treating wastewater in small communities also. In this technology, there is no sludge formation which needs extra spending on landfill disposal. Also, it is an odor-free process and the treated water is clean enough to be used again in parks, gardens and even for farm irrigation.

The major objectives of the study is: To evaluate the performance of Vermifiltration technology for the treatment of sewage through sampling & data analysis, to compare the pros and cons of the technology along with the research work done and to find the future scope of the novel Vermifiltration technology.

II. VERMIFILTRATION TECHNOLOGY

Vermifiltration was primarily coined by researchers of the University of Chile in the year 1992 as cost-effective, no sludge producing and sustainable technology. In the process filters known as a Vermifilter is a type of

ISSN- 2394-5125 VOL 7, ISSUE 14, 2020

biofilter for the wastewater treatment. Vermifilters works in an aerobic environment and wet substrate which provides the growth of microorganism as a biofilm known as slime layer. Microorganisms such as bacteria takes part in the degradation of the organic matter biochemically which are present in wastewater. Earthworms synchronize microbial biomass and activity by feeding on micro-organisms. Organic matter and Biofilm which is consumed by the earthworms is then digested into the organic matter turns into humus which is called as a vermicompost. This vermicompost is rich in minerals and nutrients as the earthworms consume the organic matter into their gut and we get degraded humus. It can be applied to soil to enhance the fertility of soil and structure.

Earthworms used in Vermifiltration technology

Earthworms known as the farmers best friend are segmented and tubular worms. The earthworm's digestive system is all over the length of its body and it respires through its skin. Earthworms excavate underpasses in the soil which as a result provides aeration to the soil thereby making way for hydrogen, oxygen, and nutrients to reach deep down the soil. They liberate waste known as castings from their bodies after organic matter is digested. Castings have various nutrients that the plant consume for their growth and nourishment. In Vermifiltration technology, earthworms play the major part in treating the wastewater by degrading the organic matter from it. Earthworms can also consume toxic heavy metals to make the industrial wastewater. All the earthworm species are best suited under all climatic conditions. Niche diversification of earthworms used for Vermifiltration are as follows:

S.no.	Earthworm species	Moisture content	Weight of adult earthworm	Temperature tolerance of earthworms	Active phase of earthworms	Existence of earthworms over the world (distribution)
1.	Eisenia fetida (Red worms)	20 – 40%	0.3 – 0.7g	15 – 30^{o}C	Active throughout the year	Found in temperate regions of Europe, India and North America
2.	Eudrilus Eugeniae (African night crawler)	20 – 40%	1.5 – 2.5g	18 – 35^{o}C	Active throughout the year	Found in South America and Tropical Africa
3.	Perionyx Excavatus (Indian blue worm)	30 – 50%	0.8 – 1.2g	8 – 30 ^{o}C	Active throughout the year	Exist in Tropical countries like India
4.	Drawida Barweli	40 – 50%	0.2 – 0.5g	20 – 30 ^{o}C	Active in months August to November	Found in Tropical countries
5.	Lampito Mauritii	20 – 40%	0.8 – 1.5g	18 – 28 ^{o}C	Active in months June to August	Exist in Plains of Indian Peninsula

Table 1: Detailed Description of Earthworm Species

Figure 1. Red Tiger Worms

Figure 2. African Night Crawler Worms

Formation of Vermifilter bed

ISSN- 2394-5125 VOL 7, ISSUE 14, 2020

Vermifilter is a cylindrical shaped biofilter having depth of 1-1.5 m and is divided into four parts in which at the bottom gravel of size 10-20 mm is placed, then another layer of gravel of size 2-4 mm is placed, sand of size 1-2 mm and garden soil bed for earthworm are placed in the vermifilter. A homogeneous mixture of Garden soil and Sawdust at a Volume Ratio of 3:1 is adopted usually to form the vermifilter bed. Natural ingredients such as wood chips, mud balls etc. can be used as filter media in place of garden soil or the mixture of the two can be taken can be in which earthworms resides. Vermifilters are naturally ventilated and is equipped with polypropylene perforated pipe to provide equal distribution of the wastewater and for aeration, a couple of pipes are inserted. The earthworms are given about a week of settling time in the soil bed so that they can acclimatize in the new environment.

Figure 3. Vermifilter bed with wood chips

Performance analysis of Vermifilter on the basis of water quality parameters

The performance evaluation has been done on the basis of testing parameters of wastewater quality and the quality parameters of treated effluent coming from the STP (1MLD) in Delhi Technological University Campus, Rohini, Delhi (India) and the data have been taken from Noida testing Laboratories.

Testing of these quality parameters decides whether the quality of water is up to mark for land irrigation or not. Various parameters have been listed along with the values of wastewater before and after treatment with Vermifiltration Technology.

SAMPLING AND ANALYSIS DATA

(Noida Testing Laboratories)

Sample received on:07/01/2020

Sample description: Treated sewage from STP

Project name: DTU Bio-filter STP, Bawana Road, New Delhi

Sample Quantity/Packing detail: 1 lit/Plastic cane

Analysis Duration:07/01/2020 to 13/01/2020

S.no.	Parameters	Test Method	Results	Unit	Limit asper MoEF/CPCB Norms
1	pH	IS:3025(Part -11)	7.68	-	6.5-9.0
2	TSS	IS:3025(Part -17)	18.6	mg/l	100 (max)
3	COD	IS:3025(Part -58)	39.0	mg/l	250 (max)
4	BOD_3	IS:3025(Part -44)	9.0	mg/l	30 (max)
5	Oil & grease	IS:3025(Part -39)	<1.0	mg/l	10 (max)
6	TDS	IS:3025(Part -16)	998	mg/l	No limit specified

ISSN- 2394-5125 VOL 7, ISSUE 14, 2020

7	Faecal coliform (MPN/100ML)	IS: 1622:1981	152	mg/l	<1000 MPN/100ml

Table 2. Wastewater Parameters Analysis

Figure 4. Color variation in raw and treated sewage

Pros and Cons of Vermifiltration technology

PROS OF VERMIFILTRATION

(1) Vermifiltration is a sludge free and odorless technology.

(2) A smaller number of mechanical equipments are required.

(3) It is an economically affordable technology with environmentally sustainable features.

(4) The nutrients and filtered water are used for irrigating useful vegetation.

(5) Wastewater is treated in few days as the conventional treatment processes may take up several days to treat the wastewater.

(6) High treatment efficiency is given with low space requirement.

CONS OF VERMIFILTRATION

(1) It requires a proper culturing of earthworms before wastewater is allowed to trickle over them. Moreover, Installation cost may be high.

(2) Clogging of pores in the soil may occur due to the solids present in organic matter due to which fly nuisance may be a problem.

(3) Wastewater going for the treatment in the vermifilter may have objectionable smell.

III. RECENT RESEARCH WORK DONE ON VERMIFILTRATION TECHNOLOGY

Y.S. Li et al., (2008) reviewed that organically polluted water can be treated using earthworms in this new technology called Vermifiltration. In this experiment, piggery (swine facility) associated pilot was developed with 66 swine for the treatment of swine manure and later preliminary results were provided. Results concluded that the population of the earthworms were grown by 30% within 4 weeks which showed the acclimation of the worms. Moreover, about 50% reduction was observed in ammonia emission for the whole system. Meiyan Xing et al., (2010) performed some experiments and studied the Vermifiltration technology. In the experiment, a filter was developed which was packed with ceramsite and quartz sands to treat Domestic wastewater. Results were concluded that the removal efficiencies of Vermifilter were BOD_5 = 54.78% to 66.36%, COD = 47.3% to 64.7%, Suspended Solids = 57.18% to 77.90%, TN = 7.63% to 14.90% and NH_4-N = 21.01% to 62.31% respectively.

Another study was conducted to treat Wastewater from Petroleum industry which contained mixture of “Aliphatic” & “Aromatic” volatile petroleum hydrocarbons and “organochlorines” which originates from Cooling liquids, gear oil and petrol spill, waste transmission and Diesel oil (Rajiv K. Sinha et al., 2012). The chemicals taken into consideration were the total Petroleum hydrocarbon, Dichloroethane, Dichloromethane and t-butyl methyl ether. About 1,000 red worms were used for the treatment of industrial wastewater in the Vermifilter. Earthworms survived and tolerated toxic wastewater and rectified contamination as odorless and pale-yellow effluent. The hydrocarbons [C 10-C 14] were reduced by 99%, the [C 15-C 28] by 99.8% and the [C 29-C 36] by 99.7% respectively. Tarun Kumar et al., (2015) studied the Vermifiltration technology in which an experiment was carried out using various natural ingredients like wood coal, riverbed material, mud balls and

ISSN- 2394-5125 VOL 7, ISSUE 14, 2020

glass balls. Earthworm species used was *Eisenia fetida*. Performance of Vermifilter with various natural ingredients as media was studied. Results showed that average BOD removal was observed for different media as 81.2% (riverbed material), 74.5% (wood coal), 72.7% (glass balls), 70.9% (mud orb). Average COD removal was 72.3% (riverbed material), 64.6% (wood coal), 61.5% (glass orb), 59.8% (mud orb). Total suspended solids removal was 75% (riverbed material), 64% (wood coal), 59% (glass orb), 55% (mud orb) respectively. Moreover, Divya C Das et al., (2015) developed a treatment process for ayurvedic effluents by combining Vermifiltration and microbial pre-treatment. Initially, the effluent was pretreated with a consortium of microbes and later fed to a Vermifilter. Solids, organic wastes, and heavy metals are absorbed and ingested through the body wall of the earthworms which was found to eliminate COD, BOD, TSS and TDS from the wastewater. The treated water was disinfected and clean enough to be utilized for irrigation.

IV. SCOPE OF THE NOVEL TECHNOLOGY – VERMIFILTRATION

The micro-organisms are playing an integral part in the industrial field of organic solid waste management. The active participation of earthworms in balancing the pH, aeration, moisture is observed in various studies and is becoming more riveting to research about the various native species of earthworms in filter mud treatment as well as in other industrial solid wastes.

Various studies have been done on the micro-organisms which has been secluded from Vermicompost and from skin with definite characteristics and enzymatic activities has divulged that the same could be used for treating other industrial wastes. Thus, in waste management research by using micro-organisms should be studied more and study about enzymatic activities such as cellulose, amylase, lipase, etc. for the possible utilizations in treatment of organic waste and treatments of wastewater.

As the field of industrial waste management is latest, there is a brighter scope for new experimental designs so as to acquire a practical answer to the biodegradation process using bacteria, yeasts, fungi, etc. or in combination of these micro-organisms in a specific proportion. Moreover, Vermifiltration technology is used and has a good scope in treating wastewater from petroleum industry. Chemicals such as the total petroleum hydrocarbon (TPH), dichloroethane (DCE), Dichloromethane (DCM) and t-butyl Methyl ether which are highly toxic in nature can be easily removed from the wastewater.

Treatment of Wastewater coming from ayurvedic industries in nowadays a major concern. As this industry has grown more years due to high demand of the products, the unwanted matter including toxic wastewater and solid waste disposal is a problem. However, Vermifiltration technology is one such eco-friendly which can be used to treat contaminated water from such industry. This technology can remove BOD, COD, TSS and total coliform to a great extent and can balance the pH in the treated effluent, so that it can be used in irrigation and other purposes. Dairy industry is one of the high wastes producing industry. The wastewater produced from dairy industries have high total solids as well as organic loading. This contamination can be easily removed by Vermifiltration technology. Treated effluent can be used for horticulture and other purposes.

V. CONCLUSION

From the sampling analysis data acquired from the laboratory and the various successful research work done under Vermifiltration technology, it can be concluded that this technology is excellent for treating wastewater coming from domestic, industrial, institutional as well as agricultural sites. It is an eco-friendly technology in which the treated water is sludge-free as well as odor-free. The technology uses earthworms which are known as the 'Unexpected warriors of mankind & companion of farmers' which gets multiplied in few days and therefore can treat great quantity of wastewater without any addition of chemicals. It is uncomplicated to run, and the installation & maintenance cost is also inexpensive. Moreover, the percentage removal of BOD, COD, TSS, turbidity and total coliform is up to a great extent. Thus, this technology proves to be the best novel technology from which the treated effluent can be used in horticulture and other purposes.

VI. REFERENCES

[1]. Sinha, R. K., Chandran, V., Soni, B. K., Patel, U., & Ghosh, A. (2012). Earthworms: nature's chemical managers and detoxifying agents in the environment: an innovative study on treatment of toxic wastewaters from the petroleum industry by Vermifiltration technology. The Environmentalist,32(4), 445–452. doi:10.1007/s10669-012-9409-2

[2]. Nie, E., Wang, D., Yang, M., Luo, X., Fang, C., Yang, X., … Zheng, Z. (2014). Tower bio-vermifilter system for rural wastewater treatment: bench-scale, pilot-scale, and engineering applications. International Journal of Environmental Science and Technology, 12(3), 1053–1064. doi:10.1007/s13762-013-0479-6

ISSN- 2394-5125 VOL 7, ISSUE 14, 2020

[3]. Kumar, T., Bhargava, R., Prasad, K. S. H., & Pruthi, V. (2015). Evaluation of Vermifiltration process using natural ingredients for effective wastewater treatment. Ecological Engineering, 75, 370–377. doi: 10.1016/j.ecoleng.2014.11.044

[4]. Timmer G, Albrecht JJ, Richter HThe biofiltration as method for wastewater treatment and pesticides removal (in German).KorrespondenzAbwasser. 1993; 5:764–780.

[5]. Sequencing batch reactor image is taken from http://www.ethicsinfinity.com/EthicsProduct-sequencing-batch-re

[6]. Boon, N., Pycke, B.F.G., Marzorati, M. and Hammes, F. (2011) Nutrient gradients in a granular activated carbon biofilter drives bacterial community organization and dynamics. Water Research 45(19),6355-6361.

[7]. Li, Y. S., Robin, P., Cluzeau, D., Bouché, M., Qiu, J. P., Laplanche, A., … Callarec, J. (2008). Vermifiltration as a stage in reuse of swine wastewater: Monitoring methodology on an experimental farm. Ecological Engineering, 32(4), 301–309. doi: 10.1016/j.ecoleng.2007.11.010

[8]. H. H. J. Cox, "Performance of a styrene degradation biofilter containing the yeast, Exopholialajeanselmei",Biotechnol. Bioeng, (53), pp. 259-266, 1997.

[9]. https://shodhganga.inflibnet.ac.in/bitstream/10603/12461/12/12_chapter%207.pdf

[10]. YanLing M., JingLian Z., BoLun Y. Removal of H2S in waste gases by an activated carbon bioreactor. Int. J. Biodeterior. Biodegrad. 2006; 57:93–98.

[11]. Meiyan, X., Xiaowei, L., & Jian, Y. (2010). Treatment performance of small-scale vermifilter for domestic wastewater and its relationship to earthworm growth, reproduction and enzymatic activity. African Journal of Biotechnology, 9(44), 7513–7520. doi:10.5897/ajb10.811.

[12]. Sinha Rajiv K., Bharambe Gokul and Chaudhari Uday, "Sewage treatment by vermifiltration with synchronous treatment of sludge by earthworms: a low-cost sustainable technology over conventional systems with potential for decentralization." Environmentalist (2008), vol. 28, pp. 409-420.

[13]. Samal, K., Dash, R. R., & Bhunia, P. (2017). Treatment of wastewater by vermifiltration integrated with macrophyte filter: A review. Journal of Environmental Chemical Engineering, 5(3), 2274–2289.

[14]. Bimal Jain. (2019). A Project report on, "Decentralised wastewater treatmentsystem (DEWATS): A case studyof an educational institution."

[15]. Sinha R. K., Bharambe G. and Bapat P., "Removal of high BOD and COD loadings of primary liquid waste products from dairy industry by vermi-filtration technology using earthworms." IJEP (2007), vol. 27 (6), pp. 486-501.

Prediction of an optimum engine response based on various input parameters on common rail direct injection diesel engine- A Response Surface Methodology Approach

Manish Kumar[a], Naushad Ahmad Ansari[a], Abhishek Sharma[b], Varun Kumar Singh[a], Raghvendra Gautam[a], Yashvir Singh[c*]

[a]Department of Mechanical Engineering, Delhi Technological University, Delhi, India

[b]Department of Mechanical Engineering, G L Bajaj Institute of Technology and Management, Greater Noida, UP, India

[c]Department of Mechanical Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India.

Abstract

As there is a growth in industrialization globally, the reduction of harmful pollutants is of considerable significance, especially in the transportation sector. In this paper, the impacts of various engine operating parameters are investigated on the performance and emissions of a CRDI (common rail direct injection) diesel engine. The primary aim of this study is to optimize emissions and efficiency parameters to achieve the optimal configuration parameters for the engine by employing the desirability approach of RSM (response surface methodology) techniques which is a suitable optimization approach that can save a lot of repetitive testing. After RSM modelling it was observed that the optimized engine settings of input factors are diesel/linseed blend concentration 8.10%, FIP (fuel injection pressure) 600 bar, EGR (exhaust gas recirculation) level 4.667%, and load on engine 9.33kg. On these constant hold values, the optimized output torque, BTE (brake thermal efficiency), BMEP (brake mean effective pressure), mechanical efficiency, HC (hydrocarbon), and CO_2 carbon dioxide have obtained 20.04 Nm, 26.035%, 3.474 bar, 52.503%, 28.14ppmv, and 7.319 %vol. respectively. Aforesaid predicted values were validated experimentally and errors in predicted values are in a limited range.

Keywords: Linseed, Biodiesel, Blend, Performance, Emission, Common rail direct injection, Response surface methodology

Abbreviation

ASTM	American Standard Test Method

ANOVA	Analysis of Variance
BMEP	Brake Mean Effective Pressure
BP	Brake Power
bTDC	Before Top Dead Centre
BTE	Brake Thermal Efficiency
CI	Compression Ignition
CO_2	Carbon Dioxide
CN	Cetane Number
CV	Calorific Value
CRDI	Common Rail Direct Injection
DOE	Design of experiment
F/A ratio	Fuel Air ratio
FIP	Fuel Injection Pressure
HC	Hydrocarbon
LB5	Diesel 95% + Linseed Biodiesel 5%
LB10	Diesel 90% + Linseed Biodiesel 10%
LB15	Diesel 85% + Linseed Biodiesel 15%
LB20	Diesel 80% + Linseed Biodiesel 20%
KOH	Potassium Hydroxide
NOx	Oxides of Nitrogen
NaOH	Sodium Hydroxide
η (Mech)	Mechanical Efficiency
PM	Particulate Matters
RCCD	Rotable central composite design

1. Introduction

With the growing shortage of world petroleum supplies and the rising consumption of petroleum products, especially transportation fuels, the need for liquid fuel replacement has become necessary for all countries. Singh et al. [1] explored that in the automobile industry, non-renewable diesel fuel is one of the primary fossil fuels and its burning has a detrimental effect on the atmosphere. Because of the gradual and rapid depletion of fossil fuels, the search for renewable fuel sources is quite significant. Gautam and Kumar [2] observed that the emissions of carbon monoxide (CO), unburnt hydrocarbon (UHC), carbon dioxide (CO_2), smoke particles, and oxides of nitrogen (NOx) from the CI engines have a severe impact on human beings and the environment. A. Saravanan et al. [3] stated that vegetable oil can be treated as a distinctive sustainable source for the production of biodiesel among various alternatives due to multiple advantages, such as readily accessible, non-toxic, non-hazardous and bio-degradable characteristics, which would not only be cost-effective but also environmentally sustainable. L. Karikalan et al. [4] found that biodiesel is selected as a replacement fuel that develops the ecological situation and adds to the supply of energy. The use of a biodiesel blend does not require any engine adjustment and in fewer cases, slight modifications are required for the use of pure 100% biodiesel. Rajesh and Saravanan [5] observed that in diesel engines when fuelled with vegetable oil, comparable power output and slightly lower thermal performance have been identified. Also, Gautam et al. [6] explored the energy fuels from bio-based sources which are the response to energy protection, jobs, reducing the resources of the world used in consuming fossil fuels, growth in the agricultural economy, utilizing wasteland to produce oil-rich crops. Ghadge et al. [7] stated that biodiesel is obtained by a transesterification process in which the fatty acid (triglycerides) reacts with alcohol in the presence of a base catalyst in a specific ratio and forms alkyl fatty acid esters (biodiesel) and glycerol.

Various edible, non-edible oil, animal fats, etc., can be used to produce biodiesel. In this work, the investigation is done on linseed biodiesel. Savita et al. [8] studied that linseed oil has an outstanding capacity for producing by-products and has a strong oil standard. Linseed seeds will yield between 33–47% oil content under ideal conditions. Modification of the profile of fatty acids and improved yield will render it an excellent candidate for usage in the biodiesel industry. And our country's climatic and geographical conditions are efficient for the linseed (Linum Usitatissimum) crop production.

Considering the above factors, Ansari et al. [9] have performed experiments to improve the diesel engine output and reduce pollutants such as HC, NOx, particulates, CO, etc., whereas maximizing the CO_2 by utilizing various biodiesel formulations, using additives, enhancing the timing of fuel injection, higher fuel injection pressure, EGR techniques, pilot injection, design adjustment of the combustion system, varying CR (compression ratio), etc. Maximization of CO_2 is advantageous because it shows that complete fuel combustion happens within the cylinder and mostly CO molecules are transformed into CO_2. Each of the aforementioned approaches will benefit to some extent. Harish et al. [10] have investigated the impacts of EGR on PB30 (70% diesel+30% Palm biodiesel) dispersed with 25 ppm TiO_2 (Titanium oxide) nanoparticles (PBN) in the CI (compression ignition) engine. The results of the experiment obtained for PBN-EGR were excellent compared to the conventional PB-EGR system, as there is a significant decline in BSFC (brake specific fuel consumption), HC and CO emissions with a rise in EGT (exhaust gas temperature) over PBN. PBN-EGR NOx emissions were on par with PB-EGR throughout the entire load. Kanana and Anand [11] analysed that when the FIP is decreased, the surface area of the fuel droplet will increase and the evaporation of the droplet will take more time to raise the HC and CO amounts. Thus, FIP has a significant impact on ignition delay and combustion characteristics. A K Agarwal [12] conducted an experiment to evaluate the impacts of Karanja biodiesel at different FIP and different injection start times. The period of fuel injection was observed to decrease slightly with the greater biodiesel blend ratio and to drop rapidly with the increasing FIP. Without any need for hardware improvements or ECU (Electronic control unit) recalibration, 10% Karanja biodiesel blends boost BTE and minimize pollution. It is found from the survey of several researchers that various parameters like the FIP, EGR rate, CR, IT (injection timing), engine load and type of fuel with different concentrations affect the performance and emission characteristics of the engines.

All of these approaches have proven to some level of benefit but the issues of poor efficiency and biodiesel pollution were still unresolved. Therefore, researchers like Tosun et al. [13] have attempted to boost the efficiency of the biodiesel-fuelled engine and its blends by utilizing different strategies of optimization. Many researchers have contributed to boost the performance of an engine by using biodiesel as a substitute for fossil diesel fuel through various optimization techniques to minimize pollutants. Researchers Pathak et al. [14] have switched to computer-based approaches as conventional approaches are very time-consuming and expensive, so they can produce the same output faster and cheaper with a minimum number of trials. There are several methods generally applied to an optimized engine's output like the Taguchi method, RSM (Response surface methodology), ANN (Artificial Neural Network),

IGA (Isogeometric Analysis), GA (genetic algorithm), Grey rational analysis, Computer design methodology, and Non-linear regression. Ghasemi et al. [15] investigated that most of the models need large quantities of the experimental algorithm to develop the networks. They also will require large computing power and a more computational period. Apart from this, RSM is a commonly used technique to solve many industrial challenges. Hoda Hamze et al. [16] explored that RSM is a collection of mathematical and statistical approaches used for simulation and problem solving, wherein many running or process variables affect the response. When experiments conducted to optimize performance and minimize emissions levels, RSM becomes very efficient.

Lingling Ma et al. [17] explaining the key benefit of the RSM that the nature of RSM-based studies requires fewer tests and is less time-consuming than a true experimental study. In many research reports, this technique is commonly used and has been implemented. Researchers have published experiments exploring the optimal blend ratios for diesel/biodiesel mixtures. Vinay and Yadav [18] optimize the parameters of efficiency and pollutants of a CI engine powered by Mahua biodiesel mixed with diesel using RSM. The analysis showed that for Mahua oil methyl esters blended with diesel, BSFC (brake specific fuel consumption) (kg/kWh): 0.2252, BTE (%): 29.2885, CO (vol. %): 0.00757, HC (ppm): 5.7195, NOx (ppm): 319.78, smoke (vol.%): 4.50 were observed at optimized input parameters (40% fuel blend and 15 Nm load torque), performance and pollution parameters at optimum solutions. Uslu and Celik [19] examined the impact on engine efficiency and pollution of i-amyl alcohol (i-AA)/gasoline fuel mixture on SI engine predicted by ANN and optimized with RSM. The RSM analysis showed that the optimum engine operating conditions are the i-AA ratio of 15 % at 8.31 CR and 2957.58 rpm engine speed. In this way, an efficient method for forecasting and maximizing engine outputs with minimal testing was found to be the ANN model with RSM support. Many research, like Esonye et al. [20] studied the strengths and precision of RSM techniques in collecting the best outputs in multi-purpose data sets to maximize a CI engine's emission and performance outputs. Molina et al. [21] studied a CI engine's fuel efficiency and NOx emissions utilizing response surface techniques to establish a control-oriented process and found that the approximate average BSFC and NOx errors were 2% and 6% over 5.5ms duration. M. Krishnamoorthi et al. [22] studied the cumulative influence of CR, IP (injection pressure) and IT (injection timing) on the engine performance by utilizing all test fuels. For F(1) (80% diese+15% bael oil+5% diethyl ether) fuel at 230 bar IP and 18 CR with 23° bTDC (before top dead centre) IT a maximum BTE of 30.05% were observed. At 230 bar IP and CR

18 with 25° bTDC IT, minimal CO of 0.41% was observed and NOx of 205.7 ppm was identified at 250 bar IP and 16 CR with 25° bTDC IT for F(1) fuel. Yatish et al. [23] implemented CRD (Central composite design) based on the RSM model for efficient synthesis of Bauhinia Variegata oil with biodiesel. Also, the engine output was tested for an optimized blend of biodiesel at variable engine loads during which a reduction in exhaust emissions (CO and UHC) was reported. Najafi et al. [24] used the application of ANN-RSM approaches to maximize biodiesel/diesel blends with energy and exergy analysis. Better engine performance was achieved with 80% engine load in the presence of B10 and B20 biodiesel blends. Parida et al. [25] experimented on CI engines by utilizing the RSM model to optimized performance and emission variables by selecting engine load, CR, and fuel blend as input parameters. BSFC, BTE, HC, NOx, and CO enhanced outcomes were 0.284 kg/kWh, 26.77%, 9.6ppm, 114.84ppm, and 0.0059% respectively, with desirability 0.97. On comparing experimental results and the predicted value of RSM mode, the errors come below 5% except CO. Singh et al. [26] examined the impact of various engine input factors like engine load, the timing of fuel injection, FIP on a CI engine when powered with the blend of Pongamia biodiesel and diesel. The performance and exhaust variables are optimized through the RSM approach. Optimum operational variables were calculated as 15° before the peak dead centre spray time, 196.36 bar fuel spray pressure, 40% biodiesel blend, and 53% engine load, according to optimization controls. Among checked findings and expected values, less than 5% of errors were found. Therefore they inferred that RSM with fair efficiency can optimize the factors influencing engine output. Sakthivel et al. [27] developed an RSM model for estimating optimum input variables from waste biomass on a diesel engine fuelled by pyrolysis oil. The fuel blend ratio, CR, and engine load were considered input variables and BSFC, BTE, and exhaust pollutants like CO, NOx, smoke, and HC were taken as output variables via the RSM model. They considered the optimum solution 18:1 CR, 20% fuel blend, and 100% load with less than 5% error ratio respectively. Veinblat et al. [28] conducted comparative work on a CI (compression ignition) engine for the analysis of combustion properties and performance by biodiesel from linseed oil. It was observed that the linseed biodiesel displayed a decreased NOx emission relative to diesel at maximum engine loads, in comparison to soybean biodiesel. A. K. Aggrawal [29] explores the use of linseed biodiesel in CI engines to establish a stable engine fuel system having low pollution parameters.

From the above literature, it was noticed that the engine operating variables viz. EGR rates, FIP, biodiesel blends, and load on the engine have a great impact on the performance and

exhaust pollutants of an engine and optimization of these inputs parameters to get a specific operating configuration of an engine is advantageous for producing maximum performance with minimal emission levels through RSM techniques as it reduces cost and number of trials. However the correlation between the FIP, EGR speeds, engine load and linseed biodiesel content in diesel on the efficiency and emissions of the CRDI diesel engine using optimization techniques RSM has not yet been investigated. To complete this gap, the novelty of the current experiment is, therefore, to apply an acceptable RSM model as a reasonable step to forecast the output and exhaust pollutants of diesel/linseed biodiesel blends, to effectively find the optimal operating configuration of the CRDI diesel engine. The RSM-based desirability approach was used in this research paper to refine engine operating variables such as FIP, EGR rates, engine load and linseed biodiesel content for the superior response of torque, BTE, BMEP, mechanical efficiency along with CO_2 and least HC emissions. For optimization evaluation by dimensionless desirability effect, Minitab'17 software was used.

2. Material and methodology

2.1 Production of Linseed biodiesel

The research was carried out on a CRDI diesel engine fuelled with linseed biodiesel blended with neat diesel oil. The linseed biodiesel was made from raw linseed oil bought from a nearby New Delhi vendor. Later, it was converted into linseed biodiesel (linseed methyl ester) through the transesterification process. In this process, 100ml of linseed oil with FFA (free fatty acid) less than 2% was taken in a beaker and heated to 110^{o}C to eliminate water content as shown in Figure 1. Then, the oil is allowed to cool around 60°C and added a mixture of methanol (20ml) and KOH (potassium hydroxide) (1 gram) slowly to the oil and stirring it constantly for around 30-40 minutes and the beaker was covered with hard cardboard. Then it was permitted to cool down in a separating funnel for about 24 hours without any agitation. The mixture formed into two layers due to the gravity effect. The heavy component settled down i.e. glycerol, and the upper element was fatty acid ester i.e. linseed biodiesel. 'Water washing' on methyl ester was performed to eliminate impurities of glycerol and KOH, and pure linseed biodiesel was obtained from heating it at 110°C for around 20minutes to remove any water and methanol evidence. Dixit and Raheman [8] found that the process of transesterification has a role in reducing the viscosity of oil and makes it suitable for use in a diesel engine. 5%, 10%, 15%, and 20% of linseed biodiesel was mixed with neat diesel (%age by volume) and marked as LB5, LB10, LB15, and LB20 for testing in the engine. The thermophysical properties of the blend of linseed biodiesel are shown in Table 1.

2.2 Experimental setup and procedure

The engine used to test the blending fuel is a single-cylinder CRDI VCR (variable compression ratio) diesel engine with 4-strokes running at a constant speed of 1500rpm. The engine was attached to an eddy current Dynamometer with a water cooling mechanism to evaluate the engine load as shown in figure 2. Table 2 displays the technical specification for the engine test setup. The engine was adjusted at 23° bTDC fuel injection timing (recommended by the manufacturer) and compression ratio at 18:1. Five selected research fuels are prepared for this study, such as neat diesel, LB5, LB10, LB15, and LB20. The performance parameters such as BTE, BMEP, torques, and mechanical efficiency were measured through the VCR engine using Lab view oriented software "EngineSoft" attached to the engine. Emission variables such as CO_2 and HC were analyzed through an AVL gas analyzer. The specifications, accuracy, and resolution of an AVL Gas Analyzer "AVL DIX 650" are shown in Table 3.

The tests were carried out at different settings in the engine by utilizing five specific fuel injection pressures (500, 525, 550, 575, and 600 bar) and five distinct EGR levels (0%, 3.5%, 7%, 10.5%, and 14%) and engine load (0kg, 3kg, 6kg, 9kg, and 12kg) and linseed biodiesel/diesel blends (0%, 5%, 10%, 15% and 20%). The magnitude of EGR rates varies progressively by raising the EGR valve and keeping the records. The cold EGR-method was used for this analysis. Saravanan [31] noted that the cold EGR technique decreases intake charge and maximizes in-cylinder temperature which is a favourable condition for the formation of NOx pollutants. The re-circulated flue gas temperature must be held colder than that of the exhaust engine, and warmer than those of the intake air charge.

Each test was executed three times and the mean was considered for analysis. HC and CO_2 emissions were measured in tests, and the readings counted for torque, BTE, BMEP, and mechanical efficiency.

2.2.1 Uncertainty analysis

E. Samaniego et al. [32] stated that due to variables such as testing, operator observation, operating environments, and environmental conditions, many chances of defects and uncertainties are likely in the experimental studies. The tests were conducted under similar operation conditions for all fuels, such as engine temperature and environmental conditions (ambient temperature, 26 ± 4 °C and relative humidity, 62 ± 2%), to improve the confidence level of the tests.

To prove the validity of the experiments, uncertainty analysis is required. The end outcome is determined from the key measurements in every experiment. The percentage

uncertainties of different variables such as torque, BTE, BMEP, and mechanical efficiency were determined using the percentage uncertainties of different methods used in the experiment. The uncertainties of the different methods used are presented in Table 4.

2.3 Response surface methodology (RSM)

RSM is a collection of mathematical methods that can be used to model and evaluate problems where multiple variables affect interest response and the purpose is to maximize the response. In practice, the necessity of the response surface is for the sample point to be chosen such that with the minimal number of experiments a reasonably reliable model can be produced. These fitting points can be selected through experiment design. A complete factorial modelling methodology was used to conduct the test matrix of the present inquiry. The experimental effects of the response variables derived from the maximum factorial design matrix were used to shape the quadratic response surface models. Relationships are formulated using the least square method. In RSM, the correlation between the response variable y and the parameters x_1, x_2, x_3.....x_n is commonly represented as.

$$Y = f\left(x_1, x_{2,} x_3, x_4 \dots \dots . x_n\right) + \varepsilon \qquad (1)$$

Where x_1, x_2, x_3, and x_4.....x_n is the input variable of the engine and y is the engine output needed to boost it, ε is error or noise found in response y and f is the response surface.

In expression (1), the function f corresponds to the correlation between a response variable and a factor that is normally unknown. Thus in RSM, the initial step is to find an adequate approximation between the unknown factors and the response variable. The linear model is the simplest approximation function (linear model) is used to represent the relationship between a response variable and the unknown factors, which can be described as follows:

$$Y = \beta_\circ + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{j\geq 1}^{k} \beta_{ij} x_i x_j + \varepsilon \qquad (2)$$

Where Y is the response, x_i the magnitude of the factors, β_0, β_i, β_{ii} and β_{ij} are the coefficient of regression, i and j are linear and quadratic coefficients, and ε is the experimental error. Expression (2) can be assumed to be the extension of a Taylor expansion (of degree 2) of any real underlying function.

In this model of RSM, the linseed biodiesel concentration in the blend, FIP, EGR, and the load are taken as input variables that could potentially affect the output response such as torque, BTE, BMEP, mechanical efficiency, HC and CO_2. The input parameters are design variables and their limiting range are the design constraints. The levels of different input factors are

shown in Table 5. Here mid-value is taken as 0 for input variables. The RCCD values assigned are -2, -1, 0, 1, and 2.

The objective for using RSM techniques here was to reach the proximity of the optimum quickly and effectively. The simplest version of the model was then used, i.e., the linear model. All association between variables was also considered to be negligible, and only primary impacts were considered. The flowchart for the RSM approach is shown in Figure 3.

2.4 RSM modelling

The combination of control factors leading to an ideal solution has been considered in the study of engine performance. To assess the value of each parameter within the defined range, the study of the variance factors was implemented to various parameters of the inputs. These results were developed based on a linear regression model driven by the software. The tests were conducted for different combinations of factors according to the design matrix run order as seen in Table 6, and the corresponding response values are noted. This includes a minimum of 31 tests with multiple input parameter combinations. Optimization was achieved using RSM's Desirability method by Derringer, of which the optimum solution was considered to be the one with the highest desirability. The resulting combination of variables for the optimal solution is known as the right variables for the operation of the engine. Analysis of Variance (ANOVA) generates empirical knowledge regarding future benefits.

A total of 31 runs of 2^p factorial runs, 2p axial runs, and 7 centre point runs consist of the control variables, numerical and coded values used in the RCCD matrix. In the present case, the 4 control factors (p) were taken into consideration.

Therefore, total number of trails $= 2^p + 2p + 7$

$$= 2^4 + (2 * 4) + 7 = 31$$

2.4.1 Sensitivity analysis

N. Vu-Bac et al. [33] stated that the intricacy of the model is rising due to increasing computing power. In general, the more complex the models are because of randomness in the input parameters, the greater the variability in the model outputs. The variance in input parameters, as well as the calibration and evaluation of the mathematical models, must determine how much the model output is altered. For these reasons, sensitivity analysis (SA) is a great aid. Therefore,

in many disciplines, such as material simulation and structural architecture, ambiguity and sensitivity analysis have recently gained widespread attention from researchers.

The goal of using RSM as a method for sensitivity analysis is to enhance the analyst's understanding of the sensitivity between the model output (response) and its relevant input variables and then use this data to enhance the decision model. Kenneth w. bauer jr. et al. [34] explored the Using RSM for sensitivity analysis on a decision analysis topic has two main advantages. The first is a major reduction in the overall number of cycles of effect diagram solution (decision tree) needed for a detailed study of sensitivity. The reduction can be considered as the model grows in complexity. The second major advantage is the additional knowledge from the use of RSM open to policymakers. By calculating basic coefficients (effects) of significant variables and defining the relationship between variables, these insights are obtained. These insights are invisible to approaches to one-way sensitivity research.

RSM has been used widely in computer modelling to do sensitivity analysis. Typically, simulation models have multiple input variables and only a fraction of them are typically important in the modelling area.

2.5 RSM modelling results

The ANOVA results of torque, BTE, mechanical efficiency, BMEP, HC, and CO_2 are shown in Tables 7 and 8. From the ANOVA results, the p-value is considered the most important factor. Awad et al. [35] stated that the p-value of various parameters must be not more than 0.05 and the parameters having a p-value greater than 0.05 are considered as not important. The variables having p-values less than 0.05 indicates that the variable has a great impact on the developed model. From Table 7, it clear that for BTE, the p-value of blend content, FIP, and EGR are greater than 0.05 and the engine load's p-value is less than 0.05 for the load. It shows that the load on the engine has a high impact on the efficiency of the engine. Similarly, for torque, the p-value of all parameters is less than 0.05. For mechanical efficiency, the p-value of FIP and EGR is higher than 0.05 but the p-value of load and blend content is less than 0.05. It indicates that load and blend % has a great impact on mechanical efficiency. The p-value of all parameters for BMEP and HC emissions is less than 0.05 as seen in Table 8. It shows that all input parameters affect BMEP and HC emissions significantly. For CO_2 emissions, the p-value of blend ratio and FIP is more than 0.05 and the p-value of load and EGR is less than 0.05 which indicates that the EGR and engine load have a more effect on the CO_2 emission.

The models have also been tested using a computational approach using the decision coefficient R^2. Sarıkaya and Güllü [36] believed that that R^2 values below 1 suggest that the experimental effects are consistent with the findings of the model indicates that the reliability is very strong form Table 9, it is showing that the finding of R^2 is 99.57%, 80.85%, 84.97%, 96.54%, 88.57%, 96.58% for torque, BTE, mechanical efficiency, BMEP, HC, and CO_2 respectively. These values of R^2 are nearly close to 1 which indicates that this mode gives reliable results.

2.6 Optimization response

For the present analysis, the RSM optimizer is implemented to boost the engine's operating characteristics. The RSM optimizer is a method used to configure multi-targets such as IC engine cases. The maximum values of BMEP, BTE, torque, and mechanical efficiency along with CO_2 and the least value of HC are taken as objectives from the array of the tested conditions.

BMEP, BTE, torque, mechanical efficiency, HC and CO_2 optimum values are 3.576 bar, 26.64%, 20.04 Nm, 54.343%, 28.71ppmv, and 7.117% vol. respectively, with the respective optimum engine operating variables at 8.10% linseed blend ratio with diesel, FIP 600 bar, EGR 4.667%, and engine load 9.33 kg as shown in Figure 4. The desirability approach for this optimization was obtained at 0.8032. In addition, RSM may be used to estimate and modify the engine output parameters for variables within the test value series. It reduces the study research substantially and thereby saves all time and costs.

3. Results and discussion

The discussion covers the influence of 4 CRDI diesel engine input factors on five engine output responses (BTE, mechanical efficiency, BMEP, torque, HC and CO_2). All these four-engine input factors (the blend ratio, FIP, EGR rates and the engine load) cannot be varied simultaneously during surface plotting.

Therefore, in this experiment, a MINITAB 17 program was used to produce a surface plot between two engine controls, while two other engine control variables remain unchanged, known as hold values. These hold values are the central value of the RCCD table described on the upper right corner of each surface plot.

3.1 Impact of control factors on Brake thermal efficiency η(Thermal)

This segment demonstrates the influence of various parameters like FIP, EGR rates, blend ratio and the engine load on brake thermal efficiency as shown in Figure 5. The surface plot in

Figure 5(i) shows the effect of BTE with blend ratio and EGR% when the hold values are (FIP, the engine load) constant (590 bar and 9.3kg). The response surface plot reveals that as the concentration of linseed biodiesel rises in the blend, the engine's output subsequently decreases. Thus the BTE of LB20 was found to be the least and diesel has a maximum among all test fuel. It could be attributed to the high viscosity and low calorific value of linseed biodiesel relative to a neat diesel resulting in poor atomization and improper combustion of fuel. The other reason for this may be a boost in BSFC as its related density always decreases as the content of the blend rises. The introduction of EGR in the engine reduces its efficiency as its levels increase. Maybe the explanation behind this is because the rates of EGR are rising; it impedes the normal cycle of combustion and reduces the efficiency of combustion as the volume of fresh oxygen required for combustion is substituted by flue gas in the cylinder which contributes to inadequate combustion of fuel and decreases its performance. Rao et al. [37] have shown a similar trend of BTE with Cottonseed biodiesel.

Figure 5(ii) shows the interactive impact of EGR rates and FIP on the BTE when the hold values are (blend ratio, the load) constant (6.1% and 9.3kg). The surface plot of response shows that for the rise in FIP, the BTE rises with all tested fuels. This is mainly owing to the greater degree of atomization at greater FIPs; full combustion is assured even if the volume of fuel injected is the same at varying fuel injection pressures. The response surface plot in Figure 5(iii) explores the impact of blend ratio and FIP on BTE when hold values are (EGR rates, load) constant (4.7% and 9.3kg). Figure 5(iv) shows the combined effect of engine load and blend ratio on BTE when hold values are (FIP, EGR) constant. The surface plot of response shows that as the load increases on the engine its efficiency also increases. This may be because it’s torque increase and a reduction in BSFC with loads cause a rise in BTE.

3.2 Impact of control factors on mechanical efficiency η (Mech.)

Figure 6(i) shows the combined effect of the blend ratio of linseed in diesel and EGR% on mechanical efficiency while holding the values of FIP and the engine load. The ratio of brake power and IP indicated power is known as mechanical efficiency. A similar trend was shown by castor biodiesel investigated by Md. Saiful Islam et al. [38]. It was noticed that the η(Mech.) reduces with a rise in linseed in the blend because the low CV of linseed biodiesel correspondingly results in poor mechanical performance. As EGR rates increase, the corresponding value of mechanical efficiency is reduced. This was mainly because EGR promotes improper combustion of fuel due to the replacement of fresh oxygen with flue gases inside the cylinder.

Figure 6(ii) shows the interactive effect of EGR rates and FIP on mechanical efficiency while holding the value of blend ratio and load at (6.1% and 9.3kg). It was observed that mechanical efficiency gets improved with the rise in FIP. It may be because of reduced viscosity, increased atomization, and better combustion of fuel. The response surface plot in Figure 6(iii) explores the impact of blend ratio and FIP on BTE on mechanical efficiency when hold values are (EGR rates, load) constant (4.7% and 9.3kg). Figure 6(iv) shows the combined effect of engine load and blend ratio on mechanical efficiency when hold values are (FIP, EGR) constant. It was found that as the load increases the corresponding value of mechanical efficiency also gets raised mainly due to increments in BP.

3.3 Impact of control factors on BMEP (brake mean effective pressure)

Figure 7 shows the response surface plot of BMEP with all input factors. Figure 7(i) shows the combined effect of blend% and EGR on BMEP when the hold values are (FIP, load) constant (590bar and 9.3kg). The BMEP is the cumulative pressure in an engine cylinder on the piston head over a period, which determines the engine's ability to do the work. Slight differences in BMEP were observed with blend material. This may be due to a decrease in BP with the blend, the BMEP also reduces. With a rise in EGR rates, the BMEP gets reduced due to the improper burning of fuel inside the cylinder.

Figure 7(ii) shows the interactive impacts of input variable EGR rates and FIP on BMEP while holding the value of blend ratio and the engine load. It was found that BMEP gets increased with an increase in FIP. It may be due to the fact that as FIP increases, it promotes better atomization of fuel inside the cylinder which in turn provides a more surface closure to enhance complete combustion of fuel. Jun cong Ge et al. [39] show a similar trend in the CRDI engine when powered with canola biodiesel. The combined impact of the blend content and FIP on BMEP are shown in Figure 7(iii) when hold values are constant at 4.7% and 590bar for EGR level and FIP respectively. Figure 7(iv) shows the effect of load and blend% on BMEP. The surface plot of response shows that as the load on engines increases the corresponding BMEP gets increased due to an increase in BP with the load.

3.4 Impact of control factors on the torque

The combined effects of linseed biodiesel in the blend, EGR rates, FIP, and engine loads on the torque are shown in Figure 8. The surface plots in Figure 8(i) shows the impact of blend concentration and EGR rates on torque when the hold values are (FIP, load) constant (590bar and 9.3kg). The quantity of torque shall be calculated by the amount of fuels supplied to the cylinders. The engine torque has been observed to decrease as the volume of linseed biodiesel

in the blend rises. This may be due to higher viscosity and the density of biodiesel blend fuel as a similar trend was illustrated by Márcio et al. [40] in a 54 kW mechanical fuel injection diesel engine. As the heating value of biodiesel is smaller than neat diesel it results in higher torque and smaller power generation because the engine speed is steady. As the EGR rates increase the corresponding value of torque decreases due to improper burning of fuel inside the cylinder.

The interactive effects of EGR % and FIP on the torque are shown in Figure 8(ii) while holding the value of blend ratio and load at (6.1% and 9.3kg). The surface plot reveals that, for all the test fuels, engine torque rises with FIP. This is because as the FIP increases, it provides a higher degree of atomization of the fuel droplets which creates a more surface area for combustion and proper combustion of fuel takes place generating high torques on the engine. Figure 8(iii) shows the effect of torque with blend content and FIP when the hold values are constant at (4.7% and 9.3kg) for EGR rates and engine load. The combined effects of load and blend content on torque are shown in Figure 8(iv) while holding the value of FIP and EGR rates. It was observed that as the load increases on the engine their corresponding value of torque increases. A load is an external force that is applied to the engine based on the type of dynamometer used. Since the load increases on the engine and the speed of the engine must be constant at 1500rpm, which results in the increases in torque to stable the speed of the engine.

3.5 Impact of control factors on HC emission

This section discusses the combined impact of all the input factors blend ratio, FIP, EGR rates, and load on the emission of HC as shown in Figure 9. The effect of blend ratio and EGR rates are shown through the surface plot in Figure 9(i) while holding the value of FIP and load. The emission of HC was reduced slightly when the ratio of the blend increased. Fewer HC emissions were due to an increase in the blending volume leading to more oxygen than diesel causes complete combustion of fuel. HC emissions increase with higher levels of EGR. It may be because the introduction of EGR in an engine allows some flue gases inside the combustion chamber instead of fresh oxygen causes poor combustion of fuel.

Figure 9(ii) shows the FIP and EGR response surface plot for HC emissions. This shows that as the FIP raises, the HC decreases due to improve oxygen consumption, contributing to stronger fuel combustion. Channapattanaa et al. [41] show a similar trend of HC on Honne biodiesel. The interactive effects of blend ratio and FIP on HC emission whose response surface plot are shown in Figure 9(iii) when hold values are (EGR rates, load) constant (4.7% and 9.3kg). The combined effect of load and blend ratio is shown in Figure 9(iv) by holding the

value of FIP and the EGR rates. It was found that as the load on the engine increases, the corresponding HC emission reduces. This may have been related to the addition of a rich mixture of air and fuel inside the combustion chamber owing to the higher fuel consumption resulting in premature combustion contributing to higher HC emissions.

3.6 Impact of control factors on the emission of CO_2

From the response surface plot, the impacts of various input parameters such as blend ratio, FIP, EGR levels, and engine load on CO_2 emissions are shown in Figure 10. It is advantageous to maximize the CO_2 emission as it implies that the fuel combustion in the cylinder takes place properly and the CO molecules are converted into CO_2 gas due to the availability of fresh oxygen. The surface plot in Figure 10 (i) shows the effect of CO_2 with blend ratio and EGR% when the hold values are (FIP, the engine load) constant (590 bar and 9.33kg). It was reported that as the quantity of linseed biodiesel in the blend increased the CO_2 emissions substantially reduced. Carbon dioxide is produced when the fuel is completely combusted in oxygen. Here CO_2 formation is small since biodiesel is typically a low carbon product and has a lower ratio of elemental carbon and hydrogen relative to diesel fuel. EGR reduces CO_2 emissions as the exhaust gas injected instead of fresh air results in incomplete burning in the combustion chamber. Thus insufficient oxygen supply does not convert CO into CO_2 and reduces CO_2 emissions. P. Kanthasamy et al. [42] shows a similar trend on CRDI engine powered with tallow methyl ester biodiesel blends.

Figure 10(ii) shows the interactive impact of EGR rates and FIP on the emission of CO_2 when the hold values are (blend ratio, the load) constant (6.1% and 9.3kg). It was found that with the increase in FIP the CO_2 emission increases. This was probably due to the fact that at higher FIP the fuel gets easily atomized which promotes proper combustion of fuel and CO gets converted into CO_2. Thus the level of CO_2 increases. The response surface plot in Figure 10(iii) explores the impact of blend ratio and FIP on CO_2 emission when hold values are (EGR rates, load) constant (4.7% and 9.3kg). Figure 10(iv) shows the combined effect of engine load and blend ratio on CO_2 emission when hold values are (FIP and EGR level) constant. It has been found that the CO_2 emission rises with increased load due to improved combustion and more fuel burns at high loads.

3.7 Experimental tests and validations

The need for optimization is to improve the performance of engine BTE, BMEP, torque, and mechanical efficiency with CO_2 and minimize the emission of HC. The desirability value obtained is 0.8032, and is recommended since it is higher and nearest to the optimization

criteria. This needs the assurance and evaluation of the optimized output. For this, the experimental test was performed at the obtained input variables configuration through RSM techniques and the test was repeated 3 times. From the outcomes, their mean value and the standard deviation were calculated and mentioned in Table 10.

The means values of the outcomes of these experimental tests were correlated with the RSM model's predicted values and the error was estimated as seen in Table 11. It was found that there is a slight variation between the experimental results and the predicted value of the output variables. The error % age for BTE, BMEP, torque, mechanical efficiency, HC, and CO_2 is 3.32%, 2.36%, 2.14%, 4.27%, 6.60%, and 5.175% which are in a limited range. The graph shown in figure 11 indicates the difference in engine outcomes values between the predicted value and experimental value.

4. Conclusions

The performance and pollutants parameters of a 4-stroke CRDI diesel engine powered by linseed biodiesel/diesel blends were analysed in the current research. The innovation of this research work is the optimization of the operating factors of the engine operated by DOE (design of experiment)/RSM run with linseed biodiesel and diesel blends. The range of linseed biodiesel ratio (5%, 10%, 15%, and 20%), engine load (0kg, 3kg, 6kg, 9kg, and 12kg), the EGR rates (0%, 3.5%, 7%, 10.5%, and 14%), and the fuel injection pressure (500kPa, 525kPa, 550kPa, 575kPa, and 600kPa) are selected as input parameters for this analysis. To evaluate the optimal engine operating conditions to achieve improved efficiency and emission parameters, the RSM approach was implemented. The following conclusions can be drawn based on the findings derived from this study.

1. The software-designed experiments helped to determine the exact responses.
2. The best desirability of 0.8032 is achieved by using the desirability method of RSM.
3. The study reveals that 600 bar FIP, 4.667% EGR, and 9.33kg engine load and the 8.10% linseed/diesel blend ratio is the best fit for achieving optimum output and emissions from the CRDI engine. In addition, at these optimal configurations of FIP, EGR, linseed blend and engine load the measured output parameters showing a BTE of 25.17 %, BMEP of 3.392 bar, the torque of 19.61 Nm, η (Mech) of 50.26 %, HC of 30ppmv and CO_2 of 7.697 % vol., which is relatively close to experimental values i.e. 26.035 %, 3.474 bar, 20.04 Nm, 52.503 %, 28.14ppmv, for BTE, BMEP, torque, η (Mech), HC and CO_2 measured at the optimized combination of input variables.

4. It is observed that BTE and BMEP decrease with an increment in linseed quantity in the blend.
5. The finding of R^2(correlation coefficient) is 99.57%, 80.85%, 84.97%, 96.54%, 88.57%, 96.58% for torque, BTE, mechanical efficiency, BMEP, HC and CO_2 respectively.
6. It was also found that with the increase of FIP, HC emission reduces while the CO_2 emission increases. An increment in CO_2 is desirable because it shows that the combustion of fuel takes place properly and most of the CO molecules converted into CO_2.

RSM is therefore considered to be an important tool for multi-objective engine optimization. By arranging blend ratio, FIP, EGR rates, and the engine load while using a linseed biodiesel blend, superior efficiency and controlled emission characteristics were achieved. In order to discover the ideal factors ratios for the CRDI diesel engine powered by linseed biodiesel blends, RSM and desirability approach were used effectively.

Acknowledgments

The author would be grateful to the faculty members of the mechanical engineering department, Delhi Technological University, Delhi to make this experiment useful by giving the required equipment and machine.

References

[1] Singh, V., Agarwal, T., Saroha, N. et al. "Performance Emissions and Combustion Analysis of CI Engine Using Ethyl Ester Kusum Oil and Butanol Blends"(No. 2019-01-0568), *SAE Technical Paper,* (2019).

[2] Gautam, R. and Kumar, N. "Performance emission and combustion studies of diesel engine on Jatropha ethyl ester and its higher alcohol blends", *International Journal of Global Warming*, 14(2), pp.159-169 (2018).

[3] Saravanan, A., Murugan, M., Reddy, M.S. et al. "Performance and emission characteristics of variable compression ratio CI engine fueled with dual biodiesel blends of Rapeseed and Mahua", *Fuel, 263(1),* pp.116751 (2020).

[4] Karikalan, L., Jacob, S., Baskar, S. et al. "Analyzing the influence of varied fuel injection pressure on diesel engine fueled with Karanja biodiesel", *Materials Today: Proceedings, 21(1),* pp.291-294 (2020).

[5] Kumar, B.R. and Saravanan, S. "Use of higher alcohol biofuels in diesel engines: A review", *Renewable and Sustainable Energy Reviews, 60,* pp.84-115 (2016).

[6] Gautam, R., Ansari, N., Sharma, A. and Singh, Y. "Development of the Ethyl Ester from Jatropha Oil through Response Surface Methodology Approach*", Pollution,* 6(1), pp.135-147 (2020).

[7] Yongphet, P., Wang, J., Wang, D., et al. "Optimization of operation conditions for biodiesel preparation from soybean oil using an electric field", *Biomass Conversion and Biorefinery*, pp.1-11 (2020).

[8] Dixit, S. and Rehman, A. "Linseed oil as a potential resource for bio-diesel: a review", *Renewable and Sustainable Energy Reviews, 16*(7), pp.4415-4421 (2012).

[9] Ansari, N.A., Sharma, A. and Singh, Y. “Performance and emission analysis of a diesel engine implementing polanga biodiesel and optimization using Taguchi method”, Process *Safety and Environmental Protection*, 120, pp.146-154 (2018).
[10] Venu, H., Subramani, L. and Raju, V.D. “Emission reduction in a DI diesel engine using exhaust gas recirculation (EGR) of palm biodiesel blended with TiO_2 nano additives”, *Renewable energy*, *140*, pp.245-263 (2019).
[11] Kannan, G.R. and Anand, R. “Effect of injection pressure and injection timing on DI diesel engine fuelled with biodiesel from waste cooking oil”, *Biomass and bioenergy*, 46, pp.343-352 (2012).
[12] Agarwal, A.K., Dhar, A., Gupta, J.G. et al. “Effect of fuel injection pressure and injection timing of Karanja biodiesel blends on fuel spray, engine performance, emissions and combustion characteristics”, *Energy Conversion and Management*, 91, pp.302-314 (2015).
[13] Tosun, E., Aydin, K. and Bilgili, M. “Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures”, *Alexandria Engineering Journal*, 55(4), pp.3081-3089 (2016).
[14] Pathak, A., Choudhury, P.K. and Dutta, R.K. “Taguchi-grey relational based multi-objective optimization of process parameters on the emission and fuel consumption characteristics of A VCR petrol engine”, *Materials Today: Proceedings*, 5(2), pp.4702-4710 (2018).
[15] Ghasemi, H., Park, H.S. and Rabczuk, T. “A level-set based IGA formulation for topology optimization of flexoelectric materials”, *Computer Methods in Applied Mechanics and Engineering*, 313, pp.239-258 (2017).
[16] Hamze, H., Akia, M. and Yazdani, F., “Optimization of biodiesel production from the waste cooking oil using response surface methodology”, *Process Safety and Environmental Protection*, *94*, pp.1-10 (2015).
[17] Ma, L., Han, Y., Sun, K., Lu, J. and Ding, J.,“Optimization of acidified oil esterification catalyzed by sulfonated cation exchange resin using response surface methodology”, *Energy Conversion and Management*, *98*, pp.46-53 (2015).
[18] Vinay, Singh, B. and Yadav, A.K. “Optimisation of performance and emission characteristics of CI engine fuelled with Mahua oil methyl ester–diesel blend using response surface methodology”, *International Journal of Ambient Energy*, 41(6), pp.674-685 (2020).
[19] Uslu, S. and Celik, M.B. “Performance and exhaust emission prediction of a SI engine fueled with I-amyl alcohol-gasoline blends: an ANN coupled RSM based optimization”, *Fuel*,265, p.116922 (2020).
[20] Esonye, C., Onukwuli, O.D. and Ofoefule, A.U. “Optimization of methyl ester production from Prunus Amygdalus seed oil using response surface methodology and Artificial Neural Networks”, *Renewable energy*, 130, pp.61-72 (2019).
[21] Molina, S., Guardiola, C., Martín, J. et al. “Development of a control-oriented model to optimise fuel consumption and NOX emissions in a DI Diesel engine”, *Applied energy*, 119, pp.405-416 (2014).
[22] Krishnamoorthi, M., Malayalamurthi, R. and Shameer, P.M.,“RSM based optimization of performance and emission characteristics of DI compression ignition engine fuelled with diesel/aegle marmelos oil/diethyl ether blends at varying compression ratio, injection pressure and injection timing”, *Fuel*, *221*, pp.283-297 (2018).
[23] Yatish, K.V., Lalithamba, H.S., Suresh, R. et al. “Optimization of bauhinia variegata biodiesel production and its performance, combustion and emission study on diesel engine:, *Renewable energy*, 122, pp.561-575 (2018).

[24] Najafi, B., Faizollahzadeh Ardabili, S., Mosavi, A. et al. "An intelligent artificial neural network-response surface methodology method for accessing the optimum biodiesel and diesel fuel blending conditions in a diesel engine from the viewpoint of exergy and energy analysis", *Energies*, 11(4), p.860 (2018).
[25] Parida, M.K., Joardar, H., Rout, A.K., et al. "Multiple response optimizations to improve performance and reduce emissions of Argemone Mexicana biodiesel-diesel blends in a VCR engine", *Applied Thermal Engineering*, 148, pp.1454-1466 (2019).
[26] Singh, Y., Sharma, A., Singh, G.K., et al. "Optimization of performance and emission parameters of direct injection diesel engine fuelled with pongamia methyl esters-response surface methodology approach", *Industrial crops and products*, 126, pp.218-226 (2018).
[27] Sakthivel, R., Ramesh, K., Marshal, S.J.J. et al. "Prediction of performance and emission characteristics of diesel engine fuelled with waste biomass pyrolysis oil using response surface methodology", *Renewable Energy*, 136, pp.91-103 (2019).
[28] Veinblat, M., Baibikov, V., Katoshevski, D., et al. "Impact of various blends of linseed oil-derived biodiesel on combustion and particle emissions of a compression ignition engine–A comparison with diesel and soybean fuels" *Energy Conversion and Management*,178, pp.178-189 (2018).
[29] Agarwal, A.K.,"Experimental investigations of the effect of biodiesel utilization on lubricating oil tribology in diesel engines. Proceedings of the Institution of Mechanical Engineers, Part D", *Journal of Automobile Engineering*, 219(5), pp.703-713 (2005).
[30] Experiment Engine set up Automobile Lab (24/02/2020) of Delhi Technological University, Delhi (2020).
[31] Saravanan, S. "Effect of exhaust gas recirculation (EGR) on performance and emissions of a constant speed DI diesel engine fueled with pentanol/diesel blends", *Fuel*, 160, pp.217-226 (2015).
[32] Samaniego, E., Anitescu, C., Goswami, S., et al. "An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications", *Computer Methods in Applied Mechanics and Engineering*, *362*, p.112790 (2020).
[33] Vu-Bac, N., Lahmer, T., Zhuang, X., et al. "A software framework for probabilistic sensitivity analysis for computationally expensive models", *Advances in Engineering Software*, *100*, pp.19-31 (2016).
[34] Bauer Jr, K.W., Parnell, G.S. and Meyers, D.A. "Response surface methodology as a sensitivity analysis tool in decision analysis", *Journal of Multi-Criteria Decision Analysis*, *8*(3), pp.162-180 (1999).
[35] Awad, O.I., Mamat, R., Ali, O.M., et al. "Response surface methodology (RSM) based multi-objective optimization of fuel oil-gasoline blends at different water content in SI engine", *Energy Conversion and Management*, 150, pp.222-241 (2017).
[36] Sarıkaya, M. and Güllü, A. "Taguchi design and response surface methodology based analysis of machining parameters in CNC turning under MQL", *Journal of Cleaner Production*, 65, pp.604-616 (2014).
[37] Rao, K.S. "Effect of EGR on Diesel Engine Performance and Exhaust Emission Running with Cotton Seed Biodiesel", *International journal of mechanical and mechatronic engineering*, 16(02), pp.64-69 (2016).
[38] Islam, M., Ahmed, A.S., Islam, A., et al. "Study on emission and performance of diesel engine using castor biodiesel", *Journal of Chemistry*, (*2014)*.
[39] Ge, J.C., Yoon, S.K., Kim, M.S. et al. "Application of canola oil biodiesel/diesel blends in a common rail diesel engine", *Applied Sciences*, *7*(1), p.34 (2017).

[40] de Carvalho, M.A., Achy, A.R., Junior, L.C., et al. "Mechanical and emissions performance of a diesel engine fueled with biodiesel, ethanol and diethyl ether blends", *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *42*(4), pp.1-10 (2020).
[41] Channapattana, S.V., Pawar, A.A. and Kamble, P.G. "Optimisation of operating parameters of DI-CI engine fueled with second generation Bio-fuel and development of ANN based prediction model", *Applied energy*, 187, pp.84-95 (2017).
[42] Kanthasamy, P., Selvan, V., Mozhi, A. et al. "Investigation on the performance, emissions and combustion characteristics of CRDI engine fueled with tallow methyl ester biodiesel blends with exhaust gas recirculation", *Journal Of Thermal Analysis And Calorimetry (2020)*.

Figure 1: The transesterification process

Figure 2: The common rail direct injection (CRDI) diesel engine setup

Figure 3: Flowchart for the RSM approach

Figure 4: RSM optimizer

Figure 5: The effect of various engine input factors on Brake thermal efficiency

Figure 6: The effect of various engine input factors on mechanical efficiency

Figure 7: The effect of various engine input factors on brake effective mean pressure

Figure 8: The effect of various engine input factors on torque

Figure 9: The effect of various engine input factors on the emission of HC

Figure 10: The effect of various engine input factors on the CO_2 emission

Figure 11: The variation between the predicted values and experimental values of engine outcomes (refer Table 11)

Table 1: Thermo-Physical Property of different blends of Linseed biodiesel

Table 2: The technical specification of the CRDI VCR engine

Table 3: The technical specification of AVL gas analyzer

Table 4: Uncertainty percentage

Table 5: Input parameters with their levels

Table 6: Experimental design matrix

Table 7: ANOVA results for torque, BTE, and mechanical efficiency

Table 8: ANOVA results for BMEP, HC, and CO_2

Table 9: Model summary

Table 10: The experimental values with their mean and standard deviation of various engine output parameters.

Table 11: Validation of output variables

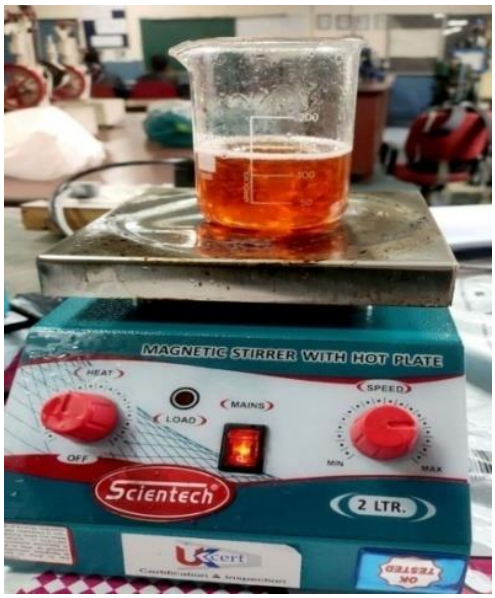


Figure 1: The transesterification process [30].

Figure 2: The CRDI (Common rail direction injection) diesel engine setup [30].

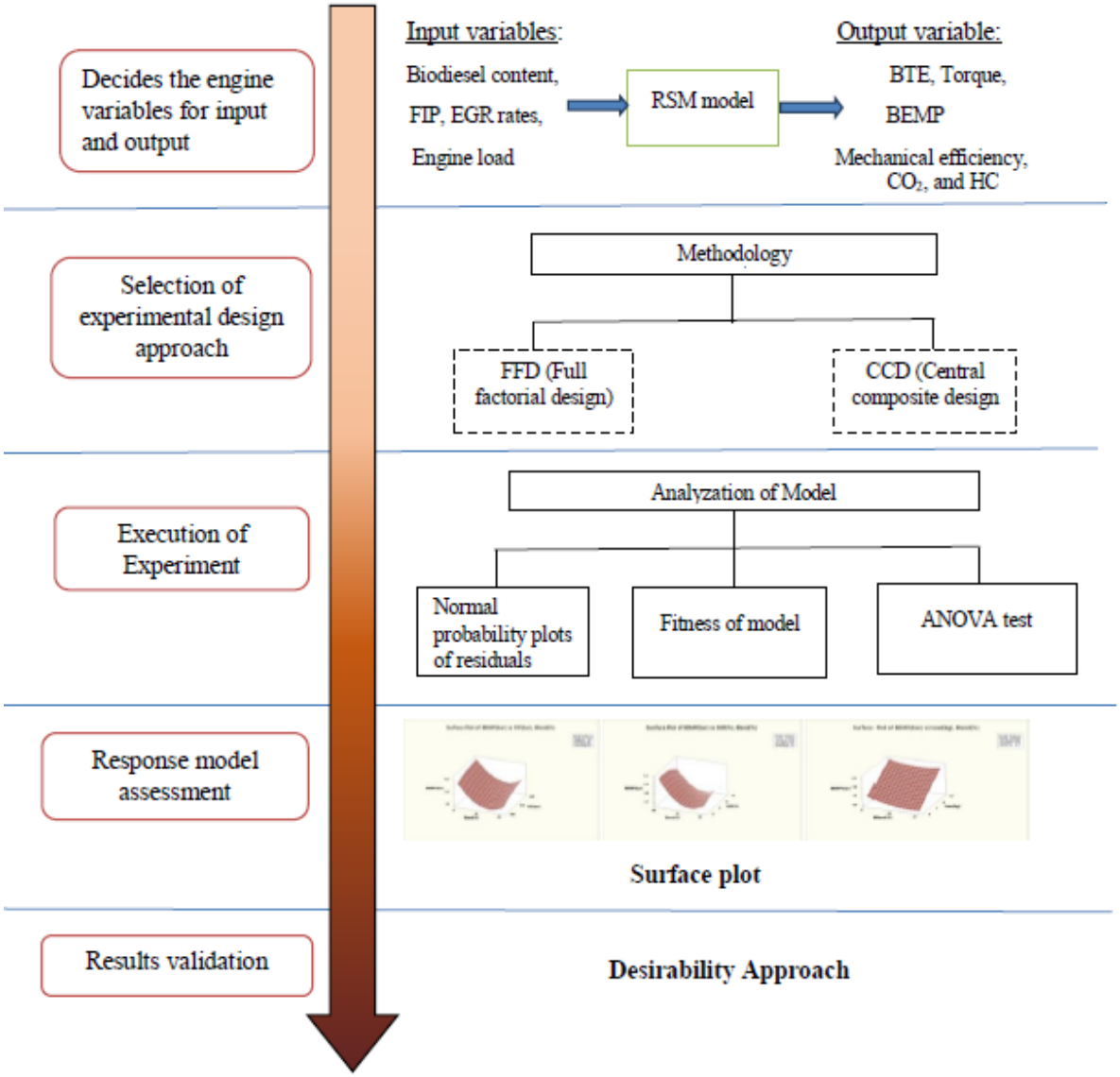


Figure 3: Flowchart for the RSM (Response surface methodology) approach.

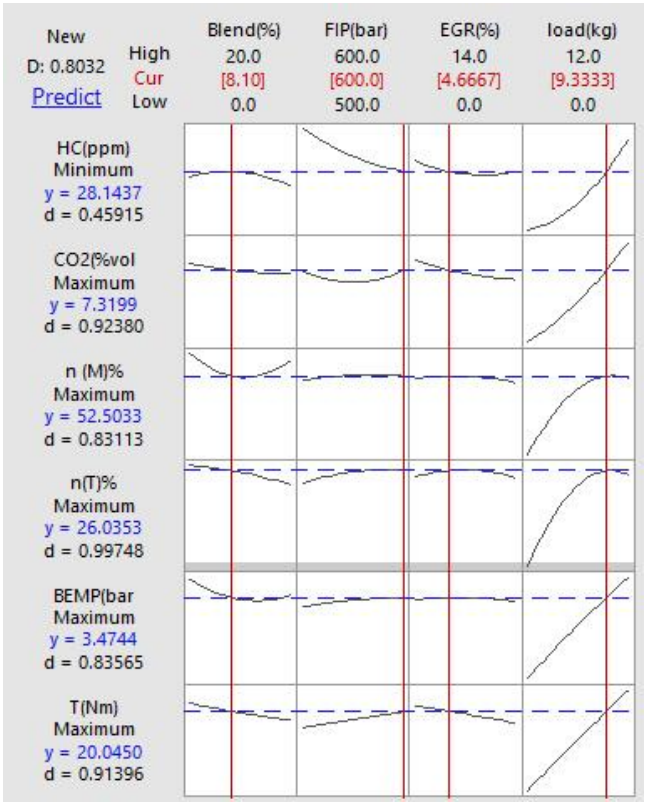
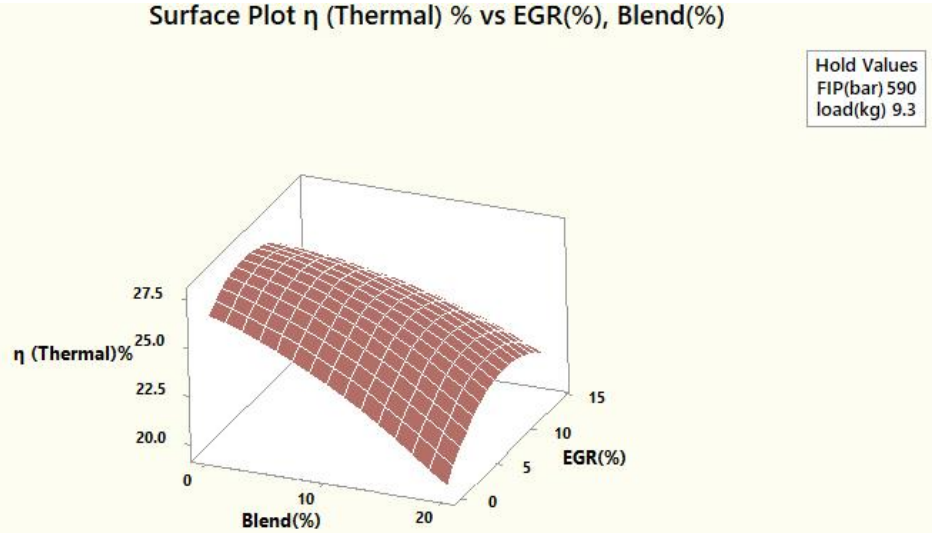
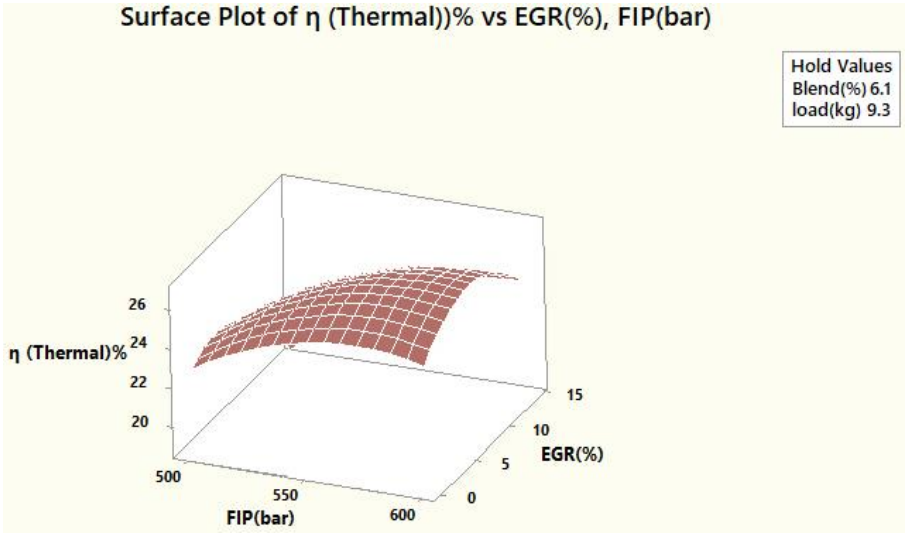


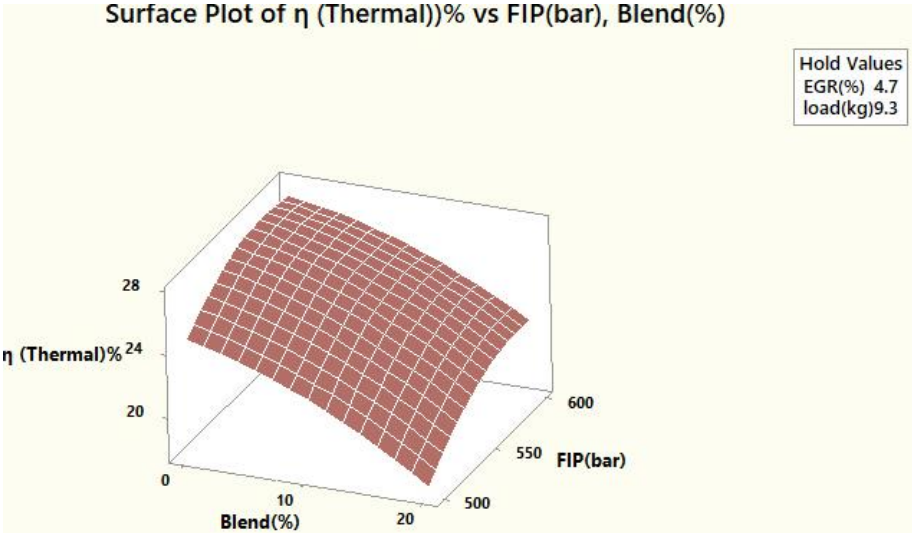
Figure 4: RSM (Response surface methodology) Optimizer.



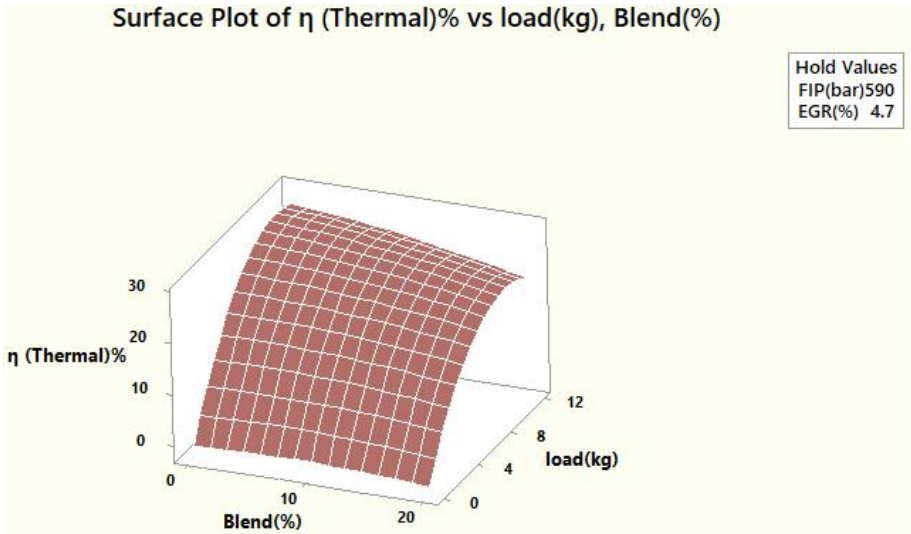
5(i)


Surface Plot of η (Thermal))% vs EGR(%), FIP(bar)
Hold Values
Blend(%) 6.1
load(kg) 9.3
η (Thermal)%
26
24
22
20
500
550
600
FIP(bar)
0
5
10
15
EGR(%)

5(ii)

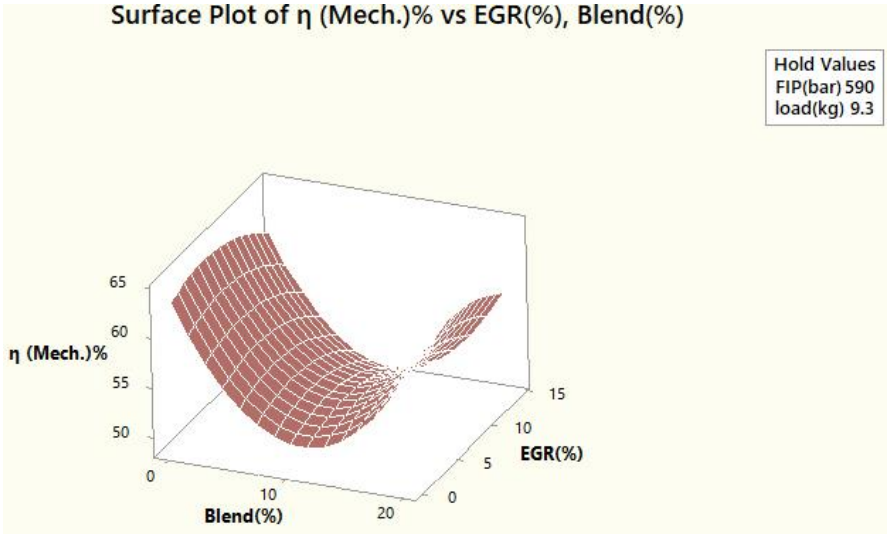

Surface Plot of η (Thermal))% vs FIP(bar), Blend(%)
Hold Values
EGR(%) 4.7
load(kg)9.3
η (Thermal)%
28
24
20
0
10
20
Blend(%)
500
550
600
FIP(bar)

5(iii)

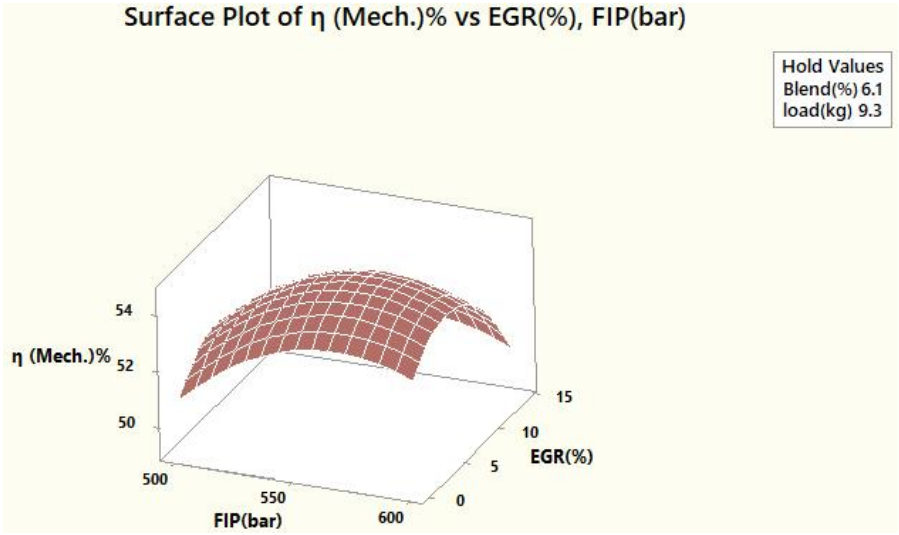

Surface Plot of η (Thermal)% vs load(kg), Blend(%)
Hold Values
FIP(bar)590
EGR(%) 4.7
η (Thermal)%
30
20
10
0
0
10
20
Blend(%)
0
4
8
12
load(kg)

5(iv)

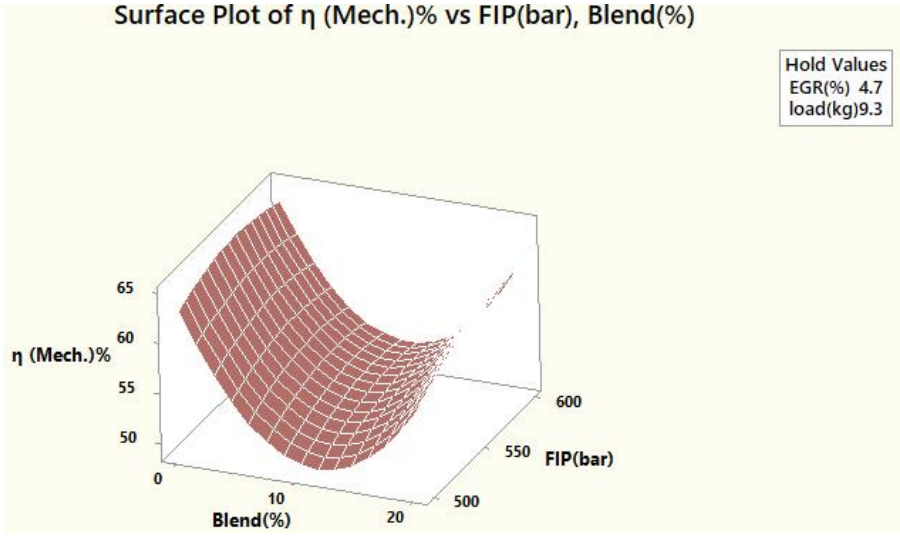
Figure 5: The effect of various engine input factors on brake thermal efficiency.



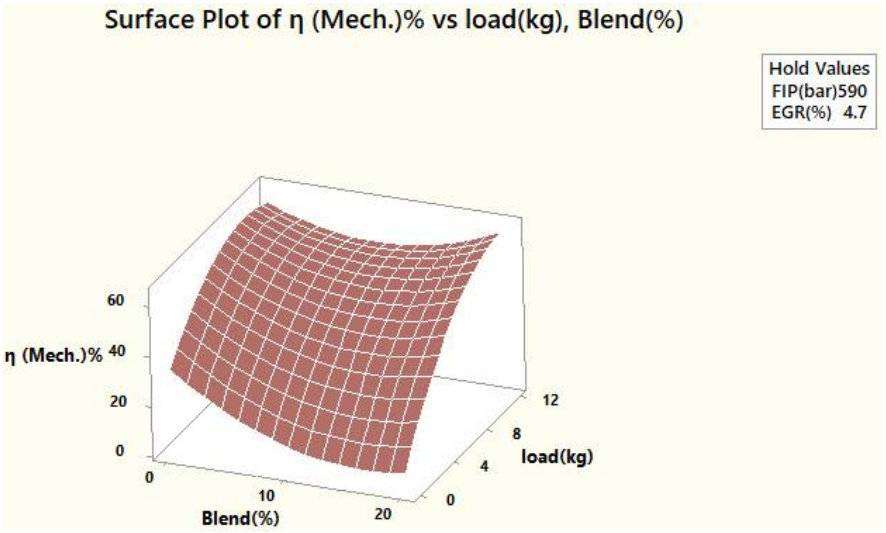
6(i)



6(ii)

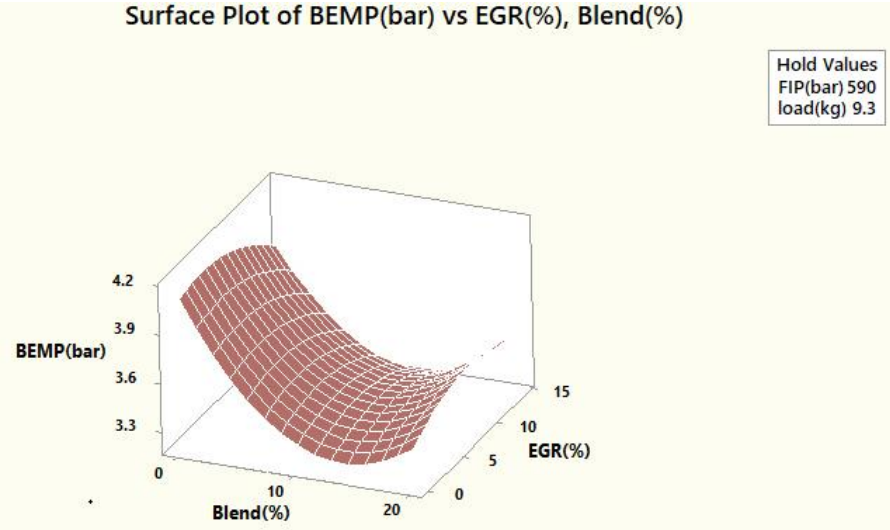


6(iii)

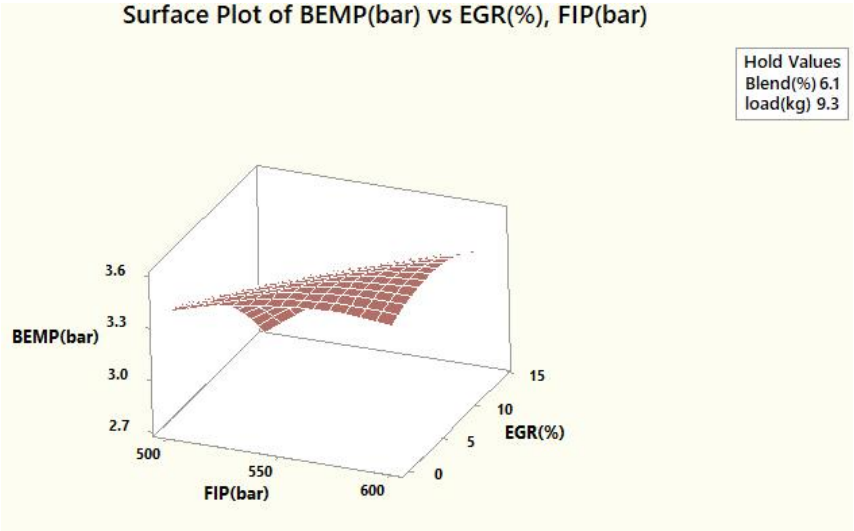


6(iv)

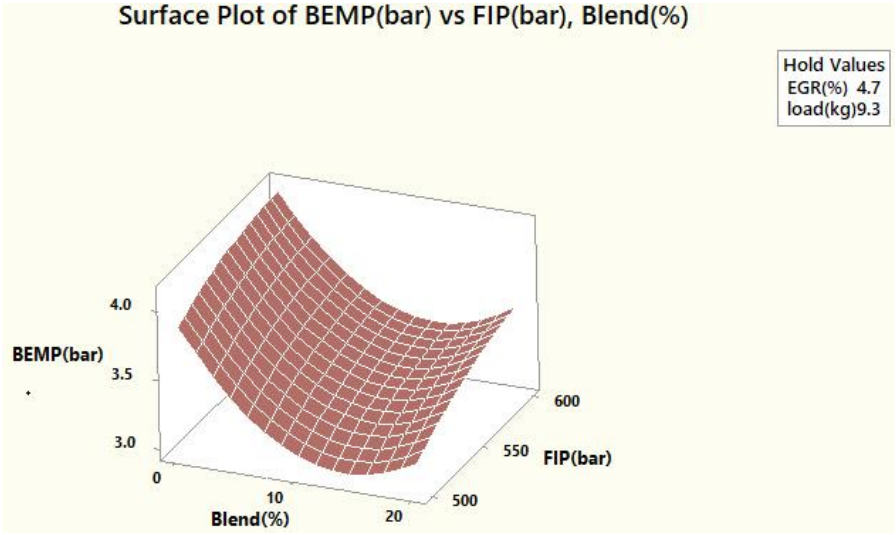
Figure 6: The effect of various engine input factors on mechanical efficiency η (Mech).


Surface Plot of BEMP(bar) vs EGR(%), Blend(%)
Hold Values
FIP(bar) 590
load(kg) 9.3
BEMP(bar)
4.2
3.9
3.6
3.3
0
10
20
Blend(%)
0
5
10
15
EGR(%)

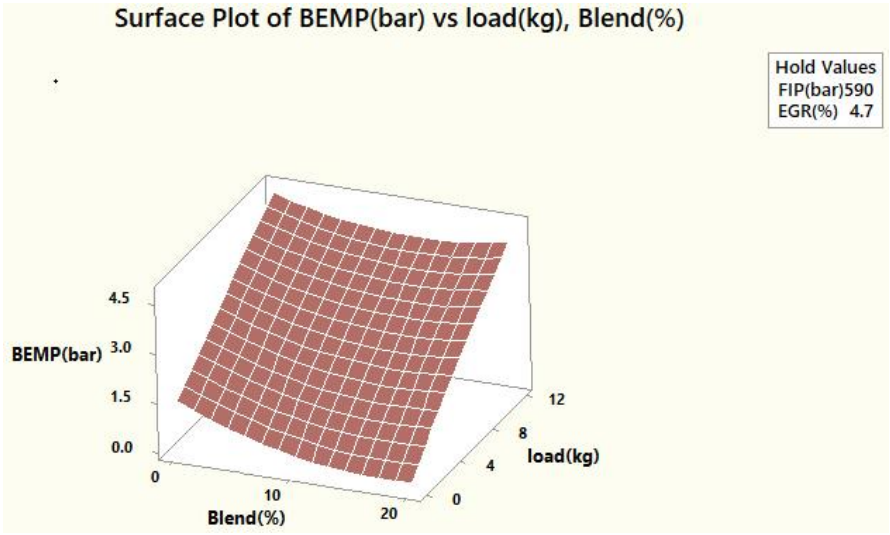
7(i)


Surface Plot of BEMP(bar) vs EGR(%), FIP(bar)
Hold Values
Blend(%) 6.1
load(kg) 9.3
BEMP(bar)
3.6
3.3
3.0
2.7
500
550
600
FIP(bar)
0
5
10
15
EGR(%)

7(ii)

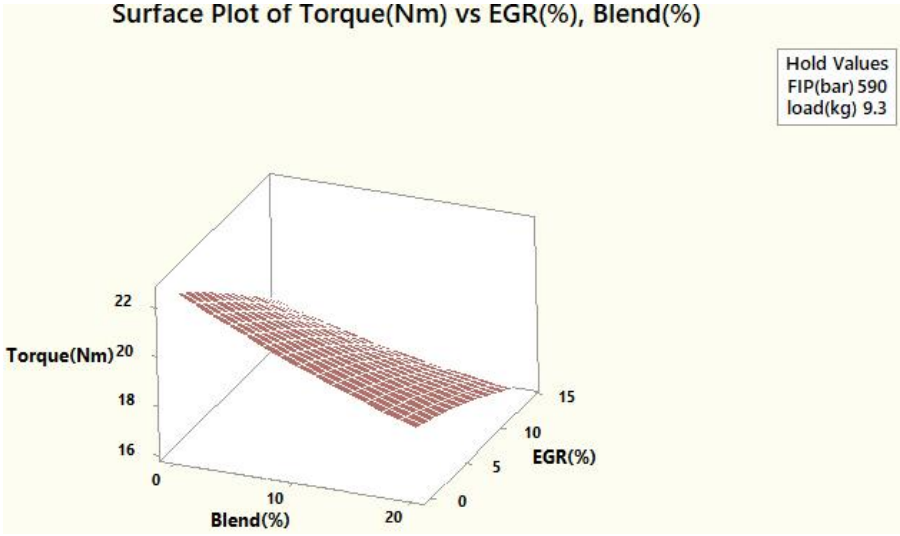


7(iii)

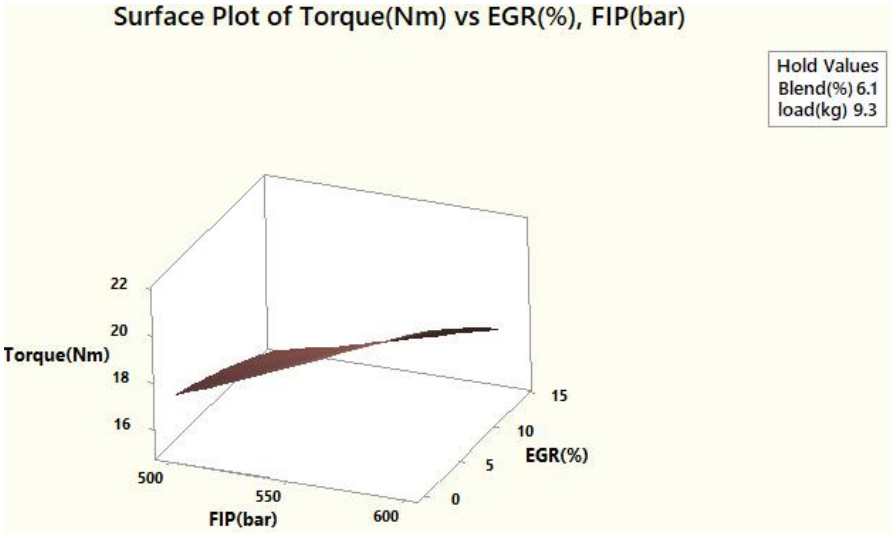


7(iv)

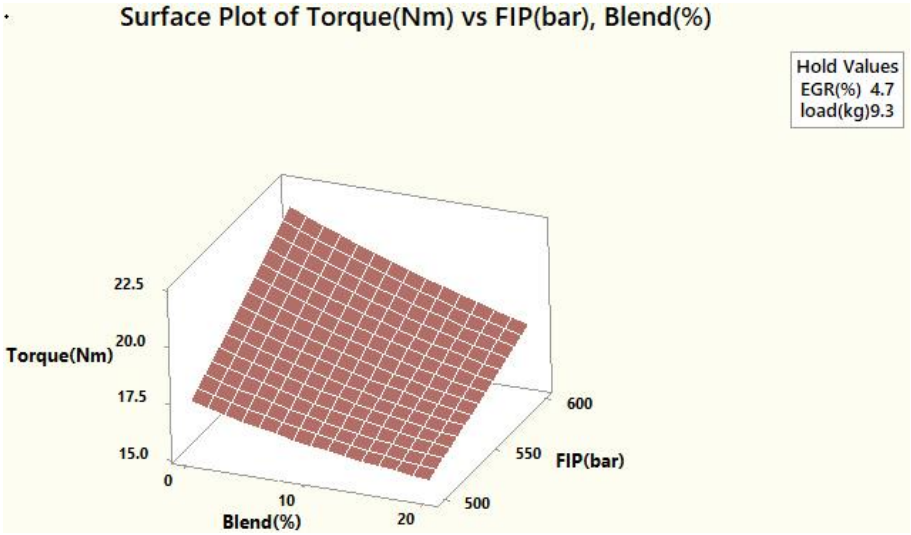
Figure 7: The effect of various engine input factors on brake mean effective pressure.


Surface Plot of Torque(Nm) vs EGR(%), Blend(%)
Hold Values
FIP(bar) 590
load(kg) 9.3
Torque(Nm)
22
20
18
16
0
10
20
Blend(%)
0
5
10
15
EGR(%)

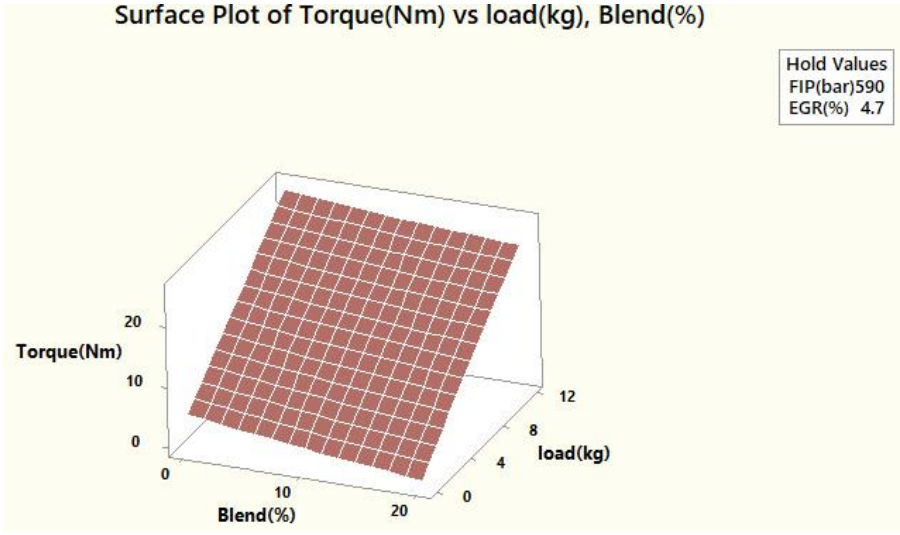
8(i)


Surface Plot of Torque(Nm) vs EGR(%), FIP(bar)
Hold Values
Blend(%) 6.1
load(kg) 9.3
Torque(Nm)
22
20
18
16
500
550
600
FIP(bar)
0
5
10
15
EGR(%)

8(ii)

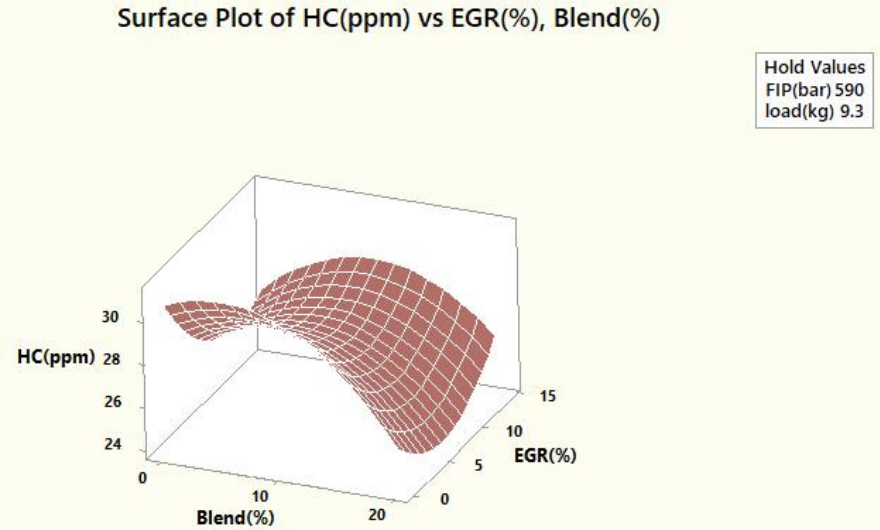


8(iii)

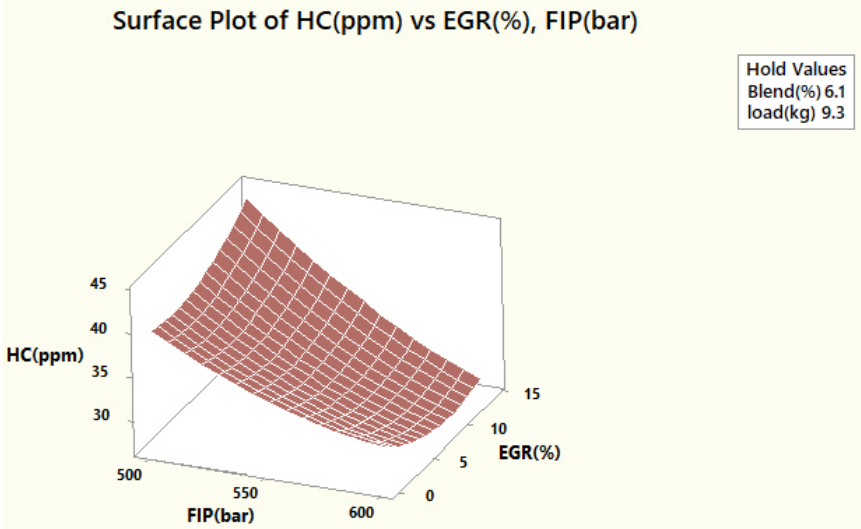


8(iv)

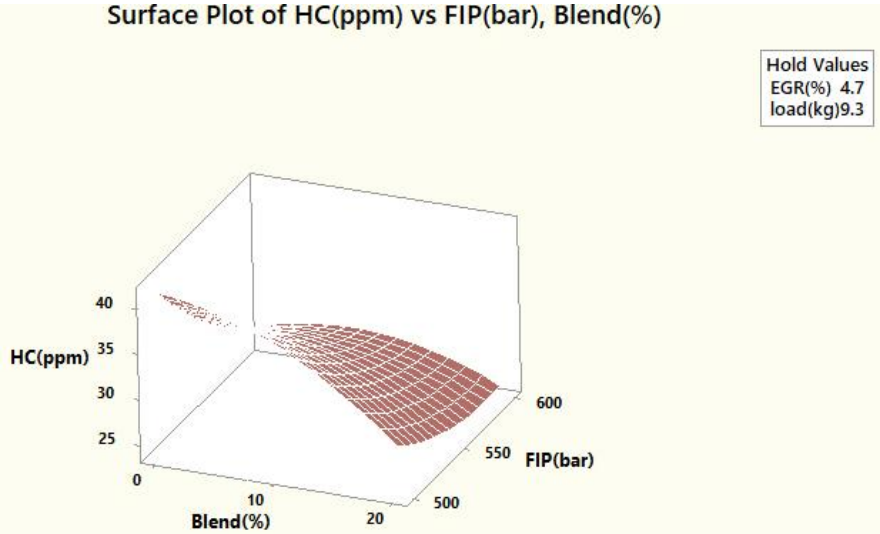
Figure 8: The effect of various engine input factors on torque.


Surface Plot of HC(ppm) vs EGR(%), Blend(%)
Hold Values
FIP(bar) 590
load(kg) 9.3
HC(ppm)
30
28
26
24
0
10
20
Blend(%)
0
5
10
15
EGR(%)

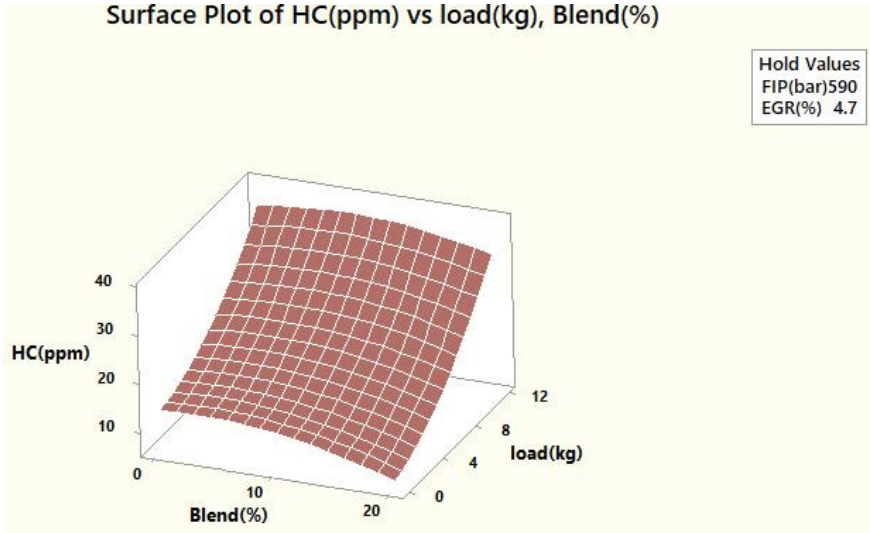
9(i)


Surface Plot of HC(ppm) vs EGR(%), FIP(bar)
Hold Values
Blend(%) 6.1
load(kg) 9.3
HC(ppm)
45
40
35
30
500
550
600
FIP(bar)
0
5
10
15
EGR(%)

9(ii)

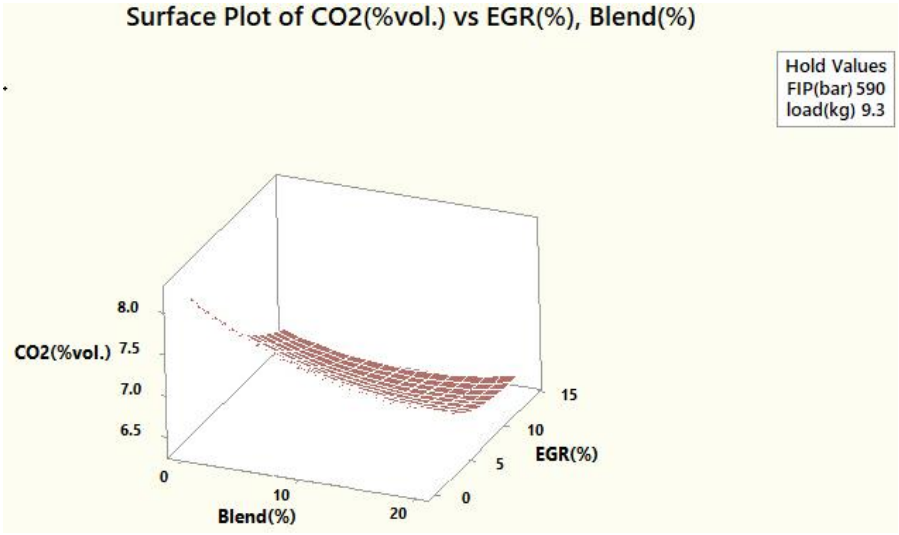


9(iii)

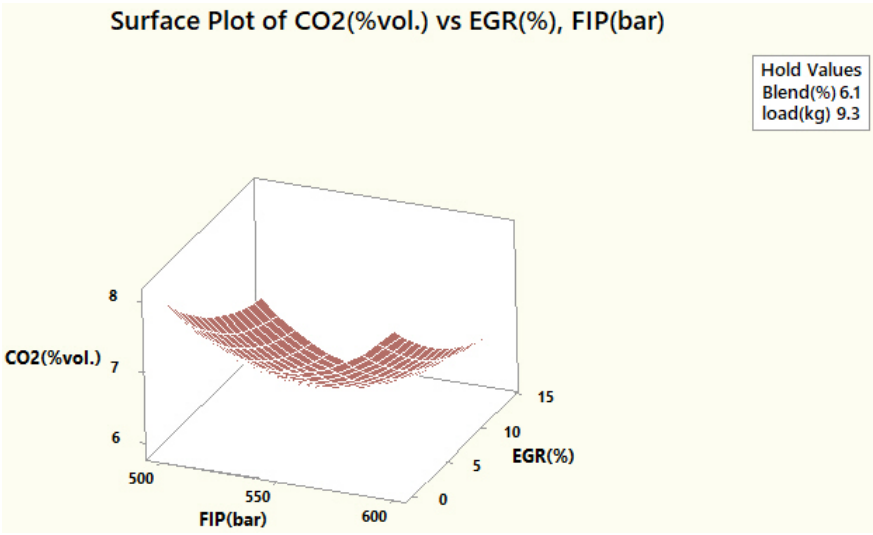


9(iv)

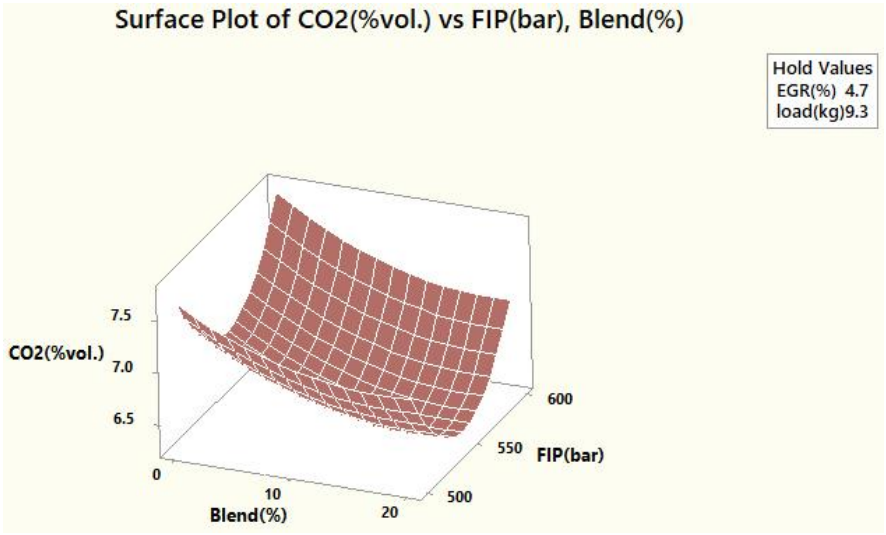
Figure 9: The effect of various engine input factors on the emission of HC (Hydrocarbon).


Surface Plot of CO2(%vol.) vs EGR(%), Blend(%)
Hold Values
FIP(bar) 590
load(kg) 9.3
CO2(%vol.)
8.0
7.5
7.0
6.5
0
10
20
Blend(%)
0
5
10
15
EGR(%)

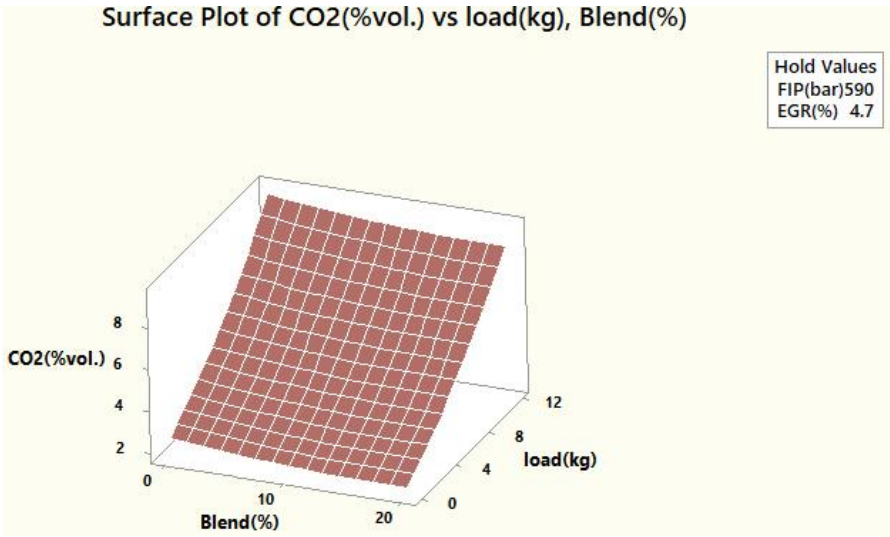
10(i)


Surface Plot of CO2(%vol.) vs EGR(%), FIP(bar)
Hold Values
Blend(%) 6.1
load(kg) 9.3
CO2(%vol.)
8
7
6
500
550
600
FIP(bar)
0
5
10
15
EGR(%)

10(ii)



10(iii)



10(iv)

Figure 10: The effect of various engine input factors on the CO_2 (Carbon dioxide) emission.

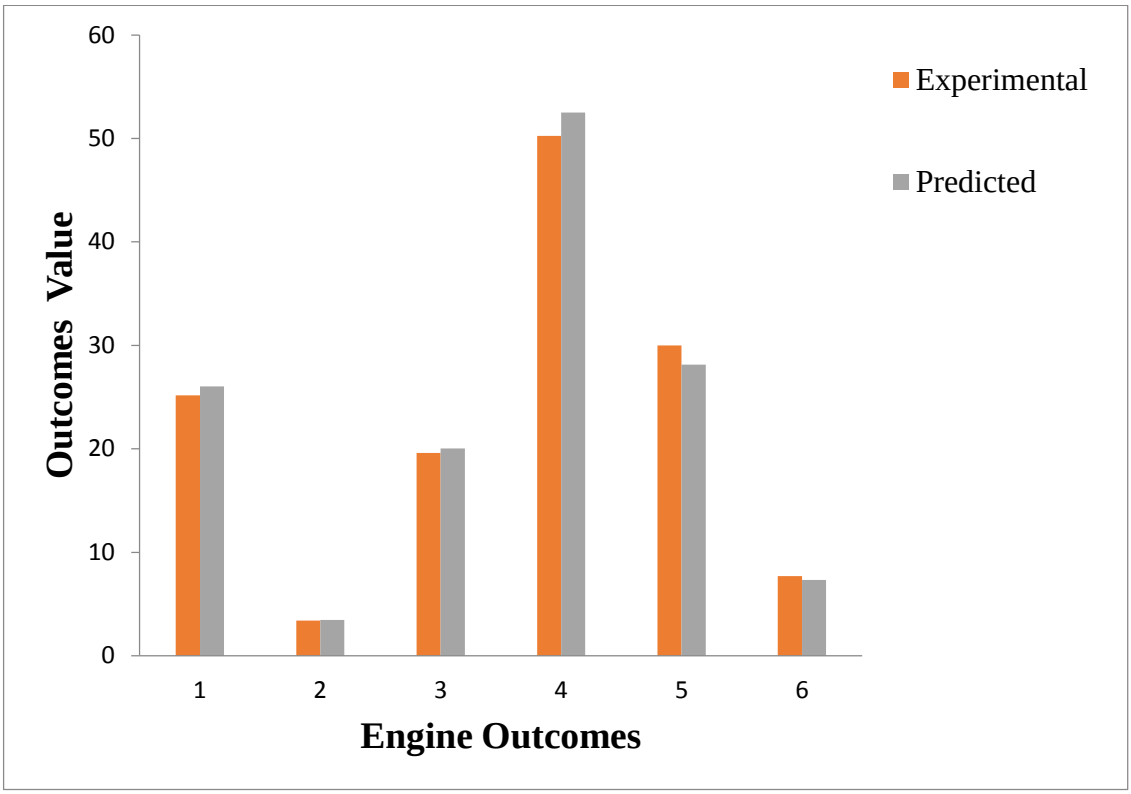


Figure 11: The variation between the predicted values and experimental values of engine outcomes (refer Table 11).

Table 1: Thermo-Physical Property of different blends of Linseed biodiesel

Properties	ASTM Method	Diesel	LB5	LB10	LB15	LB20	Linseed methyl ester
Kinematic viscosity(mm^2/sec)@40° C	ASTM D7042	2.9783	3.0185	3.0774	3.1058	3.1648	3.8894
Density(g/cm^3)	ASTM D7042	0.830	0.8338	0.8382	0.8424	0.8461	0.8757

Calorific value(kJ/kg)	D6751 37	42100	41750	41400	41050	40600	39100

Table 2: The technical specification of the CRDI VCR (Common rail direct injection Variable compression ratio) engine

Make	**KIRLOSKAR**
Product	(computerized) CRDI VCR Engine test setup 244
Engine	4-stroke, Single cylinder, stroke 110mm, 661 cc, bore 87.5 mm, water-cooled. Power 3.5 KW @ 1500 rpm, range of CR 12-18:1.
Connecting rod length	234 mm
Dynamometer	Eddy current, water-cooled type with loading unit
Injector	Type Solenoid driven
EGR	Water cooled, Range 0-15%
Piezo sensor	Combustion: Range 350Bar with low noise cable
Load indicator	Digital, Range 0-50 Kg, Supply 230VAC
Load sensor	Load cell, type strain gauge, range 0-50 Kg
Fuel flow transmitter	DP transmitter, Range 0-500 mm WC
Temperature sensor	Thermocouple, Type K
Software	"Enginesoft" Engine performance analysis software
Rota-meter	Engine cooling 40-400 LPH; Calorimeter 25-250 LBH

Table 3: The technical specification of AVL gas analyzer

Parameters	**Specification**
Type	AVL DIX 650
Operating temperature	5...40^{o}C
Warm up time	Approx... 2 min
Dimension	W 344 x D 252 x H 85
Interface	USB, Bluetooth Class 1, RS 232 (AK Protokoll)
Voltage supply	Via AVL DITEST CDS Basic Unit: 11.25 V DC

Measurement range	CO_2 = 0 ... 20% vol.
	HC = 0 ... 30.000ppm vol.
Resolution	CO_2 = 0.01 % vol.
	HC = ≤ 2.000: 1 ppm vol.
Accuracy	CO_2 = < 16.0 % vol.: ± 0, 3 % vol., ± 3 % o.M.
	≥ 16.0 % vol: ± 5 % o.M.
	HC = < 2000ppm vol.: ±4ppm vol., ±3% o.M.
	≥ 5000ppm vol.: ±5% o. M.
	≥10000ppm vol.: ±10% o. M

Table 4: Uncertainty percentage

Measuring quantity	Instrument	Manufacturer (model)	Range	Uncertainty
CO_2 emission	Gas analyzer	AVL DIX 650	0- 20% vol.	± 5%
HC emission	Gas analyzer	AVL DIX 650	0-30.000 ppm vol.	± 5%

Table 5: Engine input parameters with their levels

Coded Levels	Blend (%)	FIP (bar)	EGR (%)	Load (kg)
-2	Diesel	500	0	0
-1	LB5	525	3.5	3
0	LB10	550	7	6
1	LB15	575	10.5	9
2	LB20	600	14	12

Table 6: Experimental design matrix

Exp. No.	(Linseed) Blend (%)	FIP (bar)	EGR (%)	Load (kg)	Torque (Nm)	BTE (%)	η (Mech)(%)	BMEP (bar)	HC (ppmv)	CO2 (%vol)
1.	15	575	10.5	9	16.19	22.37	50.406	3.068	29	6
2.	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
3	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
4.	10	550	14	6	9.49	20.41	37.247	1.799	29	3.7
5.	15	525	10.5	3	2.62	9.06	14.451	0.497	20	2.8
6.	5	525	10.5	3	4.33	9.43	21.718	0.820	25	2.85
7.	15	575	10.5	3	4.77	11.26	23.472	0.903	16	2.6

8	5	525	3.5	9	17.55	23.53	52.206	3.326	35	7.05
9.	5	525	10.5	3	4.33	9.43	29.851	0.820	25	3.05
10.	10	550	7	0	0.127	0.4	0.808	0.024	13	1.5
11.	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
12.	10	550	0	6	12.69	21.91	44.457	2.404	20	4.8
13.	5	575	3.5	9	19.7	25.73	55.044	3.372	31	6.85
14.	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
15.	10	550	7	12	21.92	26.1	57.450	4.153	41	7.8
16.	15	575	3.5	3	6.57	12.18	29.596	1.245	13	3.2
17.	5	575	10.5	9	17.9	24.64	52.802	3.392	25	6.25
18.	0	550	7	6	12.79	24.9	63.007	2.424	24	4.4
19.	5	575	10.5	3	6.48	11.63	31.144	1.304	21	2.85
20.	10	500	7	6	9.89	20.51	38.532	1.875	28	4.9
21.	10	600	7	6	12.79	23.31	44.615	2.423	20	4.7
22.	5	525	10.5	9	15.75	22.44	49.629	2.984	38	6
23.	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
24.	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
25.	15	525	3.5	9	15.84	21.06	49.750	3.001	30	6.35
26.	15	575	3.5	9	17.99	23.23	52.911	3.409	26	6.6
27.	15	525	10.5	9	15.84	20.17	52.880	3.001	30	6.2
28.	5	575	3.5	3	8.28	12.64	34.537	1.569	18	3.45
29.	10	550	7	6	11.29	21.41	41.688	2.139	23	4.2
30.	15	525	3.5	3	4.42	9.98	22.062	0.838	17	3.4
31.	20	550	7	6	10.19	19.85	43.827	1.931	16	4

Table 7: ANOVA results for torque, BTE, and mechanical efficiency

		Torque				**BTE**				**Mechanical efficiency**			
Sources	DF	Adj. SS	Adj. MS	F-value	P-value	Adj. SS	Adj. MS	F-value	P-value	Adj. SS	Adj. MS	F-value	P-value
Regression	4	846.635	211.659	1495.60	0.000	976.08	244.021	27.44	0.000	4693.40	1173.35	36.74	0.000
Blend (%)	1	11.797	11.797	83.36	0.000	19.25	19.254	2.17	0.153	221.22	221.22	6.93	0.014
FIP (bar)	1	18.936	18.936	133.80	0.000	21.63	21.633	2.43	0.131	88.19	88.19	2.76	0.109
EGR (%)	1	15.073	15.073	106.51	0.000	10.07	10.074	1.13	0.297	66.04	66.04	2.07	0.162
Load (kg)	1	775.822	775.822	5482.04	0.000	901.89	901.886	101.42	0.000	4202.62	4202.62	131.58	0.000
Lack-of-fit	19	3.680	0.194	*	*	23.21	12.169	*	*	797.35	41.97	8.88	0.003

NOTE: BTE: Brake thermal efficiency

Table 8: ANOVA results for BMEP, HC, and CO_2

		BMEP				**HC**				**CO_2**			
Sources	DF	Adj. SS	Adj. MS	F-value	P-value	Adj. SS	Adj. MS	F-value	P-value	Adj. SS	Adj. MS	F-value	P-value
Regression	4	28.8027	7.4507	181.49	0.000	1170.82	292.705	51.91	0.000	67.6269	16.9067	183.41	0.000
Blend (%)	1	0.9670	0.9670	23.56	0.000	104.70	104.704	18.57	0.000	0.2696	0.2696	2.93	0.099
FIP (bar)	1	0.5922	0.5922	14.42	0.001	122.02	122.017	21.64	0.000	0.0301	0.0301	0.33	0.573
EGR (%)	1	0.4834	0.4824	11.78	0.002	42.36	42.362	7.51	0.011	1.7860	1.7860	19.38	0.000
Load (kg)	1	26.9597	26.9597	656.70	0.000	902.44	902.444	160.05	0.000	63.4019	63.4019	687.80	0.000
Lack-of-fit	19	1.0674	0.0562	*	*	146.60	7.716	*	*	2.3767	0.1251	43.78	0.000

NOTE: BMEP: Brake mean effective pressure; HC: Hydrocarbon; CO_2: Carbon dioxide

Table 9: Model summary

Model	**S**	**R^2 (%)**	**Reg. R^2 (%)**	**Pred. R^2 (%)**

Torques	0.376192	99.57	99.50	99.32
BTE	2.98206	80.85	77.90	72.35
Ƞ (Mech.)	5.65152	84.97	82.65	76.68
BMEP	0.202615	96.54	96.01	94.61
HC	2.37453	88.87	87.16	82.95
CO_2	0.303612	96.58	96.05	95.26

NOTE: R^2: Correlation coefficient

Table 10: The experimental values with their mean and standard deviation of various engine output parameters.

Output parameters	**Experimental Outcomes (No. of trials)**			**Mean Value**	**Standard deviation**
	1	**2**	**3**		
BTE (%)	25.23	24.66	25.62	25.17	0.4662
BMEP (bar)	3.382	3.415	3.379	3.392	0.0008
Torque (Nm)	19.91	19.16	19.76	19.61	0.315
η (Mech)	50.78	48.94	51.06	50.26	2.65
HC(ppmv)	30	30	30	30	0
CO_2 (%vol.)	7.549	7.921	7.621	7.697	0.0777

Table 11: Validation of output variables

Engine setting at blend ratio linseed biodiesel 8.10%, FIP 600 bar, EGR 4.66 %, and engine load 9.33 kg

Engine outcomes	BTE (%)	BMEP (bar)	Torque (Nm)	η (Mech)	HC(ppmv)	CO_2 (%vol.)
	1	2	3	4	5	6
Predicted value (RSM)	26.035	3.474	20.04	52.503	28.14	7.319
Experimental value	25.17	3.392	19.61	50.26	30	7.697
Error (%)	3.32	2.36	2.14	4.27	6.60	5.1

Biographies

Manish Kumar is a research Scholar (M.Tech). He has completed his M.Tech from Delhi Technological University (formally DCE), Delhi in Thermal engineering and B.Tech in

Mechanical Engineering from Dev bhoomi institute of technology, Dehradun. His Research interest includes renewable energy, bio fuels, IC engine etc. He has a research paper in SCI journal and a international conference paper.

Naushad Ahmad Ansari is presently working as an Assistant Professor in the Department of Mechanical Engineering, Delhi Technological University, Delhi. He has more than 20 years of teaching experience. His specialization is Thermal Engineering. He has published various research articles in reputed International journals and also reviewed various articles of the journals.

Abhishek Sharma has 17years of International Involvement in Academic Teaching & Research across recognized Indian and African Universities. He Publishes over fifty research articles in SCI/SCIE & SCOPUS indexed Journals; other peer-reviewed Journals; ISBN Book; IEEE Conference proceedings; and Serves as the Reviewer and Editorial Member in peer-reviewed Journals and Conferences.

Varun Kumar Singh is a research Scholar (M.Tech). He completed his M.Tech from Delhi Technological University (formally DCE), Delhi in Thermal engineering and B.Tech in Mechanical Engineering from SCRIET (CCS University), Meerut. His Research interest includes Renewable energy, Bio fuels and IC engine. He has a research paper in SCI journal and one international conference paper.

Raghvendra Gautam is presently working as an Assistant Professor in the Department of Mechanical Engineering, Delhi Technological University, Delhi. He has more than 11 years of teaching experience. His specialization is Thermal Engineering. He has published various research articles in reputed International journals and also reviewed various articles of the journals.

Yashvir Singh is presently working as an associate professor in the Department of Mechanical Engineering, Graphic Era Deemed to be University, Dehradun. He has more than 12 years of teaching experience. His specialization includes areas like tribology, biofuels, lubrication, manufacturing, etc. He published various research papers in reputed journals and also reviewed various articles of the journals.

Electrical Engineering
https://doi.org/10.1007/s00202-021-01263-z

ORIGINAL PAPER

Quaternions-based detection of power quality events

Rajiv Kapoor[1] · **Swapnil Verma**[1] · **Aarchishya Kapoor**[2]

Received: 9 May 2020 / Accepted: 4 March 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
This paper presents a new method for the detection of voltage disturbances caused by various reasons in households and industries. The aim of the paper is to improve the power quality disturbance methods and implement it in real-time with a very high accuracy and use existing mitigation techniques after classification of the signal. The method involves a novel idea to use the quaternions for the purpose of signal tracking. The envelope formed by Hilbert envelope detection method is used as an input for the quaternions. The quaternions produce a spike whenever there is a deviation from the pure sine wave giving very accurate results in a very short period of time. Combination of SVM and DSmT classifier is used for recognizing and classification of different signals such as sags, swells, transients and other power quality disturbances, hence can be used to monitor the power quality.

Keywords Signal tracking · Power disturbance · DSmT · ESPRIT · Power quality detection · Sag · Swell · SVM · Quaternions

1 Introduction

The monitoring of power quality disturbances has become important in the new era of heavy nonlinear solid-state devices, microprocessor-based equipment and power electronics-based loads. For the purpose of reliability and better servicing of power, analysis of power quality disturbances must be done carefully. The power quality disturbances comprise voltage sag, swell, harmonics, transient, flickers, spikes, interruption and multiples disturbances such as sag with harmonics and swell with harmonics, these multiple disturbances comes into picture in combo. The main reasons of disturbances are degraded power factor, capacitor, fast switching inductor, motor drives with unbalanced hefty loads. To deal with all these concerns, recognition and classification of disturbances is the major assignment so that decisive measures can be applied in very short duration of disturbances. In earlier state of art numerous power quality analysis techniques were suggested in the last ten years, by reviewing the literature it has been observed that there are two main steps involved in analysis of power quality disturbances, first step comprises of separation of relevant features using short time Fourier transform (STFT), S-Transform, wavelet transform (WT) and Hilbert–Huang transform (HHT) are main techniques for spectrum, analysis of non-stationary power disturbances [1, 2]. STFT suffered with fixed window size and under heavy noise conditions efficiency of WT [4] severely affected; however the role of S-transform is to extract features but computational complexity makes it unsuitable for real-time application. In case of HHT mainly implemented for recognition of power quality events but having problem of end points perturbation of empirical mode decomposition [3]. Deep learning approach is also explored for detection of power quality events but has a problem of too much computation [15, 21].

In earlier state of art numerous classification techniques were explored commonly based on neural networks such as radial basis feed forward neural network (RBFFNN) and probabilistic neural network (PNN) [5]. A Volterra series-based technique is also utilized for power signal disturbance detection; it uses the features of Power spectral entropy and standard deviation and classification is done with type-II fuzzy logic [6, 7, 9]. The extracted optimized features act as

✉ Rajiv Kapoor
rajivkapoor.dtu@gmail.com

Swapnil Verma
swapnilve@gmail.com

Aarchishya Kapoor
akapoor6_be19@thapar.edu

[1] Delhi Technological University, New Delhi, India

[2] Thapar Institute of Engineering, Patiala, India

Published online: 13 April 2021

input to above-mentioned methods originally developed for single disturbance and also having the problem of complex time-consuming computation restricting the implementation in real-time application. These issues generate the need of efficient, accurate, reliable and express novel methods for solving the issue of power quality disturbance. Multi-objective problem has been used and NSGA-II has been compared extensively with NSGA-III [24] where the idea is to minimize three objectives simultaneously. The first one considers the minimization of the total flow time of jobs in the production. In [25] authors, classified complex power quality disturbance using optimized S-Transform and kernel SVM, where optimized S-transform decomposes the frequency components in two parts, whereas in [26], authors make the cluster of power quality index data, then using features like weather, active power, reactive power are taken and based upon this they train the model using SVM methods.

This paper proposes a new algorithm of enhanced NSGA-III procedure, termed as adaptive €-NSGA-III, for revolutionary multiple power quality event optimization. Adaptive € -NSGA-III succeed the attributes key point in the actual NSGA-III, covering adaptive realization and diversification-sustentation encourage using uniquely selected points. Though, slightly change from original, a new alternative alliance, labeled as €-dominance, is explored rather than Pareto dominance (PD) in the original. The encouragement for proposed novel technique stands on the subsequent discussion. Original method accentuates population set those are showing non-dominance until now and near to the reference boundary of each and every reference mark. Even though, when disturbance in power signal is serve, PD wants alternative tool to select the multiple disturbance points of the population towards PF, it means NSGA-III mainly focuses on diversity in comparison of convergence in multiple disturbance cases. It generates the need to enhance the confluence of original method for multiple disturbance scenario, with no any change in divergence detection capability. The suggested dominance has ability to achieve this goal. In the €-dominance, each power disturbance is allocated into distinct clusters denoted by well-defined reference points. The solutions of multiple power quality disturbances in the matching clump have the opposing alliance, role of fitness function similar to penalty-based boundary intersection (PBI) function. While performing random detection of multiple power quality events, the non-dominated sorting scheme with dominant function results in detection values in certain clump, additionally make sure that results lie within specified clump. In case of 20 dB SNR wavelet transform and Volterra series explored for feature extraction in ref [7, 12] and classifier utilized in this cases are type-1 fuzzy and type-2 fuzzy with classification potency of 92.56% and 93%, respectively. In Ref. [16, 20] discrete wavelet transform (DWT) and wavelet transform (WT) for feature extraction and probabilistic neural networks (PNN) are utilized for classification of PQ events which gives out efficiency of 96.47 and 98.64%, respectively, with SNR of 20 dB. In previous research work of [14] WT for detection of Power quality disturbances and classical fuzzy as classifier which gave the classification which gives out efficiency of 96.47% & 98.64%, respectively, with SNR of 20 dB. In previous research work of [14] WT for detection of Power quality disturbances and classical fuzzy as classifier which gave the classification efficiency of 95% with the noise of 20 dB SNR. In earlier work of [18, 19] feature extraction was done by s-transform but classifiers explored are support vector machine (SVM) and decision tree with efficiencies of 98.11% and 98.5%, respectively. In ref [13, 17] Gilbert transform (GT) and WT implemented for features extraction of PQ events but neural network-based classifiers of field probing neural network (FPNN) and radial basis function neural networks (RBFNN) applied with classification efficiencies of 94.56% and 97.85% percent. Again, WT applied in [15] for detection and features extraction and classifier used is SVM with classification accuracy of 96.37%.

In this paper, we have focused on single phase power quality improvement using a novel method involving quaternions and Hilbert transform. This is achieved by taking the pure sinusoidal signal as a reference signal and computing the deviations in a disturbed signal using the quaternion values.

Brief description about quaternions and its properties along with its application in detecting the disturbance in signal is given in Section II. In section III, the proposed methodology along with the parameters used in the process are discussed in detail. Section IV shows the simulation results of different power quality disturbances and behavior of the quaternions at the onset of any disturbance. Section V provides the conclusion of the results obtained by this approach.

2 Quaternions

2.1 Quaternion introduction

Quaternions were first invented by the mathematician William Hamilton. They form a number system that is an extension of complex numbers. A quaternion is defined as follows:

Let Q be the four-dimensional vector space over R with an ordered basis e, i, j and k, then a quaternion can be expressed as a vector:

$$q = we + xi + yj + zk \tag{1}$$

where $(w, x, y, z) \in R, q \in Q$.

Quaternions are non-commutative in nature and hence the algebra involved in understanding them is complex. The concept of reduced quaternions has helped overcomes this drawback. To reduce a quaternion, it is required that the product of any two of the basis vectors e, i, j, k is such that e acts as identity. The reduced quaternion q_r may be expressed as:

$$q_r = w + xi + yj + zk + q_r \in Q \tag{2}$$

such that, $i^2 = j^2 = k^2 = ijk = -1$ and $ij = -ji = k$, $jk = -kj = i, ki = -ik = j$.

Equation (2) represents the sum of a scalar quantity w and a vector $xi + yj + zk$, that is, $q_r = a + l$, where a is some scalar quantity belonging to the set of real numbers and l represents a vector. Note that this representation can also be used to represent quaternions as the sum of a real and an imaginary part, i.e., q_r can be written as

$$q_r = w + xi + yj + zk \tag{3}$$

where

1. Real part of

   $q_r, Re(q_r) = w$

2. Imaginary part of

   $q_r, Im(q_r) = xi + yj + zk$

3. Conjugate of

   $q_r = q_r^- = w - xi - yj - zk$

4. Magnitude of

   $q_r = |q_r| = \sqrt{q_r q_r}$

The matrix representation of Eq. (2) is given by,

$$q_r = (wxyz) \tag{4}$$

A quaternion can also be represented as,

$$q_r = |q_r| e^{i\varphi_x} e^{j\varphi_y} e^{k\varphi_z} \tag{5}$$

where $(\varphi_x, \varphi_y, \varphi_z)$ are reduced quaternion phase angles calculated as:

$$\varphi_x = \tan^{-1}\left(\frac{A}{B}\right) \tag{6}$$

where $A = 2(wx + yz)$

$B = w^2 - x^2 - y^2 + z^2$

$$\varphi_y = -\left[\sin^{-1}(2(xz - wy))\right] \tag{7}$$

$$\varphi_z = \tan^{-1}\left(\frac{2(wz + xy)}{w^2 + x^2 - y^2 - z^2}\right) \tag{8}$$

The most important properties of quaternions include rotation. The next sub-section discusses this property in detail.

2.2 Quaternion rotation

Any rotation in three dimensions can be expressed with respect to a set of three fixed axes, say, i, j, k. The fixed axes constitute an inertial frame of reference.

A quaternion has four elements that encode the information of orientation represented by these angles. Let q_i be a unit quaternion such that $q_i = (wxyz)$, in the matrix form as previously defined in Eq. (4). Here w, the scalar part, defines the amount of rotation that has been performed about the vector represented by $xi + yj + zk$.

If φ is the amount of rotation and the unit vector $t = li + mj + nk$ defines the axis of rotation, then the quaternion elements are calculated as

$$(wxyz) = \left(\cos\frac{\varphi}{2} l\sin\frac{\varphi}{2} m\sin\frac{\varphi}{2} n\sin\frac{\varphi}{2}\right) \tag{9}$$

Therefore, on comparing Eqs. (2) and (9), we see that q_r can be written as follows:

$$q_r = e^{\frac{\varphi}{2}(li+mj+nk)} = \cos\frac{\varphi}{2} + (limjnk)\sin\frac{\varphi}{2} \tag{10}$$

Quaternions provide a better measure of rotation than Euler angles because of the gimbal lock problem experienced due to the latter. To convert from quaternion representation to Euler Representation use the set of formula defined in Eqs. 5, 6 and 7 for $\varphi_x, \varphi_y, \varphi_z$. Now, once again consider Fig. 1. We can represent the orientation of the vector v using the direction cosines. The direction cosine matrix for the given vector is defined as left multiplication of three rotation matrices X, Y and Z, where

$$X = \left[1000\cos\varphi_x - \sin\varphi_x \sin\varphi_x \cos\varphi_x\right] \tag{11}$$

$$Y = \left[\cos\varphi_y\, 0\, \sin\varphi_y 010 - \sin\varphi_y\, 0\, \cos\varphi_y\right] \tag{12}$$

$$Z = \left[\cos\varphi_z - \sin\varphi_z 0\, \sin\varphi_z \cos\varphi_z 0001\right] \tag{13}$$

The complete direction cosine matrix on combining the three rotations is expressed as:

$$DCM = \left(A_1 - A_2 + A_3 + A_4 - A_5 - A6\right) \tag{14}$$

where

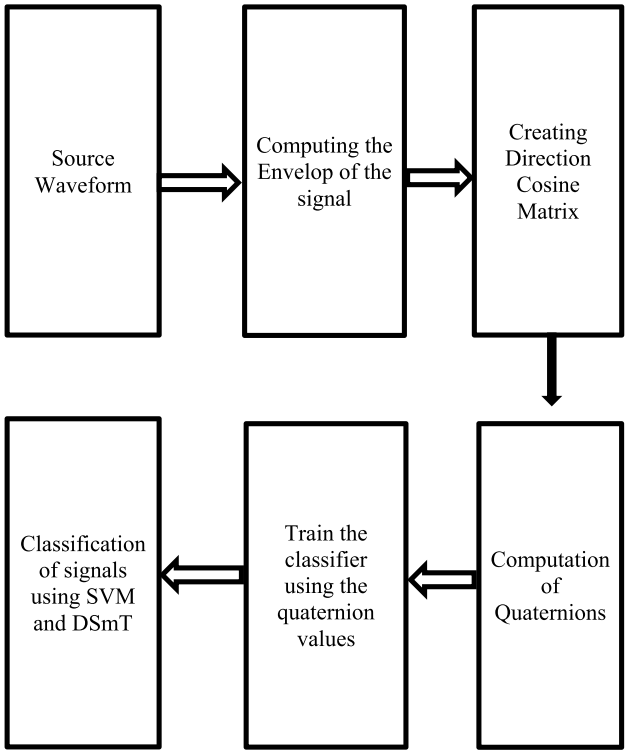


Fig. 1 Illustration of the method proposed

$$A_1 = \cos\varphi_y \cos\varphi_z \sin\varphi_x \sin\varphi_y \cos\varphi_z$$

$$A_2 = \cos\varphi_x \sin\varphi_z \cos\varphi_x \sin\varphi_y \cos\varphi_z$$

$$A_3 = \sin\varphi_x \sin\varphi_z \cos\varphi_y \sin\varphi_z \sin\varphi_x \sin\varphi_y \sin\varphi_z$$

$$A_4 = cos\varphi_x cos\varphi_z cos\varphi_x sin\varphi_y sin\varphi_z$$

$$A_5 = \sin\varphi_x \cos\varphi_z$$

$$A_6 = \sin\varphi_y \sin\varphi_x \cos\varphi_y \cos\varphi_x \cos\varphi_y$$

In the proposed method, we have considered a freely rotating vector that rotates in i,j plane, that is, the vector rotates only about k axis. This means that $\varphi_x = \varphi_y = 0$. On substituting $\varphi_x = \varphi_y = 0$ in Eq. (14), we obtain,

$$DCM = \left[\cos\varphi_z \; -\sin\varphi_z \; 0 \; \sin\varphi_z \; \cos\varphi_z \; 0 \; 0 \; 1\right] \tag{15}$$

which is also the DCM representing rotation about Z-axis as shown in Eq. (13). A DCM in terms of quaternion elements can be represented as follows:

$$DCM = \left(w^2 + x^2 - y^2 - z^2 \; 2(xy + wz) \; 2(xz - wy) \; 2(xy - wz) \; w^2 - x^2 + y^2 - z^2 \; 2(yz + wx) \; 2(xz + wy) \; 2(yz - wx) \; w^2 - x^2 - y^2 + z^2\right) \tag{16}$$

Using equating Eqs. (15) and (16), the value of $q_i = (wxyz)$ can be computed.

To apply a desired rotation on any vector $p = p_x i + p_y j + p_z k$, in 3D space, we consider a quaternion,q, with real component equal to zero and perform the conjugation of p by q, i.e.,

$$p' = qpq^{-1} \tag{17}$$

where p' represents the position vector of the new position obtained after rotating, i.e., if $p = li + yj + zk$, $q = w + xi + yj + zk$, and $q^{-1} = w - xi - yj - zk$, then,

$$p' = \left(0\, l'\, m'\, n'\right) \tag{18}$$

$$\begin{aligned} &= \left(0(w^2 + x^2 - y^2 - z^2)l + 2(xy - wz)m + 2(xz + wy)n2(xy + wz)l\right. \\ &+ (w^2 - x^2 + y^2 - z^2)m + 2(yz + wx)n2(xz - wy)l \\ &\left. + 2(yz + wx)m + (w^2 - x^2 - y^2 + z^2)n\right) \end{aligned} \tag{19}$$

Quaternions are known to be good trackers. They are used for the correct prediction of orientation. They find wide application in vector signal processing, in the field of computer vision and communication systems [13–17]. The simulation results obtained detects changes in waveforms accurately whenever a power quality disturbance occurs, thereby providing enough information about the accuracy and speed of the method. The method is explained in the next section.

3 Proposed method

Many of the previous discussed methods have a number of disadvantages which result in slow output and detection of the disturbances present in the signal. The method proposed in this paper tries to overcome these problems by providing fast and accurate results. In particular, many of the previously proposed methods for power quality improvement and detection are not very accurate and practically realizable in real-time. In this paper, the method proposed aims to fulfill these requirements. The use of Quaternions and Hilbert envelope detection method has been used to track the signal and produce the desired output whenever the disturbances occur.

The types of PQ disturbances that we face in our industries and homes should be familiar to us before the detection and mitigation takes place. The database of all the single and multiple PQ disturbances are listed in Table 1

Table 1 Numerical modeling of PQ disturbances

PQ event	Numerical model	Parameters
Pure sine	$V(t)=A\sin(t)$	$A = 1p.u., \omega = 2\pi 50 rad/sec$
Sag	$V(t) = \left[1-\beta\left\{u\left(t-t_1\right)-u\left(t-t_2\right)\right\}\right]\sin(t)$	$0.1 \le \beta \le 0.9, T \le t_2 - t_1 \le 9T$
Swell	$V(t) = \left[1+\beta\left\{u\left(t-t_1\right)-u\left(t-t_2\right)\right\}\right]\sin(t)$	$0.1 \le \beta \le 0.8, T \le t_2 - t_1 \le 9T$
Interruption	$V(t) = \left[1-\beta\left\{u\left(t-t_1\right)-u\left(t-t_2\right)\right\}\right]\sin(t)$	$0.9 \le \beta \le 1.0, T \le t_2 - t_1 \le 9T$
Flicker	$V(t)=\left[1+\beta_f sin(\sigma t)\right]\sin(\omega t)$	$0.1 \le \beta_f \le 0.2{,}5 \le \sigma \le 20Hz$
Oscillatory transients	$V(t)=\sin(t)+\beta e^{-(t-t_1)/\tau}\sin\left(t-t_1\right)\{u\left(t_2\right)-u(t_1)\}$	$0.1 \le \beta \le 0.8{,}0.5T \le t_2 - t_1 \le 3T, 8ms \le \tau \le 40ms, 300 \le f_n \le 900Hz$
Notch	$V(t)=\sin(\sin(t)-\text{sign}(\sin(t))$ $\times[\sum_{n=0}^{9}k\times\{u\left(t-\left(t_1+0.02n\right)\right)-u\left(t-t_2+0.02n\right))\}]$	$0.1 \le K \le 0.4{,}0 \le t_1, t_2 \le 0.5T, 0.01T \le t_2 - t_1 \le 0.05T$
Spike	$V(t)=\sin(t)+sign(\sin(t))$ $\times[\sum_{n=0}^{9}K\times\{u\left(t-\left(t_1+0.02n\right)\right)-u\left(t-t_2+0.02n\right))\}]$	$0.1 \le K \le 0.4{,}0 \le t_1, t_2 \le 0.5T, 0.01T \le t_2 - t_1 \le 0.05T$
Harmonics	$V(t)=\beta_1\sin(\omega t)+\beta_3\sin(\omega t)+\beta_5\sin(\omega t)+\beta_7 sin(\omega t)$	$0.05 \le \beta_3, \beta_5, \beta_7 \le 0.15, \sum \beta_i^2 = 1$
Sag with harmonics	$V(t)=\left(1-\beta\left(u\left(t-t_1\right)-u\left(t-t_2\right)\right)\right)\left(\beta_1\sin(\omega t)+\beta_3\sin(\omega t)\right.$ $\left.+\beta_5\sin(\omega t)+\beta_7\sin(\omega t)\right)$	$0.1 \le \beta \le 0.9, T \le t_2 - t_1 \le 9T$ $0.05 \le \beta_3, \beta_5, \beta_7 \le 0.15, \sum \beta_i^2 = 1$
Swell with harmonics	$V(t)=\left(1+\beta\left(u\left(t-t_1\right)-u\left(t-t_2\right)\right)\right)\left(\beta_1\sin(\omega t)\right.$ $\left.+\beta_3\sin(\omega t)+\beta_5\sin(\omega t)+\beta_7\sin(\omega t)\right)$	$1.1 \le \beta \le 1.8, T \le t_2 - t_1 \le 9T$ $0.05 \le \beta_3, \beta_5, \beta_7 \le 0.15, \sum \beta_i^2 = 1$
Flicker with harmonics	$V(t) = (1+\beta_f\sin(\sigma t))$ $\left(\beta_1\sin(\omega t)+\beta_3\sin(\omega t)+\beta_5\sin(\omega t)+\beta_7\sin(\omega t)\right)$	$0.1 \le \beta_f \le 0.2{,}5 \le \beta \le 20Hz$ $0.05 \le \beta_3, \beta_5, \beta_7 \le 0.15, \sum \beta_i^2 = 1$
Interruption with harmonics	$V(t)=\left(1-\beta\left(u\left(t-t_1\right)-u\left(t-t_2\right)\right)\right)(\beta_1\sin(\omega t)$ $+\beta_3\sin(\omega t)+\beta_5\sin(\omega t)+\beta_7\sin(\omega t))$	$0.9 \le \beta \le 1.0, T \le t_2 - t_1 \le 9T$ $0.05 \le \beta_3, \beta_5, \beta_7 \le 0.15, \sum \beta_i^2 = 1$

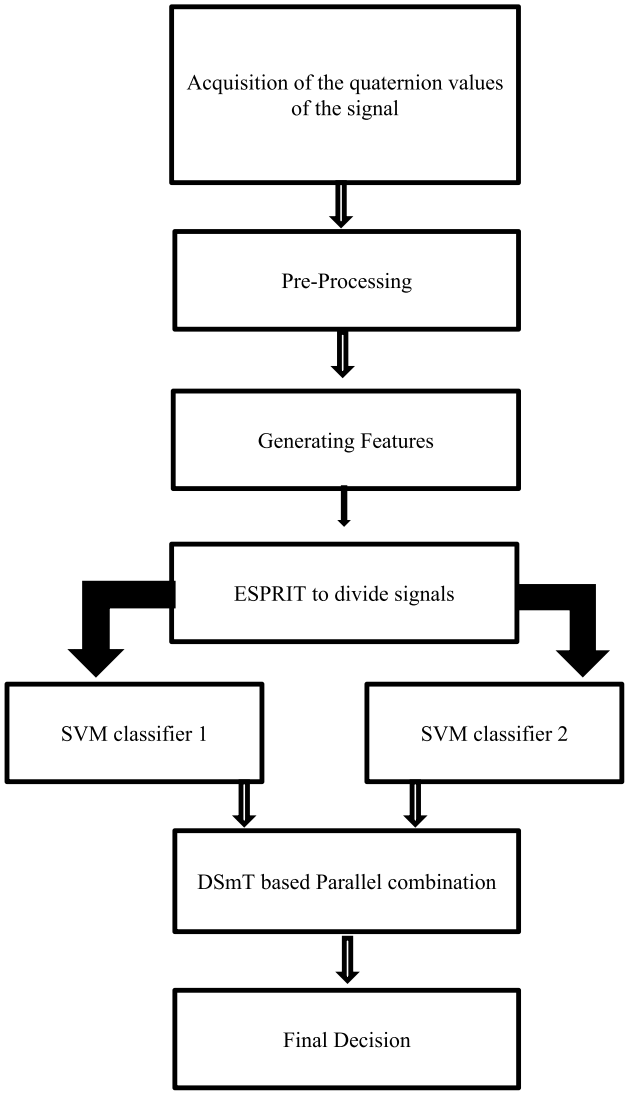


Fig. 2 Blocks diagram algorithm of proposed method

which will be advantageous in estimating the generalization ability of the classifier. The proposed method first creates and computes the envelope of the voltage signal. Envelope of a signal can be described as a smooth curve outlining the extremes of an oscillating signal. The envelope of the signal is calculated using the Hilbert transform. In this method the analytical signal of the input signal is created using the Hilbert transform. An analytic signal is a complex signal, where the real part is the original signal and the imaginary part is the Hilbert transform of the original signal. Mathematically, the envelope Y(t) of a signal X(t) is defined as the magnitude of the analytic signal as shown by the following equation.

$$Y(t) = \sqrt{X(t)^2 + X^*(t)^2} \tag{20}$$

where $X^*(t)$ is the Hilbert transform of X(t).

Figure 2 shows the block diagram of the algorithm.

Hilbert transform of the signal is found out by using a 32-point Parks-McClellan FIR filter. The analytical signal is formed by multiplying the Hilbert transform of the signal by sqrt(−1) (the imaginary unit) and add it to the time-delayed original signal. The Hilbert transform is implemented by FIR filter hence it is necessary to delay the input as it will introduce a delay of half the filter length.

Absolute value of the analytical signal is taken to find the envelope. The envelope is a low frequency signal compared to the original signal. Down-sampling and passing the signal through the low pass filter is required to reduce its sampling frequency, to eliminate ringing and to smooth the envelope. Hilbert envelope detection method is used over other existing methods, because other methods like Square-law detectors suffer from distortion. Peak detectors work with continuous signals and most samples will not be near the carrier peak in a sampled system unless the oversampling ratio is quite high relative to the carrier or IF frequency. The envelope obtained is then fed to the direction cosine matrix or DCM. The components or elements in the DCM are then converted to quaternion values or, in other words, a vector tangent to the signal in the direction of movement. Equations (3) and (15) are used to calculate the DCM and quaternion used in the process. The real and imaginary components of quaternions are obtained for different signals containing various types of voltage disturbances. The values of one imaginary component, in this case the k axis, generate a spike on the onset of any kind of voltage disturbance. It is necessary to know more about the parameters of the disturbance signals that occur due to various reasons. Estimating the frequencies would benefit us in mitigating the disturbances present in the signal. Hence, to estimate such frequencies ESPRIT algorithm is used. ESPRIT (Estimation of Signal Parameters via Rotational Invariance Technique) is an algorithm used for estimating frequencies of a sum of sinusoids by the use of an Eigen-decomposition approach. ESPRIT exploits an underlying rotational invariance of signal subspaces spanned by two temporally displaced data vectors. Consider the estimation of r complex-valued sinusoids in a noise. The sequence is represented by the vector –

$$s(n) = [s(n), s(n+1), \ldots\ldots\ldots\ldots .s(n+M-1)]' \tag{21}$$

$$= x(n) + y(n) \tag{22}$$

where $x(n)$ is the signal vector and $y(n)$ is the noise vector.

We now define a time-displaced vector $t(n) = s(n+1)$ to exploit the deterministic character of the sinusoids.

$$t(n) = [t(n), t(n+1) \ldots \ldots \ldots \ldots t(n+M-1]^t \quad (23)$$

$$= [s(n+1), s(n+2).....s(n+M)]^t \quad (24)$$

With the above description $t(n)$ and $s(n)$ can also be expressed as –

$$s(n) = Ka + y(n) \quad (25)$$

$$t(n) = K\varphi a + y(n) \quad (26)$$

where $a = [a_1, a_2, a_3 \ldots \ldots \ldots \ldots \ldots a_h]^t$, $a_i = A_i e^{j\varphi i}$ and φ is a diagonal $r \times r$ matrix which is given by,

$$\varphi = diag\left[e^{j2\pi f_1}, e^{j2\pi f_2}, \ldots \ldots \ldots .e^{j2\pi f_r}\right] \quad (27)$$

φ is called a rotation operator. The K matrix is the $M \times r$ Vandermonde matrix specified by the column vectors

$$k_i = \left[1, e^{j2\pi f_1} e^{j4\pi f_i}, \ldots e^{j2\pi(M-1)f_i}\right], i = 1,2, \ldots h \quad (28)$$

Auto-covariance matrix for the data vector $s(n)$ is given by-

$$\zeta_{ss} = E[s(n)s^H(n)] \quad (29)$$

$$= KRK^H + \gamma_\omega^2 I$$

$$\zeta_{ss} - \gamma_\omega^2 I = KRK^H = W_{ss} \quad (30)$$

where R is $r \times r$ is a diagonal matrix consisting of the powers of the complex sinusoids.

The Cross-variance matrix of the signal vectors $s(n)$ and $t(n)$ is –

$$\zeta_{st} = E\left[s(n)t^H(n)\right] = KR\varphi^H K^H + \gamma_\omega^2 \zeta_\omega \quad (31)$$

$$\zeta_{st} - \gamma_\omega^2 \zeta_\omega = KR\varphi^H K^H = W_{st} \quad (32)$$

where $\zeta_\omega = E[y(n)y^H(n+1)]$. It should be mentioned that both ζ_{ss} and ζ_{st} are Toeplitz matrices.

Based on the above mathematical relationship we can formulate the ESPRIT algorithm for estimating frequencies (f_i). From the data, autocorrelation values $p_{ss}(m), m = 1,2 \ldots \ldots .M$ is computed and P_{ss} and P_{st} matrices are formed corresponding to the estimates of ζ_{ss} and ζ_{st}. The next step involves computation of eigenvalues of P_{ss}. Further, we need to compute $W_{ss}^* = P_{ss} - \gamma_\omega^* I$ and $W_{st}^* = P_{st} - \gamma_\omega^* D$ where,

$$D = [000 \cdots 00100 \ldots 00010 \ldots 00 : : \ : 000 \ldots 10] \quad (33)$$

Eigenvalues of the matrix pair $[W_{ss}^*, W_{st}^*]$ is further calculated. The r generalized eigenvalues of the matrices that lie on the unit circle determine the elements of φ and hence the sinusoidal frequencies. Further, Support Vector Machine (SVM) is used for the classification of different disturbances into different categories such as sag, swell, harmonics, interruptions and transients. SVM is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines are systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [18]. These classifications are divided into linearly and nonlinearly separable binary classification. The case we present here can be solved using nonlinearly separable binary classification.

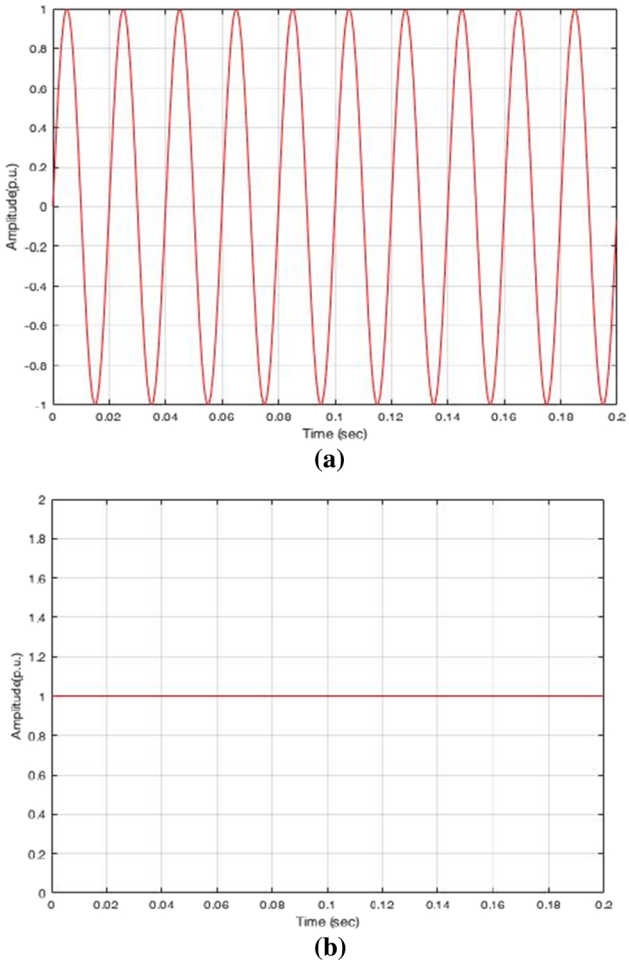


Fig.3 **a** Pure Sine Wave, **b** Quaternion Output of the pure sine wave

Dezert–Smarandache theory (DSmT) is used along with the outputs of the SVM to enhance the classification performance. Initially the SVM classifier is trained with the features generated by different methods that operate independently of each other. Degrees of imprecision for the recognition task is the

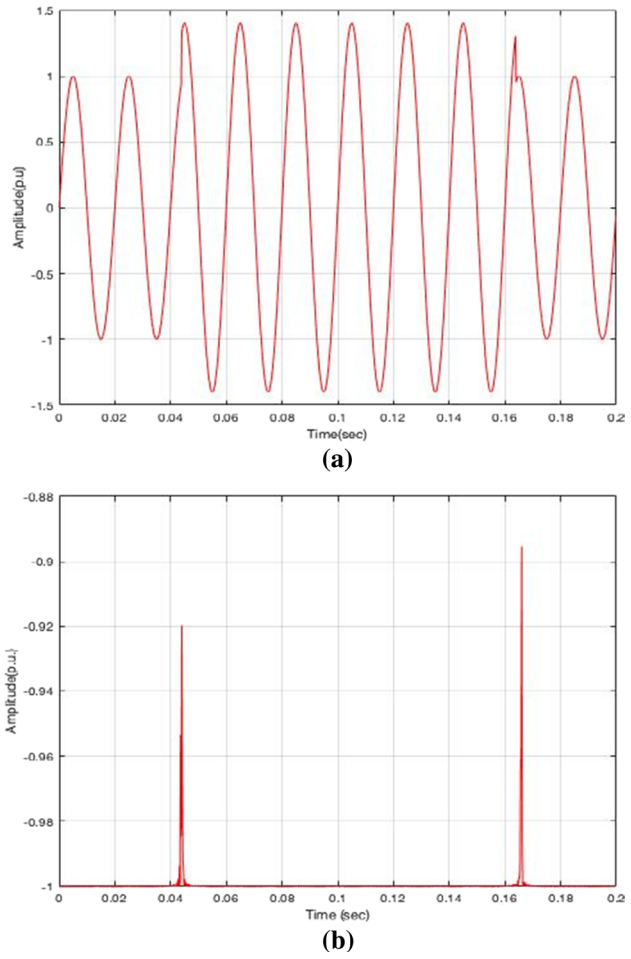


Fig. 4 **a** Voltage Swell, **b** Quaternion output of the signal

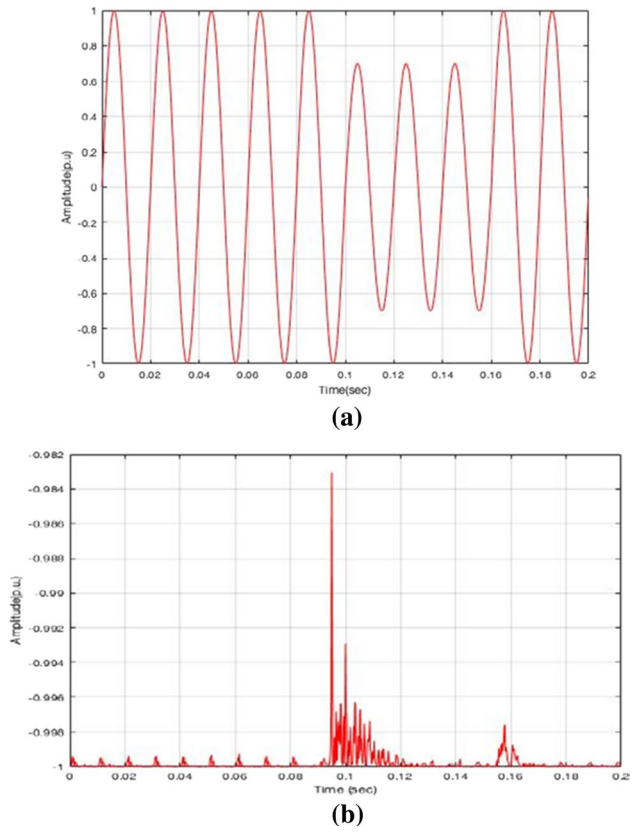


Fig. 5 **a** Voltage Sag, **b** Quaternion output of the signal

provided by the outputs of SVMs classifiers. These are then transformed in posterior probabilities using a sigmoid transformation. Hence, in order to enhance the performances of signal classification system, a supervised model based on DSmT for managing significantly the conflict provided from the two SVMs classifiers is used. The DSmT contains two models: the free model and the hybrid model. First model deals with the limits of the size of $\varkappa$, whereas the second model has the capability to minimize the size. Hence the second model is preferred in this case. PCR6 combination rule based on the DSmT is used for the parallel combination of two classifiers, let us consider them to be S1 and S2. For multiple-class problem, the frame of discernment, a reference domain, is defined for performing the combination. This is composed of a finite set of exhaustive and mutually exclusive hypotheses. The frame of discernment, in the context of probabilistic theory is composed of multiple elements as:

$$\varkappa = \{\varnothing_1, \varnothing_2, \varnothing_3 \ldots \ldots \varnothing_i\} \tag{34}$$

and a mapping function $n \in [0,1]$ is associated for each class which defines the corresponding mass verifying $n(\theta) = 0$ and $\sum_{i=0}^{k} n(\varnothing_i) = 1$.

Distribution of unitary mass of certainty over all the composite propositions built from the elements of $\varkappa$ with the union ($\cup$) and intersection ($\cap$) operators instead making this distribution over the elementary hypothesis is the main concept of DSmT. A decision of membership of a disturbed signal to one of the simple classes of $\varkappa$ is made using the statistical classification technique. Therefore, the hyper-powerset $P^{\varkappa}$ is defined as:

1. $\theta, \varnothing_1, \varnothing_2, \varnothing_3 \ldots \ldots \varnothing_i \in P^{\varkappa}$.
2. If $I, J \in P^{\varkappa}$, then $I \cap J \in P^{\varkappa} and I \cup J \in P^{\varkappa}$.
3. Except for the elements obtained by using rules 1 and 2, no other elements belong to $P^{\varkappa}$.

The DSmT uses generalized basic belief mass (gbbm) or generalized basic belief assignment computed on hyper-powerset of $\varkappa$ and map which can be defined as:

$$n(.) : P \rightarrow [0,1] \tag{35}$$

associated with a given source of evidence as

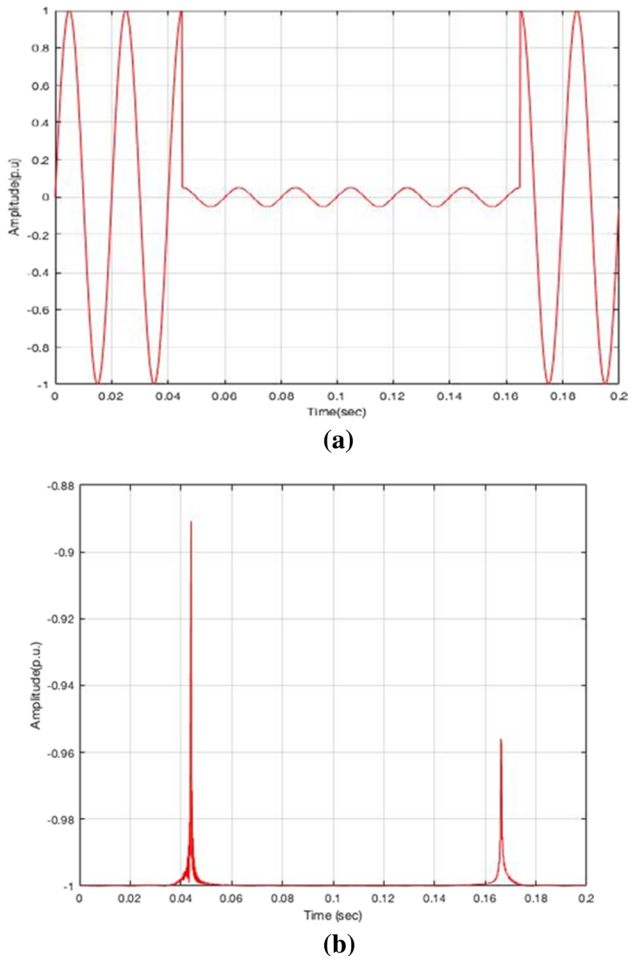


Fig. 6 **a** Voltage Interruption **b** Quaternion output of the signal

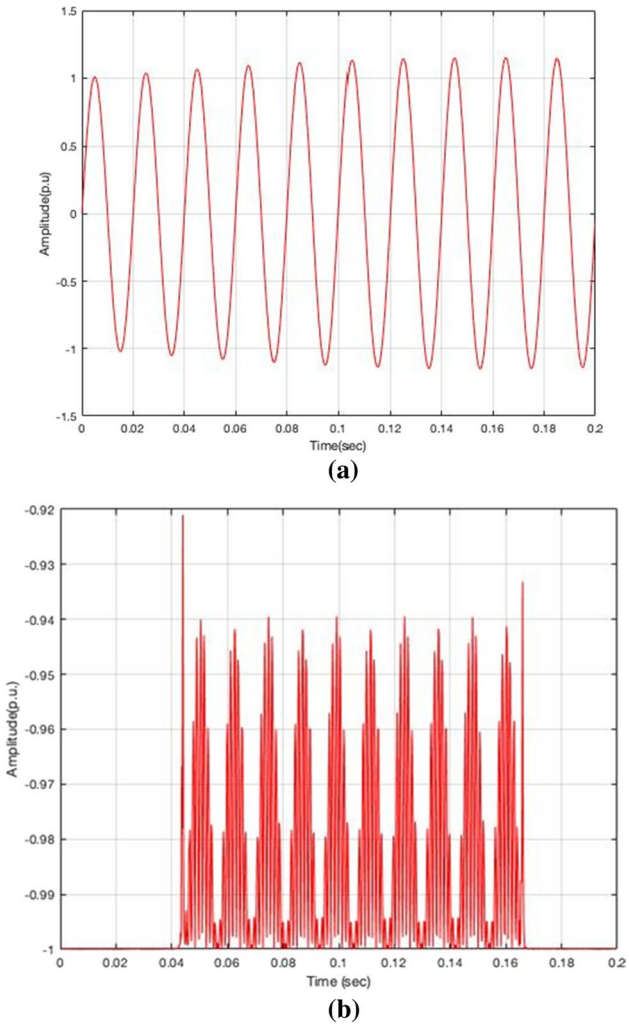


Fig. 7 **a** Voltage Flicker **b** Quaternion output of the signal

$$n(\theta) = 0$$

and

$$\sum\nolimits_{I\in\varkappa} n(I) = 1 \tag{36}$$

The combined masses n_{PCR6} obtained from $n_1(.)$ *and* $n_2(.)$ are given as:

$$n_{\mathrm{PCR6}}(I_i) = \{\, 0 \text{ if } I_l \in \gamma n_{\wedge}(I_l) + \textstyle\sum_{a=1}^{2} n_a^2(I_l) H_a \text{ otherwise } \} \tag{37}$$

whereas

$$H_a = \sum\nolimits_{D_{\alpha_r(1)}\cap I_i \in \gamma D_{\alpha_r(1)}\in\varkappa} \frac{n_{\alpha_r(1)} D_{\alpha_r(1)}}{n_r(I_i) + n_{\alpha_r(1)} D_{\alpha_r(1)}} \tag{38}$$

$\gamma = \{\gamma_N, \theta\}$ is the set of all relatively and absolutely empty elements. γ_N is the set of all elements of $\varkappa$ that have been forced to be empty in the hybrid model N which is defined by the exhaustive and exclusive constraints. The denominator of Eq. 38, i.e., $n_r(I_i) + n_{\alpha_r(1)} D_{\alpha_r(1)}$ is different to zero, where $\alpha_r(1)$ counts from 1 to 2 avoiding r, i.e., $\alpha_1(1) = 2$ and $\alpha_2(1) = 1$. Apart from this, θ is an empty set. Hence, the term $n_{\wedge}(I_i)$ represents a conjunctive consensus called DSm Classic (DSmC) combination rule. This is defined by:

$$n_{\wedge}(I_l) = \left\{ 0 \quad if \quad I_l \in \theta \sum_{J,H\in\chi, J\cap H = I_l} n_1(J) n_2(H) \quad otherwise \right\} \tag{39}$$

Let T set of K learning samples which are separable in s classes $\{\varnothing_0, \varnothing_1 \ldots\ldots \varnothing_{s-1}\}$, such that $\mathrm{T} = \{(a_i, b_i), a_i \in R^p; i = 1,2 \ldots., K, b_i \in \{0,1 \ldots k\}\}$.

Combination of binary SVM is performed using the multi class implementation. It is based on One Against All method. This method each SVM is designed to separate a class from others. The DSmT module in the combination of SVM takes up different steps to accomplish the task.

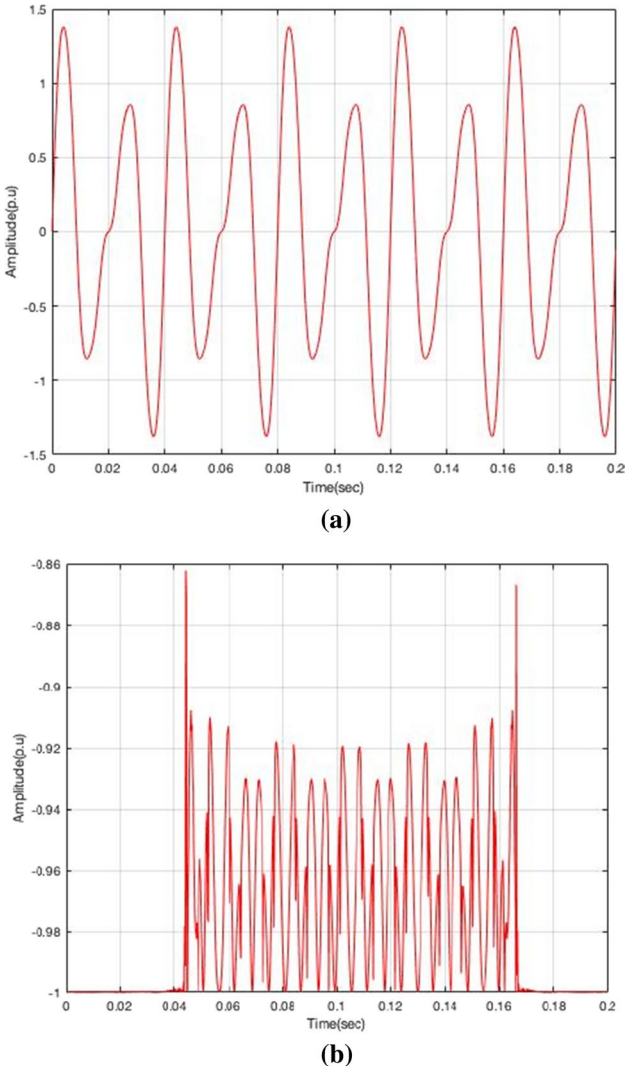


Fig. 8 **a** Voltage harmonics **b** Quaternion output of the signal

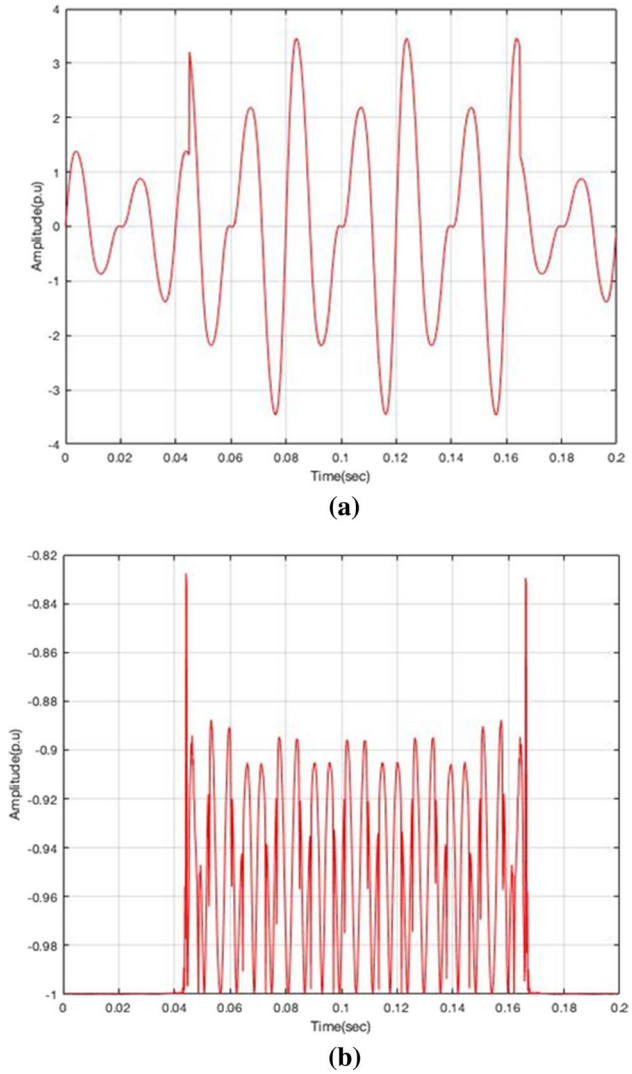


Fig. 9 **a** Voltage Swell with harmonics **b** Quaternion output of the signal

The first task is performed by the SVM as it its outputs are calibrated using a sigmoid transformation and the masses of simple and complementary classes are estimated using a supervised model, respectively. In case of a conflict generated by the information sources, the combined masses are computed as follows:

$$G_c = G_1 \oplus G_2 \tag{40}$$

where $\oplus$ defines the PCR6 combination rule.

The classification of the signals to one of the simple classes of $\varkappa$ is made using statistical classification technique. Initially probabilistic measure is used by converting the combined beliefs using probabilistic transformation. This measure is called Dezert–Smarandache probability (DSmP) which maps a belief measure to a subjective probability measure defined as:

$$DSmP_\tau(\varnothing_l) = G_c(\varnothing_l) + \left(G_c(\varnothing_l) + \tau\right) * v \tag{41}$$

$$v = \sum_{o*} \frac{G_c(I_j)}{\sum_{I_a \in 2^\varkappa I_a \subset I_j Q_n(I_a)=1} G_c(I_a) + G_c(I_a) + \tau Q_n(I_j)} \tag{42}$$

where $o^* = I_j \in 2^\varkappa I_j \supset \varnothing_l Q_n(I_i) \geq 2$ and $l = \{0,1,2 \ldots \ldots .k\}, \tau \geq 0$ is a tuning parameter, N is the Shafer's model for $\varkappa$ and $Q_n(I_a)$ denotes the DSm cardinal. Hence, maximum likelihood test is taken to make the final decision. It is given by:

$$f \in \varnothing_i \;\; if\; DSmP_\tau(\varnothing_j) = max\{DSmP_\tau(\varnothing_j), 0 \leq j \leq k\} \tag{43}$$

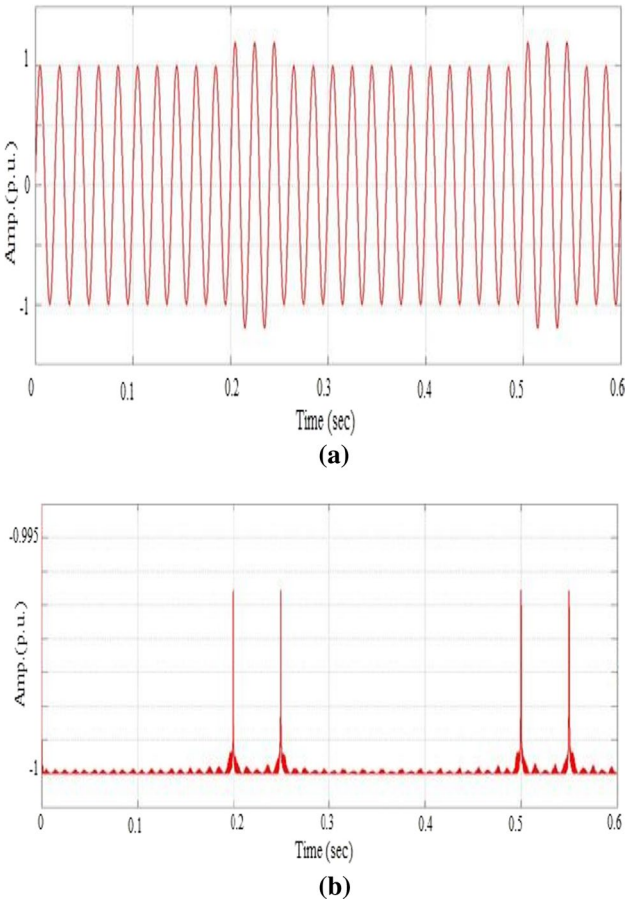


Fig. 10 **a** Voltage Swell at two instances, **b** Quaternion output of the signal

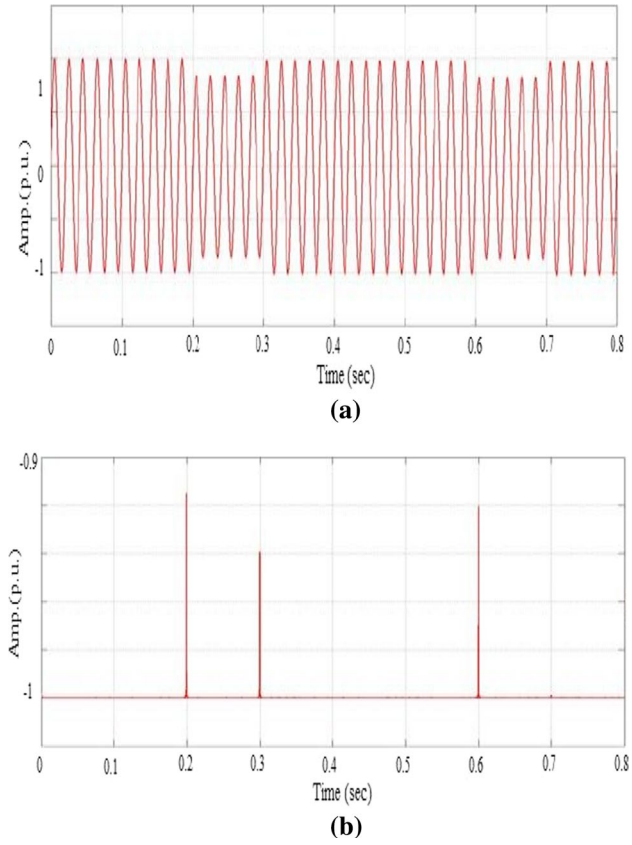


Fig. 11 **a** Voltage Sag at two instances, (**b**) Quaternion output of the signal

where f is the signal characterized by both the descriptors and τ is fixed to 0.001.

The proposed method is to be executed as shown in Fig. 2. In which, first the quaternion value of the signal is acquired and after doing following operations like preprocessing, generating features and ESPRIT to divide signals, two types classifier i.e., classifier 1 and classifier 2 is applied. The finial outputs of proposed method are recorded after applying the DSmT-based parallel combination method to the classified output of the two different classifiers.

Table 1 shows the numerical modeling of the PQ disturbances with parameters.

4 Simulation results

The signals used for simulation have frequency of 50 Hz and a peak value of 1 p.u. The quaternion elements w, x, y, z are computed from the DCM, as explained in the section above. Since we have considered the rotation vector to be in the i, j plane, i.e., the rotation is taking place only about the k axis, we obtain the DCM as specified by Eq. (15). Upon equating Eqs. (15) with (16), we see that $x = y = 0$.

Initially the quaternion values for a pure sine wave are taken into account to eliminate any possibilities of any erroneous result. Quaternion output for a pure sine wave is shown in Fig. 3b. The quaternion output, as discussed earlier, is the magnitude of the imaginary part of the quaternion vector. This imaginary part has the components $x = y = 0$ while the z axis have some definite value. The figure shown in this section contains two parts. The first part shows the nature of the signal, whereas the other part shows the z component of the quaternion vector.

As clear from Fig. 3a and b, no spike is generated in case of a pure sine wave. This output is the frame of references while comparing the quaternion output in case of other signals containing disturbances (Fig. 4).

A spike is observed to be generated by the z component at the same instant when swell occurs indicating the presence of such disturbance in the signal.

Similar output can be seen in the case of the sag component in the signal in Fig. 5. A spike can be seen at the same

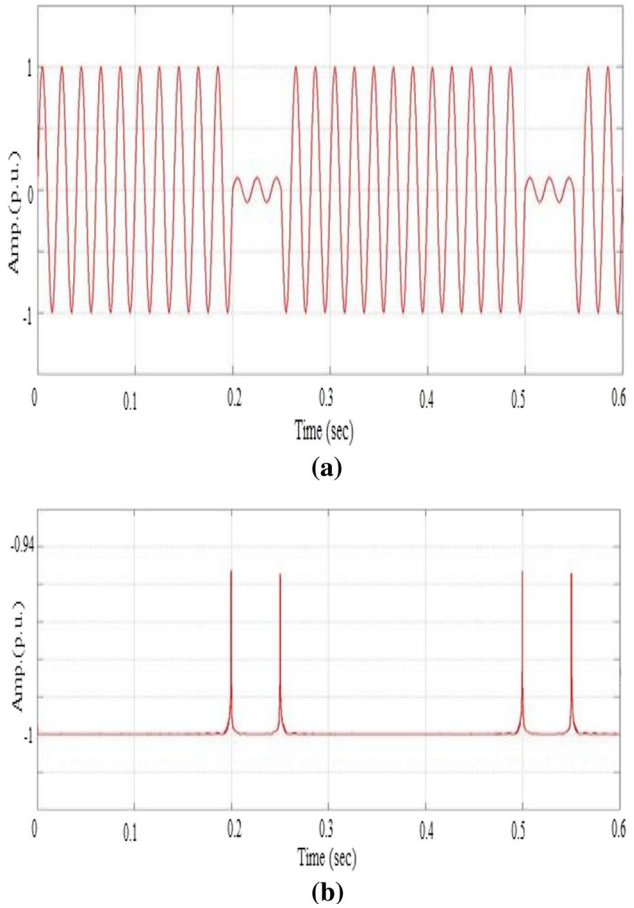


Fig. 12 **a** Voltage Interruption at two instances, **b** Quaternion output of the signal

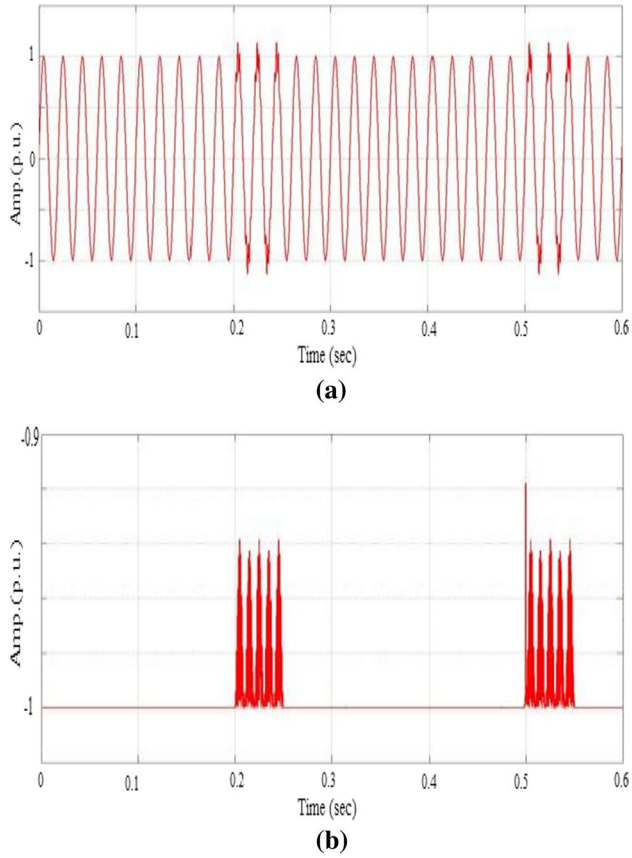


Fig. 13 **b** Voltage Flicker at two instances, **b** Quaternion output of the signal

instant when the sag occurs but with a lower amplitude of the quaternions (Fig. 6).

Figure 7 represents an interruption a pure sinusoidal signal. Interruption in a signal occurs when the amplitude of the signal falls below 0.1 p.u. As seen by the quaternion output or Fig. 7b, the moment the interruption occurs the quaternion generates a spike.

Flicker signal is defined the presence of a swell component of a very high frequency sinusoid in the combination of a pure sine wave. As seen in Fig. 7b, the quaternion response of the flicker signal is different than previously encountered disturbances.

The signal shown in Fig. 8a is a signal containing different harmonics components. These include 3rd, 5th and 7th harmonics for a small duration of the signal. The quaternion response can be observed in Fig. 8b which suggests that the quaternion spikes exist for the complete duration of such disturbances.

Sometimes swell or sag occurs along with the harmonics in a signal. This method is able to detect such cases in a similar fashion as explained earlier. A spike is generated at the instant where the swell occurs and the pattern is continuous in the harmonic component region of the signal. Hence, the quaternion result obtained is a combination of two quaternion responses (Fig. 9).

There are cases where the same disturbances occur immediately at two different instants in a given period of time. The method proposed is able to detect such disturbances indicating different time instants at which they occur. Figure 10a shows a similar case of two swell components in a signal in a given period of time. The quaternion response clearly indicates the two instances where the swell occurs without any conflicting results.

The similar case is presented for the sag signal in Fig. 11a to check the accuracy of method under different circumstances.

Figure 12a represents two instances of interruption in a signal. The quaternion result is very accurate in detecting different instants of interruption as can be seen in Fig. 12b.

Figure 13a represents a flicker signal having a frequency of 500 Hz in disturbance region. The quaternion result forms a similar pattern in every cycle when the flicker occurs as shown in Fig. 13b.

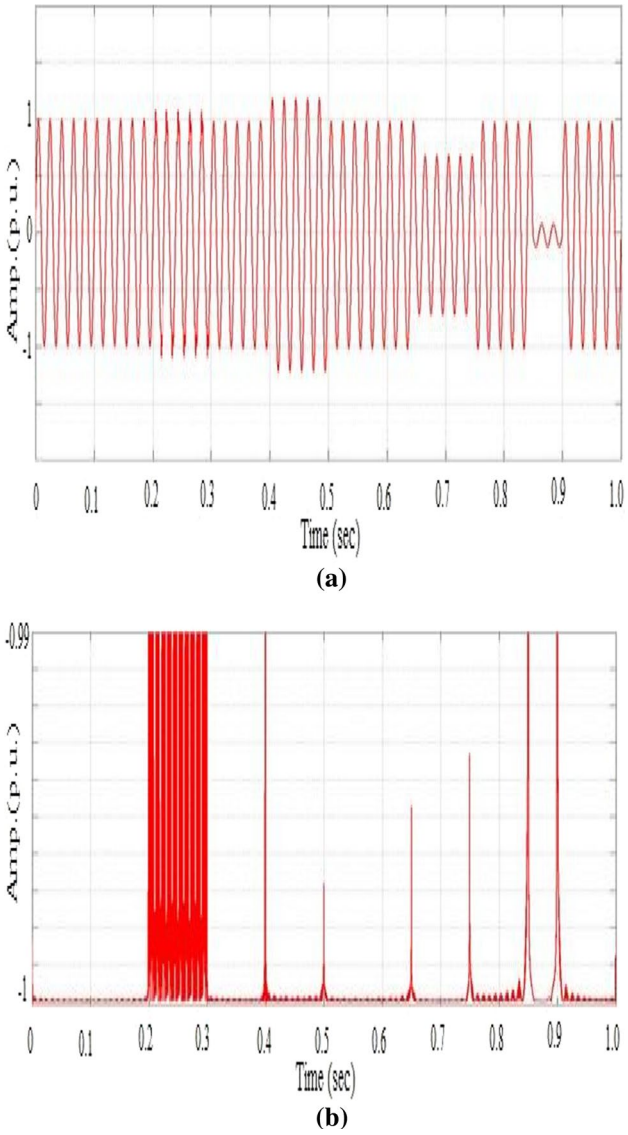


Fig. 14 **a** Multiple disturbances without harmonics; **b** Quaternion output of the signal

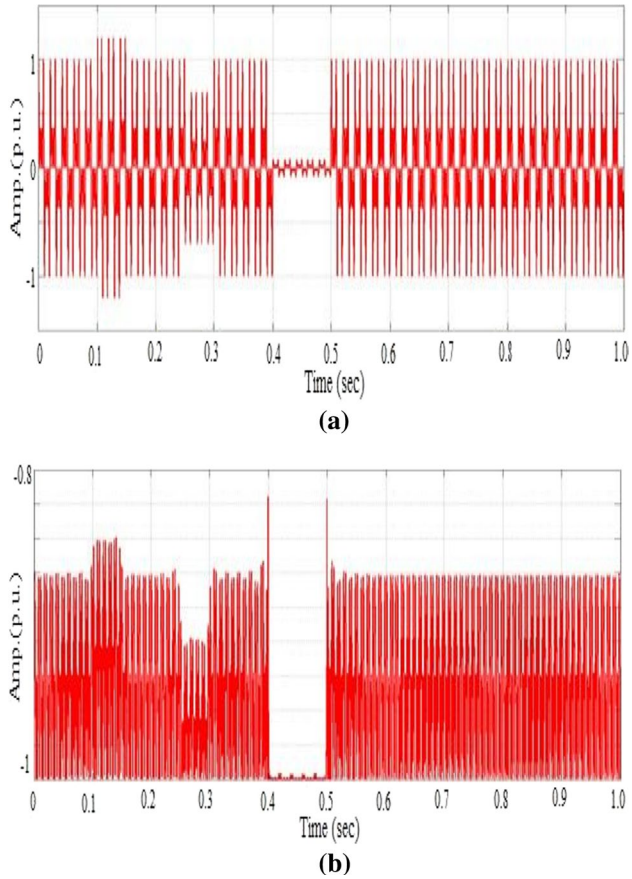


Fig. 15 **a** Multiple disturbances with harmonics; **b** Quaternion output of the signal

Figure 14a represents different kinds of disturbances in a duration of 1 s. The disturbances present in the signal, in order, include flicker, swell, sag and interruption. The quaternion response is generated for all the disturbances irrespective of their impact on the signal. Thus, the method gives appropriate results in case multiple disturbances occur in a very short period of time.

Harmonics may occur along with the other kind of disturbances which is shown in Fig. 15a. When this situation occurs, the quaternion response does not provide a very clear result which can be useful for the classification and recognition of the disturbances present in the signal. In such conflicting cases, the combination of SVM-DSmT classifier provides very accurate results. These results hold true even when there are different disturbances present in the signal in a given period of time. The multiple events should occur at different instances to differentiate between different disturbances. In case, two events occur at the same instant, the algorithm proposed will be able to detect the disturbance but will be difficult to differentiate between the two. Hence, for such purposes, SVM-DSmT combination is used to classify the signal with very high accuracy. ESPRIT algorithm provides appropriate results for the frequency estimation. The r generalized eigenvalues of the matrices explained in Section III that lie on or near the unit circle estimates the frequency components present in the signal. The ESPRIT output of sag and swell signal is represented on Fig. 16a and b, respectively. Since the signal contains only the fundamental frequency hence only one frequency is plotted in the figure.

The eigenvalues lie slightly inside the circumference of the circle representing the decrease in the amplitude of the signal. For swell signal the eigenvalues lie slightly outside the circumference of the unit circle. Since the flicker signal contains only two frequencies, the fundamental frequency of 50 Hz and a high frequency component, Fig. 16c presents

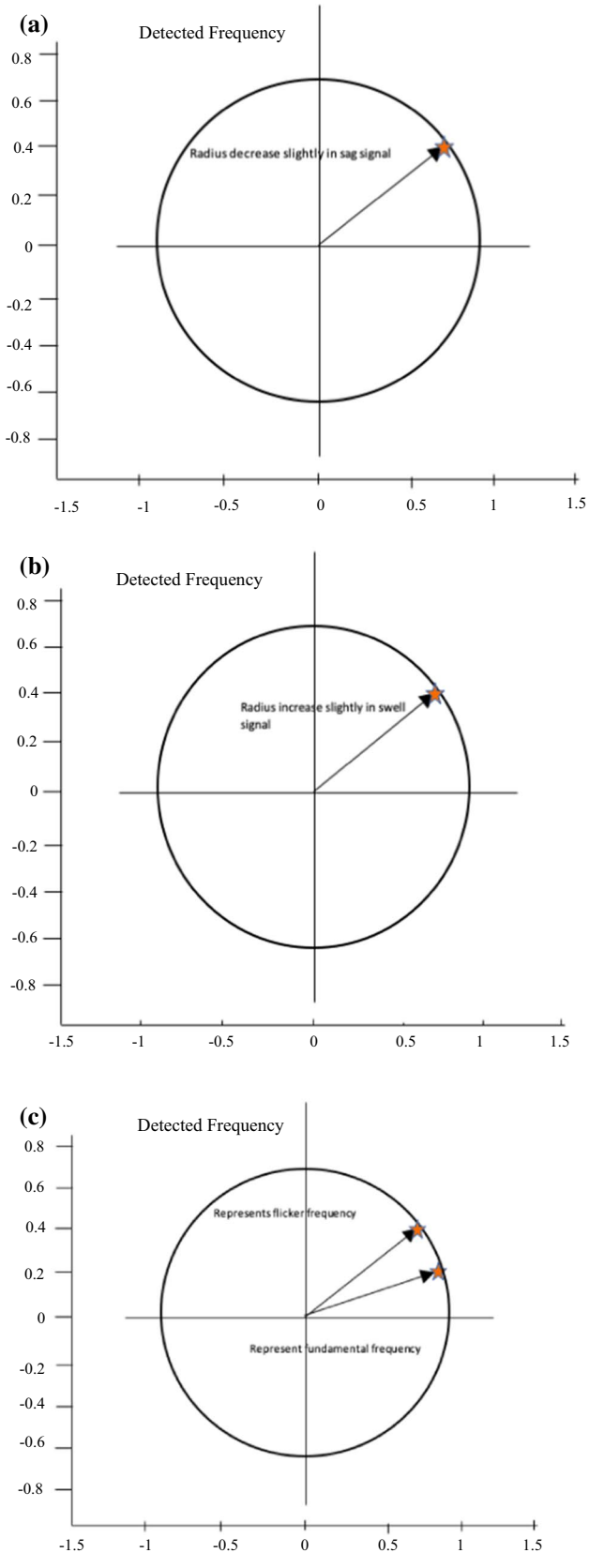


Fig. 16 **a** ESPRIT output for sag signal (**b**) ESPRIT output for swell signal (**c**) ESPRIT output for flicker signal (**d**) ESPRIT output for Interruption signal (**e**) ESPRIT output for signals with harmonics (**f**) ESPRIT output for signal having multiple disturbances with harmonics

two eigenvalues representing the different frequencies. On the other hand, in case of an interruption the frequency of the signal is considered to be around 0, hence the eigenvalue lies on the x-axis.

The ESPRIT algorithm is used for detecting different harmonic components in the signal. Figure 16e shows the same output. The different plots of eigenvalues represent 3rd and 5th harmonic components in the signal apart from the fundamental frequency. Figure 16f shows the ESPRIT output of the multiple disturbances present in the signal. It can be seen that when multiple disturbances occur, the eigenvalues are plotted separately where each eigenvalue indicates the kind of disturbance present in the signal. SVM-DSmT combination is used for the classification of signals with ease and better accuracy. Thus after training the SVM with the signal features, different signals containing various disturbances are used as an input for the classification whose details are given in Table 2.

Signals mentioned in Table 2 are used as input for the classification of the signals whose results are shown in Table 4. Table 4 represents a confusion table showing the classification of the signals according to the disturbances when the inputs are provided to the SVM-DSmT system. From Table 4 we can see that, for the pure sine wave, the classification done by the combination gives 100% accurate result as zero quaternion output is easy to recognize. The recognition of sag signal also gives 100% recognition results. This indicates that the proposed algorithm works extremely well in detecting and classifying sag signals. The algorithm works equally well in the case of detecting the same noise multiple times over a given period of time. The classification of swell signals gives the accuracy of 98.6%. The confusion matrix shows that one of the inputs were recognized as swell with harmonics.

This proposed method was implemented in real time for detecting the PQ disturbances. The developed implementation was focused on providing detailed analysis of events. Regarding the designing phase computational complexity of the system, that comprises the construction of the Quaternion-based approach for the classification of the disturbances and then expressing the lower and upper limits for the detection stage. This was performed off-line. For real-time execution of the classifier, the parameters required were saved in the Labview. This real-time classification process requires 6021 sums, 10,233 multiplications. The computational complexity of the detector is 512 sums, 1024 multiplications and few comparison as shown in Table 3.

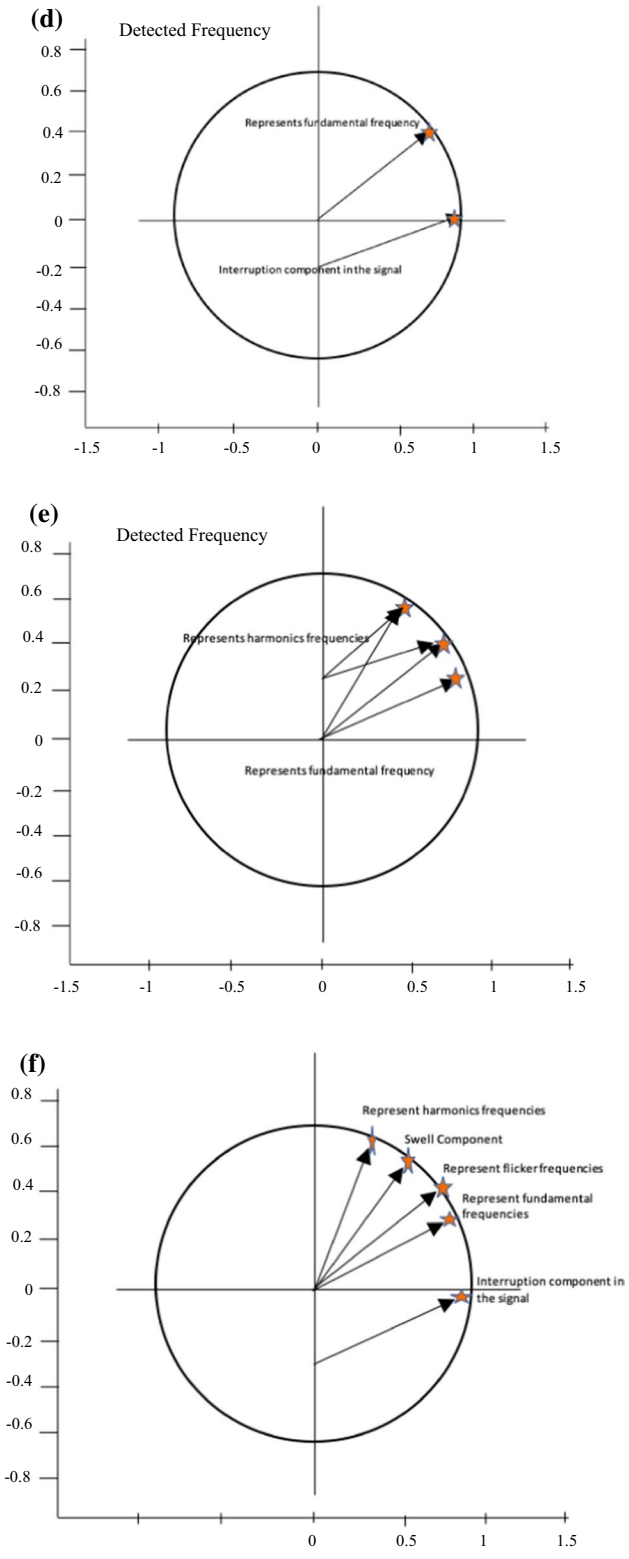


Fig. 16 (continued)

Table 2 Event data and their numbers

PQ events	Number of input signals
(a) Pure sine	80
(b) Sag	60
(c) Swell	75
(d) Interruption	95
(e) Flicker	87
(f) Harmonics	67
(g) Swell with Harmonics	90
(h) Same noise twice	93
(e) Flicker	72
(f) Harmonics	101
(g) Swell with Harmonics	818

Table 3 Computational time

PQ events	Proposed methods (microsecond)	[25] (microsecond)	[26] (microsecond)
(a) Pure Sine	70	71.98	71.30
(b) Sag	80	83.32	83.40
(c) Swell	80.23	80.21	80
(d) Interruption	81.21	81.30	81.17
(e) Flicker	86.08	86.34	86.27
(f) Harmonics	78.9	80.0	81.21
(g) Swell with Harmonics	79.67	81.23	82.45
(h) Same noise twice	79.99	82.01	83.89
(e) Flicker	79.40	79.39	80.01
(f) Harmonics	79.07	79.98	83.42
(g) Swell with Harmonics	83.76	84.32	85.0
(h) Same noise twice	84.56	89.0	89.34

Table 4 Performance comparison of different classifiers proposed earlier

S. no.	Classifiers or methods used	Accuracy (%)
1	FCA [19]	95.86
2	DGT-MLP [20]	97.66
3	Fuzzy logic [21]	98.00
4	ANN [22]	98.10
5	SVM [23]	98.33
6	WT-SVM [24]	98.51
7	Proposed method	98.91

Table 5 Performance of the classifier to recognize PQ disturbance

		Output class									
	PQ events	Pure Sine	Sag	Swell	Interruption	Flicker	Harmonics	SWH	SNT	MNWH	MNWOH
Target class	Pure Sine	80	0	0	0	0	0	0	0	0	0
	Sag	0	60	0	1	0	0	0	0	0	0
	Swell	0	0	74	0	0	0	0	0	0	0
	Interruption	0	0	0	94	0	0	0	0	0	0
	Flicker	0	0	0	0	86	0	1	0	0	0
	Harmonics	0	0	0	0	0	66	1	0	0	1
	SWH	0	0	1	0	1	1	88	0	1	1
	SNT	0	0	0	0	0	0	0	93	0	0
	MNWH	0	0	0	0	0	0	0	0	71	0
	MNWOH	0	0	0	0	0	0	0	0	0	99
	Accuracy (%)	100	100	98.6	98.9	98.8	98.5	97.7	100	98.6	98.02

This may be because of the amplitude of the signal increases in both the cases. The input must be having a very high order harmonics which might not have a great impact on the signal. The performance in the case of interruption signal was recorded to be 98.9% accurate. One of the cases was recorded under the sag signal category. Similar results were obtained in the case of flicker, harmonics, swell with harmonics (SWH), multiple noise with harmonics (MNWH) and multiple noise without harmonics (MNWOH) giving the accuracy of 98.8%, 98.8%, 97.7%, 98.6% and 98.02%, respectively. The accuracy for the proposed method comes to the previous proposed methods. Table 3 shows the comparison of the accuracy of classification between previously proposed methods with the method given in this paper.

5 Conclusion

Signal disturbances like sags and swells in voltages and currents have a highly undesirable effect on the power quality in a system while noise and transients can lead to deterioration of electrical or electronic equipment connected on the system buses. The method proposed in this paper detects the presence of undesirable components in the signal using quaternion values. These values produce a spike or form a pattern at the onset of any disturbance. The presence of multiple harmonic components along with different types of disturbances present in the signal can be estimated by ESPRIT algorithm. The classification of the detected disturbances is done with the help of SVM-DSmT combination of recognition and classification giving a very high accuracy of 98.91% as shown by the confusion matrix in Table 3. A comparison from the previously proposed methods shown in Table 4 shows the improved accuracy of the proposed method. Finally, the simulation results show good performance, power quality disturbance detection and classification of the whole system under different conditions (Table 5).

References

1. Abdoos AA, Mianaei PK, Ghadikolaei MR (2016) Combined VMD-SVM based feature selection method for classification of power quality events. Appl Soft Comput 38:637–646
2. Abid FB, Zgarni S, Braham A (2018) Distinct bearing faults detection in induction motor by a hybrid optimized SWPT and aiNet-DAG SVM. IEEE Trans Energy Convers. https://doi.org/10.1109/TEC.2018.2839083
3. Ahmadi H, Akbari Foroud A (2016) Design of joint active and reactive power reserve market: a multi-objective approach using NSGA II. IET Gener Transm Distrib 10(1):31–40
4. Ben Abid F, Zgarni S, Braham A (2018) Distinct bearing faults detection in induction motor by a hybrid optimized SWPT and aiNet-DAG SVM. IEEE Trans Energy Convers 33(4):1692–1699
5. Bin J, YuanYuan XY (2017) Orthogonal design-based NSGA-III for the optimal lockage Co-scheduling problem. IEEE Trans Intell Transp Syst 99:1–11
6. Ciro GC, Dugardin F, Farouk Y, Kelly R (2016) A NSGA-II and NSGA-III comparison for solving an open shop scheduling problem with resource constraints. IFAC Pap Online 49(12):1272–1277
7. Hashemi F, Mohammadi M (2016) Islanding detection approach with negligible non-detection zone based on feature extraction discrete wavelet transform and artificial neural network. Int Trans Electr Energ Syst. https://doi.org/10.1002/etep.2197
8. JLee IWC, Dash PK. (2003) S-transform based intelligent system for classification ofpower quality disturbance signals. IEEE Trans Ind Electr 50:800–5
9. Jamil M, Singh R, Sharma SK (2015) Fault identification in electrical power distribution system using combined discrete wavelet transform and fuzzy logic. J Electr Syst Inf Technol 2:257–267
10. Kanirajan P, Kumar VS (2015) Power quality disturbance detection and classification using wavelet and RBFNN. Appl Soft Comput 35:470–481

11. Kapoor R, Kumar R, Tripathi MM (2018) Volterra bound interval type-2 fuzzy logic-based approach for multiple power quality events analysis. IET Electr Syst Transp 8(3):188–196. https://doi.org/10.1049/iet-est.2017.0054
12. Kapoor R, Tripathi MM (2017) Detection and classification of multiple power signal patterns with Volterra series and interval type-2 fuzzy logic system. Prot Control Mod Power Syst. https://doi.org/10.1186/s41601-017-0039-z
13. Khokhar S, Zin AAM, Memonb AP, Mokhtar AS (2017) A new optimal feature selection algorithm for classification of power quality disturbances using discrete wavelet transform and probabilistic neural network. Measurement 95:246–259
14. Mahela OP, Shaik AG (2017) Recognition of power quality disturbances using S-transform based ruled decision tree and fuzzy C-means clustering classifiers. Appl Soft Comput 59:243–257. https://doi.org/10.1016/j.asoc.2017.05.061
15. Mahela OP, Shaik AG, Gupta N (2015) A critical review of detection and classification of power quality events. Renew Sustain Energy Rev 41:495–505
16. Martínez J, Iglesias C, Matías JM, Taboada J, Araújo M (2014) Solving the slate tile classification problem using a DAGSVM multiclassification algorithm based on SVM binary classifiers with a one-versus-all approach. Appl Math Comput 230:464–472
17. Mehera SK, Pradhan AK (2010) Fuzzy classifiers for power quality events analysis. Electr Power Syst Res 80:71–76
18. Moravej Z, Abdoos AA, Pazoki M (2009) Detection and classification of power quality disturbances using wavelet transform and support vector machines. Electr Power Compon Syst 38:182–196
19. Moravej Z, Pazoki M, Niasati M, Abdoos AA (2013) A hybrid intelligence approach for power quality disturbances detection and classification. Euro Trans Electr Power 23(7):914–929
20. Rajiv K, Tripathi MM (2019) Long short-term memory-convolution neural network based hybrid deep learning approach for power quality events classification. In: Saini H, Singh R, Kumar G, Rather G, Santhi K (eds) Innovations in electronics and communication engineering. Lecture notes in networks and systems. Springer, Singapore
21. Silva KM, Souza BA, Brito NSD (2006) Fault detection and classification in transmission lines based on wavelet transform and ANN. IEEE Trans Power Deliv 21:2058–63
22. Song J, Xie Z, Zhou J, Yang X, Pan A (2017) Power quality indexes prediction based on cluster analysis and support vector machine. CIRED-Open Access Proc J 2017:814–714
23. Tang Q, Qiu W, Zhou Y (2020) Classification of Complex Power Quality Disturbances Using Optimized S-Transform and Kernel SVM. IEEE Trans Ind Electr 67:9715–9723
24. Uyar M, Yildirim S, Gencoglu MT (2008) An effective wavelet based feature extraction method for classification of power quality disturbance signals. Electr Power Syst Res 78:1747–55
25. Zhan Wang, Xiangjun Zeng, Xiaoxi Hu, Jingying Hu. (2012) The multi-disturbance complex power quality signal HHT detection technique. In: IEEE PES conference on innovative smart grid technologies, Tianjin
26. Zhu Y, Liang J, Chen J, Ming Z (2017) An improved NSGA-III algorithm for feature selection used in intrusion detection. Knowl-Based Syst 116:74–85

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

Resilient Protection of Medium Voltage DC Microgrids against Cyber Intrusion

Vaibhav Nougain, Sukumar Mishra, *Senior Member, IEEE,* Sushree Subhasmita Jena

***Abstract*—Differential current protection has been a reliable approach for line protection in DC systems. However, due to its dependence on communication for terminal current information, resiliency of such a protection approach is poor against external cyber intrusions (ECI) of different types, i.e., false data injection (FDI) and time synchronization attacks (TSA). As a result of which, the system is prone to collapse due to line tripping in the presence of coordinated ECIs. The scope of this paper is to propose a resilient protection approach for medium voltage DC (MVDC) microgrids differentiating between the real faults and the ECIs in a system. The approach also detects ECIs on an adjacent line occurring simultaneous to a real fault where the system deploys current limiting reactors (CLR) which have the function of limiting the peak of fault current. Modal transformation is carried out to obtain the line-mode and zero-mode voltage across these CLRs. The mode voltages across the faulty mode networks are utilised along with differential current protection to propose a resilient protection approach against ECIs, which is sensitive to high impedance faults (HIFs), working well in the presence of White Gaussian Noise (WGN) in measurement. Further, a sensitivity function is formulated which gives the variation of decisive variables i.e., line-mode voltage and zero-mode voltage with respect to change in decisive parameters i.e., fault resistance and fault location in the system.**

***Index Terms*—Cyber Security, Cyber Intrusion, DC Microgrids, False Data Injection, Differential Current Protection.**

I. Introduction

MEDIUM Voltage DC (MVDC) systems have found their application in shipboard electrical systems (supplying to the propulsion and ship service loads) [1]-[2], onboard DC marine vessels [3]-[4] and remote area mine sites [5] due to ease of integration of multiple energy storage systems, due to relaxed synchronisation requirements, lesser losses, complexity with lesser number of conversion stages in the system [6]. Such multi-terminal MVDC microgrids, when subjected to a fault contingency experience a high rate of rise of current due to discharging converter capacitor. The interfacing converter is prone to get damaged in case of improper control logic [7]. Even if the control is proper, the converter control is bypassed to limit the current [8] affecting the power transfer in the system. Either way, there is an indispensable requirement for a fast, selective and reliable protection approach which ensures continuous operation of power transfer even under a fault. The most effective and popular fault identification schemes in the literature for DC microgrids is differential current relay (DCR) for line protection [5]-[6], [9]-[10]. The reasons for its acceptance as primary protection for DC microgrids are:

- *High Sensitivity to HIFs*: DCR is sensitive to high impedance faults (HIFs). Schemes like overcurrent relays, directional overcurrent relays have low sensitivity towards HIFs [9] making their application questionable for LVDC and MVDC microgrids where the fault resistance can be up to 100 times the line impedance. Hence, it is difficult to decide if the system is subjected to a HIF or a load change with the latter schemes.
- *High Selectivity*: Distinguishing between internal and external faults in a system [11].

As a result of these attributes, DCR has been used in various LVDC [12]-[13] and MVDC based microgrid setups [5]-[6], [11]. Authors in [12]-[13] explores the application of DCR in photovoltaic based LVDC microgrid whereas the authors in [5] implements DCR at a remote area mine working on an islanded MVDC microgrid. With innumerable dynamic DC systems, DCR has been proven to be a fast, selective and reliable protection approach independent of system configuration. However, operation of DCR requires communication infrastructure which makes it prone to external cyber intrusions (ECI). ECI hampers the normal operation of DCR functionality in different ways, i.e., denial of service (DoS), false data injection (FDI) and time synchronisation attacks (TSA). DoS [14] is described as interruption of real time data inactivating the relay operation. FDI is described as injection of false data to trigger relay operation even under healthy system condition [15]-[18]. This type of cyber intrusion can occur by penetrating into local area network (LAN) of a substation [15] or by penetrating into the wireless communication channel used for DCR operation. The attack on Ukrainian power system [16] is a practical example of such intrusion. TSA is similar to FDI and includes manipulation of time stamped samples by penetrating into synchronisation protocols [19], [20]. Like FDI, TSA also trigger relay operation even under healthy system conditions [20]. The impact of DoS can somehow be mitigated using backup protection and redundant communication links as pointed out in [23]. However, the impact of coordinated FDI and TSA leading to false triggering of relay can cause system collapse disrupting the objective of effective protection for continuous power transfer. Hence, there is a need for a resilient protection approach ensuring immunity against FDI and TSA.

However, the related literature exploring resiliency of protection approaches against ECIs is in its early stages. Au-

Vaibhav Nougain and Sukumar Mishra are with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, New Delhi, 110016 India. e-mail: (nougainvaibhav@gmail.com, sukumariitdelhi@gmail.com).

Sushree Subhasmita Jena is with the Department of Electrical Engineering, National Institute of Technology, Hamirpur, Himachal Pradesh, 177005 India. e-mail: subhasmita.023@googlemail.com

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

thors in [21]-[22] have studied the effects of FDI and TSA on DCR in AC grids using phasor-based methods. However, these methods can not be extended to DC microgrids. Authors in [23] explore the detrimental effects of FDI and TSA on MVDC microgrids, concluding that coordinated ECIs can lead to system collapse. The authors then install passive oscillator circuit (POC) (which is a parallel combination of current limiting reactor (CLR) and a capacitor) at each converter terminal. The idea used by the authors is that in case of fault contingency, the POC circuit resonates and generates a specific damped frequency, f_d. This frequency is only generated for a physical fault and not for any ECI. Fast Fourier Transform (FFT) is used to obtain voltage across POC corresponding to frequency, f_d. The resiliency in protection is ensured using *AND* logic on DCR and POC voltage. However, as pointed by authors in [23] itself, the method cannot detect ECIs on adjacent line occurring simultaneous to a real fault in the system.

The scope of the proposed paper is to offer a resilient protection approach which can differentiate between real faults and ECIs in a MVDC microgrid setup. The approach also detects ECIs on an adjacent line occurring simultaneous to a real fault which has not been discussed in the literature so far. The proposed work deploys CLR at each converter terminal, using modal transformation to analyse the faulty network of the MVDC microgrid under different fault contingencies i.e., PTP and PTG. The line-mode and zero-mode voltages across CLR are utilised along with the DCR to propose a method which is resilient to ECIs such as FDI and TSA. The proposed scheme is fast, sensitive to HIFs and robust in the presence of WGN in measurement.

II. Test MVDC microgrid: Operation and Control

A six bus multi-terminal MVDC microgrid is taken as the test system operating at ±10kV [23] based on IEEE Recommended Practice for 1kV to 35kV Medium-Voltage DC Power Systems on Ships [31]-[33]. Bus 1 has system diesel generators (DGs) which are running in droop mode to control the voltage of system [23]. Bus 2 has the induction loads i.e., cranes, winches and propulsion system. Bus 3 has the non-vital AC and DC loads like air conditioners, humidifiers and auxiliary loads. Bus 4 has the pulse loads whereas bus 5 has energy storage system (ESS) battery interfaced with a bidirectional DC/DC converter. The control approach for ESS is constant current mode for grid supportive mode whereas constant voltage mode for the case the system is subjected to (n-1) contingency [24]. Fuel cell at bus 6 is used with a boost DC/DC converter operating in constant current control for high state of charge (SOC) of ESS and grid supportive mode and constant voltage control for (n-1) contingency and low SOC of ESS [5]. The multi-terminal MVDC microgrid consists of seven lines as shown in Fig. 1, modelled as single-core 1.9/3.3 kV 1000 mm^2 cables with copper conductor, XLPE insulation (2.8 mm thick) and PVC sheathing (2.5 mm thick) [23]. A bipolar structure for cable connections is considered for power transfer. The grounding scheme considered is TN-S which provides midpoint grounding at each converter terminal [23]. The fault

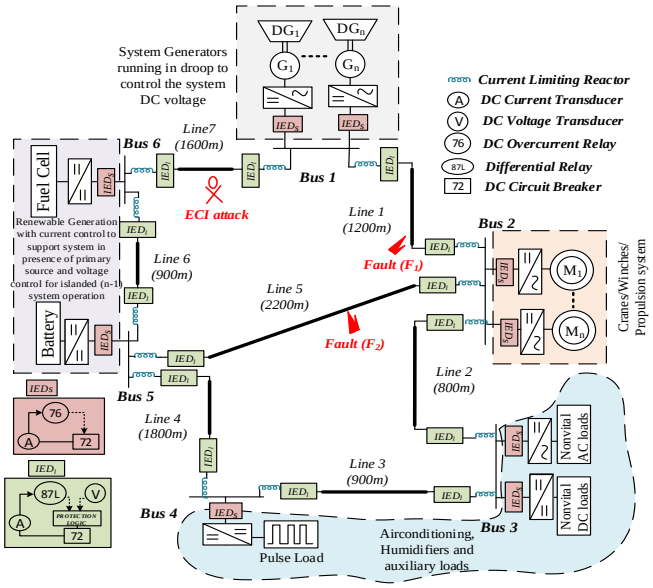


Fig. 1: MVDC microgrid test system

contingencies considered in the test system are of different types i.e., pole to pole (PTP), positive pole to ground (P-PTG) and negative pole to ground (N-PTG). Current limiting reactors (CLR) are deployed in Fig. 1 to limit the rate of rise of current during a fault contingency. The sources and loads in the setup are protected using IED_s which is a combination of DC current transducer, DC overcurrent relay and DC circuit breakers as shown in Fig. 1. The lines in the system are protected using IED_l which integrates the DCR and mode (line-mode and zero-mode) voltages across CLR proposing a resilient protection scheme.

III. Resiliency analysis of DCR against ECIs

DCR employs terminal line current to make a trip decision in case of a fault contingency in the system. In case of a fault contingency, the differential current feeds the fault and hence increases above a certain threshold, I_{set}. Hence, the trip condition is defined as an *OR* condition for the following equation:

$$\begin{cases} \left|I_{m_+n_+}(k)\right| = \left|I_{mn_+}(k) + I_{nm_+}(k)\right| > I_{set} & \text{(1a)} \\ \left|I_{m_-n_-}(k)\right| = \left|I_{mn_-}(k) + I_{nm_-}(k)\right| > I_{set} & \text{(1b)} \end{cases}$$

where $I_{mn_+}(k)$ and $I_{nm_+}(k)$ give the m and n bus terminal current values of positive polarity whereas $I_{mn_-}(k)$ and $I_{nm_-}(k)$ give the same for negative polarity.

A. Vulnerability to FDIs

FDI attacks are executed by penetrating into the communication channels (LAN or wireless) and manipulating the real time sampled values of data. To explain how FDIs can lead to false tripping of DCR under healthy system condition, we consider eq. (1a) where terminal current, $I_{nm_+}(k)$ has been manipulated to a false data variable, defined as $\zeta_{nm_+}(k)$. Hence, eq. (1a) is modified to:

$$\left|I_{m_+n_+}(k)\right| = \left|I_{mn_+}(k) + \zeta_{nm_+}(k)\right| \quad (2)$$

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

Based on eq. (2), the trip region even under normal operation of system due to FDI attack is defined for ζ_{nm_+} as shown in Fig. 2.

B. Vulnerability to TSAs

Focused on targeting the synchronisation of terminal current information by spoofing the IEEE 1588 Precision Timing Protocol in a substation [25], TSAs can manipulate the time stamped samples [23] of remote current measurement. Assuming TSA changes the time stamped samples of $I_{nm_+}(k)$ by either leading or lagging nT_s samples, eq. (1a) is modified to either eq. (3a) or (3b) :

$$\begin{cases} \left|I_{m_+n_+}(k)\right| = \left|I_{mn_+}(k) + I_{nm_+}(k+n)\right| & \text{(3a)} \\ \left|I_{m_+n_+}(k)\right| = \left|I_{mn_+}(k) + I_{nm_+}(k-n)\right| & \text{(3b)} \end{cases}$$

TSA can lead to DCR tripping even under normal power transients in system. Fig. 2 shows the trip region for FDIs and TSAs.

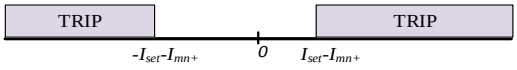


Fig. 2: Region for ζ_{nm+}(k) for FDIs; I_{nm+}(k+n) for TSAs

IV. Fault Analysis for Different Fault Contingencies Based on Line and Zero Mode Voltages Across CLR

The approach used to mitigate the poor resiliency of DCR against ECIs is based on analysis of faulty network of MVDC microgrid. Unlike [23], the proposed method deploys just the CLR of the POC circuit at each converter terminal i.e., the proposed method does not require parallel capacitor of the POC circuit as used in [23]. A modal transformation (shown in eq. (4)) is used in the proposed analysis [27], [34] to obtain the zero-mode and line-mode equivalent networks. The transformation decouples the dependency of poles of bipolar line configuration under PTG faults.

$$\begin{bmatrix} x_l \\ x_0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_p \\ x_n \end{bmatrix} \quad (4)$$

where x_l and x_0 are defined as the line-mode and zero-mode variables whereas x_p and x_n are the positive-pole and negative-pole variables. The proposed analysis works to obtain the line-mode and zero-mode voltage across different CLRs in the faulty mode networks for different fault contingency i.e., PTP, PTG. The proposed analysis works independent of different structure of MVDC microgrid classifying the fault as internal or external based on the mode voltages across internal and external CLRs. Since the proposed analysis is localised in nature unlike DCR which uses communication for decision making, the proposed scheme (using DCR as triggering relay and mode voltage based analysis as the tripping relay) can not be penetrated by ECIs.

A. Equivalent network during a fault contingency

Fig. 3 shows the equivalent network of the MVDC microgrid during a PTP fault, F_1 (see Fig. 1). Fig. 3 is used to understand the generic equivalent network of a multi-terminal bipolar structured MVDC microgrid. L_{mn} and R_{mn} are the simplified R-L equivalent of the line configuration between bus m and bus n. The fault location, d is the percentage of unit length for the PTP fault with fault resistance R_f. L_{Tmn} is the current limiting reactor between bus m and n whereas C_m is the output capacitor at bus m in mid-point grounded configuration. The proposed analysis works on a simplified equivalent network shown in Fig. 4 for PTP fault. The fault occurs between *bus a* and *bus b* where CLR between these buses are internal CLRs (L_{Tab}, L_{Tba}) whereas other CLRs are external CLRs (L_{Tad},L_{Tbc}). For a generic multi-terminal structure, the CLRs which are not related to *bus a* and *bus b* are external CLRs. Hence, the simplified equivalent network can be used for the proposed analysis. The analysis assumes a constant pre-fault voltage throughout the MVDC system which means $V_{dc_i}\big|_{i=a}^{i=n} = V_{dc}$. This assumption has been justified in Section VIIB.

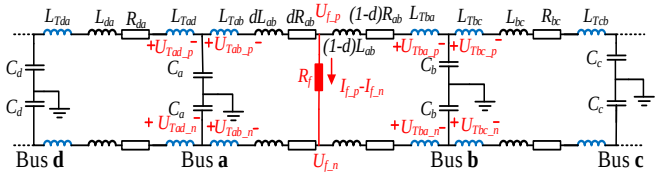


Fig. 4: Simplified equivalent network for a PTP fault

B. Fault analysis for PTG fault contingency

For a positive-pole to ground (P-PTG) fault contingency shown in Fig. 4, U_{f_p}=$I_{f_p}R_f$ and I_{f_n}=*0*. Using the modal transformation, the conditions are transformed to the form:

$$\begin{cases} U_{f_l} + U_{f_0} = 2I_{f_l}R_f & \text{(5a)} \\ I_{f_l} = I_{f_0} & \text{(5b)} \end{cases}$$

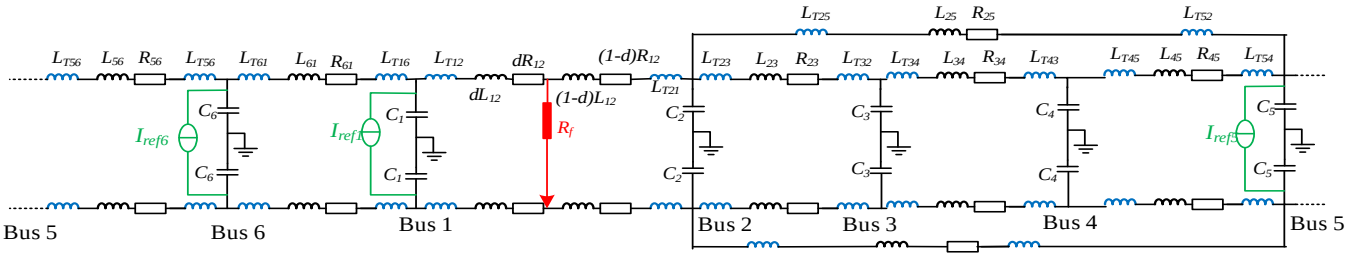


Fig. 3: Equivalent network of MVDC microgrid during a *PTP* fault contingency

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

Considering eq. (5), the combined equivalent mode network for a P-PTG is shown in Fig. 5. The line-mode voltage, U_{Tmn_l} across the CLR, (L_{Tmn}) for a P-PTG fault is derived as:

$$\begin{cases} U_{Tab_l} = -\dfrac{sL_{Tab}V_{dc}}{Z_2 + Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_1} \\ \qquad + \dfrac{sL_{Tab}V_{dc}}{Z_1 + Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]} & (6a) \\ U_{Tad_l} = \dfrac{sL_{Tad}V_{dcd}}{Z_3} - \dfrac{sL_{Tad}V_{dca}}{Z_3} \approx 0 & (6b) \\ U_{Tba_l} = \dfrac{sL_{Tba}V_{dc}}{Z_1 + Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_2} & (6c) \\ \qquad - \dfrac{sL_{Tba}V_{dc}}{Z_2 + Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \\ U_{Tbc_l} = \dfrac{sL_{Tbc}V_{dcb}}{Z_4} - \dfrac{sL_{Tbc}V_{dcc}}{Z_4} \approx 0 & (6d) \end{cases}$$

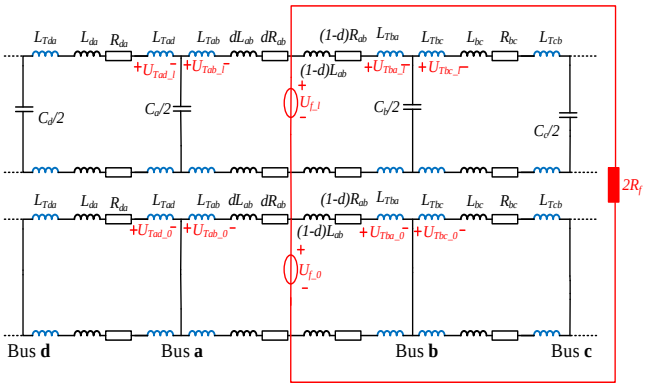


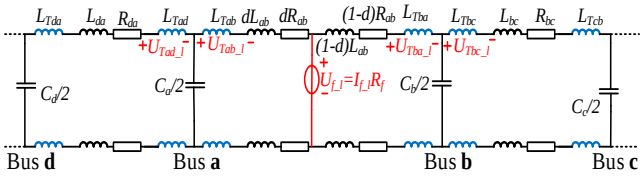
Fig. 5: Simplified equivalent mode network for a P-PTG fault

The impedances, Z_1-Z_4 are mathematically defined as:

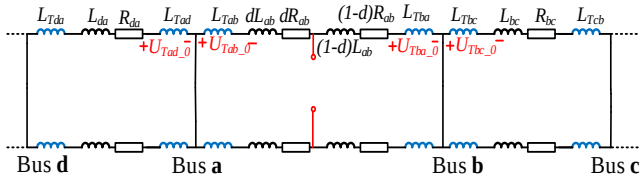
$$\begin{cases} Z_1 = 2s\left(L_{Tab} + dL_{ab}\right) + 2dR_{ab} & (7a) \\ Z_2 = 2s\left(L_{Tba} + (1-d)L_{ab}\right) + 2(1-d)R_{ab} & (7b) \\ Z_3 = 2s\left(L_{Tad} + L_{da} + L_{Tda}\right) + 2R_{da} & (7c) \\ Z_4 = 2s\left(L_{Tbc} + L_{bc+L_{Tcb}}\right) + 2R_{bc} & (7d) \end{cases}$$

Similarly, the zero-mode voltage, U_{Tmn_0} across L_{Tmn} for a P-PTG fault is derived as:

$$\begin{cases} U_{Tab_0} = \dfrac{sL_{Tab}V_{dc}}{Z_1 + Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_1} & (8a) \\ \qquad + \dfrac{sL_{Tab}V_{dc}}{Z_2 + Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_1} \\ U_{Tad_0} = 0 & (8b) \\ U_{Tba_0} = \dfrac{sL_{Tba}V_{dc}}{Z_1 + Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_2} & (8c) \\ \qquad + \dfrac{sL_{Tba}V_{dc}}{Z_2 + Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_2} \\ U_{Tbc_0} = 0 & (8d) \end{cases}$$



(a) Line-mode equivalent network



(b) Zero-mode equivalent network

Fig. 6: Simplified equivalent mode network for a *PTP* fault

For a negative-pole to ground (N-PTG) fault contingency, the line-mode voltage, U_{Tmn_l} is same as eq. (6a)-(6d). However, the zero-mode voltage, U_{Tmn_0} is derived as:

$$\begin{cases} U_{Tab_0} = -\dfrac{sL_{Tab}V_{dc}}{Z_1 + Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_1} & (9a) \\ \qquad - \dfrac{sL_{Tab}V_{dc}}{Z_2 + Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_1} \\ U_{Tad_0} = 0 & (9b) \\ U_{Tba_0} = -\dfrac{sL_{Tba}V_{dc}}{Z_1 + Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_2 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_2} & (9c) \\ \qquad - \dfrac{sL_{Tba}V_{dc}}{Z_2 + Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]} \times \dfrac{Z_1 \| \left[(Z_1 \| Z_2) + 2R_f \right]}{Z_2} \\ U_{Tbc_0} = 0 & (9d) \end{cases}$$

C. Fault analysis for PTP fault contingency

For a *PTP* fault contingency in the system shown in Fig. 4, U_{f_p}-U_{f_n}=$I_{f_p}R_f$ and I_{f_p}+I_{f_n}=*0*. Applying modal transformation, the conditions are transformed to the form:

$$\begin{cases} U_{f_l} = I_{f_l}R_f & (10a) \\ I_{f_0} = 0 & (10b) \end{cases}$$

Considering eq. (10), the equivalent mode network for a *PTP* fault is shown in Fig. 6. The zero-mode voltage, U_{Tmn_0}=*0* across every L_{Tmn} for a PTP fault as evident from Fig. 6(b). The line-mode voltage, U_{Tmn_l} for a *PTP* fault is derived as:

$$\begin{cases} U_{Tab_l} = -\dfrac{sL_{Tab}V_{dc}}{Z_2 + Z_1 \| R_f} \times \dfrac{Z_1 \| R_f}{Z_1} + \dfrac{sL_{Tab}V_{dc}}{Z_1 + Z_2 \| R_f} & (11a) \\ U_{Tad_l} = \dfrac{sL_{Tad}V_{dcd}}{Z_3} - \dfrac{sL_{Tad}V_{dca}}{Z_3} \approx 0 & (11b) \\ U_{Tba_l} = \dfrac{sL_{Tba}V_{dc}}{Z_1 + Z_2 \| R_f} \times \dfrac{Z_2 \| R_f}{Z_2} - \dfrac{sL_{Tba}V_{dc}}{Z_2 + Z_1 \| R_f} & (11c) \\ U_{Tbc_l} = \dfrac{sL_{Tbc}V_{dcb}}{Z_4} - \dfrac{sL_{Tbc}V_{dcc}}{Z_4} \approx 0 & (11d) \end{cases}$$

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

D. Classification of different fault contingencies

Based on the discussion in previous subsections, line-mode and zero-mode voltages are utilised to classify the different fault contingencies in a system i.e., PTP, P-PTG, N-PTG or external. Fig. 7 shows the basis for classification where U_{lset} and U_{0set} are the line-mode and zero-mode voltage thresholds. Section VI explains defining thresholds, U_{lset} and U_{0set}.

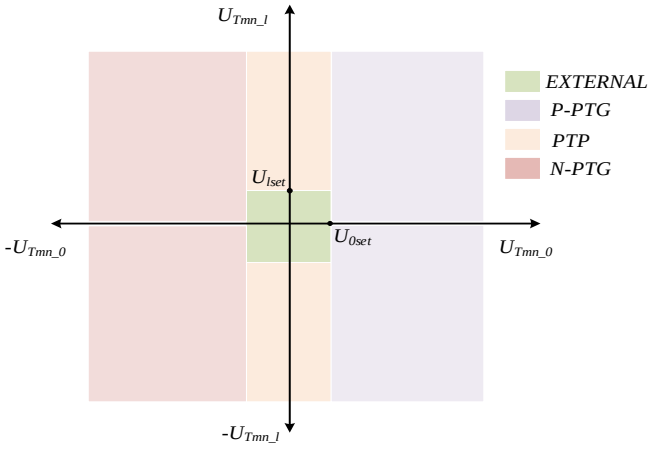


Fig. 7: Classification of different fault contingencies based on U_{Tmn_l} and U_{Tmn_0}

V. Proposed Resilient Fault Identification Approach against ECIs

The proposed fault identification approach ensuring resiliency against ECIs is shown in Fig. 8. The scheme uses DCR operation as the trigger signal for the evaluation of line-mode and zero-mode voltages (U_{Tmn_l} and U_{Tmn_0}) using discrete wavelet transform (DWT). The mother wavelet is *symlets* with order 8 i.e., *sym8* and level 1 with sampling frequency 20kHz. If the differential current of either the positive or the negative polarity of any two buses m and n violate the differential current threshold, I_{set}, the proposed approach proceeds forward to calculate U_{Tmn_l} and U_{Tmn_0}. Based on the findings of fault classification of section IV and Fig. 6, the approach is further used to differentiate if the system is actually subjected to a fault contingency or there is an ECI or some external fault. For the case of ECI or external fault, no trip signal is generated whereas for any internal fault (PTP or PTG), a trip signal is generated. The approach also helps to differentiate between the exact type of fault that the system is subjected to. The approach does not generate a trip signal for any external fault, which is the reason that the approach detects ECIs on the adjacent line occurring simultaneous to a real fault. To elaborate, let's assume that line 1 is subjected to a fault whereas at the same time, there is an ECI on line 6 as shown in Fig. 1. In both the cases, DCR triggers evaluation of U_{Tmn_l} and U_{Tmn_0}. Since, the mode voltages only violate the threshold in case of an internal fault, U_{T12_l} and U_{T21_l} violate U_{lset} whereas U_{T16_l} and U_{T23_l} do not violate U_{lset}. Hence, relay at line 1 generates a trip signal whereas the relay at line 6 does not generate a trip signal.

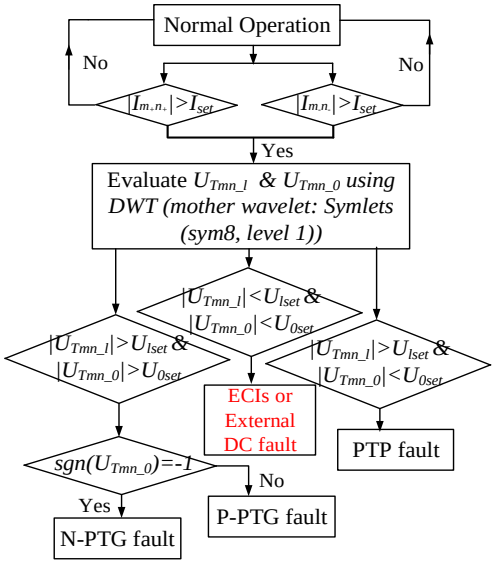


Fig. 8: Layout of the proposed resilient protection approach based on DCR and mode voltages

VI. Sensitivity Analysis

The decision of thresholds (U_{0set} and U_{lset}) is made using the logarithmic partial derivative based sensitivity analysis [30]. The function *G(s)* is the mode-voltage, whereas the variables p_i are the critical parameters (R_f, d). The sensitivity of function *G(s)* for a small variation in p_i is defined as:

$$S_{p_i}^{G(s)} \triangleq \left.\frac{\partial \ln G(s)}{\partial \ln p_i}\right|_{p_{i0}} = \left.\frac{\frac{\partial \ln G(s)}{G(s)}}{\frac{\partial \ln p_i}{p_i}}\right|_{p_{i0}} = \left.\frac{\partial G(s)}{\partial p_i}\right|_{p_{i0}} \frac{p_{i0}}{G_0(s)} \quad (12)$$

where p_{i0} represents nominal value of the *ith* parameters and $G_0(s)$ represents the nominal value of mode-voltage with p_{i0} parameter. For a small change (Δp_i), the mode-voltage function is evaluated as:

$$G(s,p_i) = G(s,p_{i0}) + \frac{\Delta p_i}{p_{i0}} S_{p_i}^{G(s)} \quad (13)$$

where $\Delta p_i = p_i - p_{i0}$ and for i=1, $\sum_{i=1}^{n} S_{p_i}^{G(s)} = S_{p_1}^{G(s)}$.

1) Sensitivity with fault resistance, R_f: *PTP* line-mode voltage is considered as the sensitivity function (U_{Tab_l}). The sensitivity of line-mode voltage with fault resistance is defined as:

$$S_{R_f}^{U_{Lab_l}} = \left.\frac{\partial \ln U_N}{\partial \ln R_f}\right|_{R_{f0}} \overset{0}{\cancel{}} - \left.\frac{\partial \ln U_D}{\partial \ln R_f}\right|_{R_{f0}} \quad (14)$$

$$S_{R_f}^{U_{Lab_l}} = -\frac{\gamma R_{f0}}{Z_2 + \gamma R_{f0}} \quad (15)$$

where $U_{Lab_l} = \frac{U_N}{U_D}$ and $1 + \frac{Z_2}{Z_1} = \gamma$. For a small change ($\Delta R_f$), the line-mode voltage, U_{Tab_l} is evaluated as:

$$U_{Tab_l}(s,R_f) = U_{Tab_l}(s,R_{f0}) - \frac{\gamma \Delta R_f}{Z_2 + \gamma R_{f0}} \quad (16)$$

where $\Delta R_f = R_f - R_{f0}$. With increasing R_f, the signal damping is increasing which means voltage drop due to fault is lesser for higher R_f. The variation of mode voltages

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

with R_f is shown in Fig. 8 for PTP and PTG faults. Based on the analysis, the thresholds U_{0set} and U_{lset} are defined as 5kV considering a safety margin of 3.09kV.

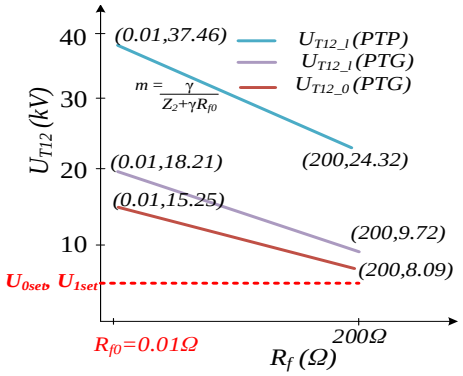


Fig. 9: Variation of mode voltages with R_f

2) Sensitivity with fault location, d: The sensitivity of line-mode voltage with fault location is defined as:

$$S_d^{U_{Lab_l}} = \frac{\partial \ln U_N}{\partial \ln d}\Big|_{d_0}^{0} - \frac{\partial \ln U_D}{\partial \ln d}\Big|_{d_0} \tag{17}$$

$$S_d^{U_{Lab_l}} = -2sL_{12}\frac{(Z_1 - Z_2)d_0}{Z_2 + \gamma R_f} \tag{18}$$

For a small change ($\Delta d = d - d_0$), the line-mode voltage, U_{Tab_l} is evaluated as:

$$U_{Tab_l}(s,d) = U_{Tab_l}(s,d_0) - 2sL_{12}\frac{(Z_2 - Z_1)\Delta d}{Z_1 + \gamma R_f} \tag{19}$$

Fig. 10 shows the variation of mode voltages with fault location, *d*. The drop in voltage is maximum for *d=50%* as evident from eq. (19) and Fig. 10. The thresholds, U_{0set} and U_{lset} defined as 5kV gives a safety margin of 12.17kV. The threshold for the trigger event, I_{set} is defined for a

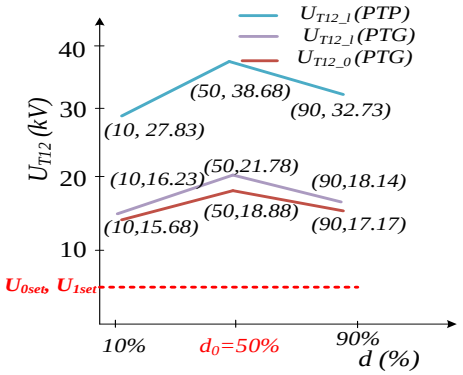


Fig. 10: Variation of mode voltages with fault location, *d*

DC fault contingency in the system with the objective that normal system transients should not be identified as a fault. Also, considering that HIFs violate I_{set} and the trigger should be activated in time keeping margin for the complete algorithm. Hence, I_{set} is defined as 0.5kA.

VII. Performance Evaluation

The section simulates the test system using PSCAD/EMTDC based electromagnetic transient simulations. The parameters employed for the simulation validation are mentioned in Table I. The sampling

TABLE I: Parameters in PSCAD/EMTDC

DC Voltage *(kV)*	±10
Smoothing Reactor *(mH)*	10
Converter terminal Capacitance, C_n *(μF)*	1000-5000
Line length *(km)*	**Line 1**: 4.2 **Line 2**: 3.5 **Line 3**: 4.5 **Line 4**: 2.1 **Line 5**: 3.3 **Line 6**: 1.5
Parameters of UGC	Resistance [Ω/km]: 0.098 Inductance [mH/km]: 0.074 Capacitance [μF/km]: 0.272

frequency, f_s used for the simulation results is 100 kHz for clear results. The variation of mode voltage with f_s is included in Section VII(C3) showing that a sampling frequency of 20 kHz is sufficient for the proposed method. The section validates the performance of the proposed scheme for different faults (PTP, P-PTG, N-PTG) and resiliency against FDIs and TSAs. Further, the performance is validated for power transients, AC external faults, detecting ECIs on adjacent line occurring simultaneous to a real fault. Finally, the maximum fault detection time is included for the method.

A. Performance for different faults

1) Validation for P-PTG faults: A P-PTG fault (F_1 in Fig. 1) occurs at t=0.3s at a distance of 50% and R_f=10Ω in line 1. The decisive parameters are analysed in Fig. 11. Fig. 11(a) shows positive pole differential current, I_{1+2+} violating I_{set} while negative pole current, I_{1-2-} does not violate the threshold (Fig. 11(b)). Similarly, differential currents of other buses (other than bus 1-2) do not violate I_{set}. Fig. 11(d) shows DC link voltage at bus 1 (V_{dc1}), positive pole voltage with CLR (V_{L1p}) and negative pole voltage with CLR (V_{L1n}). It can be seen that V_{L1p} varies instantaneously while V_{dc1} is around 20kV for 5ms upon fault inception. This is due to sluggish drop of voltage across capacitor. As a common observation for all fault contingencies, DC link voltage does not vary much in comparison to the drop in voltage across CLR. Therefore, the assumption of a constant pre-fault voltage (V_{dc}) is justified for the proposed scheme (section IV(A)). As evident from Fig. 11(f), external buses (other than bus 1-2) do not see instantaneous variation of V_{Lmp} where $m \in [3,6]$. Fig. 11(g)-(i) analyse the tripping parameter i.e., mode voltages at different buses. As evident, mode voltages for bus 1 and bus 2 violate U_{0set} and U_{0set}.

2) Validation for N-PTG faults: Similar to previous sub-section (P-PTG), Fig. 12 shows the fault identification for a N-PTG fault at t=1s at a distance of 30% and R_f=50Ω. Similar conclusions can be drawn as for P-PTG fault. However, the zero-mode voltage, U_{Tmn_0} for N-PTG fault is negative in comparison to P-PTG fault as evident from Fig. 12(g)-(h).

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

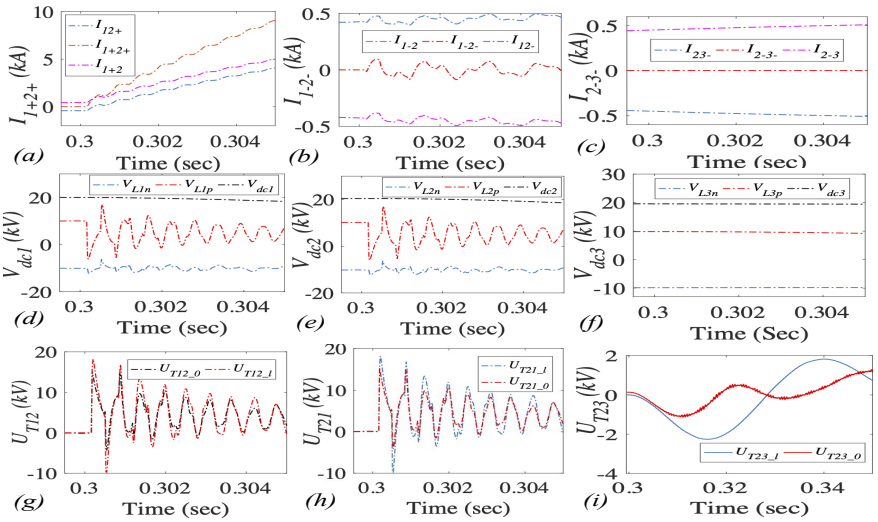


Fig. 11: Simulation results for P-PTG fault at t=0.3s for (a), (b) DCR at Bus 1, (c) Negative polarity DCR at bus 3, (d) Voltage at Bus 1 (kV), (e) Voltage at Bus 2 (kV), (f) Voltage at Bus 3 (kV), (g), (h) Mode voltage at Bus 1, (i) Mode voltage at Bus 3 (kV) vs Time (sec)

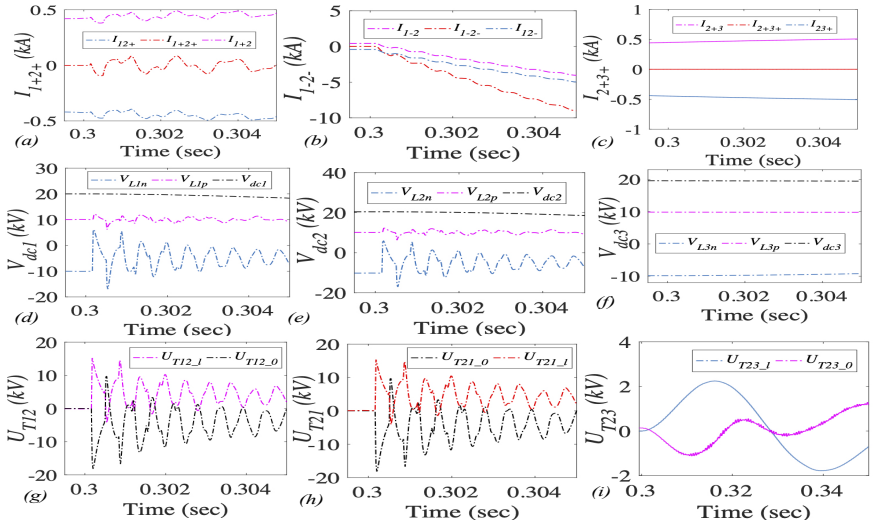


Fig. 12: Simulation results for N-PTG fault at t=0.3s for (a), (b) DCR at Bus 1, (c) Negative polarity DCR at bus 3, (d) Voltage at Bus 1 (kV), (e) Voltage at Bus 2 (kV), (f) Voltage at Bus 3 (kV), (g), (h) Mode voltage at Bus 1, (i) Mode voltage at Bus 3 (kV) vs Time (sec)

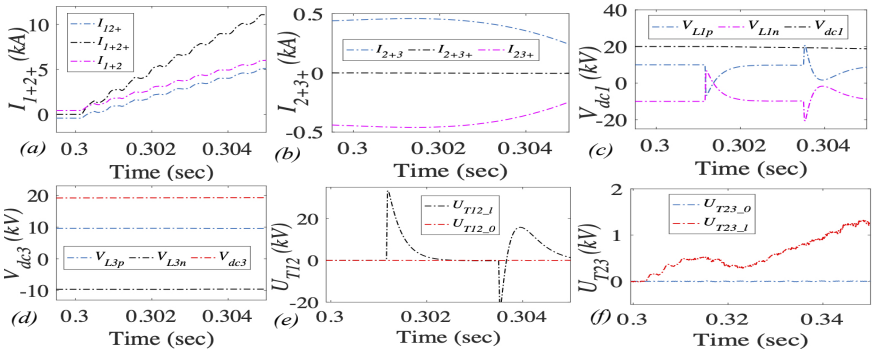


Fig. 13: Simulation results for PTP fault at t=0.3s for (a) DCR at Bus 1, (b) Negative polarity DCR at bus 3, (c) Voltage at Bus 1 (kV), (d) Voltage at Bus 3 (kV), (e) Mode voltage at Bus 1, (f) Mode voltage at Bus 3 (kV) vs Time (sec)

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

3) Validation for PTP faults: Fig. 13 shows the fault identification for a PTP fault. V_{L1p} and V_{L1n} both vary instantaneously for bus 1-2 (Fig. 13(c)). The zero-mode voltage, U_{Tmn_0} is nearly 0 for PTP faults as evident from Fig. 13(e).

B. Validation of proposed protection approach against ECIs

1) Validation against FDIs: A FDI attack takes place at line 1 where ζ_{1+2} is the false data injected in place of I_{1+2} as shown in Fig. 14(a). Fig. 14(c)-(d) show mode voltages for the intrusion. It can be seen that since the mode voltage calculation is based on local parameters, cyber intrusion cannot cause variation in mode voltages. The mode voltages only vary for physical fault in the system.

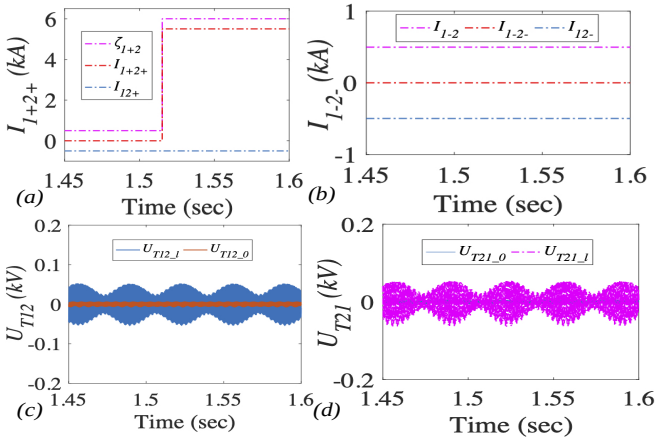


Fig. 14: Simulation results for FDI attack at t=1.52s for (a), (b) DCR at Bus 1-2, (c), (d) Mode Voltage at Bus 1-2 (kV) vs T (sec)

2) Validation against TSAs: A TSA takes place at line 1 where $I_{12+}(k+n)$ is the parameter under synchronization attack. For a load change in the system at t=0.75s, the positive pole differential current, I_{1+2+} violates I_{set}. The mode voltages used as the tripping signal do not violate U_{0set} and U_{lset} as shown in Fig. 15(c)-(d).

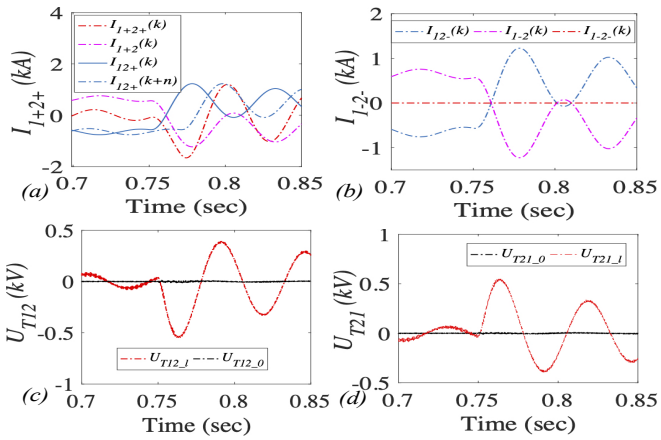


Fig. 15: Simulation results for TSA due to load change at t=0.75s for (a), (b) DCR at Bus 1-2, (c), (d) Mode Voltage at Bus 1-2 (kV) vs T (sec)

3) Validation against FDI attacks simultaneous to fault: A FDI attack takes place at line 2 simultaneous to a PTP fault at line 1 at t=1s. As evident from Fig. 16, parameters of bus 1, subjected to an actual fault violate their respective thresholds while bus 3, subjected to FDI attack violates I_{set}. However, mode voltage, U_{T23} does not violate U_{0set} and U_{lset}. Hence, a fault generates a trip signal whereas an FDI attack generates a trigger signal (violation of I_{set}) but does not generate a trip signal.

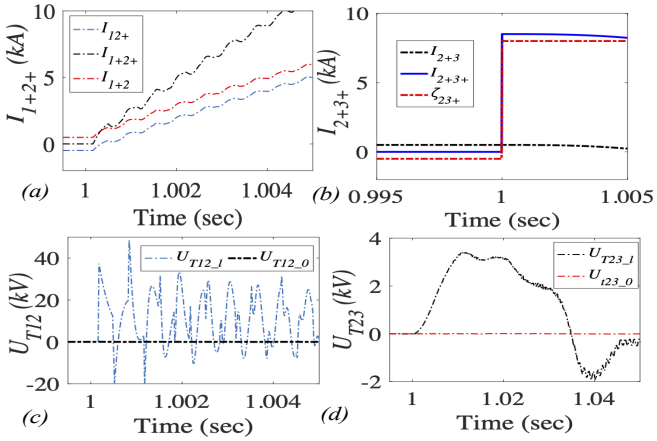


Fig. 16: Simulation results for fault at line 1 and simultaneous FDI attack at line 2 for (a) DCR at Bus 1, (b) DCR at Bus 3, (c) Mode Voltage at Bus 1 (kV), (d) Mode Voltage at Bus 3 (kV) vs T (sec)

4) Validation against TSA attacks simultaneous to fault: A TSA takes place at line 2 simultaneous to a PTP fault at line 1 at t=1s. As evident from Fig. 17, parameters of bus 1, subjected to an actual fault violate their respective thresholds while bus 3, subjected to TSA violates I_{set}. However, mode voltage, U_{T23} does not violate U_{0set} and U_{lset}. Hence, a fault generates a trip signal whereas a TSA generates a trigger signal (violation of I_{set}) but does not generate a trip signal.

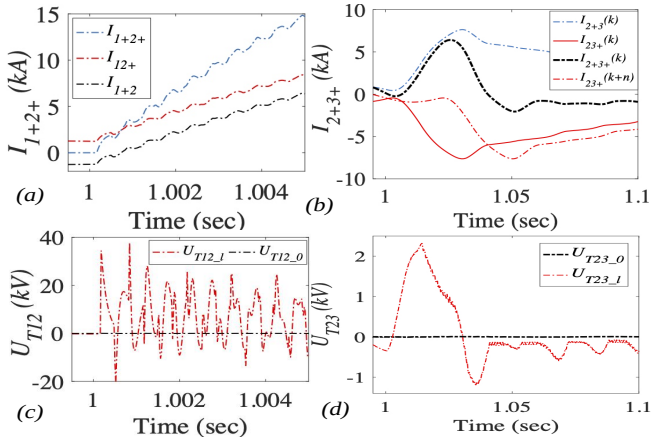


Fig. 17: Simulation results for fault at line 1 and simultaneous TSA at line 2 for (a) DCR at Bus 1, (b) DCR at Bus 3, (c) Mode Voltage at Bus 1 (kV), (d) Mode Voltage at Bus 3 (kV) vs T (sec)

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

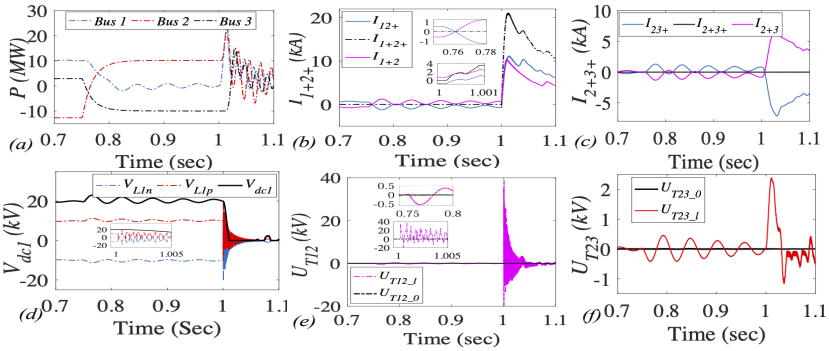


Fig. 18: Simulation results for power transients at t=0.75s and PTP fault at t=1s for (a) P (MW) vs Time (sec), (b) DCR at Bus 1, (c) DCR at Bus 3, (d) Voltage at Bus 1 (kV), (e) Mode voltage at Bus 1 (kV), (f) Mode voltage at Bus 3 (kV) vs Time (sec)

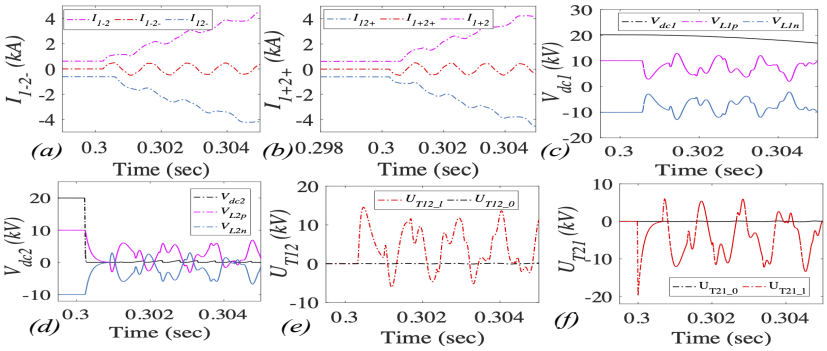


Fig. 19: Simulation results for AC external fault at bus 2 at t=0.3s for (a), (b) DCR at Bus 1, (c) Voltage at Bus 1 (kV), (d) Voltage at Bus 2 (kV), (e) Mode voltage at Bus 1 (kV), (f) Mode voltage at Bus 2 (kV) vs Time (sec)

C. Robustness Validation

1) Power Transients: To check the performance of the scheme under system power transients, the P_{ref} command at bus 1-3 are changed at t=0.75s (P_1 from +10.2 to 0MW, P_2 from -12.6MW to +10.4MW, P_3 from +2.4MW to -10.4MW). Further, a PTP fault occurs at t=1s at line 1.

As evident from Fig. 18(b)-(c), differential current, I_{1+2+} does not violate I_{set} at t=0.75s (power transients) but it does that at t=1s (PTP fault contingency). Fig. 18(d) shows bus voltage at bus 1. As expected, the voltage fluctuates around 20kV for a power transient and collapses for a fault. Fig. 18(e) shows mode voltages of bus 1-2 not violating U_{0set} and U_{lset} for power transient but violating the respective thresholds for a fault contingency. Bus 3-6 mode voltages do not violate U_{0set} and U_{lset} for power transient as well as external DC fault at line 1.

2) AC external faults: To check the response of AC faults on the proposed scheme, a 3ϕ metallic AC fault is simulated on the AC side of bus 2 (Fig. 1). As a result of which, V_{dc2} collapses instantaneously along with V_{L2p} and V_{L2n}. As evident from Fig. 19(a)-(b), I_{1+2+} and I_{1-2-} do not violate threshold, I_{set}. However, the mode voltages maloperate for an external AC fault violating U_{0set} and U_{lset}. As a result of which, DCR becomes an integral part of the proposed scheme for external AC faults where DCR does not generate any triggering signal. Hence the proposed method works for external AC faults as well.

3) Effect of Sampling Frequency: The effect of different f_s on mode voltage variation is analysed in Fig. 20. For f_s=100kHz, U_{T12_l}=25.63kV, for f_s=50kHz, U_{T12_l}=24.58kV and for f_s=20kHz, U_{T12_l}=21.18kV. This shows that a sampling frequency, f_s=20kHz is sufficient for the proposed method. DCR parameter (I_{m+n+}) has a gradual increase unlike mode voltages and hence is unaffected by f_s variation.

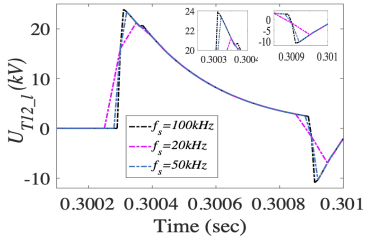


Fig. 20: Simulation results for varying sampling frequency, f_s

4) Fault Detection Time for the proposed scheme: Detection time is a result of delay of propagation time of travelling waves, delay to violate the triggering event and delay due to DWT. Table II shows the maximum fault detection time for the proposed method. As the cables

TABLE II: Fault Detection Time for the proposed scheme

Maximum Propagation time of travelling wave	*0.05ms*
Delay for HIF (*200*Ω) to violate triggering event	*0.72ms*
Delay due to 1-level wavelet transform	*0.32ms*
Maximum delay encountered	*1.09ms*

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

considered in MVDC microgrids are under 5km, the maximum propagation time of travelling wave is 0.05ms. The delay for HIF (200Ω) to violate triggering threshold, I_{set} is 0.72ms. The delay due to DWT is 0.32ms which makes the maximum fault detection time as 1.09ms. WGN in the measurement induce high frequency components in the measured data of current and voltage. As a result, a rolling mean filter is employed with a moving window of 50 sample steps [29]. WGN is random in nature with the property of zero mean [27]. Hence, the filter eliminates the effect of WGN to a great extent.

VIII. Conclusion

The contribution of the proposed work has been to address the issue of resiliency in differential current protection in Medium Voltage DC (MVDC) microgrids against external cyber intrusions (ECIs) of different types, i.e., false data injection (FDI) and time synchronization attacks (TSAs). The proposed protection approach also differentiates between a real fault and ECIs on adjacent line occurring simultaneously which has not been addressed in the literature so far. The approach deploys current limiting reactors (CLR) and analysis the mode-voltages across CLR using modal transformation. The approach includes differential current relay (DCR) and mode voltage analyses to propose a resilient protection scheme against ECIs. The proposed method has good sensitivity to high impedance faults (HIFs). To evaluate the thresholds used in the analysis, a sensitivity function has been formulated to analyse the variation of mode-voltages with respect to decisive parameters like fault resistance and fault location.

References

[1] N. Doerry, J. Amy and C. Krolick, "History and the Status of Electric Ship Propulsion, Integrated Power Systems, and Future Trends in the U.S. Navy," in Proceedings of the IEEE, vol. 103, no. 12, pp. 2243-2251, Dec. 2015, doi: 10.1109/JPROC.2015.2494159.

[2] J. D. Herbst, A. L. Gattozzi, A. Ouroua and F. M. Uriarte, "Flexible test bed for MVDC and HFAC electric ship power system architectures for Navy ships," 2011 IEEE Electric Ship Technologies Symposium, Alexandria, VA, 2011, pp. 66-71, doi: 10.1109/ESTS.2011.5770843.

[3] Z. Jin, L. Meng, J. M. Guerrero and R. Han, "Hierarchical Control Design for a Shipboard Power System With DC Distribution and Energy Storage Aboard Future More-Electric Ships," in IEEE Transactions on Industrial Informatics, vol. 14, no. 2, pp. 703-714, Feb. 2018, doi: 10.1109/TII.2017.2772343.

[4] K. Satpathi, A. Ukil, S. S. Nag, J. Pou and M. A. Zagrodnik, "DC Marine Power System: Transient Behavior and Fault Management Aspects," in IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 1911-1925, April 2019, doi: 10.1109/TII.2018.2864598.

[5] C. Yuan, M. A. Haj-ahmed and M. S. Illindala, "Protection Strategies for Medium-Voltage Direct-Current Microgrid at a Remote Area Mine Site," in IEEE Transactions on Industry Applications, vol. 51, no. 4, pp. 2846-2853, July-Aug. 2015, doi: 10.1109/TIA.2015.2391441.

[6] V. Nougain, S. Mishra and A. K. Pradhan, "MVDC Microgrid Protection Using a Centralized Communication With a Localized Backup Scheme of Adaptive Parameters," in IEEE Transactions on Power Delivery, vol. 34, no. 3, pp. 869-878, June 2019, doi: 10.1109/TPWRD.2019.2899768.

[7] R. M. Cuzner and V. Singh, "Future Shipboard MVdc System Protection Requirements and Solid-State Protective Device Topological Tradeoffs," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 5, no. 1, pp. 244-259, March 2017, doi: 10.1109/JESTPE.2016.2638921.

[8] J. Yang, J. E. Fletcher and J. O'Reilly, "Short-Circuit and Ground Fault Analyses and Location in VSC-Based DC Network Cables," in IEEE Transactions on Industrial Electronics, vol. 59, no. 10, pp. 3827-3837, Oct. 2012, doi: 10.1109/TIE.2011.2162712.

[9] A. Hooshyar and R. Iravani, "Microgrid Protection," in Proceedings of the IEEE, vol. 105, no. 7, pp. 1332-1353, July 2017, doi: 10.1109/JPROC.2017.2669342.

[10] S. D. A. Fletcher, P. J. Norman, S. J. Galloway, P. Crolla and G. M. Burt, "Optimizing the Roles of Unit and Non-unit Protection Methods Within DC Microgrids," in IEEE Transactions on Smart Grid, vol. 3, no. 4, pp. 2079-2087, Dec. 2012, doi: 10.1109/TSG.2012.2198499.

[11] M. Monadi, C. Gavriluta, A. Luna, J. I. Candela and P. Rodriguez, "Centralized Protection Strategy for Medium Voltage DC Microgrids," in IEEE Transactions on Power Delivery, vol. 32, no. 1, pp. 430-440, Feb. 2017, doi: 10.1109/TPWRD.2016.2600278.

[12] S. Dhar and P. K. Dash, "Differential current-based fault protection with adaptive threshold for multiple PV-based DC microgrid," in IET Renewable Power Generation, vol. 11, no. 6, pp. 778-790, 10 5 2017, doi: 10.1049/iet-rpg.2016.0577.

[13] S. Dhar, R. K. Patnaik and P. K. Dash, "Fault Detection and Location of Photovoltaic Based DC Microgrid Using Differential Protection Strategy," in IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 4303-4312, Sept. 2018, doi: 10.1109/TSG.2017.2654267.

[14] J. Liu, X. Lu and J. Wang, "Resilience Analysis of DC Microgrids Under Denial of Service Threats," in IEEE Transactions on Power Systems, vol. 34, no. 4, pp. 3199-3208, July 2019, doi: 10.1109/TPWRS.2019.2897499.

[15] J. Hong, C. Liu and M. Govindarasu, "Integrated Anomaly Detection for Cyber Security of the Substations," in IEEE Transactions on Smart Grid, vol. 5, no. 4, pp. 1643-1653, July 2014, doi: 10.1109/TSG.2013.2294473.

[16] D. E. Whitehead, K. Owens, D. Gammel and J. Smith, "Ukraine cyber-induced power outage: Analysis and practical mitigation strategies," 2017 70th Annual Conference for Protective Relay Engineers (CPRE), College Station, TX, 2017, pp. 1-8, doi: 10.1109/CPRE.2017.8090056.

[17] B. Moussa, M. Debbabi and C. Assi, "A Detection and Mitigation Model for PTP Delay Attack in an IEC 61850 Substation," in IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 3954-3965, Sept. 2018, doi: 10.1109/TSG.2016.2644618.

[18] Y. Yang, H. Xu, L. Gao, Y. Yuan, K. McLaughlin and S. Sezer, "Multidimensional Intrusion Detection System for IEC 61850-Based SCADA Networks," in IEEE Transactions on Power Delivery, vol. 32, no. 2, pp. 1068-1078, April 2017, doi: 10.1109/TPWRD.2016.2603339.

[19] Z. Zhang, S. Gong, A. D. Dimitrovski and H. Li, "Time Synchronization Attack in Smart Grid: Impact and Analysis," in IEEE Transactions on Smart Grid, vol. 4, no. 1, pp. 87-98, March 2013, doi: 10.1109/TSG.2012.2227342.

[20] L. Chen et al., "Remedial Pilot Main Protection Scheme for Transmission Line Independent of Data Synchronism," in IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 681-690, Jan. 2019, doi: 10.1109/TSG.2017.2750205.

[21] A. Ameli, A. Hooshyar and E. F. El-Saadany, "Development of a Cyber-Resilient Line Current Differential Relay," in IEEE Transactions on Industrial Informatics, vol. 15, no. 1, pp. 305-318, Jan. 2019, doi: 10.1109/TII.2018.2831198.

[22] A. Ameli, A. Hooshyar, E. F. El-Saadany and A. M. Youssef, "An Intrusion Detection Method for Line Current Differential Relays," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 329-344, 2020, doi: 10.1109/TIFS.2019.2916331.

[23] A. Ameli, K. A. Saleh, A. Kirakosyan, E. F. El-Saadany and M. M. A. Salama, "An Intrusion Detection Method for Line Current Differential Relays in Medium-Voltage DC Microgrids," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3580-3594, 2020, doi: 10.1109/TIFS.2020.2991892.

[24] A. Jahangir, V. Nougain and S. Mishra, "Control Topology of Hybrid Energy Storage System for AC-DC Microgrid," 2018 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Chennai, India, 2018, pp. 1-5, doi: 10.1109/PEDES.2018.8707682.

[25] D. M. E. Ingram, P. Schaub, D. A. Campbell and R. R. Taylor, "Quantitative Assessment of Fault Tolerant Precision Timing for Electricity Substations," in IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 10, pp. 2694-2703, Oct. 2013, doi: 10.1109/TIM.2013.2263673.

[26] J. Liu, N. Tai and C. Fan, "Transient-Voltage-Based Protection Scheme for DC Line Faults in the Multiterminal VSC-HVDC System," in IEEE Transactions on Power Delivery, vol. 32, no. 3, pp. 1483-1494, June 2017, doi: 10.1109/TPWRD.2016.2608986.

[27] S. Yang, W. Xiang, R. Li, X. Lu, W. Zuo and J. Wen, "An Improved DC fault Protection Algorithm for MMC HVDC Grids based on Modal Domain Analysis," in IEEE Journal of Emerging and Selected Topics in Power Electronics, doi: 10.1109/JESTPE.2019.2945200.

[28] Y. M. Yeap, N. Geddada, and A. Ukil, âAnalysis and validation of wavelet transform based DC fault detection in HVDC system,â in Appl. Soft Comput., vol. 61, pp. 127â137, 2017.

[29] V. Nougain, V. Nougain and S. Mishra, "Low-voltage DC ring-bus microgrid protection with rolling mean technique," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-6.

[30] A. MALKHANDI, N. Senroy and S. Mishra, "Return Ratio Shaping Approach to Stabilize Inverter-Weak Grid System," in IEEE Transactions on Energy Conversion, doi: 10.1109/TEC.2020.3003074.

[31] "IEEE Recommended Practice for the Design and Application of Power Electronics in Electrical Power Systems," in IEEE Std 1662-2016 (Revision of IEEE Std 1662-2008) , vol., no., pp.1-68, 9 March 2017, doi: 10.1109/IEEESTD.2017.7874058.

[32] "IEEE Recommended Practice for 1 kV to 35 kV Medium-Voltage DC Power Systems on Ships," in IEEE Std 1709-2018 (Revision of IEEE Std 1709-2010) , vol., no., pp.1-54, 7 Dec. 2018, doi: 10.1109/IEEESTD.2018.8569023.

[33] T. Ericsen, N. Hingorani and Y. Khersonsky, "Power electronics and future marine electrical systems," in IEEE Transactions on Industry Applications, vol. 42, no. 1, pp. 155-163, Jan.-Feb. 2006, doi: 10.1109/TIA.2005.861306.

[34] J. Liao, N. Zhou and Q. Wang, "DC Grid Protection Method Based on Phase Planes of Single-End Common-and Differential-Mode Components," in IEEE Transactions on Power Delivery, vol. 36, no. 1, pp. 299-310, Feb. 2021, doi: 10.1109/TPWRD.2020.2976721.

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRD.2021.3074879, IEEE Transactions on Power Delivery

Vaibhav Nougain received B.Tech degree in electrical engineering in 2017 from Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India. He is currently working toward the Ph.D. degree at the Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India. He was a Visiting Student with The Center for Electric Power and Energy (CEE), Technical University of Denmark in 2019. His research interests include coordinated protection and control of DC microgrids, and protection of HVDC systems.

Sukumar Mishra (Senior Member, IEEE) received his M.Tech and PhD in Electrical Engineering from National Institute of Technology, Rourkela in 1992 and 2000 respectively. Presently, Dr. Mishra is a Professor at the Indian Institute of Technology Delhi and has been its part for the past 17 years. He has won many accolades such as INSA Medal for Young Scientist (2002), INAE Young Engineer Award (2009, 2002), INAE Silver Jubilee Young Engineer Award (2012), Samanta Chandra Shekhar Award (2016), IETE Bimal Bose award (2019), National Mission Innovation Champion award (2019) and NASI-Reliance Industries Platinum Jubilee Award for Application Oriented Innovation in Physical Sciences (2019). He has been granted fellowships from academies like NASI (India), INAE (India), and professional societies like IET (U.K.), IETE (India), IE (India). He has also been recognized as the INAE Industry Academic Distinguished Professor. His research interests lie in the field of Power Systems, Power Quality Studies, Renewable Energy and Smart Grid.
Prof. Mishra is currently acting as the ABB Chair professor and has previously delegated as the NTPC, INAE, and Power Grid Chair professor. He has also served as an Independent Director of the Cross Border Power Transmission Company Ltd. and the River Engineering Pvt. Ltd. And has carried out many important industrial consultations with TATA Power, Microtek and others. He is the founder of SILOV SOLUTIONS PRIVATE LIMITED, a company that specifically deals in products related to renewable energy sources utilizable at household scale as well as at commercial setups. From March 2020, he has also been functioning as the Associate Dean Research and Development of IIT Delhi. Prof. Mishra has been working in close association with IEEE Delhi Section Executive Committee for the past few years and is currently serving as an Editor for the IEEE Transactions on Smart Grid, IEEE Transactions on Sustainable Energy and was an Area Editor for the IET Generation, Transmission Distribution journal.

Sushree Subhasmita Jena received her Bachelor of Technology degree in Electrical Engineering from National Institute of Technology, Hamirpur, Himachal Pradesh, India in 2020. She was selected for the Indian Academy of Sciences - Summer Research Fellowship program-2019, for a research internship at Indian Institute of Technology (IIT), Delhi. Her interest areas are Automation, Embedded Systems and Power Systems which come under the broad scope of Electrical and Computer Engineering. Currently, she works with Tata Power Solar Systems Limited in the Department of Engineering, where she is responsible for technical design, generation analysis and costing for installation and commissioning of innovative solar rooftop solutions and other sustainable small scale utility projects for the western India region.

0885-8977 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2021 at 10:53:05 UTC from IEEE Xplore. Restrictions apply.

Clinical and Translational Oncology
https://doi.org/10.1007/s12094-021-02604-x

REVIEW ARTICLE

Review of biomarker systems as an alternative for early diagnosis of ovarian carcinoma

A. Srivastava[1] · A. Gupta[1] · S. Patidar[2]

Received: 15 February 2021 / Accepted: 19 March 2021
© Federación de Sociedades Españolas de Oncología (FESEO) 2021

Abstract

Early diagnosis of ovarian carcinoma is bound to boost the long-term endurance rate of the patients. Most ovarian tumors happen post menopause when the ovaries have no vital operation and therefore irregular ovarian role causes no signs. According to Muinao T. et al. (Heliyon. 5(12):e02826, 2019), if we consider the frequency of ovarian carcinoma to be moderate, a screening technique must accomplish a base specificity of 99.6% and sensitivity of over 75%. The classification and approval of early diagnostic biomarkers explicit to ovarian carcinoma are essentially required. Prevailing methods for early diagnosis of ovarian carcinoma incorporate TVS, biological marker examination, or a blend of the two or other. In recent years, it has been revealed that a combination of at least two biomarkers has beaten single biomarkers in measures for early diagnosis of the illness. In the present document, we survey the ongoing exploration of innovative characteristic methodologies and possible panels of carcinoma biological markers for the early diagnosis of ovarian carcinoma and discuss biomarkers as the plausible apparatus for model improvement and other progressed approaches as an effective alternative to the prevailing methods for early diagnosis of this dreadful disease to evade bogus analysis and inordinate expense.

Keywords Ovarian carcinoma · CA-125 · Early diagnosis · Biomarker · Sensitivity · Specificity

Introduction

Ovarian carcinoma is the fifth leading reason for carcinoma-related death in ladies in the US and is the main source of gynecologic carcinoma deaths. Regardless of being one-tenth as prevalent as breast carcinoma, it is more deadly and conveys a 1:70 lifetime hazard, according to Chappelear A [4].

As mentioned by National Ovarian Cancer Coalition [40], ovarian carcinoma is divided into three common cell types:

- Epithelial tumor—cells covering the outer lining of the ovaries
- Germ Cell tumor—cells that are destined to form eggs
- Stromal tumor—Cells that release hormones and connect the different structures of the ovaries

As reported by Lockwood S [24], about 50% of ovarian carcinoma cases occur in women over age 60. Due to this, in combination with the anatomical area of the ovaries somewhere deep in the pelvis, ovarian carcinoma growths cause barely any indications until they are large enough or have spread to different parts. Unfortunately, only 15% of ovarian carcinomas are diagnosed at an early stage. Thus, ovarian carcinoma is normally analyzed at an advanced stage when, regardless of advances in surgical and chemotherapeutic management during the recent decade, endurance rates are poor.

Lockwood S [24] also states that early diagnosis of ovarian carcinoma increases the life expectancy by at least five years in 92% of cases as at this time the disease is confined only to the ovary and easily treatable, thus younger women, i.e., below age 65, show better rates of survival than older women. When they factored all the stages of diagnosis, they

✉ S. Patidar
sanjaypatidar@dtu.ac.in

A. Srivastava
anshikasrivastava_2k18bt002@dtu.ac.in

A. Gupta
anshulgupta_2k18bt003@dtu.ac.in

[1] Department of Biotechnology, Delhi Technological University, New Delhi, India

[2] Department of Computer Science, Delhi Technological University, New Delhi 110042, India

Published online: 11 April 2021

found that only 45% of women with the disease were able to survive five years after they were diagnosed. The chances of death increase with the stage of the disease, as only about 17% of women survived for five years when it has spread beyond the abdomen.

Momenimovahed Z. et al. [33] argues that the high death rate of ovarian carcinoma, 4.4% of the whole carcinoma-related deaths amid women, is because of the absence of a screening methodology to distinguish the early-phase illness.

Figure 1 shows how the biomarkers for ovarian carcinoma have evolved over the years while Table 1 shows their sensitivity (SN), specificity(SP), methods involved in their diagnosis, and their accuracy also [62–70]. Muinao T. et al. [38] claims that a screening technique must accomplish a base SP of 99.6% and an SN of over 75% for early-phase carcinoma to keep away from an inadmissible degree of false-positive outcomes and accomplish a positive predictive value (PPV) of 10%, considering the low commonness of ovarian carcinoma.

Related work

A few pieces of evidence have featured countless biological markers with possible applicability in the early diagnosis of ovarian carcinoma but their application has been a challenge. Consequently, the mixture of some biological markers has not functioned as a preferred option over the single biological marker to give high SN and SP. For example, separating the harmless from harmful instances of ovarian carcinoma by ROMA has delivered an AUC (Area Under the Curve) of 91.2%. Further expansion of YKL-40, transthyretin, ApoA1, Beta 2 microglobulin, transferrin, and LPA to ROMA has anyway not accomplished surprisingly better AUC, as reported by Moore RG et al. [35] Accordingly, cautious choice of calculation and precise logical methodology will improve the adequacy of biological marker panel and prognostication of early-stage ovarian carcinoma.

Statistical or mathematical approaches

While the present indicative mechanism of ovarian carcinoma analyzes the degree of CA-125, it delivers a few bogus positive outcomes. The degrees of basal biological markers that show significant changes at a moment (change point) before the expansion in CA125 that might have been a significant deciding variable in recognizing the early illness are usually missed prompting higher odds of analysis when the illness has progressed. The numerical calculation, for example, in the progressive Bayesian Model and Markov Chain Monte Carlo techniques have been used to calculate the change point of biological markers previously also or those without demonstrating the change point in CA125 and fill the void. The researchers found that ovarian carcinoma biological markers, for example, HE4 and glycodelin displayed a change point in 80% and 60% individually while the CA125 level does not show a change point. Also, the change point of these two biological markers showed

Fig. 1 Evolution of ovarian carcinoma biomarkers

Table 1 SP and SN of single biomarkers for the early diagnosis of ovarian carcinoma from numerous papers

Biomarker	Specificity (%)	Sensitivity (%)	Methods performed	Accuracy	Reference
CA-125	80	92	Electrochemilluminescence(ECLIA) technique	78%	[27]
HE4	85	71.19	Clinical chemistry and immunochemistry	–	[59]
B7-H4	90.48	81.82	Electro-chemiluminescence immunoassays	85.57%	[19]
Osteopontin	98	7.6	ELISA	–	[36]
VEGF	71	74	ELISA	–	[7]
KLK6 (hK6)	95	21–26	Immunoassay	–	[10]
IL-6	86	84.1	LabMAP assays	90%	[15]
IL-8	98	65.5	LapMAP technology	80.3%	[25]
RECAF	100	52.5	Chemiluminescence assay	–	[52]
SMRP	98	15.4	MESOMARK Assay	–	[36]
Transthyretin	95	47	Singleplex Lummex bead assays	–	[8]
Prostasin	94	51.4	Enzyme-linked immunosorbent assay	–	[32]

a higher likelihood of happening quicker than CA125, as reported by Mariño IP et al.[30] The investigation sums up in favor of the mix of CA125 with HE4 or glycodelin as opposed to CA125 alone as a biological marker to improve the productivity for early diagnosis of ovarian carcinoma. While considering a biological marker panel, essentially, compounds of biological markers in a collection are not sufficient yet picking the correct compound matters and at the same time staying a hurdle. A hypothetical finding dependent on the multivariate normal distribution and linear algorithm showed that the extension of a marker, however, of low predictive capability, to a primary marker will expand AUC if the extra marker is not well associated with the primary marker. Conversely, if the extra marker interacts well with the primary marker, at that point, the likelihood to expand the AUC is less regardless of the predictive capability of the extra marker. For instance, Prolactin with moderately lower AUC all alone attributable to its negative connection with CA125, shows a higher AUC in mixture with CA125 as compared to biological markers with higher univariant AUC than prolactin that has a positive relationship with CA125 [48]. Henceforth, it is hard to locate the most reliable biological marker mixtures in foreseeing the illness and along these lines, a statistical strategy may be useful in bringing the most ideal biological marker mixture.

Havrilesky LJ et al. [17] used a two-venture multi-panel biological marker for use in medical screening has been notified. In this, the principal assay recognizes the composition of a biological marker or a panel of biological markers to rationalize the illness enhanced populace and limits the genuine negatives. Besides, the composition of the second biological marker or biological markers panel is identified to distinguish more correctly the genuine positive outcomes. In the first step, HE4 at a higher level than 1.8 ng/ml is considered as a positive result, and in the subsequent step, Receiver Operating Characteristic (ROC) arc Best cutoff algorithm considered the result positive if CA125 is positive at > 35 U/ml, or if any two of Glycodelin, MUC-1, or Plau-R are investigated positive. The calculation recognizes the early-phase ovarian carcinoma at SN and SP of 76.7% and 97.2% individually.

Raman spectroscopy

The presence of a low degree of serum biological markers in the early phase of the illness might be difficult in handling clinical investigations. As of late, Raman spectroscopy has been utilized to measure the blood tests for the determination of early and advanced-stage ovarian carcinoma as stated in reports by Paraskevaidi M et al. [44] and Peng D et al. [47] Raman Spectroscopy depends on the phenomenon of inelastic light dissipating caused when light associates with the matter. The variation in energy of the specimen's electron to its initial level from its energized state is a quality of explicit biomolecules including proteins, nucleic acids, and lipids, which are instructive about the specimen.

Surface-enhanced Raman spectroscopy (SERS) is the improved adaptation of Raman spectroscopy that improves the Raman signal by 10^3–10^{10} times. It depends on coarse metallic coverings or nanostructures, for example, silver or gold nanoparticles, and utilizes the increment in the electromagnetic field from motions of surface electrons, named surface plasmons, states Fazio B et al. [13] This permits the identification of atoms even at low intensity, according to Kong K et al. [22] and Lyng FM et al. [28]

Utilizing Raman spectroscopy to assess the investigation of early ovarian carcinoma, as is evident from Fig. 2, has accomplished an SN of 93% and SP of 97%. Silver nanoparticle (AgNPs) of 100 nm breadth was utilized for the improvement procedure for SERS for early diagnosis of this disease and has been able to accomplish an SN of 80% and SP of 94%. To separate among healthy and ovarian carcinoma cases, the strategy was surveyed in various degrees of CA125 to preclude that the identification precision was because of CA125 level, and the two strategies gave agreeable outcomes. The two methods distinguished five spectral biological markers, which are promising as a multi-panel biological marker in diagnosing ovarian carcinoma, as reported by Paraskevaidi M. et al. [44] These investigations feature the promising part of Raman spectroscopy in the biological marker studies and its selection in multi-biological marker studies will be an imminent achievement in the field of biological marker research.

Alternative methodologies

As explained by Elias KM et al. [11], recent researches showed that autoantibodies, microRNAs, and circulating tumor DNA in the blood or liquid from the cervix, uterus, or the fallopian tube could increase the possibility of early diagnosis of ovarian carcinoma besides disclosure of CA125 biological marker.

Antibody array

As explained by Das PM et al. [9], antibody array is an emerging innovation for direct profiling of protein expression. The most utilized antibody array is the sandwich ELISA and has been utilized in the diagnosis of leukemias, breast, urinary bladder, prostate, and colon carcinomas. Sanchez-Carbayo et al. [71] could segregate bladder carcinoma victim specimens from the control specimens with an accuracy of 93.7% in their investigation. Currently, no antibody arrays have been reported for ovarian carcinoma detection but they have achieved fair success in the detection

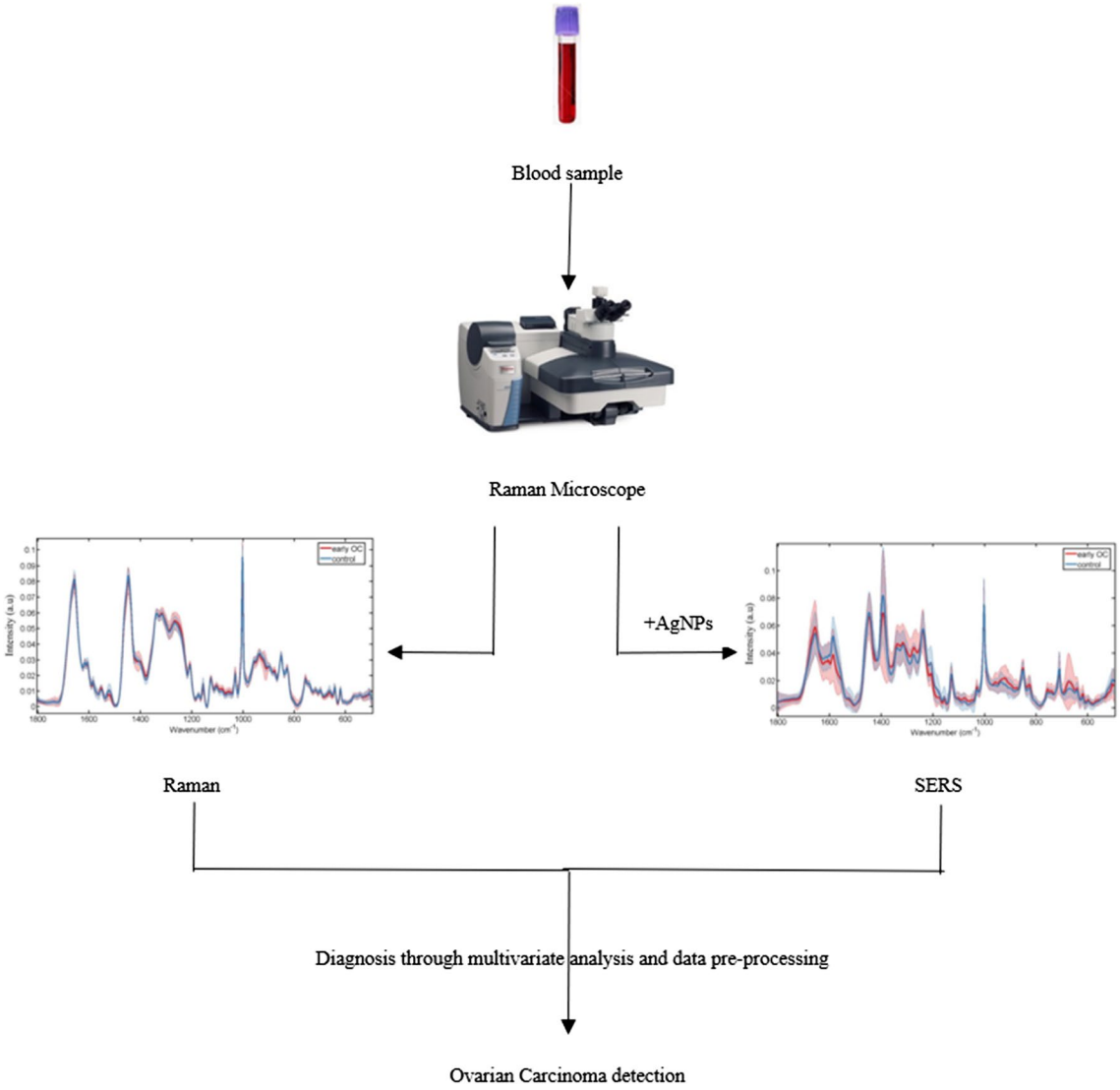


Fig. 2 Identification of ovarian carcinoma using raman spectroscopy

of other tumors. A significant disadvantage of this is the lack of good quality antibodies and their expense.

Imaging

Earlier use of transabdominal ultrasonography has largely been replaced by TVS, which provides information regarding ovarian volume, morphology, echogenicity, presence of abnormal lesions, blood flow, and any other pelvic masses. Its SN and SP can be increased with the help of morphological indices that incorporate cyst wall structure, thickness, and complexity of septae, presence of internal papillae or solid structures, in agreement with Das PM et al. [9]. In a study by Xiang et al. [56], use of 3-D microbubble contrast with TVS increased the SP and SN of TVS to 100% and 98% while differentiating benign from malignant ovarian masses. However, the TVS does not overcome the challenge of imaging the fallopian tubes or detecting small ovarian lesions and the procedure is quite expensive.

Another technique is power Doppler which can be also used to evaluate low-velocity blood flow to differentiate benign from ovarian malignancies. 3D power Doppler, an improved version of 2D power Doppler, provides imaging and measurement of blood flow even in solid areas and excrescences of complex cysts and significantly improved SP. With the use of microbubble contrast agents, it provides us information about neovascularity in some ovarian tissues. However, it cannot detect small ovarian carcinomas and its application is also limited due to the availability of instruments and the required expertise for interpretation of the results, as stated by Das PM et al. [9]

As specified by Adolphi NL et al. [1], magnetic resonance imaging (MRI) and magnetic relaxometry (MRX) can be used to differentiate and locate magnetic nanoparticles without using ionizing radiation. Nebgen DR et al. [41] mentioned that superconducting quantum interference device (SQUID) technology detects delay in magnetic relaxation when nanoparticles attached to a monoclonal antibody specific to the ovarian carcinoma cells interact with each other, differentiating those cells from the normal cells. This technique has been used for the detection of other carcinomas, such as breast carcinoma and leukemia, and holds great potential for the detection of early ovarian carcinoma.

Light-induced endogenous fluorescence (autofluorescence) has proved its utility in the diagnosis of precancerous lesions in different organ systems, such as the cervix. Autofluorescent imaging of surgically removed fallopian tubes shown a SN of 73%, a SP of 83%, and a PPV of 57% for ovarian or serous tubal intraepithelial carcinomas but does require improvement to be used as the main screening strategy for diagnosis of ovarian carcinoma, says McAlpine JN et al. [31]

Autoantibodies

The limited use of protein biomarkers as screening standard for detecting ovarian carcinoma in its early stages can be attributed to the small size of starting carcinomas and/or low levels of the biomarkers. However, autoantibodies were found to be released even in early phases of the carcinoma which show its possible utility in early diagnosis of ovarian carcinoma, as stated by Macdonald IK et al. [29]. Ahmed AA et al. [2] found that TP53, a tumor-suppressor gene (TSG), was mutated in almost all cases of high-grade ovarian carcinoma. TP53 autoantibody level was more than 20% greater in patients before treatment in comparison to their level after treatment, in both early- and advanced-stage ovarian carcinoma. Fortner RT et al. [14] showed that other autoantibodies, such as anti-homeobox gene A7 (HOXA7) [14] and anti-interleukin 8 (IL-8) [23], are found at elevated levels in case of some forms of carcinoma and thus can also be a possible candidate for early diagnosis of ovarian carcinoma. National Cancer Institute Early Detection Research Network (NCIEDRN) is investigating different autoantibody panels as a supplement to CA125 and to improve the SN of two-stage methodologies in which TVS is done only if the biomarkers levels elevate and give positive results, as reported by Nebgen DR et al. [41]

MicroRNAs (miRNAs)

As stated in Nakamura K et al. [39], miRNAs are short (~22 nucleotides) ss (single-stranded) non-coding RNAs that circulate either in association with proteins (e.g., Argonaute 2) or packed in exosomes, micro-vesicles, or apoptotic bodies and are involved in post-transcriptional gene expression. The miRNAs circulating at any instant depends on the physiological conditions of the person and thus, they pose as a potential candidate for carcinoma biomarkers. As biomarkers, several miRNAs demonstrate good results when used in the detection of EOC (Epithelial Ovarian Carcinoma). Based on Zheng H et al. [61], miR-205 and let-7f combination demonstrated excellent accuracy [AUC = 0.831 (95% CI (Confidence Interval), 0.772–0.880] for EOC with SN of 62.4% and SP of 92.9%, especially in patients with stage I disease. Yokoi A et al. [57] found that early-stage EOC could be differentiated from benign tumors with SN of 86% and SP 83% with a panel of eight miRNAs. miRNA has a great potential as a biomarker to be used in detection of this deadly disease but these are still unexplored in comparison to CA125, HE4, and TVS in terms of prospective trials.

Circulating tumor DNA (ctDNA)

According to Patel KM et al. [45], circulating tumor DNA (ctDNA) is the term used for the free DNA containing the mutations that are responsible for formation of tumors found in the circulation. Analysis of it provides us with a potential non-invasive procedure for the detection and monitoring of tumors, indirectly carcinoma. In a study, stated in Nebgen DR et al. [41], they were able to detect mutations at allelic frequencies of 2% with more than 97% SN and SP by detecting ctDNAs present in the sample with the help of targeted deep sequencing. It has advantages over others in being non-invasive technique and ability to detect carcinomas on which to do biopsies is a difficult task while detecting mutations present in even non-carcinoma patients leads to increase in number of false-positives and inability to detect the carcinoma in its early stages can be stated as its major drawbacks.

DNA methylation

It is an established fact, in accordance with Esteller M [12], that in most forms of carcinoma, DNA promoter hypermethylation of TSG occurs which represses TSG and results in the formation of the tumors. This DNA promoter generally occurs in the early phases of the carcinogenesis which make it a potential marker for the diagnosis of ovarian carcinoma in its early stages. In a study by Widschwendter M et al. [55], DNA methylation markers were used to detect the cases of the ovarian carcinoma with the accuracy of over 50% from the pool of samples which were acquired up to 2 years prior to their work. But there is still a need for more prospective studies to explore them as plausible markers and to use them as main screening technique for detection of the disease [41].

Other techniques

Microvesicle proteomics of utero-tubal lavage (UtL) fluid biopsies have an upper hand over serum biological markers giving high SN and SP in recognizing early-phase ovarian carcinoma. Lineage markers, for example, CA125 exhibited no huge distinction among patients and controls in microvesicle proteomics. Notwithstanding, a 9-protein classifier introduced 70% SN and 76.2% SP for the determination of high-class ovarian carcinoma. The utilization of MS (Mass Spectrometry)-based targeted assays in the research, based on studies by Barnabas GD et al. [3], led to an increase in accuracy and multiplexing capabilities in MS-based clinical tests in comparison to the normally utilized antibody test like ELISA.

Paraneoplastic signs are uncommon heterogeneous dysfunctions conducted with the discharge of tumor hormones or autoimmune system reactions brought about by tumor cells against onconeural antigens of the sensory system or the hidden tumor, according to Pelosof LC et al. [46] A few confirmations by Chatterjee M et al. [5] have stated the presence of various onconeural antibodies in ovarian carcinoma. The early stage of paraneoplastic side effects before the clinical indication of carcinoma increases its importance and the utilization of paraneoplastic antigens as biological markers for early diagnosis of ovarian carcinoma continue to be investigated.

Additionally, in agreement with Muinao T et al. [38], to achieve the highest SN and SP of the early diagnosis of ovarian carcinoma from an expanding number of promising biological markers, commitment from various fields of science, computation, maths, or medication is expected to build up a model to improve the apprehension performance of multi-biological markers in the clinical settings.

Proposed work

The proposed methodology through this paper is to suggest a method that will help in the detection of ovarian carcinoma in the early stages using an efficient combination of biomarkers and/or other methodologies based on the already existing studies. We have collected data on the different biomarkers currently present in the industry and on that basis identified which one will turn out to be more useful than the other. We have done comparative research on the SN, SP, and accuracy of multiple biomarkers. The advantage of this paper is that it has all the data in one place which makes it easier to identify the correct biomarker for use.

The portrayal of single biomarkers and multi-panel biomarkers regarding the healthy controls is dependent on SP and SN. In the pie charts above (Fig. 3), we have compared the average SP, SN, and accuracy values of single biomarker and multi-biomarker systems shown in Tables 1, 2, respectively, to compare them as a whole. As shown in Fig. 4, multi-panel biological markers posed as a better candidate, most of the time, for early diagnosis of ovarian carcinoma and with the cautious choice of calculations and logical methodology, they can be used as a main screening procedure to achieve the gold standards in diagnosis and also make the process affordable.

We should also keep in mind the mental trauma the person will go through in false-positive cases. The medications that the patient will take could lead to serious health issues. In the case of false-negatives, it diminishes our purpose of having a technique which can detect ovarian carcinoma in its early stages to increase the possibility of curing it to the maximum level or at least increasing the life expectancy of the patients by some years.

Literature review

There exists an urgent need to build up a screening strategy that can identify stage I and II of the ovarian carcinoma with high SN, SP and accuracy for ovarian carcinoma in a bigger populace of inmates with the disease and should also be cost effective. An extensive group of high-risk carcinoma patients is required to correctly identify the best biological marker panel for its treatment which will also help us in improving the SN, SP, and precise PPV.

Ovarian carcinoma growth does not show explicit indications that permit its recognizable evidence during early phases. As reported by Visintin I. et al. [54], an analysis is made after progression to later stages, so, all in all, the distribution of carcinoma restricts efficient treatment. Hence, the improvement of sensitive and explicit strategies for early diagnosis has been a need as a method for improving the treatment of this illness. The complication in the scientific demonstration, initiation of carcinoma, and gene articulation analysis had combined to considerable trouble in comprehension and analysis of the ailment.

A satisfactory screening test for the early diagnosis of ovarian carcinoma ought to extraordinarily boost the long-term endurance rate of the patients. Serum markers are helpful apparatuses for the analysis and development of patients with various types of carcinoma growth. Various proteins have been recognized as explicit markers for ovarian carcinoma either in blood or urine.

Examination and therapy are generally accomplished utilizing surgery and chemotherapy. However, the result relies to a great extent upon how the disease has progressed into various successive stages. Following the technological developments, genomic data including gene expressions, proteomic and metabolic profiles have been obtained. Nevertheless, as the information acquired is perplexing and

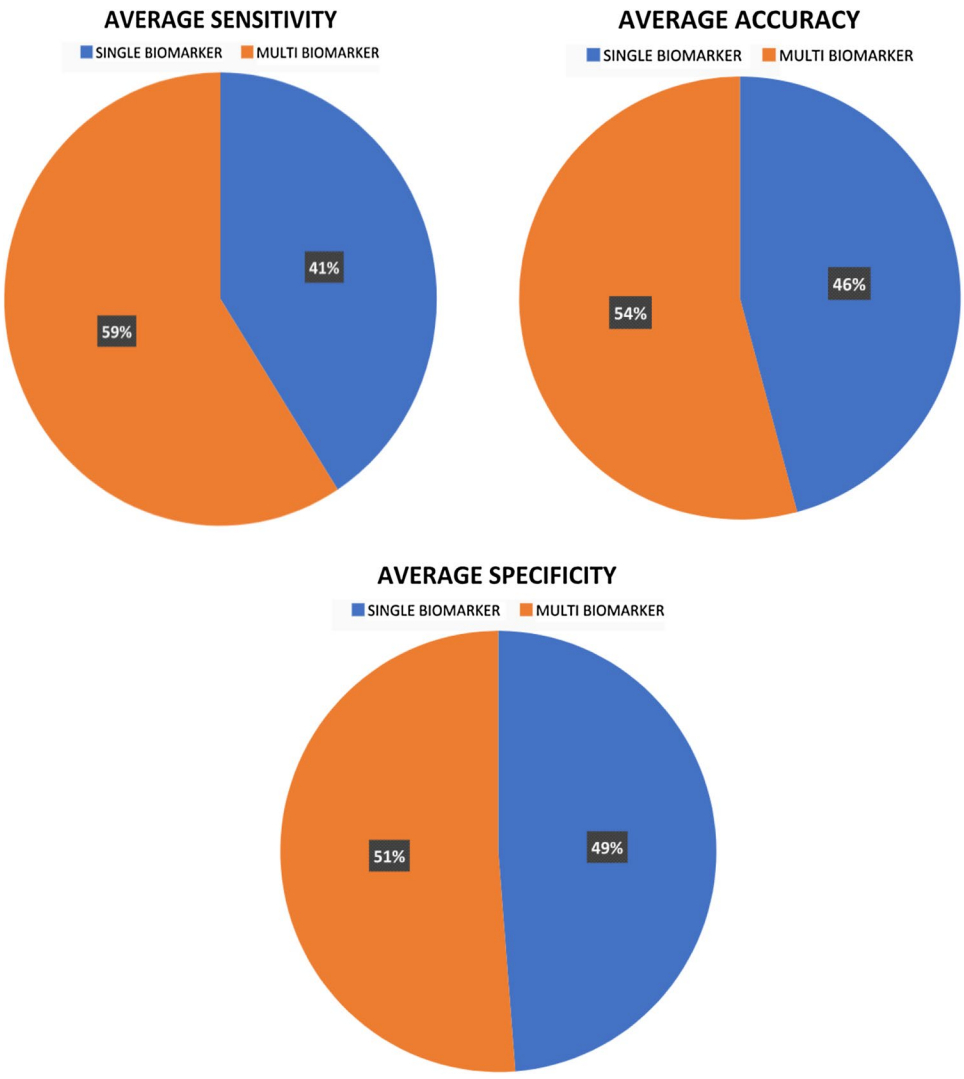


Fig. 3 Average specificity, sensitivity, and accuracy of single and multi-panel biological markers

high-dimensional, there ought to be computational ways to deal with the examination of such post-genome informational collections that help computational diagnosis of biomarkers and further comprehend the mechanism of the ailment, as stated by Sarojini et al. [49]

Currently, there are no screening procedures suggested for the initial diagnosis of ovarian carcinoma in females. Cancer antigen-125 (CA-125) is the highest-quality degree cyst marker found at strangely elevated levels in the plasma of numerous ladies with ovarian carcinoma. Notwithstanding, numerous non-carcinomatous conditions show significant levels of CA-125 and a few ladies have ordinary CA-125 level in the early phase of ovarian carcinoma, recommending CA-125 biomarker is not explicit enough for the screening of early-phase ovarian carcinoma [38].

CA-125 distinguishes a gathering of cell surface glycoproteins of dubious biological importance. Even though CA-125 is raised in 82% of ladies with advanced ovarian carcinoma, it has restricted clinical application for the recognition of early-phase illness, showing a PPV of $< 10\%$. In line with Das PM et al. [9], even the expansion of ultrasound screening to CA-125 estimation improves the PPV to just $\approx 20\%$.

All individual ovarian carcinomas growth-related serum protein biomarkers need adequate SN or SP for the identification of early-phase carcinoma. Additionally, no single protein has given satisfactory SN and SP. As appeared in a study, every marker alone ineffectively separates among carcinoma and healthy people, though the blend of these markers could fundamentally increment both the SN and SP of the test.

CA-125 has an SN of just half to 60% for early-phase disease in postmenopausal ladies when SP is set at 99%. Neither CA125 nor TVS has the required SN or SP for early diagnosis when used alone. Transvaginal sonography (TVS), CT scan, MRI, and power Doppler gives under 90% SN for early ovarian tumor growth, and their cost and generally high false-positive rates prevent yearly screening.

Table 2 SP and SN of a panel of biomarkers for the early diagnosis of ovarian carcinoma from numerous papers

Biomarker	Specificity	Sensitivity	Methods performed	Accuracy	Reference
HE4 + CA125 (HECA)	95	45.9	MESOMARK™ Assay	–	[36]
Prostasin + CA125 (ProCA)	94	92	Enzyme-linked immunosorbent assay	–	[32]
KLK6 + CA125 (KLCA)	90	42	Immunoassay	–	[10]
Apolipoprotein A1 (APOA1) and transthyretin (APOTR)	96.5	52.4	SELDI-TOF–MS Protein Chip array chromatographic assay	–	[34]
Transthyretin with CA-125, ApoA1, and transferrin (TCAT)	98	96	Chemiluminescence and immunoturbidimetry technology	–	[42]
B7-H4 + CA125 (BHCA)	97	65	ELISA	–	[50]
CA-125, HE4, CEA, and VCAM-1 (CHCV)	98	86	Bead-based xMAP immunoassays	–	[58]
CA125, HE4, E-CAD, and IL-6 (CHEI)	95.7	84.2	Simple Plex immunoassay	–	[16]
Osteopontin, leptin, prolactin, and insulin-like growth factor-II (IGF-II) (OLPI)	95	95	ELISA assays and Microarray analysis	–	[37]
CA-125, Prolactin, Osteopontin, IGF-II, macrophage inhibitory factor (MIF), and leptin (CPOI)	99.4	95.3	ELISA	98.7%	[20]
CA 125, Eotaxin, CA 19–9, cVCAM, G-CSF, IL-2R, MIF and EGFR (CECV)	98.7	98.2	LapMAP™ technology	–	[26]
Transthyretin, connective tissue-activating protein III, ApoA1 and CA-125 (TCAC)	98	84	Immunoassay, SELDI-TOF–MS, protein chip arrays	–	[6]
CA-125, transferrin, TTR + ApoA1 (CTTA)	92	89	Chemiluminescence technology and immunoturbimetry technology	–	[52]
VEGF + CA-125 + HE4 (VCH)	82	84	ELISA and chemiluminescent microparticle immunoassay	–	[23]
ApoA1 + CA-125 + TTR (ACT)	95	93.9	multiplex liquid assay system	–	[21]
β2- microglobulin (β2-M), ApoA1 and CA-125 (MAC)	98%	94	Surface Plasmon Resonance spectroscopy and Multiplexed fluorescence spectroscopy	–	[43]
CA-125, CA 72–4, CA 15–3, and M-CSF (CCCM)	98	68	Radioimmunochemicalassay	–	[51]
CA-125, a cleavage fragment of inter–alpha-trypsin inhibitor heavy chain H4, truncated form of transthyretin and apolipoprotein A1 (CHTA)	97	74	Immunoassay	–	[60]
RECAF + CA125 (RECA)	97	86	Chemiluminescence assay	–	[53]

Utilizing TVS as a second-line test, past CA-125-based screening investigations demonstrate that a first-line SP of 98% for a yearly test could guarantee required SP (> 99.6%) and PPV (> 10%) and would decrease the number of ultrasound assessments performed every year to a practical degree of 2%, claims Jacob IJ et al. [18] and Muinao T. et al. [38]

Utilizing CA125 as the main biomarker for finding will pass up a great opportunity for those who do not display this antigen. Serum biological markers apart from CA125 have not been, as of now, utilized as a discovery device for early-phase carcinoma because of their decreased SN or SP, and accordingly, the search continues on finding biological markers with increased SN and SP to forecast the happening of metastasis before it occurs in the patients. Various possible biological markers for early diagnosis of ovarian carcinoma, which are dependent on SN and SP, have been summed up in Tables 1, 2.

None of the single biomarkers for the discovery of early-phase ovarian carcinoma has accomplished the necessary SP and SN. A few investigations dependent on multi-biomarker procedures have been accounted to improve the SN over the single biological marker at comparable SP in analyzing early-phase ovarian carcinoma. CA125 in blend with a few other serum tumors biological markers have been tried. Although revealing which mixture of biomarkers will be the most effective is slightly impossible due to the non-availability of adequate data. Multivariate investigation including tumor biomarkers alongside ultrasound and a proper algorithm will be an additional advantage.

The blend of RECAF and CA125 serum gives the SN and SP important to screen for ovarian carcinoma, and specifically, to recognize early phases, where cure rates are the most elevated. Tcherkassova et al. [53] in his study mentioned that while ovarian diseases can be distinguished with a mix of CA125 and RECAF serum, not all patients recognized with

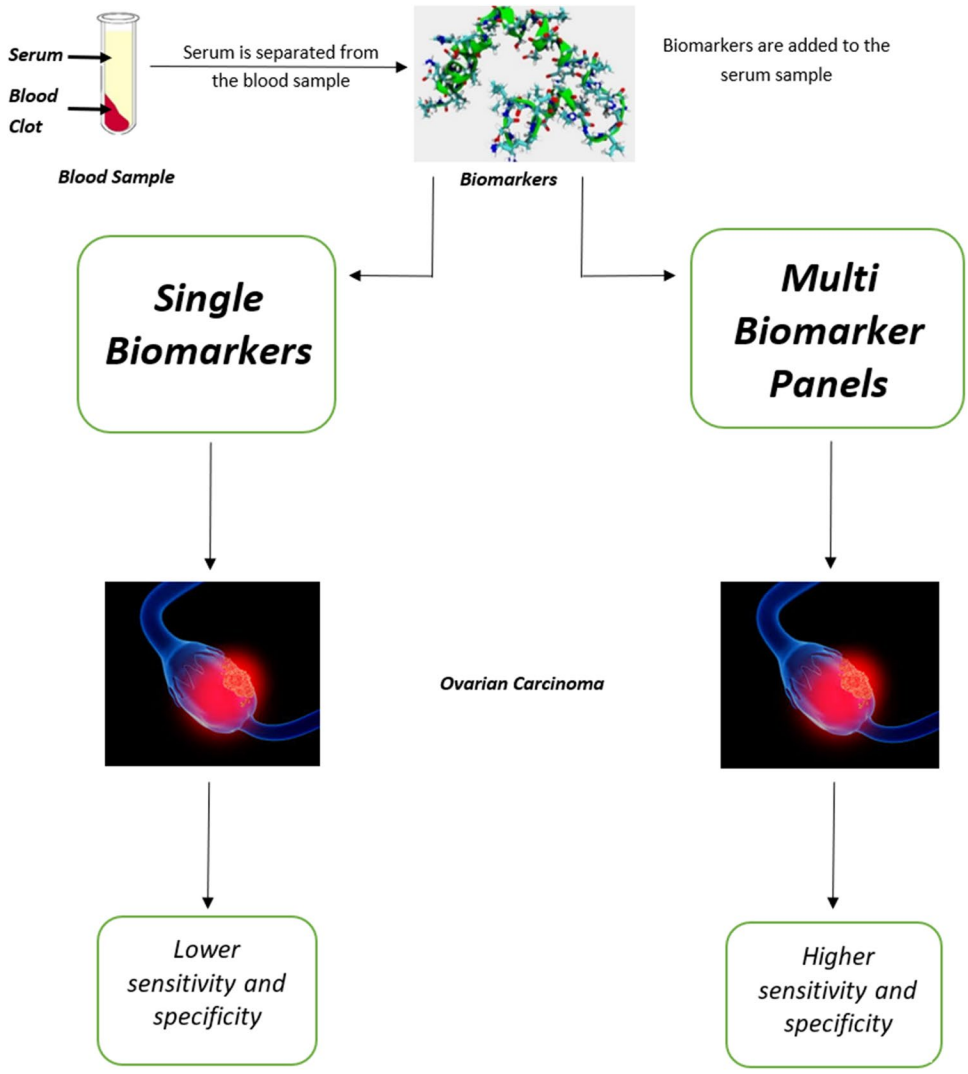


Fig. 4 Proposed flow chart

this technique will have ovarian carcinoma since RECAF is raised in numerous other carcinoma types. According to the reports of Jing et al. [19] and Simon et al. [50], the mix of CA125 with B7-H4 expanded the SN of early diagnosis from 52% for CA125 alone to 65% at 97% SP.

As is evident from Table 2 and Fig. 5, CA-125 with CA 19–9, EGFR, G-CSF, Eotaxin, IL-2R, cVCAM, MIF improved the SN with 98.2% and SP of 98.7% in early-phase diagnosis of the disease [26]. Kim K et al. [20] worked on Osteopontin, leptin, prolactin, IGF-II, MIF, and CA-125, which is also a great blend of biomarkers since it offers an accuracy of 98.7% with SN and SP at 95.3% and 99.4%, respectively.

Conclusion and future prospects

The need for an efficient biomarker is increasing day by day because the lifestyle of people has changed. Nowadays, people are more attracted to junk food and barely exercise which results in hormonal imbalance which in turn increases the chances of ovarian carcinoma. With a consistent advancement in biomarker diagnosis utilizing present-day procedures and complex bioinformatics diagnostic apparatuses, a dependable and fruitful ovarian carcinoma screening methodology will soon be accessible. A trustworthy and approved biological marker in context to SP, SN, and accuracy, is inaccessible at the moment. A definitive objective must be to utilize different serum biomarkers and/or TVS or other imaging procedures to recognize this preclinical illness in an effective manner, i.e., the procedure should not only have the required aspects like high SP, SN and accuracy but should also be economical so that it is accessible to everyone. Ideal biological markers with high accuracy that enhance the early diagnosis are urgently needed. The drawback of all the existing papers is that the data are very scattered. To find about the different methodologies that have potential application in early detection of ovarian carcinoma,

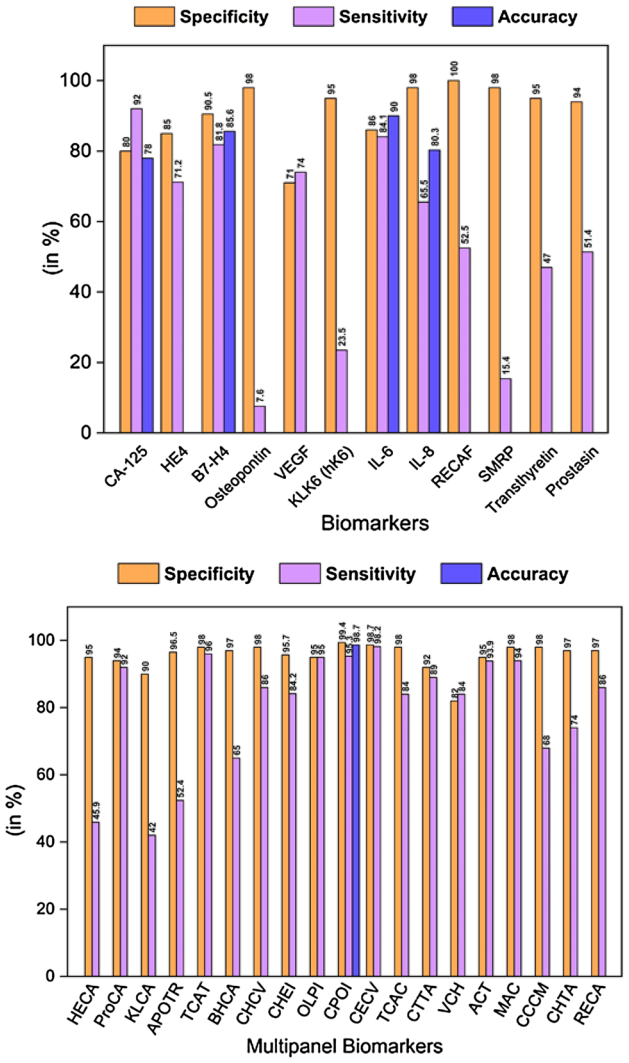


Fig. 5 Graphical representation of reported SP, SN and accuracy of single- and multi-panel biological markers shown in the Tables 1, 2, respectively

especially biomarkers, several papers have to be searched and it becomes a hectic and time-consuming task to read all of them. This investigation gives the information about advantages and disadvantages of a particular system that have been employed or has the potential to be employed individually or in combination to provide us with an accurate and economical method to deal with diagnosing early phases of ovarian carcinoma and along these lines, diminish disease-related mortality. We trust that these potential biomarkers will end up being successful biomarkers after clinical preliminaries.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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. Adolphi NL, Butler KS, Lovato DM, et al. Imaging of Her2-targeted magnetic nanoparticles for breast cancer detection: comparison of SQUID-detected magnetic relaxometry and MRI. Contrast Med Mol Imaging. 2012;7(3):308–19. https://doi.org/10.1002/cmmi.499.
2. Ahmed AA, Etemadmoghadam D, Temple J, et al. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. J Pathol. 2010;221(1):49–56. https://doi.org/10.1002/path.2696.
3. Barnabas GD, Bahar-Shany K, Sapoznik S, et al. Microvesicle proteomic profiling of uterine liquid biopsy for ovarian cancer early detection. Mol Cell Proteomics. 2019;18(5):865–75. https://doi.org/10.1074/mcp.ra119.001362.
4. Chappelear A. About ovarian cancer: johns hopkins ovarian cancer center. In: Hopkinsmedicine.org. 2017. https://www.hopkinsmedicine.org/kimmel_cancer_center/cancers_we_treat/ovarian/about/ (Accessed 15 Dec 2020)
5. Chatterjee M, Hurley LC, Tainsky MA. Paraneoplastic antigens as biomarkers for early diagnosis of ovarian cancer. Gynecol Oncol Rep. 2017;21:37–44. https://doi.org/10.1016/j.gore.2017.06.006.
6. Clarke CH, Yip C, Badgwell D, et al. Proteomic biomarkers apolipoprotein A1, truncated transthyretin and connective tissue activating protein III enhance the sensitivity of CA125 for detecting early stage epithelial ovarian cancer. Gynecol Oncol. 2011;122(3):548–53. https://doi.org/10.1016/j.ygyno.2011.06.002.
7. Cooper BC, Ritchie JM, Broghammer CL, et al. Preoperative serum vascular endothelial growth factor levels: significance in ovarian cancer. Clin Cancer Res. 2002;8(10):3193–7.
8. Cramer DW, Bast RC, Berg CD, et al. Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. Cancer Prev Res. 2011;4(3):365–74. https://doi.org/10.1158/1940-6207.capr-10-0195.
9. Das PM, Bast RC Jr. Early detection of ovarian cancer. Biomark Med. 2008;2(3):291–303. https://doi.org/10.2217/17520363.2.3.291.
10. Diamandis EP, Scorilas A, Fracchioli S, et al. Human Kallikrein 6 (hK6): a new potential serum biomarker for diagnosis and prognosis of ovarian carcinoma. J Clin Oncol. 2003;21(6):1035–43. https://doi.org/10.1200/jco.2003.02.022.
11. Elias KM, Guo J, Bast RC. Early detection of ovarian cancer. Hematol Oncol Clin North Am. 2018;32(6):903–14. https://doi.org/10.1016/j.hoc.2018.07.003.
12. Esteller M. Promoter hypermethylation and brca1 inactivation in sporadic breast and ovarian tumors. J Nat Cancer Inst. 2000;92(7):564–9. https://doi.org/10.1093/jnci/92.7.564.
13. Fazio B, Andrea C, Foti A, et al. SERS detection of biomolecules at physiological pH via aggregation of gold nanorods mediated by optical forces and plasmonic heating. Scientif Rep. 2016;6:1–24. https://doi.org/10.1038/srep26952.
14. Fortner RT, Damms-Machado A, Kaaks R. Systematic review: tumor-associated antigen autoantibodies and ovarian cancer early detection. Gynecol Oncol. 2017;147(2):465–80. https://doi.org/10.1016/j.ygyno.2017.07.13.

15. Gorelik E. Multiplexed immunobead-based cytokine profiling for early detection of ovarian cancer. Cancer Epidemiol Biomark Prev. 2005;14(4):981–7. https://doi.org/10.1158/1055-9965.epi-04-0404.
16. Han C, Bellone S, Siegel ER, et al. A novel multiple biomarker panel for the early detection of high-grade serous ovarian carcinoma. Gynecol Oncol. 2018;149(3):585–91. https://doi.org/10.1016/j.ygyno.2018.03.050.
17. Havrilesky LJ, Whitehead CM, Rubatt JM, et al. Evaluation of biomarker panels for early stage ovarian cancer detection and monitoring for disease recurrence. Gynecol Oncol. 2008;110(3):374–82. https://doi.org/10.1016/j.ygyno.2008.04.041.
18. Jacobs IJ, Menon U. Progress and challenges in screening for early detection of ovarian cancer. Mol Cell Proteomics. 2004;3(4):355–66. https://doi.org/10.1074/mcp.r400006-mcp200.
19. Jing X, Wei F, Li J, et al. Diagnostic value of soluble B7–H4 and carcinoembryonic antigen in distinguishing malignant from benign pleural effusion. Clin Respir J. 2017;12(3):986–90. https://doi.org/10.1111/crj.12615.
20. Kim K, Visintin I, Alvero AB, Mor G. Development and validation of a protein-based signature for the detection of ovarian cancer. Clin Lab Med. 2009;29(1):47–55. https://doi.org/10.1016/j.cll.2009.02.001.
21. Kim Y-W, Bae SM, Lim H, Kim YJ, Ahn WS. Development of multiplexed bead-based immunoassays for the detection of early stage ovarian cancer using a combination of serum biomarkers. PLoS ONE. 2012;7(9):e44960. https://doi.org/10.1371/journal.pone.0044960.
22. Kong K, Kendall C, Stone N, Notingher I. Raman spectroscopy for medical diagnostics — From in-vitro biofluid assays to in-vivo cancer detection. Adv Drug Deliv Rev. 2015;89:121–34. https://doi.org/10.1016/j.addr.2015.03.009.
23. Lawicki S, Będkowska GE, Gacuta-Szumarska E, Szmitkowski M. The plasma concentration of VEGF, HE4 and CA125 as a new biomarkers panel in different stages and sub-types of epithelial ovarian tumors. J Ovarian Res. 2013;6(1):45. https://doi.org/10.1186/1757-2215-6-45.
24. Lockwood S. Ovarian Cancer. HealthyWomen. Published September 16, 2009. https://www.healthywomen.org/condition/ovarian-cancer (Accessed 3 Jan 2021)
25. Lokshin AE, Winans M, Landsittel D, et al. Circulating IL-8 and anti-IL-8 autoantibody in patients with ovarian cancer. Gynecol Oncol. 2006;102(2):244–51. https://doi.org/10.1016/j.ygyno.2005.12.011.
26. Lokshin AE, Yurkovetsky Z, Nolen B, et al. Multimarker assay for early diagnosis of ovarian cancer. Cancer Res. 2006;66:155–155.
27. Lycke M, Kristjansdottir B, Sundfeldt K. A multicenter clinical trial validating the performance of HE4, CA125, risk of ovarian malignancy algorithm and risk of malignancy index. Gynecol Oncol. 2018;151(1):159–65. https://doi.org/10.1016/j.ygyno.2018.08.025.
28. Lyng FM, Traynor D, Ramos IRM, Bonnier F, Byrne HJ. Raman spectroscopy for screening and diagnosis of cervical cancer. Anal Bioanal Chem. 2015;407(27):8279–89. https://doi.org/10.1007/s00216-015-8946-1.
29. Macdonald IK, Parsy-Kowalska CB, Chapman CJ. Autoantibodies: opportunities for early cancer detection. Trend Cancer. 2017;3(3):198–213. https://doi.org/10.1016/j.trecan.2017.02.003.
30. Mariño IP, Blyuss O, Ryan A, et al. Change-point of multiple biomarkers in women with ovarian cancer. Biomed Signal Process Control. 2017;33:169–77. https://doi.org/10.1016/j.bspc.2016.11.015.
31. McAlpine JN, El Hallani S, Lam SF, et al. Autofluorescence imaging can identify preinvasive or clinically occult lesions in fallopian tube epithelium: a promising step towards screening and early detection. Gynecol Oncol. 2011;120(3):385–92. https://doi.org/10.1016/j.ygyno.2010.12.333.
32. Mok SC, Chao J, Skates S, et al. Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. JNCI J NatCancer Instit. 2001;93(19):1458–64. https://doi.org/10.1093/jnci/93.19.1458.
33. Momenimovahed Z, Tiznobaik A, Taheri S, Salehiniya H. Ovarian cancer in the world: epidemiology and risk factors. Internat J Women's Health. 2019;11:287–99. https://doi.org/10.2147/ijwh.s197604.
34. Moore LE, Fung ET, McGuire M, et al. Evaluation of apolipoprotein A1 and posttranslationally modified forms of transthyretin as biomarkers for ovarian cancer detection in an independent study population. Cancer Epidemiol Biomark Prev. 2006;15(9):1641–6. https://doi.org/10.1158/1055-9965.epi-05-0980.
35. Moore RG, Blackman A, Miller MC, et al. Multiple biomarker algorithms to predict epithelial ovarian cancer in women with a pelvic mass: Can additional makers improve performance? Gynecol Oncol. 2019;154(1):150–5. https://doi.org/10.1016/j.ygyno.2019.04.006.
36. Moore RG, Brown AK, Miller MC, et al. The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. Gynecol Oncol. 2008;108(2):402–8. https://doi.org/10.1016/j.ygyno.2007.10.017.
37. Mor G, Visintin I, Lai Y, et al. Serum protein markers for early detection of ovarian cancer. Proc Nat Acad Sci. 2005;102(21):7677–82. https://doi.org/10.1073/pnas.0502178102.
38. Muinao T, Deka Boruah HP, Pal M. Multi-biomarker panel signature as the key to diagnosis of ovarian cancer. Heliyon. 2019;5(12):e02826. https://doi.org/10.1016/j.heliyon.2019.e02826.
39. Nakamura K, Sawada K, Yoshimura A, Kinose Y, Nakatsuka E, Kimura T. Clinical relevance of circulating cell-free microRNAs in ovarian cancer. Mol Cancer. 2016;15(1):12–34. https://doi.org/10.1186/s12943-016-0536-0.
40. National Ovarian Cancer Coalition - NOCC Staff. Types and Stages - National Ovarian Cancer Coalition. Ovarian.org. Published 2021. https://www.ovarian.org/about-ovarian-cancer/what-is-ovarian-cancer/types-a-stages. (Accessed 22 Jan 2021)
41. Nebgen DR, Lu KH, Bast RC. Novel approaches to ovarian cancer screening. Curr Oncol Rep. 2019;21(8):12–53. https://doi.org/10.1007/s11912-019-0816-0.
42. Nosov V, Su F, Amneus M, et al. Validation of serum biomarkers for detection of early-stage ovarian cancer. Am J Obstet Gynecol. 2009;200(6):6391–5. https://doi.org/10.1016/j.ajog.2008.12.042.
43. Pal MK, Rashid M, Bisht M. Multiplexed magnetic nanoparticle-antibody conjugates (MNPs-ABS) based prognostic detection of ovarian cancer biomarkers, CA-125, β-2M and ApoA1 using fluorescence spectroscopy with comparison of surface plasmon resonance (SPR) analysis. Biosens Bioelectron. 2015;73:146–52. https://doi.org/10.1016/j.bios.2015.05.051.
44. Paraskevaidi M, Ashton KM, Stringfellow HF, et al. Raman spectroscopic techniques to detect ovarian cancer biomarkers in blood plasma. Talanta. 2018;189:281–8. https://doi.org/10.1016/j.talanta.2018.06.084.
45. Patel KM, Tsui DWY. The translational potential of circulating tumour DNA in oncology. Clin Biochem. 2015;48(15):957–61. https://doi.org/10.1016/j.clinbiochem.2015.04.005.
46. Pelosof LC, Gerber DE. Paraneoplastic syndromes: an approach to diagnosis and treatment. Mayo Clin Proc. 2010;85(9):838–54. https://doi.org/10.4065/mcp.2010.0099.
47. Peng D, Xu T, Mason TJ, Wu W. A study of ovarian cancer biomarker amplification using ultrasound for early stage detection. Ultrasonics. 2014;54(2):451–4. https://doi.org/10.1016/j.ultras.2013.05.014.

48. Pinsky PF, Zhu CS. Building multi-marker algorithms for disease prediction–-the role of correlations among markers. Biomarker Insights. 2011;6:BMIS7513. https://doi.org/10.4137/bmi.s7513.
49. Sarojini S, Tamir A, Lim H, et al. Early detection biomarkers for ovarian cancer. J Oncol. 2012;22:1–15. https://doi.org/10.1155/2012/709049.
50. Simon I, Zhuo S, Corral L, et al. B7–H4 is a novel membrane-bound protein and a candidate serum and tissue biomarker for ovarian cancer. Can Res. 2006;66(3):1570–5. https://doi.org/10.1158/0008-5472.can-04-3550.
51. Skates SJ, Horick N, Yu Y, et al. Preoperative sensitivity and specificity for early-stage ovarian cancer when combining cancer antigen CA-125II, CA 15–3, CA 72–4, and macrophage colony-stimulating factor using mixtures of multivariate normal distributions. J Clin Oncol. 2004;22(20):4059–66. https://doi.org/10.1200/jco.2004.03.091.
52. Su F, Lang J, Kumar A, et al. Validation of candidate serum ovarian cancer biomarkers for early detection. Biomark Insight. 2007;2:369–75.
53. Tcherkassova J, Abramovich C, Moro R, et al. Combination of CA125 and RECAF biomarkers for early detection of ovarian cancer. Tumor Biology. 2011;32(4):831–8. https://doi.org/10.1007/s13277-011-0186-1.
54. Visintin I, Feng Z, Longton G, et al. Diagnostic markers for early detection of ovarian cancer. Clin Cancer Res. 2008;14(4):1065–72. https://doi.org/10.1158/1078-0432.ccr-07-1569.
55. Widschwendter M, Zikan M, Wahl B, et al. The potential of circulating tumor DNA methylation analysis for the early detection and management of ovarian cancer. Genome Med. 2017;9:1–34. https://doi.org/10.1186/s13073-017-0500-7.
56. Xiang H, Huang R, Cheng J, et al. Value of three-dimensional contrast-enhanced ultrasound in the diagnosis of small adnexal masses. Ultrasound Med Biol. 2013;39(5):761–8. https://doi.org/10.1016/j.ultrasmedbio.2012.11.008.
57. Yokoi A, Yoshioka Y, Hirakawa A, et al. A combination of circulating miRNAs for the early detection of ovarian cancer. Oncotarget. 2017;8(52):89811–23. https://doi.org/10.18632/oncotarget.20688.
58. Yurkovetsky Z, Skates S, Lomakin A, et al. Development of a multimarker assay for early detection of ovarian cancer. J Clin Oncol. 2010;28(13):2159–66. https://doi.org/10.1200/jco.2008.19.2484.
59. Zhang L, Chen Y, Wang K. Comparison of CA125, HE4, and ROMA index for ovarian cancer diagnosis. Curr Probl Cancer. 2019;43(2):135–44. https://doi.org/10.1016/j.currproblcancer.2018.06.001.
60. Zhang Z, Bast RC, Yu Y, et al. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Can Res. 2004;64(16):5882–90. https://doi.org/10.1158/0008-5472.can-04-0746.
61. Zheng H, Zhang L, Zhao Y, et al. Plasma miRNAs as diagnostic and prognostic biomarkers for ovarian cancer. PLoS ONE. 2013;8(11):e77853. https://doi.org/10.1371/journal.pone.0077853.
62. Papanikolaou G, Gkouvatsos K, Pantopoulos K. Transferrin. Encycloped Signal Mole . 2018;4:5615–23. https://doi.org/10.1007/978-3-319-67199-4_101940.
63. Robbins J. Transthyretin from discovery to now. Clin Chem Lab Med. 2002;5:40–112. https://doi.org/10.1515/cclm.2002.208.
64. Wikipedia Contributors. ApoA-1 Milano. Wikipedia. Published December 30, 2020. https://en.wikipedia.org/wiki/ApoA-1_Milano#:~:text=5%20Subsequent%20Development-,Discovery,small%20village%20in%20northern%20Italy. (Accessed 28 Jan 2021)
65. Smith JB, Stashwick C, Powell DJ. B7–H4 as a potential target for immunotherapy for gynecologic cancers: a closer look. Gynecol Oncol. 2014;134(1):181–9. https://doi.org/10.1016/j.ygyno.2014.03.553.
66. Ribatti D. From the discovery of vascular endothelial growth factor to the introduction of Avastin in clinical trials - an interview with Napoleone Ferrara. Int J Dev Biol. 2011;55(4–5):383–8. https://doi.org/10.1387/ijdb.103216dr.
67. Simmons AR, Baggerly K, Bast RC. The emerging role of HE4 in the evaluation of epithelial ovarian and endometrial carcinomas. Oncology (Williston Park, NY). 2013;27(6):548–556. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4085777/. (Accessed 28 Jan 2021)
68. Wikipedia Contributors. Osteopontin. Wikipedia. Published January 4, 2021. https://en.wikipedia.org/wiki/Osteopontin. (Accessed 28 Jan 2021)
69. Scholler N, Urban N. CA125 in ovarian cancer. Biomark Med. 2007;1(4):513–23. https://doi.org/10.2217/17520363.1.4.513.
70. Wikipedia Contributors. Carcinoembryonic antigen. Wikipedia. Published January 2, 2021. https://en.wikipedia.org/wiki/Carcinoembryonic_antigen. (Accessed 28 Jan 2021)
71. Sanchez-Carbayo M, Socci ND, Lozano JJ, Haab BB, Cordon-Cardo C. Profiling bladder cancer using targeted antibody arrays. Am J Pathol. 2006;168(1):93–103. https://doi.org/10.2353/ajpath.2006.050601.

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

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/350377810

Scene Text Recognition in the Wild with Motion Deblurring Using Deep Networks

Chapter · March 2021

DOI: 10.1007/978-981-16-1103-2_9

CITATIONS
0

READS
43

5 authors, including:

Seba Susan
Delhi Technological University
91 PUBLICATIONS **573** CITATIONS

SEE PROFILE

All content following this page was uploaded by Seba Susan on 30 March 2021.

The user has requested enhancement of the downloaded file.

The final published version of this article is available at: 10.1007/978-981-16-1103-2_9 1

Scene Text recognition in the wild with motion deblurring using deep networks

Sukhad Anand[1], Seba Susan[2], Shreshtha Aggarwal[3], Shubham Aggarwal[4], and Rajat Singla[5]

[1] Delhi Technological University, Bawana Road, New Delhi
sukhad_bt2k15@dtu.ac.in
[2] Delhi Technological University, Bawana Road, New Delhi
sebasusan@dce.ac.in
[3] Delhi Technological University, Bawana Road, New Delhi
aggarwalshreshtha2309@gmail.com
[4] Delhi Technological University, Bawana Road, New Delhi
shubham.aggarwal.1996@gmail.com
[5] Delhi Technological University, Bawana Road, New Delhi
rajatsingla25@gmail.com

Abstract. In this paper, the problem of text detection and recognition in videos has been addressed. We address two major issues that make it difficult to extract information in a video captured by a moving vehicle. Video captured by a moving vehicle contains a lot of blurs caused by motion which is one of the major issues preventing accurate recognition of text. The second major issue is the orientation of the text being detected, which may not be in the same plane. We propose a novel end-to-end pipeline consisting of deep networks. Our pipeline consists of a a fully convolution network to detect text, Generative Adversarial Network to remove motion blur, a rectification network which makes use of Thin Spline Transformations and a Spatial Transform network to handle text which is not straight i.e perspective and curved, and a recognition network to recognize the text. We only deblur the region around text boxes instead of complete images. We also track the text boxes in each frame to avoid re-recognition of text in consecutive frames. This significantly improves the performance of the system, as proved by higher classification scores achieved as compared to state of the art.

Keywords: Scene understanding · Motion blur · Optical character recognition · Fully Convolution network · Text recognition · Text detection · Deblurring

1 Introduction

Text can be used to understand images and videos. There are various fields where textual information can be useful. Navigation in autonomous vehicles, automating industrial processes, understanding complex scenes are some of the various applications.

A lot of text detection and recognition methods have been proposed. These methods proposed in literature [1], [2], [3], [4], [5] provide considerable accuracy in images, but they fail to perform well on videos. Some challenges in static images include extremely complex backgrounds, irregular illumination around the text, different varieties of text. Videos introduce some other challenges like motion blur and focus change issues. Video compression also introduces some blocking occlusions. Blocking artifacts may be created by video compression. To extract temporal information from the videos, text tracking may be required that might be affected due to occlusions.

Therefore, an accurate pipeline is needed for robust text reading from videos. In the past few years, a lot of work has been done in the area of text detection and extraction from natural scenes, and scene text recognition has also attracted great interest from the industry in recent years due to its significance in a large number of applications. But natural scene videos are usually blurred due to motion, causing the loss of finer texture details, that results in low accuracy for scene text recognition and detection. The deblurring network we have used preserves these lost finer texture details in images and generates solutions that look convincing and are close to the real image. Detection of text boxes is done in natural scenes. Conventionally, text detection involves feature extraction and matching with some form of image contrast enhancement to highlight the text [6]. Text detection is a prerequisite and hence works as an important part in the procedure of extraction of textual information from backgrounds. Early approaches for text detection prove to be inefficient when dealing with perspective or curved text.

The specific contributions of this work are:

- An end-to-end system consisting of deep networks which is able to successfully detect and recognize text in real time, making it usable for automatic navigation in autonomous vehicles.
- To remove motion blur, we propose to use a Generative Adversarial Network because it achieves state-of-the-art performance and is 5 times faster than its closest competitor.
- To handle perspective text and curved text, we use an end-to-end trainable rectification network. We also handle in-plane rotated text by considering all positions 90 degree apart.
- The proposed pipeline significantly outperforms state-of-the-art methods in accuracy.
- We make sure that the pipeline runs in real time by deblurring only patches around the text boxes and not the complete image.

2 Related Work

In the past few years, there has been a lot of research in the domain of scene text recognition. Many methods have been proposed for this task [1], [2],[3], [4], [5]. Text recognition from videos along with text detection and localization has garnered huge interest in the industry. Various region based and connected

component based methods were developed before the arrival of deep learning methods. Local image regions have been used in region based methods for the detection and recognition of text. In connected components based methods, the text is divided into connected components and then detection and recognition is performed on these connected components [3], [4]. In [8], a novel image operator is presented that calculates the value of stroke width for every image pixel for the task of text detection in natural scenes.

In [9], maximally stable extreme regions are extracted as character candidates that are further grouped into text candidates using a single link clustering algorithm. For text recognition, various methods like [10] use part-based tree-structured character detection that models each type of character instead of using multi-scale sliding window approach to detect and recognize characters at the same time. It makes use of character specific structure information. Jon Almazan[11] uses an approach where both text strings and word images are embedded in a common vectorial subspace(label embedding) followed by learning of attributes and a common subspace regression. An example of document based text segmentation is [6] in which text is identified as a texture after application of adaptive contrast enhancement.

An integrated or end-to-end system is proposed by Wang et al. [2] in which two systems are built, where the first component is a two stage pipeline performing text detection followed by an OCR engine and the second is a system for generic object recognition. The methods used for text retrieval from images before deep learning methods were introduced, included extraction of low-level or mid level image features from the scene, which demanded repetitive preprocessing and post-processing steps. These methods cannot handle occlusions well, and hence with the appearance of deep learning methods having higher performance and efficiency, the shortcomings of traditional methods were overcome.

With the development of Convolution Neural Networks based methods[12],[13], we can see that better performances are achieved in text detection and recognition . Various end-to-end systems for text detection and recognition, merged with text tracking techniques, where candidate text regions are detected by using stroke-map and morphological operations. Antani et al. divided scene text extraction into four simpler tasks: detection, localization, segmentation, and recognition [14], presenting a unified framework for video text detection, tracking and recognition.

In this paper, we have proposed an end-to-end system for scene text recognition consisting of simple pipeline that predicts arbitrary oriented words or text lines followed by motion deblurring of videos and finally a recognition network that predicts a sequence of characters from the rectified image. We use the DeblurGAN architecture proposed by [15] for motion deblurring. This method has higher recall and F1 score than its competitors and hence is significantly more efficient and helps in detection from blurred images and videos. Existing methods in the field of scene text recognition do not include deblurring of motion videos, neither is speed taken into account for the end-to-end system. Our sys-

tem consists of four simpler tasks : text detection,video motion deblurring, text rectification and text recognition, and is a lot quicker than the earlier systems built for the same task and also removes blur caused in the scene due to motion. The previous work done on text recognition along with text tracking techniques have not used any temporal information of the videos, rather all the work is done on frames.

3 The proposed methodology

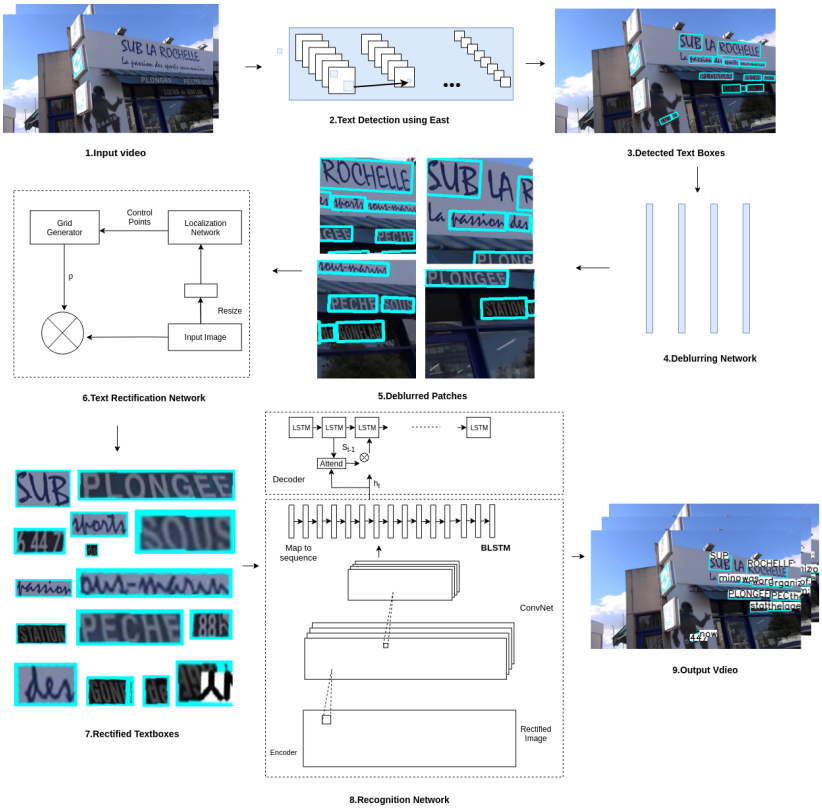


Fig. 1. Proposed end-to-end pipeline for scene text recognition from videos with motion deblurring

Our proposed architecture shown in Fig 1 consists of a network that is an end-to-end learning method for a single pipeline that detects arbitrary oriented words or text lines of quadrilateral shapes in natural scenes from a video using a single neural network followed by motion deblurring of these detected shapes.

We extract a 256x256 patch centered around the detected text boxes which is fed into the deblurring network. This not only helps to deblur the regions containing text but also provides a computational advantage as we don't have to deblur the complete image which saves a lot of time. After detecting and deblurring the text boxes in scenes, a recognition network predicts the stream of characters from a rectified image the details of which are described in next. Some challenging issues in scene text recognition include handling distorted or irregular text. So we first add a rectification network that transforms the image by rectifying the text in it and uses a flexible Thin-Plate Spline transformation to handle text irregularities. This is followed by a recognition network which uses attention mechanism to predict the text as found in the text boxes. We track the text boxes in the frames using Dlib's[29] implementation of [30]. We take a threshold of 7.0 for peak to side lobe ratio. If this ratio is below this value then we again perform recognition else we use the text recognised in previous frames.We achieve state-of-the art accuracy with processing speed of 20 FPS. We next describe each module in our end-to-end pipeline for text recognition from videos.

3.1 Textbox Detection

The first stage of the process flow is a neural network model used for text detection that generates word or text-line level predictions directly, without any redundant or slow intermediate steps, and produces the final text boxes detected. This pipeline can produce both word level or line level predictions, that can be of any geometric shape like a rotated box or a quadrangle. The model used here is a fully convolution neural network used for text detection and it outputs dense per-pixel predictions of words or text lines. This eliminates redundant or intermediate steps like text region formation, candidate proposal and word partition. The post-processing steps include only thresholding and Non Maximum Supression on the predicted geometric shapes. The detector is called EAST, an Efficient and Accurate Scene Text detection pipeline[7].

3.2 Deblurring

We then use the DeblurGAN(Generative Adversarial Network) introduced in [15] that is an approach based on conditional generative adversarial networks and a multi-component loss function. Here, a Wasserstein GAN with the gradient penalty and perceptual loss is used that is better than traditional MSE or MAE and hence helps in restoring finer texture details, which further works as a better input for the text recognition network. This model is trained on blurred outside-scene videos and hence works best for our experiments. We find that the formulation used in this model makes training faster and the resulting model generalizes better. The loss function is computed as a combination of content and adversarial loss.

$$L = \underbrace{\underbrace{L_{GAN}}_{advloss} + \underbrace{\lambda.L_X}_{contentloss}}_{totalloss} \tag{1}$$

where λ equals to 100 in all experiments; adversarial and content loss are defined as in [15].

3.3 Text Recognition

The proposed model in [16] comprises of two parts, a text rectification network and a text recognition network. A rectification network transforms the image by rectifying the text in it and uses a flexible Thin-Plate Spline (TPS)[20] transformation to handle irregularities in the text. It is much more flexible than other 2D transformations. TPS handles a vast variety of deformation on images and performs a non-rigid deformation. TPS is able to rectify both perspective and curved text that are the most common and typical types of irregular text.

A text recognition network is a trainable network which is trained with images and their ground truth text annotations. Connectionist Temporal Classification (CTC) method [21] gives a loss function which is differentiable and hence insensitive about horizontal placement of characters and spacing which makes the recognition network trainable end-to-end. Even though CTC is efficient, it fails to provide a mechanism which could model the dependencies among the output characters. Hence, [41] is based on an external language model, like a lexicon, to include language information prior to the recognition process. The recognition problem is tackled using a sequence-to sequence model with extensions to a bidirectional decoder. This network captures the character dependencies, and hence incorporates language modeling into the recognition process. The dependencies among the characters are captured by the bidirectional decoder in both directions.The model is trained end-to-end under a multi-task setting, whose objective is

$$L = \frac{-1}{2} \Sigma_{t=1}^{T} (\log p_{ltr} \frac{y_t}{I} + \log p_{rtl} \frac{y_t}{I}) \tag{2}$$

where y1, . . . , yt, . . . , yT constitutes the groundtruth text represented by a character sequence. The objective is the average of the losses on the left-to-right decoder and the right-to-left decoder, whose predicted distributions are denoted by p_{ltr} and p_{rtl}, respectively. I denotes the actual text.

4 Experiments

We train separate networks on different datasets and use them in a single system to perform end-to-end text detection and recognition. The end-to-end recognition accuracy on different datasets is calculated. The testing is also performed on some videos we collected using mobile phone camera for speed and accuracy.

4.1 Datasets

A single system is created which is able to perform text recognition in videos in real time.The deep networks are trained on different datasets and are then combined to boost the accuracy. We use GoPro dataset to train DeblurGAN. GoPro dataset [22] consists of 2103 pairs of blurred and sharp images in 720p quality, taken from various scenes.

For training the text detection network, we combine both ICDAR 2013 [23] and ICDAR 2015 [24] dataset and combine them to create trainset. ICDAR 2013 dataset contains most of the images from ICDAR 2003 [26] dataset. The dataset contains 1015 images. It contains no lexicon. ICDAR 2015 [24] is the baseline dataset for end-to-end scene text recognition. These are taken by a pair of Google Glasses without careful positioning and focusing. The dataset contains a lot of images containing irregular text including perspective text and curved text. This dataset helps to train a robust text detector which is able to detect perspective, curved and rotated text which may be found in videos.

We train recognition and rectification network on a dataset created by combining Street View Text[27] and ICDAR 2015 [24] dataset. Street View Text (SVT) is collected from the Google Street View. The test set contains 647 images of cropped words. Many images in SVT are severely corrupted by noise, blur, and low resolution. Each image is associated with a 50-word lexicon. Training on these datasets makes the model capable of detecting text in complicated scene images which are blurred and contain perspective and curved text.

All the trained models are combined into one system to work on videos. Our system performs better detection as shown by results of our approach on ICDAR 2013 test set.

To compare our system with other recognition systems, we have also performed experiments on the test sets of SVT and IIIT5K dataset[43]

4.2 Training Details

We use OpenCV [28] and Dlib [29] image processing for text boxes tracking and processing videos. All the experiments have been conducted on Google Compute Engine [31] with 1.2 GHz CPU, 64 GB RAM, Tesla K80 GPU and running Ubuntu 16.04. We use tensorflow for the networks implementation. We follow the training procedure for each of the deep networks as follows:

Detection network The network is trained on 512x512 corpus from images of ICDAR 2013 dataset with 24 as the batch size. The network is trained end-to-end using ADAM [32] optimizer. Initially learning rate is 1e-3 and is decayed to one-tenth with every 27000 mini-batches and is stopped at 1e-5. The network stops training when the performance stops improving.

DeblurGAN The model was trained on a random corpus of number of images of size 256x256 extracted from 1000 Gopro dataset training images. The learning

rate is fixed initially with the value 10-4 till 150 epochs. After this, the learning rate is decayed to zero over the next 150 epochs. The model is trained with batch size = 1, as cited by the authors in their research paper. The network is trained end-to-end using ADAM optimizer. The training phase took 6 days for training the DeblurGAN network.

Rectification and Recognition network The rectification and recognition network is trained on all the images in the SVT dataset[27].The dataset contains 9 million images generated from a set of 90k common English words. Words are rendered onto natural images with random transformations and effects.

4.3 Results

We have tested our end-to-end text recognition method on the test of ICDAR 2013 scene image dataset to evaluate the performance of our method.

Table 1. TEXT RECOGNITION RESULTS ON ICDAR 2013 SCENE IMAGE DATASET(end-to-end)

Method	Precision	Recall	F score
Our method	0.94	0.80	0.88
Multi Frame Tracking	0.93	0.81	0.87.
VGGMaxBBNet [34]	0.89	0.82	0.86
TextProposals [35]	-	-	0.81
Stradvision-1	0.88	0.75	0.81
TextSpotter [36]	0.85	0.69	0.77

Table 1 shows the results for end-to-end text recognition of our end-to-end system and various other state of the art methods. In this table, TextSpotter network is the baseline in the end-to-end text recognition task of ICDAR 2015 RRC Challenge 3 [25]. Multi Frame tracking network is a superior method than these methods for end-to-end text recognition because of higher tracking accuracy. Our method has better performance in terms of precision and F-score percentages than all of the state-of-the-art methods given in the table and hence it outperforms the best methods for the text recognition task.

Table 2 compares the recognition accuracies across a number of methods. Our method achieves best results out of all the previous approaches. Particularly, on IIIT5k[43] and SVT, our method significantly reduces recognition errors comparing to the previous bests.Considering that the tested datasets have covered an extremely wide range of real-world scenarios , it is clear that our method generally works better.

We also test the speed of our complete system on various videos captured. We achieve state-of-the art accuracy with processing speed of 20 FPS. Table 3 compares the recognition speed of our system as compared to others. The speed and accuracy of our system make it suitable for various autonomous systems.

Table 2. RECOGNITION RESULTS COMPARISON ON SVT AND IIIT5K DATASET (ACCURACY)

Method	SVT	IIIT5K
Our method	0.9831	0.9917
Cheng et al.[37]	0.971	0.992
Yang et al.[38]	0.952	0.977
Lee et al.[39]	0.963	0.979
Shi et al.[40]	0.955	0.983
Shi et al.[41]	0.975	0.987
Jaderberg et al.[42]	0.932	0.978

Table 3. RECOGNITION SPEEDS PER IMAGE

Method	Processing Time(milliseconds)
Our method	17
Cheng et al.[37]	60
Yang et al.[38]	64
Lee et al.[39]	83
Shi et al.[40]	104
Shi et al.[41]	121
Jaderberg et al.[42]	243

5 Conclusion and future work

In this paper, we have proposed a pipeline of networks for an end-to-end scene text recognition from natural scene videos after blind motion deblurring. Our system consists of four simpler tasks: text detection, video motion deblurring, text rectification and text recognition, and is lot quicker than the earlier systems built for the same task and also removes blur caused in scene due to motion. The proposed pipeline performs better than other state-of-the-art methods in terms of accuracy, with a processing speed of 20 FPS. The system can be used in various autonomous systems and self driving cars. There can be further improvements in the processing speed by using tracking between the frames.

References

1. Chen, X., Yuille, A.L.: Detecting and Reading Text in Natural Scenes. In: CVPR (2). pp. 366–373 (2004).
2. Wang, T. et al.: End-to-end text recognition with convolutional neural networks. In: ICPR. pp. 3304–3308 IEEE Computer Society (2012).
3. Yin, X.-C. et al.: Robust Text Detection in Natural Scene Images. CoRR. abs/1301.2628, (2013).
4. Neumann, L., Matas, J.: Real-Time Lexicon-Free Scene Text Localization and Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 38, 9, 1872–1885 (2016).
5. Jaderberg, M. et al.: Reading Text in the Wild with Convolutional Neural Networks. International Journal of Computer Vision. 116, 1, 1–20 (2016).

6. Susan, S., Devi, K.M.R.: Text area segmentation from document images by novel adaptive thresholding and template matching using texture cues. Pattern Anal. Appl. 23, 2, 869–881 (2020).
7. Zhou, X. et al.: EAST: An Efficient and Accurate Scene Text Detector. In: CVPR. pp. 2642–2651 IEEE Computer Society (2017).
8. Epshtein, B. et al.: Detecting text in natural scenes with stroke width transform. In: CVPR. pp. 2963–2970 IEEE (2010).
9. Yin, X.-C. et al.: Robust Text Detection in Natural Scene Images. CoRR. abs/1301.2628, (2013).
10. Shi, C. et al.: Scene Text Recognition Using Part-Based Tree-Structured Character Detection. In: CVPR. pp. 2961–2968 IEEE Computer Society (2013).
11. Almazán, J. et al.: Word Spotting and Recognition with Embedded Attributes. IEEE Trans. Pattern Anal. Mach. Intell. 36, 12, 2552–2566 (2014).
12. Ciresan, D.C. et al.: Multi-column deep neural network for traffic sign classification. Neural Networks. 32, 333–338 (2012).
13. Krizhevsky, A. et al.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012).
14. Antani, S.K. et al.: Robust Extraction of Text in Video. In: ICPR. pp. 1831–1834 IEEE Computer Society (2000).
15. Kupyn, O. et al.: DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. CoRR. abs/1711.07064, (2017).
16. Shi, B. et al.: ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. IEEE Trans. Pattern Anal. Mach. Intell. 41, 9, 2035–2048 (2019).
17. Mishra, A. et al.: Top-down and bottom-up cues for scene text recognition. In: CVPR. pp. 2687–2694 IEEE Computer Society (2012).
18. Jaderberg, M. et al.: Deep Features for Text Spotting. In: Fleet, D.J. et al. (eds.) ECCV (4). pp. 512–528 Springer (2014).
19. Yang, X. et al.: Learning to Read Irregular Text with Attention Mechanisms. In: Sierra, C. (ed.) IJCAI. pp. 3280–3286 ijcai.org (2017).
20. Bookstein, F.L.: Principal warps: thin-plate splines and the decomposition of deformations. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 11, 6, 567–585 (1989).
21. Graves, A. et al.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Nets. In: ICML '06: Proceedings of the International Conference on Machine Learning. (2006).
22. Nah, S. et al.: Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. In: CVPR. pp. 257–265 IEEE Computer Society (2017).
23. Karatzas, D. et al.: ICDAR 2013 Robust Reading Competition. In: ICDAR. pp. 1484–1493 IEEE Computer Society (2013).
24. Karatzas, D. et al.: ICDAR 2015 competition on Robust Reading. In: ICDAR. pp. 1156–1160 IEEE Computer Society (2015).
25. Yao, C. et al.: Incidental Scene Text Understanding: Recent Progresses on ICDAR 2015 Robust Reading Competition Challenge 4. CoRR. abs/1511.09207, (2015).
26. Lucas, S.M. et al.: ICDAR 2003 Robust Reading Competitions. In: ICDAR. pp. 682–687 IEEE Computer Society (2003).
27. Goodfellow, I.J. et al.: Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks. (2013).
28. https://opencv.org/
29. http://dlib.net/
30. Danelljan, M. et al.: Accurate Scale Estimation for Robust Visual Tracking. In: Valstar, M.F. et al. (eds.) BMVC. BMVA Press (2014).

31. https://cloud.google.com/
32. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization, http://arxiv.org/abs/1412.6980, (2014).
33. Jaderberg, M. et al.: Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. CoRR. abs/1406.2227, (2014).
34. Jaderberg, M. et al.: Reading Text in the Wild with Convolutional Neural Networks. International Journal of Computer Vision. 116, 1, 1–20 (2016).
35. i Bigorda, L.G., Karatzas, D.: TextProposals: A text-specific selective search algorithm for word spotting in the wild. Pattern Recognition. 70, 60–74 (2017).
36. Neumann, L., Matas, J.: Real-Time Lexicon-Free Scene Text Localization and Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 38, 9, 1872–1885 (2016).
37. Cheng, Z. et al.: Focusing Attention: Towards Accurate Text Recognition in Natural Images. In: ICCV. pp. 5086–5094 IEEE Computer Society (2017).
38. Yang, X. et al.: Learning to Read Irregular Text with Attention Mechanisms. In: Sierra, C. (ed.) IJCAI. pp. 3280–3286 ijcai.org (2017).
39. Lee, C.-Y., Osindero, S.: Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. In: CVPR. pp. 2231–2239 IEEE Computer Society (2016).
40. Shi, B. et al.: Robust Scene Text Recognition with Automatic Rectification. CoRR. abs/1603.03915, (2016).
41. Shi, B. et al.: An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. CoRR. abs/1507.05717, (2015).
42. Jaderberg, M. et al.: Deep Structured Output Learning for Unconstrained Text Recognition. In: Bengio, Y. and LeCun, Y. (eds.) ICLR. (2015).
43. Mishra, A. et al.: Scene Text Recognition using Higher Order Language Priors. In: Bowden, R. et al. (eds.) BMVC. pp. 1–11 BMVA Press (2012).

View publication stats

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

Segmented Rotor Mutually Coupled Switched Reluctance Machine for Low Torque Ripple Applications

Siddharth Mehta, *Student Member, IEEE,* Md Ashfanoor Kabir, *Member, IEEE,* Prerit Pramod, *Senior Member, IEEE,* and Iqbal Husain, *Fellow, IEEE*

***Abstract*—Mutually coupled switched reluctance motors (MC-SRMs) are alternatives in the family of reluctance machines that can overcome switched reluctance motor's (SRM's) system-level integration and control challenges by utilizing standard three-phase voltage source inverters. In this paper, a novel segmented-rotor, fractional-slot, tooth-wound MCSRM with notched rotor design is presented, which has very low torque ripple compared to other machine types in the SRM family. The torque ripple is reduced to 3.6% without utilizing any current profiling techniques or torque sharing functions. The ripple minimization is achieved primarily through rotor segment shaping, which has a strong influence on stator flux densities, flux linkages, and torque harmonics. The design strategy, along with the optimization details, is presented for a 120 W, 12 slot-8 pole MCSRM. The designed MCSRM is prototyped for experimental verification and validation of the finite element analysis results and design methodology.**

***Index Terms*—MCSRM, concentrated winding, flux barrier, rotor shaping, segmented rotor, SRM, torque ripple**

I. Introduction

TORQUE ripple reduction has been a prime research focus for switched reluctance machines (SRMs) for several decades [1]. With a robust mechanical structure, fault-tolerant operation and ease in manufacturing, SRMs are prime contenders for variable speed drives. However, for high-performance applications, such as automotive (traction, electric power steering (EPS) etc.), SRMs are unable to make the case because of high torque ripple, noise and structural vibrations.

Researchers have reported various ripple reduction techniques through the current profiling, which can reduce the ripple to as low as 2% [2]. However, these current profiling methods are not applicable for the entire torque-speed range and require intense optimization process to profile the currents [3]. With this optimized current profiling, the SRMs can still have torque ripple as high as 20-30% depending on the electromagnetic design [4]. The demerits of the unconventional power converter and control techniques for SRMs can be overcome with the use of mutually coupled SRM (MCSRM) drives that utilize the standard 3-phase voltage source inverters (VSIs) [5]–[8]. Furthermore, MCSRMs retain the positive features of SRMs, such as low-cost construction, robust and permanent magnet free design, and high-temperature operation.

The improvement of SRMs has been approached from two aspects: winding variations and innovative rotor designs. From the winding aspect, MCSRMs were designed with full pitched winding which utilized the mutual flux linkage for torque production [5]. The design of full pitched MCSRMs (FP-MCSRMs) with segmented and conventional toothed SRM rotors are illustrated in [9]. Compared to the full pitched winding, fractional slot winding reduces the end-winding length, increases slot fill factor and thermal conductivity, reduces copper loss and provides a compact design [10]–[12]. However, the penalty is paid in terms of increased harmonic content in the air gap flux density and reduced power factor. The authors in an earlier publication have shown through an extensive analysis on four different combinations of full slot and fractional slot winding, with bipolar and unipolar excitations, that the fractional slot configuration is the better alternative because of the reduced end-winding length leading to smaller housing as compared to FP-MCSRMs [13].

On the machine geometry aspect, the effect of rotor segmentation has shown improvement in the machine torque density [14]–[16]. Segmented rotor designs with fractional slot MCSRMs (FrS-MCSRM) have been analyzed and presented in [8], [14], [17] where it is shown that with the help of rotor segmentation, the torque production capability increases by maintaining a low unaligned inductance. Similarly, the research on optimization of the segmented rotor SRMs with a higher number of rotor poles is presented in [18], [19], whereas a more focused rotor tooth design optimization is presented in [20], [21].

In this work, electromagnetic design of a segmented rotor FrS-MCSRM with a novel rotor design for minimizing the torque ripple is presented. The proposed ripple minimization concept is from the machine design aspect alone, prior to any control based current profiling approach which is reported in our previous work [12], [13]. The FrS-MCSRM can be operated with a standard six-switch VSI with bipolar-balanced phase currents. The rest of the paper is structured in the following way: Fundamentals and features of the proposed MCSRM topology are covered in Section II; analysis and

This work was accomplished through the support of the ABB Distinguished Professorship endowment and Nexteer Automotive Corporation for research on electric machines and drives at the FREEDM Center, North Carolina State University.

S. Mehta and I. Husain are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606 USA (e-mail:smehta5@ncsu.edu; ihusain2@ncsu.edu).

M. A. Kabir is with Turntide Technologies (e-mail: mkabir@ncsu.edu).

P. Pramod is with the Advanced Systems and Software Engineering group of Nexteer Automotive Corporation, Saginaw, Michigan, USA (email: prerit.pramod@nexteer.com).

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

rotor segment design strategy of FrS-MCSRM for torque ripple reduction is illustrated in Section III; finite element analysis (FEA) based design optimization of FrS-MCSRM is presented in Section IV; Section V presents the details of the prototyping stages of the 12 slot - 8 pole (12/8), 120 W FrS-MCSRM; and analysis and simulation results are validated with experiments in Section VI.

II. Fractional Slot Mutually Coupled Switched Reluctance Machine

The focus of this paper is on torque ripple mitigation but to provide the readers with more clarity on the presented topology authors are listing the distinctions between FrS-MCSRM and synchronous reluctance machines (SynRMs) below.

The FrS-MCSRM presented in this work consists of concentrated tooth-wound copper coils similar to conventional SRMs and with SynRMs concentrated windings [10], [22], [23]. Similar to its counterparts, the FrS-MCSRM utilizes the principle of reluctance variation around the air-gap to produce torque. However, there are distinct features which present a clear distinction between the FrS-MCSRM and the SynRMs which are mentioned below.

- Unlike SynRMs which are sinusoidally excited [10], [22], [23], trapezoidal motional-EMF and torque characteristics of FrS-MCSRMs requires pulsed or trapezoidal excitation for optimal torque generation [23].
- The phase-flux characteristics as shown in Fig. 1(a) have an unaligned value where there is minimal flux engagement and an aligned value where there is maximum flux engagement between the stator and rotor similar to conventional SRMs [23]. In comparison, SynRM has sinusoidal flux linkages as shown in [23]
- Only the SRMs have a dc quantity present in their phase-flux. Neither SynRM nor the brushless PM machines have a dc quantity in their flux characteristics [10], [22]. Similar to SRMs, the FrS-MCSRM proposed in this research has a strong dc quantity as observed from Fig.1(b).
- MCSRMs are well represented using self inductance and mutual inductance as shown in Fig. 1(c), rather than direct axis and quadrature axis inductances like SynRM as presented in [10], [22]. Similar to SRMs the self inductance is bell shaped and a strong function of rotor position and phase current.

The major difference between SRMs and presented topology is the strong contribution of mutual inductance in torque production which also varies significantly with the saturation and rotor position (Fig. 1(c)). Through time and frequency domain analysis and contrast against its SynRM counterparts it was shown that the proposed topology belongs to the family of MCSRM [24].

MCSRMs can have both full slot and fractional slot winding configuration and can be excited using both unipolar and bipolar currents. In [12] and [13], the authors have shown the advantage of the fractional slot winding configuration compared to other winding configurations in terms of reduced end-winding and copper loss; this finding served as the motivation to explore innovations for torque ripple minimization for FrS-MCSRM. To utilize the conventional VSI, the winding is configured with a Y configuration. The rotor segments which allow short flux paths are attached to rotor support for structural integrity. A figure of a single rotor segment and the complete assembled rotor is shown in Fig. 2(a) and (b), respectively. The rotor segment's shape is optimized to minimize torque ripple using finite element analysis tools.

A rectilinear representation of the machine for the aligned and the unaligned positions is shown in Fig. 3(a) and (b), respectively. During the aligned position, the rotor segment shorts two stator teeth, providing a minimum reluctance path for maximum flux engagement between the stator and rotor, as shown in Fig. 3(a). In the other scenario, when the rotor and stator teeth completely overlap the flux lines encounter a path with high reluctance and hence, the flux engagement between the stator and rotor is minimal. Since the short flux loops do not require the rotor back iron to complete the path, a robust non-magnetic material such as aluminium or structural steel can be utilized to form the rotor support. The short flux paths in FrS-MCSRM helps in reducing the core loss [12]. Unlike FP-MCSRMs and conventional SRMs, both self and mutual inductances of FrS-MCSRM are bell shaped and vary with rotor position and phase currents, as shown in Fig. 1(c), which makes this machine a variant of SRM. As a result both the self and mutual inductance contribute significantly towards torque production [6], [25]. The torque can be expressed as

$$\begin{aligned} T = &\frac{1}{2}\frac{dL_a}{d\theta}i_a^2 + \frac{1}{2}\frac{dL_b}{d\theta}i_b^2 + \frac{1}{2}\frac{dL_c}{d\theta}i_c^2 \\ &+ \frac{dM_{ab}}{d\theta}i_a i_b + \frac{dM_{bc}}{d\theta}i_b i_c + \frac{dM_{ca}}{d\theta}i_c i_a \end{aligned} \tag{1}$$

where T is the electromagnetic torque, L_x is the self inductance and M_{xy} is the mutual inductance between phases x and y, where $\{x, y\} \in \{a, b, c\}$. θ is the mechanical rotor position and i_a, i_b, and i_c are the phase currents.

III. Torque Ripple: Sources and Mitigation

Sources of torque ripple for SRMs/MCSRMs can be identified through the study of their flux characteristics. The sources of torque ripple in segmented FrS-MCSRM is qualitatively and quantitatively evaluated in this section through Fourier series expansion of the position and current dependent flux characteristics.

A. Conventional Segmented Rotor

The conventional rotor of a segmented SRM/MCSRM is trapezoidal-shaped laminated steel, as shown in Fig. 4(a). Ideally, flat torque characteristics are desired to avoid torque fluctuations for SRMs/MCSRMs excited using trapezoidal excitations. The torque characteristics of the conventional segmented rotor have large fluctuations, as shown in Fig. 5(a), which result in increased torque ripple. The torque is a derivative, of the integral, of the flux-linkage with respect to the rotor position and any sudden changes in the slope of the flux characteristics will result in fluctuation in the torque

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

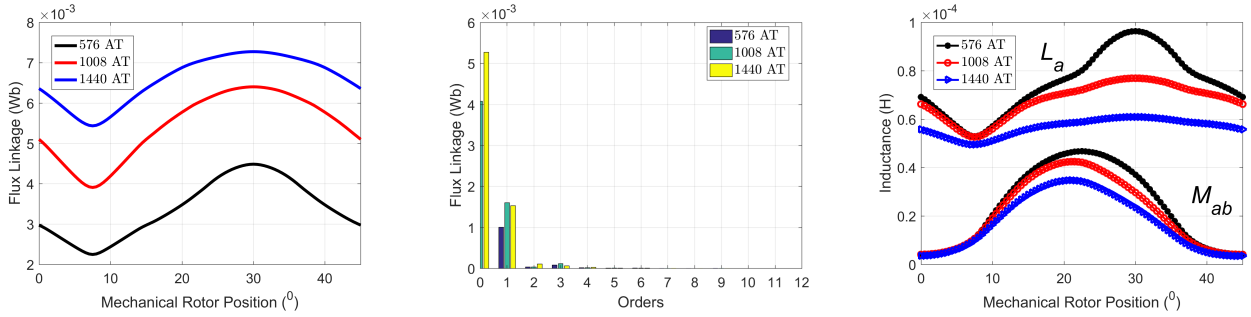
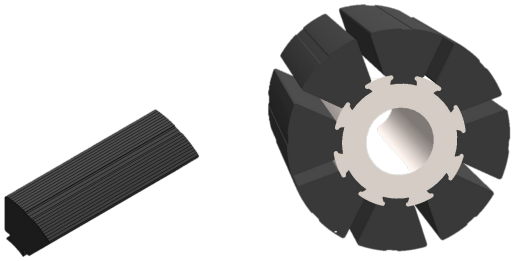


Fig. 1. (a) Flux characteristics, (b) harmonic decomposition of flux characteristics, and (c) inductance characteristics of 12/8 FrS-MCSRM.



(a) Single rotor segment (b) Segmented rotor assembly

Fig. 2. 3D representation of the segmented rotor.

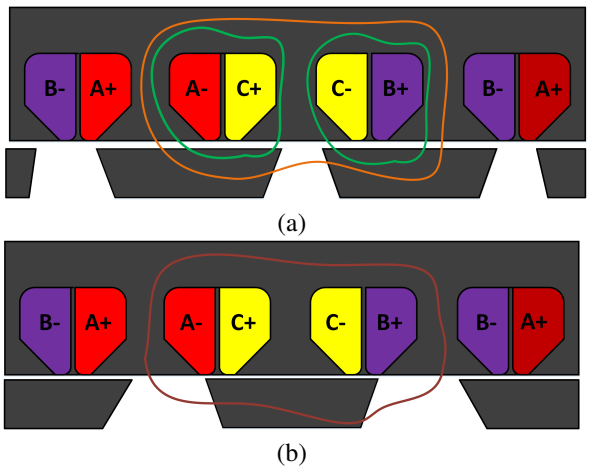


Fig. 3. Flux paths for (a) aligned (b) unaligned positions of FrS-MCSRM.

(a) (b)

Fig. 4. (a) Conventional (without notch) and (b) proposed (with notch) rotor segments.

characteristics. The slope variations in the flux are highlighted in Fig. 5(b).

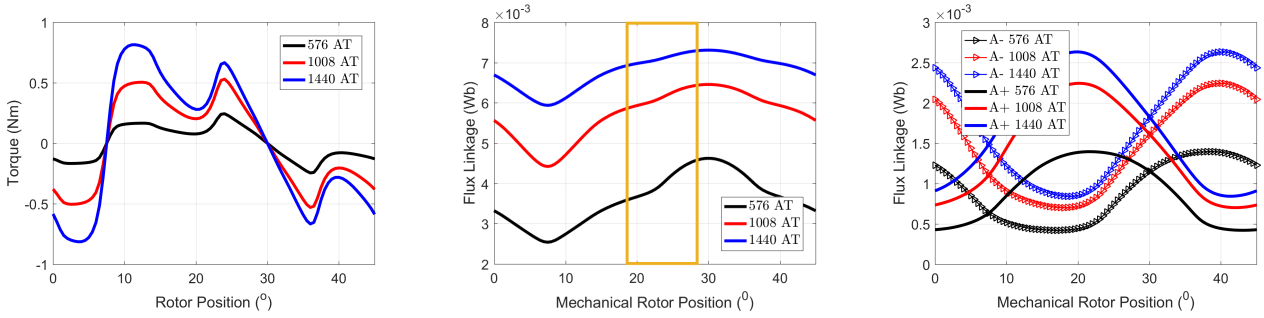
To identify the source of torque fluctuations, let us consider a single coil of phase A consisting of two conductors, A+ and A-. When only phase A is excited with dc current, the flux linkage will be the summation of flux linked with the two conductors, namely, A+ and A-. The flux characteristics of the two conductors, A+ and A-, are shown in Fig. 5(c). A key observation is that the flux linkage among the two conductors varies with the rotor position. The fluctuations in the phase flux, see Fig. 5(b), occur during the transitioning period, that is when the flux of the conductor A+ begins to decrease, and the flux of conductor A- begins to build up. This scenario is linked to the relative position of the rotor pole and the stator poles at three discrete rotor positions, as shown in Fig. 6. The flux lines indicate how the flux distribution varies around A+ and A- conductors.

In Fig. 6(a), at $\theta = 22°$, the flux lines link rotor pole 1 and conductors A+ and C- while minimal flux lines encircle A-. As the rotor moves ahead, the flux distribution gradually varies and attains an equal distribution around A+ and A- conductors at $\theta = 30°$, as shown in Fig. 6(c) and the same is also observed in Fig. 5(c). During this transitioning phase, the increase in the flux density around A- is not sufficient to counter the decrease in the flux density around A+. From Fig. 5(c) for $576AT$ between rotor position of $22°$ and $26°$, the flux linkage of A+ decreases by 57.40 μWb but the flux linkage of A- increases by 340 μWb. As a result, the total flux linked with phase A suffers from dips, as observed in Fig. 5(b). These dips result in sudden variation in the slope of the flux, and that gives rise to torque ripple.

The analysis of flux characteristics reveals, that to effectively produce a smooth torque, the transition of flux from A+ to A- needs to be as smooth as possible and that the total flux linked with phase A doesn't have sudden jumps. A smooth transition of flux between the conductors can be achieved by introducing flux barriers to alter the reluctance of the flux paths such that the total flux linked with each coil has a constant slope. The reluctance is altered by introducing notches in the center of the rotor poles. The effect of these notches on the flux paths is presented next.

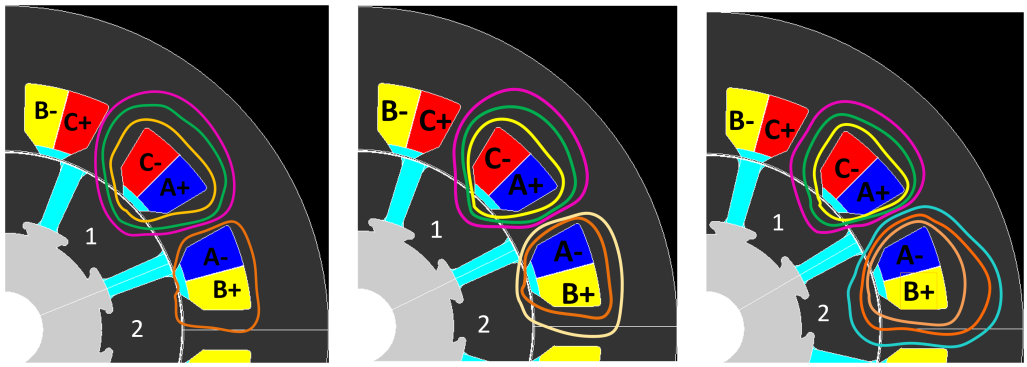
0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications



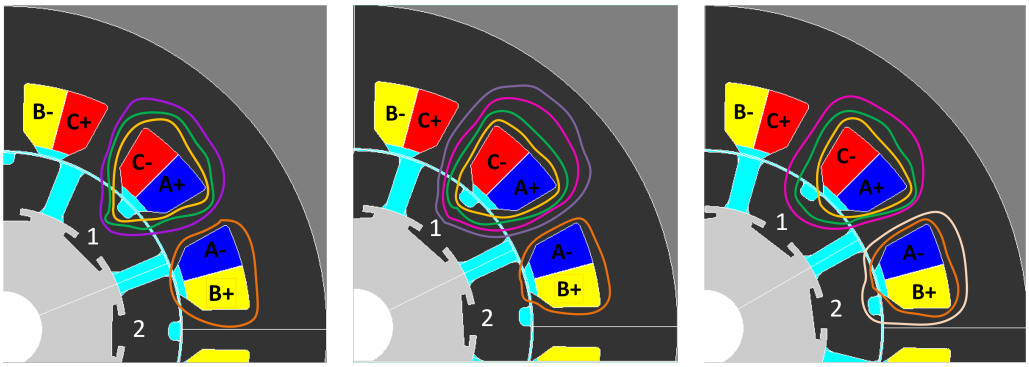
(a) (b) (c)

Fig. 5. (a) Torque characteristics, (b) flux linkage characteristics, and (c) flux linkage of individual conductor for conventional rotor with only one phase excited.



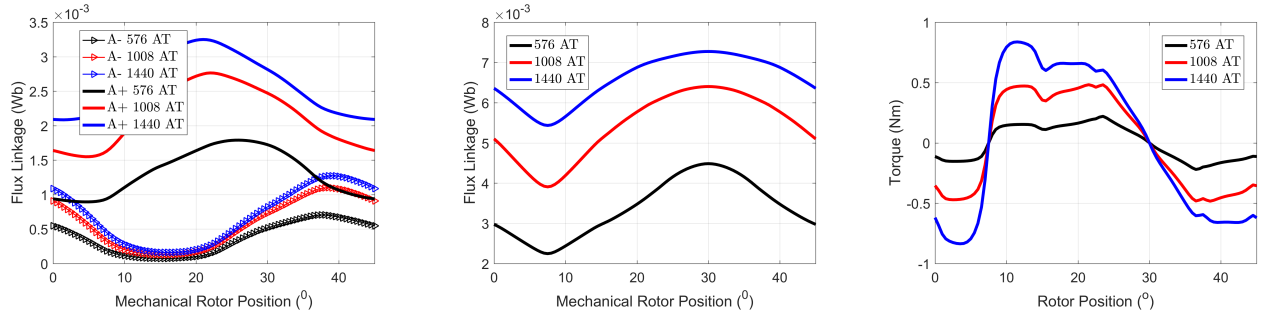
(a) (b) (c)

Fig. 6. Flux distribution for conventional rotor design at (a) $\theta = 22°$, (b) $\theta = 26°$, and (c) $\theta = 30°$.



(a) (b) (c)

Fig. 7. Flux distribution for proposed rotor design at (a) $\theta = 22°$, (b) $\theta = 26°$, and (c) $\theta = 30°$.



(a) (b) (c)

Fig. 8. (a) Flux linkage for individual conductor, (b) flux linkage for a single phase, and (c) torque characteristics of a single phase for proposed design.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

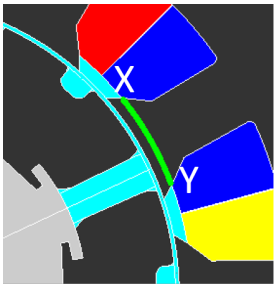


Fig. 9. Illustration of XY path perpendicular to which stator flux density is analyzed.

B. Notched Rotor Design

The proposed rotor design with notches in the center is shown in Fig. 4(b). The addition of the notch in the rotor introduces varying reluctance of the flux paths as can be seen in the flux distribution plots in Fig. 7. At $\theta = 22^\circ$ in Fig. 7(a), the flux distribution is similar to Fig. 6(a) as the flux is linking conductor A+, C- and rotor pole 1. As the rotor moves from 22° to 26°, the flux lines encircle conductor A+ through rotor pole 1 instead of migrating to the rotor pole 2 and encircling conductor A-. At 26°, the flux density around A+ increases as the notch of rotor pole 1 coincides with the slot opening. Whereas, the notch of rotor pole 2 is in front of the stator pole and acts as a flux barrier by increasing the reluctance and forcing the flux lines to complete the path around A+ through rotor pole 1. This scenario is depicted in Fig. 7(b) and is also seen in Fig. 8(a). However, this was not the case at 26° with the conventional rotor design. The flux lines started to build around conductor A- as there is no flux barrier to prevent the flux from entering rotor pole 2, as depicted in Fig. 6(b). As the rotor advances to 30° and beyond the flux lines, start to circulate coil conductor A- through rotor pole 2 as rotor pole 2 now offers a path of less reluctance because the notch of the rotor pole 2 now starts coinciding with the slot opening. This results in a more gradual migration of flux lines around A+ to A- and a smoother total flux linked with the phase coil, as observed from Fig. 8(b). The end result is a torque profile that can deliver reduced torque ripple, as shown in Fig. 8(c).

To further validate the above analysis, the stator pole flux density, B_s, normal to the path XY (see Fig. 9), is analyzed for both rotor designs for the same magneto-motive force (MMF) of $576AT$. All the geometrical dimensions are the same, and the only difference between the designs is the presence of a notch in the proposed design. The stator flux densities for different rotor positions for both the rotor designs are illustrated in Fig. 10(a) and (b). The flux in the stator pole is the surface integral of the stator flux density, i.e. area under the curve. The stator flux for both the rotor designs is shown in 10(c). If the area under the curve of the stator pole's flux density increases abruptly with the change in rotor position, it will result in a sudden increase in the flux of the stator pole, as observed from Fig. 10(c), thus a sudden increase in the flux-linkage of the phase coil. These sudden changes will manifest themselves as fluctuations in torque, resulting in torque ripple.

Fig. 10(a) and (b) also indicate that the proposed rotor design has decreased stator flux density which implies an increase in reluctance. As a consequence, the flux linked between the rotor and stator for the proposed design decreases. However, the average torque in doubly salient reluctance machines is dependent on the ratio of the unaligned to the aligned flux instead on the flux value itself. With this design improvement, both the aligned and unaligned flux values decrease, thereby maintaining the saliency ratio. Therefore, designing the flux barrier, a trade-off is needed between torque quality and average torque.

The effect of the notch on average torque and torque ripple is analyzed using harmonic decomposition of flux and torque characteristics of the machine. To perform the spectrum analysis of flux and torque characteristics two phases are excited with equal and opposite polarity dc currents in the finite element (FE) model. A flat torque profile, which is the ideal choice for reduced torque ripple for doubly salient reluctance machines, is possible if there are only odd-order harmonics present in flux. However, the flux characteristics typically have both odd and even order harmonics, see Fig. 11(a). The phase flux characteristics is given by

$$\lambda_x(i_x, \theta_e) = \sum_{n=1}^{m} \lambda_{nax}(f(i_x)) \sin(n\theta_e) + \lambda_{nbx}(f(i_x)) \cos(n\theta_e) \tag{2}$$

where λ_x is the flux of a phase, I is the current magnitude, θ_e is the electrical rotor position, n is the harmonic order, and λ_{nax} and λ_{nbx} are the Fourier series coefficients. In the linear region (low saturation), for a fixed current magnitude, the Fourier coefficients can be assumed to be independent of the phase current, and the phase torque can be estimated from the phase flux characteristics and is given by

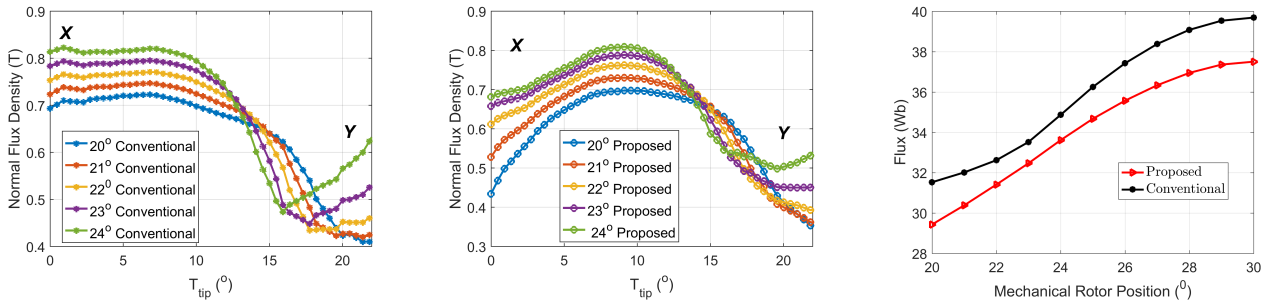
$$\begin{aligned} T_x(i_x, \theta_e) &= \frac{\partial}{\partial \theta_e} \int_0^{i_x} \lambda_x(i_x, \theta_e) di_x \\ &\approx nI\Big(\sum_{n=1}^{m} \lambda_{nax} \cos(n\theta_e) - \lambda_{nbx} \sin(n\theta_e) \Big) \end{aligned} \tag{3}$$

where T_x is the torque of a phase. The torque characteristics obtained using the analytical Fourier series based expression, while considering the first nine orders, is shown in Fig. 12 along with the torque characteristics from the FE model to validate the harmonic decomposition approach. From (3), it is evident that the orders appearing in flux will also appear in the torque characteristics but increased in magnitude by a factor of nI [26]. Therefore, both even and odd-order harmonics appear in torque characteristics as well. The consequence of the even-order harmonics is abrupt slope changes in the flux characteristics which produce uneven peaks in torque characteristics and result in torque ripple as observed in Fig. 5(a). The harmonic spectrum for the conventional rotor, showing 2^{nd} order harmonic and above, is shown in Fig. 11(b).

To mitigate the problem of torque ripple, the ratio of the magnitude of the harmonics need to be maintained, such that the cumulative effect results in a flat torque profile. Each harmonic adds its signature in the torque that is unique.

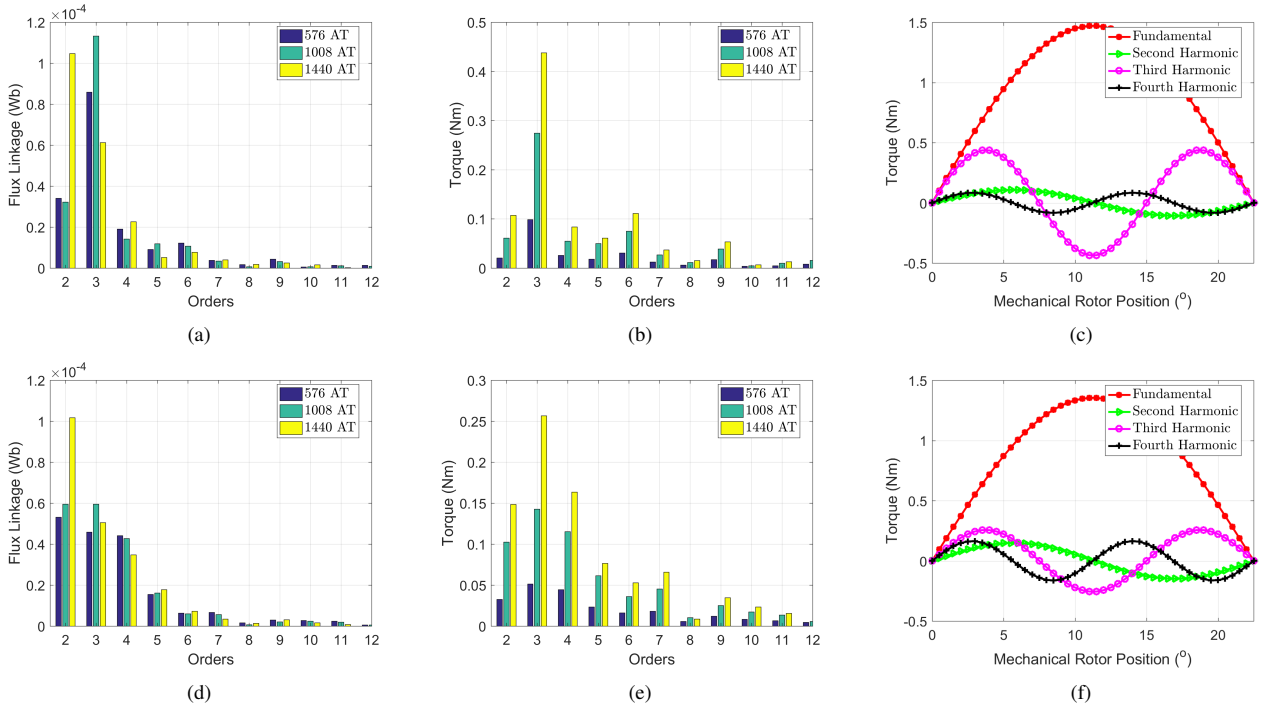
0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications



(a) (b) (c)

Fig. 10. Stator tooth flux density at 576 AT for (a) conventional design, (b) proposed design, and (c) stator pole flux comparison.



(a) (b) (c)

(d) (e) (f)

Fig. 11. Harmonic spectrum of (a) flux linkage, (b) torque and (c) significant orders of torque for the conventional design. Harmonic spectrum of (d) flux linkage, (e) torque and (f) significant orders of torque for the proposed rotor design.

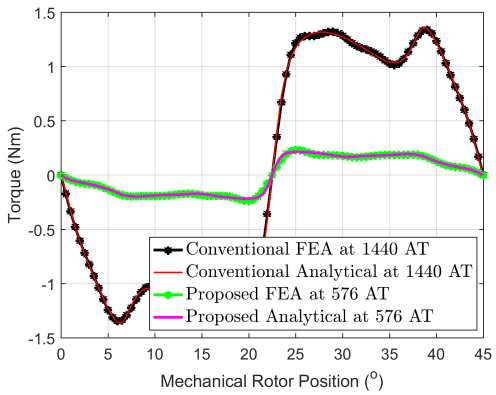


Fig. 12. Comparison of torque characteristics obtained from FEA and using Fourier series based analytical equation presented in (3).

For example, the 2^{nd} order harmonic results in a peak on the increasing side of the torque characteristics and the 4^{th} order results in a peak on the falling side of the torque characteristic. Whereas, the 3^{rd} order harmonic will in result in two equal peaks on the increasing and falling sides of the torque characteristics and a dip in the center of each half-cycle. Similarly, the 5^{th} order harmonic will result in a sharp peak in the center of each half-cycle that can compensate for the dip created by the 3^{rd} order. Each order for both the designs are shown in Fig. 11(c) and (f) to provide a better visualization of each order.

In the proposed rotor design, the 3^{rd} harmonic is decreased to reduce the dip in the middle of each half-cycle. Whereas, the 5^{th} order is increased to compensate for the dip. The magnitudes of the 2^{nd} and 4^{th} order harmonics are increased to maintain a ratio with respect to the 3^{rd} order to avoid the formation of uneven peaks in the torque characteristic.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

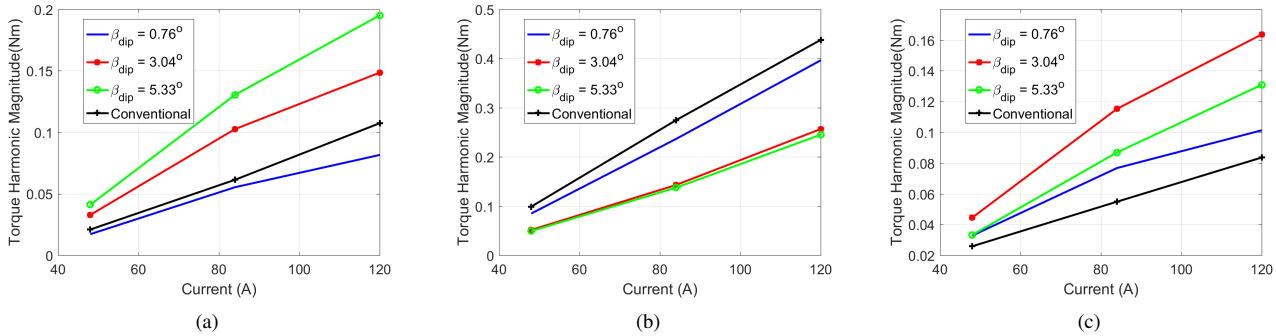


Fig. 13. Sensitivity analysis of (a) second order, (b) third order, and (c) fourth order harmonics in torque with variation in β_{DIP}

Besides, higher-order odd harmonics are also increased to provide a flatness to the torque characteristics.

C. Parametric Sensitivity Analysis

Three main parameters can control the size of the notch: Angle of opening, β_{DIP}, notch height, H_{DIP}, and notch base radius, Arc_{DIP}. The flux and torque profiles are the most sensitive to β_{DIP}. By introducing this design feature, the objective is to make the torque as flat as possible, by increasing the higher-order odd harmonics and also balancing out even order harmonics to avoid the formation of uneven lobes. For sensitivity analysis, β_{DIP} is varied from $0.76°$ to $5.33°$, and the harmonic contents in the torque characteristics are analyzed. The variation of 2^{nd}, 3^{rd}, and 4^{th} order harmonics with β_{DIP} and saturation is presented in Fig. 13. The variation of average torque, T_{avg}, of a half-cycle and peak-to-peak torque T_{ppk}, along with the analysis of significant harmonic orders, for 1440 AT are summarized in Table I. From the above sensitivity analysis, it is evident that the torque's fundamental and the third harmonic components reduce with an increase in the β_{DIP}. From this, it can be concluded that with an increase in β_{DIP}, average torque will also reduce and peak-to-peak torque does not vary linearly with β_{DIP}. Peak-to-peak torque is a cumulative result of several harmonics, which depend on multiple geometrical parameters. The optimum design of the rotor geometry to minimize peak-to-peak torque is obtained through a multi-dimensional optimization in this research.

IV. Application and Design Optimization

A high-performance automotive application of EPS has been targeted for design, optimization and prototype development since this application has the most stringent torque quality requirements. The initial design of the FrS-MCSRM based on the sizing equations is calculated for a 120 W, 12/8 configuration [12]. The minimization of torque ripple, T_r, is defined as the main objective for the optimization process. The torque ripple is defined as

$$T_r = \frac{T_{max} - T_{min}}{T_{avg}} \tag{4}$$

where T_{max} is the maximum torque and T_{min} is the minimum torque. The machine is designed for an existing EPS frame size of stator outer diameter, OD, stack length, L_{stk}, air gap length, g, and shaft diameter, D_{sh}. Besides, the dc-link voltage, V_{dc}, and peak current, I_{max} are also constrained. These constraints are given in (5). The design parameters considered for the optimization are shown in Fig. 14. The optimization of the proposed MCSRM is performed using Altair Flux 2D and the Altair Hyperstudy optimization package. The global surface response method (GSRM) is used to optimize the machine. The GSRM option available in Hyperstudy utilizes the concept of adaptive response surface based optimization to provide the ability to handle multi objective formulation; the method has a faster and efficient response which improves the accuracy of the response surface [27]. The optimization process is constrained by the geometrical constraints along with a constraint on the average torque $\geq$ 1.3 Nm, 10% more than the rated torque to account for non-idealities. The optimization problem can be summarized as

$$\begin{aligned} \min \quad & T_r = f(\boldsymbol{X}) \\ \text{Subject to constraints:} \quad & OD = 85 \ mm \\ & L_{stk} = 32 \ mm \\ & T_{avg} \geq 1.3 \ Nm \\ & g = 0.3 \ mm \\ & D_{sh} = 10 \ mm \\ & I_{max} \leq 200 \ A \end{aligned} \tag{5}$$

where

$$\boldsymbol{X} = \begin{bmatrix} \beta_{DIP}, H_{DIP}, Arc_{DIP}, \beta_{SEG}, T_{tip} \\ \beta_{SEGB}, H_{SEG}, LSI, \beta_s, R_1, R_2, R_3 \end{bmatrix}.$$

To reduce the optimization time, one-fourth geometry is used while optimizing the machine. The variation of torque ripple and average torque with three main rotor parameters, β_{DIP}, H_{DIP}, and Arc_{DIP} are shown in Fig. 15 and Fig. 16, respectively. Iteration number 52 is selected as it gives the lowest torque ripple while satisfying all the constraints. Additional details on the optimization process is presented in [8], [12]. The final optimized geometry for the ripple minimized FrS-MCSRM are given in Table II. The flux density distribution at the rated condition is shown in Fig. 17. Net flux and torque characteristics are obtained from the FE model of the optimized design by exciting two phases together and

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

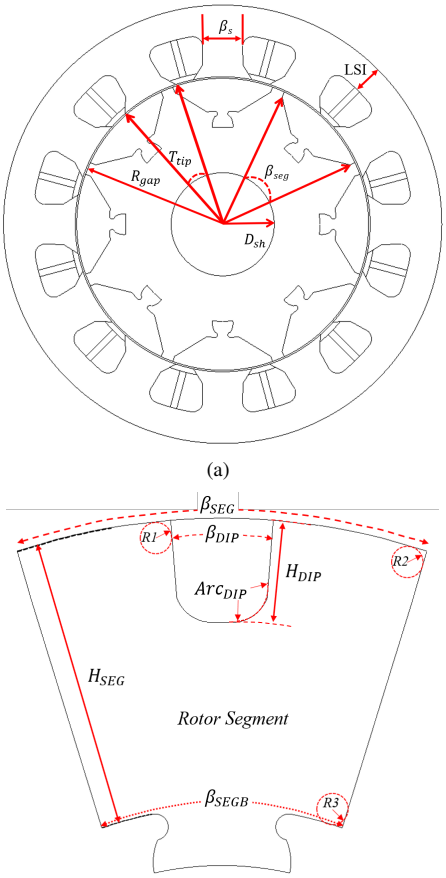
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

TABLE I
PARAMETRIC SENSITIVITY ANALYSIS AT 1440 AT

Order	Conventional	β_{DIP}= 0.76°	β_{DIP}= 3.04°	β_{DIP}= 5.33°
1^{st}(Nm)	1.47	1.46	1.35	1.18
2^{nd} (Nm)	0.1074	0.081	0.148	0.195
3^{rd} (Nm)	0.438	0.396	0.256	0.2453
4^{th} (Nm)	0.083	0.101	0.163	0.1309
5^{th} (Nm)	0.061	0.056	0.767	0.08
7^{th} (Nm)	0.037	0.051	0.065	0.032
9^{th}(Nm)	0.0535	0.0814	0.0348	0.072
T_{avg} (Nm)	1.0224	1.0142	0.9154	0.812
T_{ppk} (Nm)	0.32	0.34	0.07	0.253

TABLE II
PARAMETERS OF THE OPTIMIZED DESIGNS

Parameters	Value
LSI	10.29 mm
L_{stk}	32 mm
OD	45 mm
g	0.3 mm
D_{sh}	10 mm
Slot Opening	2.92 mm
β_s	5.35 mm
R_{gap}	23.21 mm
H_{seg}	7.36 mm
H_{DIP}	1.52 mm
β_{DIP}	3.047 °
β_{seg}	37.3 °
T_{tip}	22.76 °



(a)

(b)

Fig. 14. Machine design parameters for optimization.

a comparison with the conventional rotor is presented in Fig. 18 [25]. After optimization of the machine's geometrical parameters, the turn number is calculated as 12 per slot per phase. Each phase has two conductors in series and two parallel paths considering the dc-link voltage constraint. With the optimized design the peak-to-peak torque is limited to 0.07 Nm at 120 A resulting in decrease in toque ripple by 75.16% as compared to the conventional design.

V. PROTOTYPE DEVELOPMENT

A 12/8, 120 W, FrS-MCSRM using optimized dimensions has been prototyped; the stages of prototyping are shown in Fig. 19 and Fig. 20. Laminations of eight rotor segments are manufactured with the rotor notch using laser cutting as shown in Fig 19(a) and the rotor support is manufactured using 6061-T6 aluminium alloy with electrical discharge manufacturing (EDM) technology for high precision. The rotor slots are glued to the rotor support using Loctite super glue. The shaft is inserted using a hydraulic process. The completed rotor assembly is shown in Fig. 19(d).

The stator has 12 slots which are wound with copper conductors of AWG 19 with insulation paper glued to the stator slots. 3D printed blocks, as shown in Fig. 20(a), are used to hold the insulation paper in place. The winding in the prototype is tooth wound with a slot fill factor of 30%. The low slot-fill factor is not a fundamental problem of the design, but rather a limitation of the automated winding apparatus available that can use wires of only specific AWG sizes at the prototype manufacturing facility. There is ample space to increase the fill factor without stator segmentation, as can be seen in Fig. 20(b). Each slot has 12 conductors per phase which give a total of 24 conductors per slot. Each phase has two parallel paths and two conductors in series. The windings are connected in a Y fashion. The stator is then glued to the housing.

VI. RESULTS AND DISCUSSIONS

The experimental test bed consists of the prototype FrS-MCSRM coupled to a dyno motor. An induction motor drive is used as the dyno-motor to rotate the test motor at a constant speed. A torque sensor with a precision of $\pm$ 1 m-Nm and a position encoder are used to obtain the data using a personal computer. The experimental verification of the prototype is conducted at an industrial facility.

A. Static Characteristics

To extract the static torque characteristics, fixed currents are supplied simultaneously through 2 phases of opposite polarity, ranging from 0 to 120 A in steps of 4 A, while rotating the motor at a fixed speed of 25 RPM. The current is regulated using a 12 V power supply with a current capability of 200 A and is operated in the constant current mode. During this test, torque, phase currents, and speed signals are recorded. The torque characteristics obtained from the experiment and the FE model are shown in Fig. 21(a). It is observed from

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

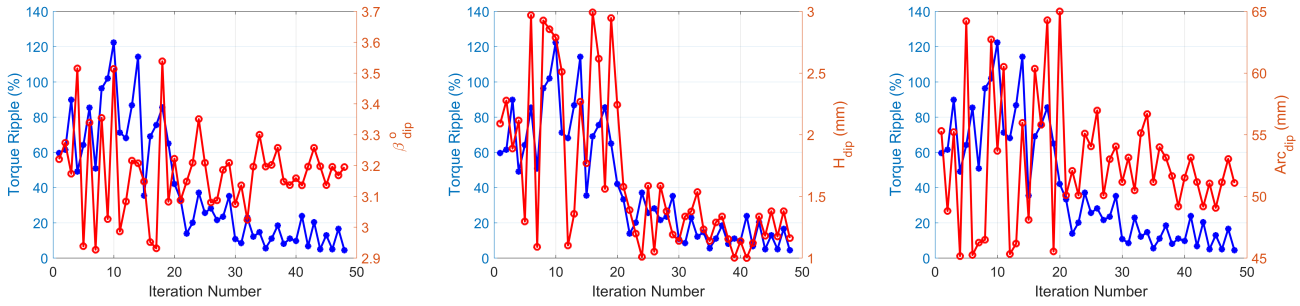


Fig. 15. Sensitivity analysis of torque ripple with variation in (a) β_{dip}, (b) H_{dip} and (c) Arc_{dip}

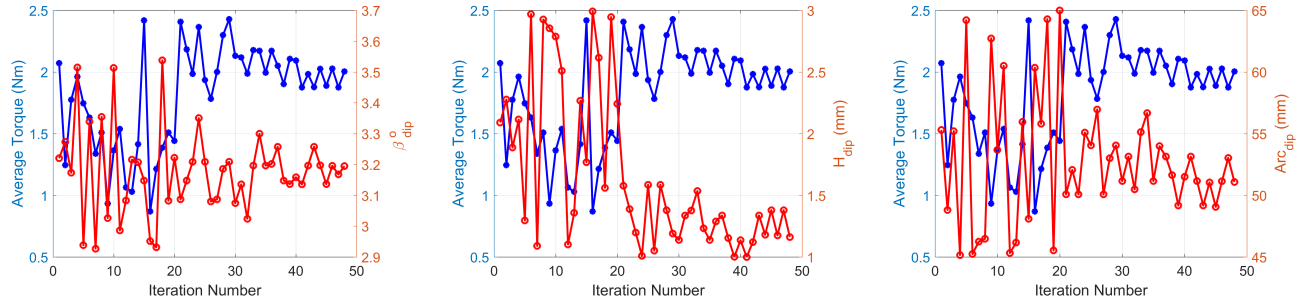


Fig. 16. Sensitivity analysis of average torque with variation in (a) β_{dip}, (b) H_{dip} and (c) Arc_{dip}

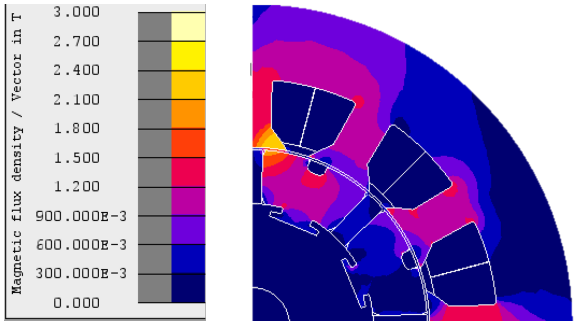


Fig. 17. Flux density distribution at the rated condition.

Fig. 21(a) that the experimental and FE characteristics have an excellent match for the lower excitation levels. However, as saturation increases, a mismatch is observed. This can be linked to the uncertainty of the magnetic characteristic of the lamination material. The mismatch can also be correlated to the manufacturing precision of the rotor segments. As the laminations are laser cut which has lower precision compared to that of EDM, any mismatch in the notch of the prototyped segment will result in differences between the experimental and FE torque characteristics.

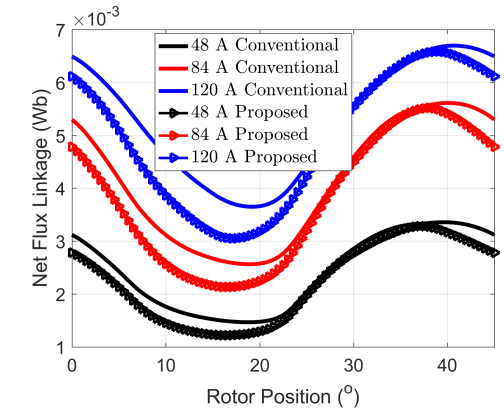
B. Steady-State Analysis

MCSRMs are generally operated with trapezoidal currents. To verify the steady-state performance, the FrS-MCSRM is operated using a 3 phase VSI drive and dSPACE Microautobox-II in torque controlled mode using balanced-trapezoidal-bipolar phase currents. A discrete proportional-integral (PI) based current controller is designed for a fixed bandwidth to regulate the current commands at a fixed switching frequency of 20 kHz [28]. The current tracking performance for three different operating points is shown in Fig. 22. The induction motor drive is operated in speed-controlled mode at 500 RPM. The test is performed for discrete current commands ranging from 20 A to 104 A. The commanded phase current and the measured currents are shown in Fig. 22(a) for a particular scenario. The average torque is estimated from the experimental tests at a speed of 500 RPM, and the experimental data is compared with the average torque obtained from the FE model as shown in Fig. 21(b).

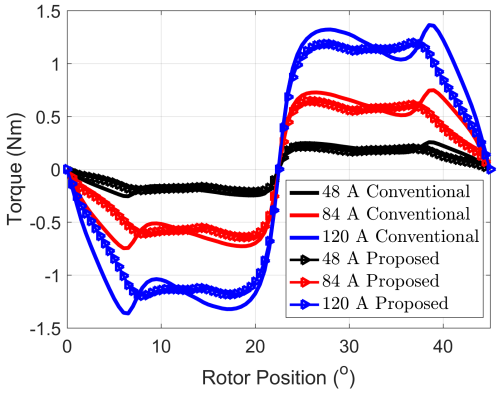
To evaluate the torque ripple performance, the prototyped machine is operated with trapezoidal currents ranging from 90 A to 120 A at a low speed of 20 RPM. The commanded and measured current at 90 A, 20 RPM is shown in Fig. 22(b). Torque ripple, in general, is a function of not just the electromagnetic design but is severely impacted by errors in position sensor, current sensor, initial rotor position, and current controller bandwidth. Hence, to isolate the torque ripple contributions from the above secondary causes, a pseudo-experimental approach is adopted, which is described in detail in [25]. The phase currents and rotor position recorded during the experiment are fed to the static characteristics obtained from the experiment, as shown in Fig. 21(a). Torque ripple is also obtained from the FE model, and a comparison is presented in Fig. 21(c) and in Table III. The torque ripple has been effectively minimized to below 5% range, and

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

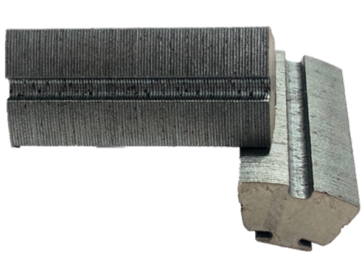


(a)

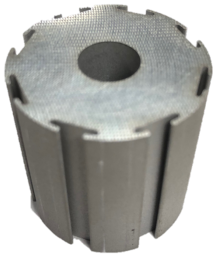


(b)

Fig. 18. Comparison of (a) flux and (b) torque characteristics between conventional and optimized proposed designs.



(a) Rotor segments



(b) Rotor support

(c) Assembled rotor

(d) Rotor with shaft

Fig. 19. Rotor assembly of 12/8 FrS-MCSRM.

(a) Unwound stator

(b) Concentrated winding

Fig. 20. Stator assembly of 12/8 FrS-MCSRM.

the agreement between the experimental and the FE results through the preceding analysis provides a proof of concept.

C. Performance Comparison

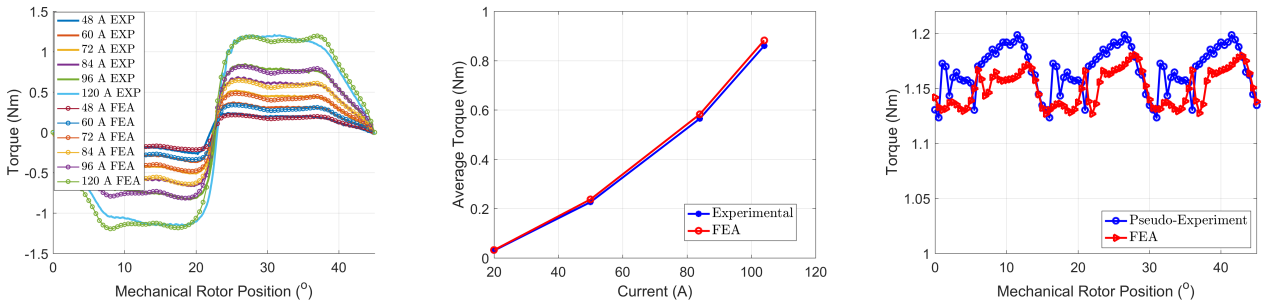
To benchmark, the performance of the proposed design, a comparative analysis of the average torque and torque ripple is presented between the proposed design and the conventional segmented rotor that doesn't include a rotor notch. The conventional segmented rotor design is also optimized for low torque ripple [12]. Both the machines have the same turn numbers and same series/parallel connection. The comparison between the average torque and torque ripple between the two designs is shown in Fig. 23(a) and (b), respectively. The average torque of the proposed sign is effectively reduced by 5-10% in the entire operating range.

For low saturation levels, the torque ripple performance is judged from the absolute peak-to-peak value, since average torque is low, which is defined by the requirements of the application. For EPS applications, the allowable absolute peak-to-peak torque ripple is in the range of 30 mNm - 40 mNm, depending on the gear ratio of the mechanical system. This condition is met by the proposed design until 440 AT. At higher saturation levels, the torque ripple is calculated using (4) and is brought as low as 3.6% with the proposed design, whereas the lowest torque ripple for the conventional segmented rotor design is 8%. The worst torque ripple reported for this design is 27.6% and 42% with the conventional rotor design. During the entire operating range, the proposed design has improved torque ripple performance as compared to the conventional design. For operating points having high percentage values, torque ripple can be further mitigated by profiling phase currents [2].

In [29], the authors have reported a torque ripple of 68% for a segmented SRM designed for low torque ripple applications. In comparison to that, the proposed rotor design with notches shows tremendous improvement. A comparison of torque ripple is also presented with a 12/8 FP-MCSRM designed and optimized with the same constraints and targets [12], [25]. The torque ripple for FP-MCSRM has been shown to fluctuate in the range of 27-38% in Fig. 23. Clearly, the proposed FrS-MCSRM has improved performance as compared to its counterparts from torque ripple stand point.

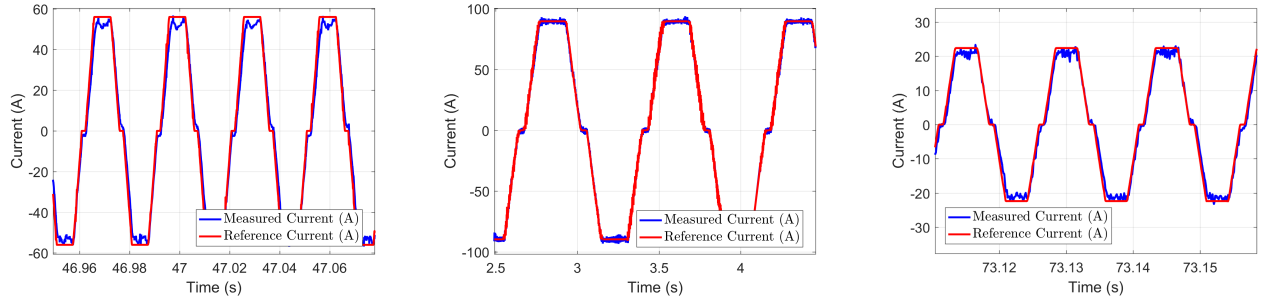
0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications



(a) (b) (c)

Fig. 21. Comparison of experimental and FE (a) static torque characteristics (b) average torque, and (c) torque ripple.



(a) (b) (c)

Fig. 22. Experimentally measured and reference phase currents at (a) 54 A, 500 RPM (b) 90 A, 20 RPM, and (c) 20A, 1500 RPM.

TABLE III
TORQUE RIPPLE PERFORMANCE COMPARISON FOR FRS-MCSRM

Parameter	Model	90 A	112 A	120 A
T_r (%)	Pseudo-Experiment	12.42	6.8	6.44
T_r (%)	FEA	11.21	6.18	4.64

VII. CONCLUSIONS

This work has proposed a new rotor design for mutually coupled switched reluctance machines (MCSRMs). The new rotor design has flux barriers that prevent sudden variations in the flux resulting in smoother flux linkage characteristics. The flux barriers have been designed as notches in the center of each rotor segment. Through harmonic decomposition, the effect of the rotor notch on the torque ripple and average torque is presented highlighting the importance of 2^{nd}, 3^{rd}, and 4^{th} harmonic in minimizing torque ripple. Through sensitivity analysis, it can be concluded that the torque ripple is not just a function of a few selective orders but a collective combination of significant orders. A prototype FrS-MCSRM of the proposed design is built, and extensive characterization of the prototype is performed to validate the proposed design. The torque ripple has been effectively brought down below the 10% range and going as low as 3.6%. Reduced end winding length, compact design, high fill factor, ability to be operated with conventional VSIs, and reduced torque ripple performance makes FrS-MCSRM an attractive alternative to SRMs and FP-MCSRMs. Hence, pushing them one step closer to industrial adoption for high-performance applications.

ACKNOWLEDGEMENT

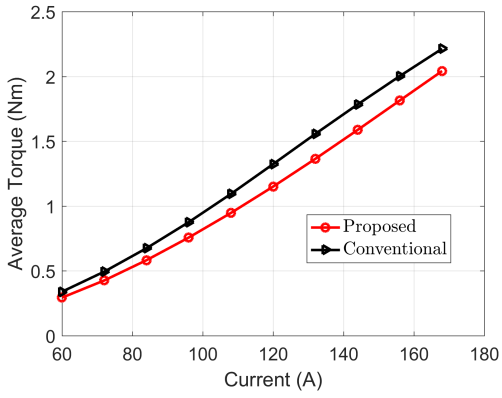
The authors would like to acknowledge the support of Nexteer Automotive Corporation for helping in experimental verification of results and Altair for providing the FE software.

REFERENCES

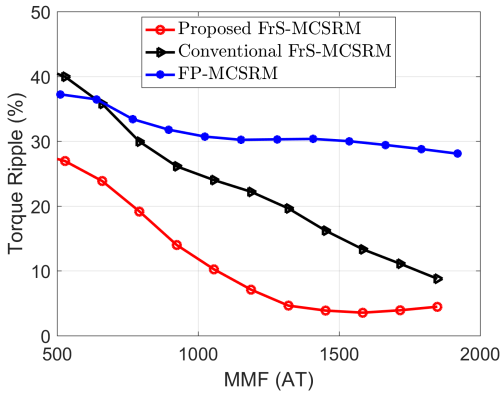
[1] I. Husain, "Minimization of torque ripple in SRM drives," *IEEE Trans. Ind. Electron.*, vol. 49, no. 1, pp. 28–39, Feb. 2002.
[2] R. Mikail, I. Husain, Y. Sozer, M. S. Islam, and T. Sebastian, "Torque-Ripple minimization of switched reluctance machines through current profiling," *IEEE Trans. Ind. Appl.*, vol. 49, no. 3, pp. 1258–1267, May 2013.
[3] S. Mehta, M. A. Kabir, and I. Husain, "Extended speed current profiling algorithm for low torque ripple SRM using model predictive control," in *2018 IEEE Energy Conversion Congress and Exposition (ECCE)*, Sept. 2018, pp. 4558–4563.
[4] J. Ye, B. Bilgin, and A. Emadi, "An offline torque sharing function for torque ripple reduction in switched reluctance motor drives," *IEEE Trans. Energy Convers.*, vol. 30, no. 2, pp. 726–735, June 2015.
[5] B. C. Mecrow, "New winding configurations for doubly salient reluctance machines," in *Conference Record of the 1992 IEEE Industry Applications Society Annual Meeting*, Oct. 1992, pp. 249–256 vol.1.
[6] B. Mecrow, "Fully pitched-winding switched-reluctance and stepping-motor arrangements," *IEE Proceedings B (Electric Power Applications)*, vol. 140, no. 1, pp. 61–70, Jan. 1993.
[7] M. A. Kabir and I. Husain, "Design of mutually coupled switched reluctance motors (MCSRMs) for extended speed applications using 3-phase standard inverters," *IEEE Trans. Energy Convers.*, vol. 31, no. 2, pp. 436–445, June 2016.
[8] M. Kabir and I. Husain, "Segmented rotor design of concentrated wound switched reluctance motor (SRM) for torque ripple minimization," in *2016 IEEE Energy Conversion Congress and Exposition (ECCE)*, Sept. 2016, pp. 1–6.
[9] B. C. Mecrow, E. A. El-Kharashi, J. W. Finch, and A. G. Jack, "Preliminary performance evaluation of switched reluctance motors with segmental rotors," *IEEE Trans. Energy Convers.*, vol. 19, no. 4, pp. 679–686, Dec. 2004.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications



(a)



(b)

Fig. 23. Comparison of (a) average torque and (b) torque ripple for 12/8 MCSRMs.

[10] C. M. Spargo, B. C. Mecrow, J. D. Widmer, and C. Morton, "Application of Fractional-Slot concentrated windings to synchronous reluctance motors," *IEEE Trans. Ind. Appl.*, vol. 51, no. 2, pp. 1446–1455, Mar. 2015.

[11] C. M. Spargo, B. C. Mecrow, and J. D. Widmer, "Application of fractional slot concentrated windings to synchronous reluctance machines," in *2013 International Electric Machines Drives Conference*, May 2013, pp. 618–625.

[12] S. Mehta, M. A. Kabir, and I. Husain, "Design and analysis of mutually coupled srms for low torque ripple applications using standard voltage source inverters," in *2019 IEEE International Electric Machines Drives Conference (IEMDC)*, 2019, pp. 1240–1247.

[13] M. A. Kabir and I. Husain, "Concentrated winding segmented rotor switched reluctance machine (SRM) using three-phase standard inverters," in *2015 IEEE Energy Conversion Congress and Exposition (ECCE)*, Sept. 2015, pp. 5567–5572.

[14] B. C. Mecrow, J. W. Finch, E. A. El-Kharashi, and A. G. Jack, "Switched reluctance motors with segmental rotors," *IEE Proceedings - Electric Power Applications*, vol. 149, no. 4, pp. 245–254, July 2002.

[15] R. Vandana and B. G. Fernandes, "Segmented switched reluctance motor with novel winding and excitation strategy," in *The XIX International Conference on Electrical Machines - ICEM 2010*, Sept. 2010, pp. 1–4.

[16] V. R and B. G. Fernandes, "Design methodology for High-Performance segmented rotor switched reluctance motors," *IEEE Trans. Energy Convers.*, vol. 30, no. 1, pp. 11–21, Mar. 2015.

[17] B. C. Mecrow, E. A. El-Kharashi, J. W. Finch, and A. G. Jack, "Segmental rotor switched reluctance motors with single-tooth windings," *IEE Proceedings - Electric Power Applications*, vol. 150, no. 5, pp. 591–599, Sept. 2003.

[18] J. D. Widmer and B. C. Mecrow, "Optimized segmental rotor switched reluctance machines with a greater number of rotor segments than stator slots," *IEEE Trans. Ind. Appl.*, vol. 49, no. 4, pp. 1491–1498, July 2013.

[19] J. D. Widmer, R. Martin, and B. C. Mecrow, "Optimisation of an 80kw segmental rotor switched reluctance machine for automotive traction," in *2013 International Electric Machines Drives Conference*, May 2013, pp. 427–433.

[20] Y. Li, S. Ravi, and D. C. Aliprantis, "Tooth shape optimization of switched reluctance motors for improved torque profiles," in *2015 IEEE International Electric Machines Drives Conference (IEMDC)*, May 2015, pp. 569–575.

[21] G. Li, J. Ojeda, S. Hlioui, E. Hoang, M. Lecrivain, and M. Gabsi, "Modification in rotor pole geometry of mutually coupled switched reluctance machine for torque ripple mitigating," *IEEE Trans. Magn.*, vol. 48, no. 6, pp. 2025–2034, June 2012.

[22] C. M. Spargo, B. C. Mecrow, J. D. Widmer, C. Morton, and N. J. Baker, "Design and validation of a synchronous reluctance motor with single tooth windings," *IEEE Trans. Energy Convers.*, vol. 30, no. 2, pp. 795–805, June 2015.

[23] M. A. Kabir, "High performance reluctance motor drives with three-phase standard inverter," Ph.D. dissertation, North Carolina State University, 2017.

[24] S. Mehta, "Design, modeling, and control of doubly salient reluctance machines," Ph.D. dissertation, North Carolina State University, Aug. 2020.

[25] S. Mehta, M. A. Kabir, I. Husain, and P. Pramod, "Modeling of mutually coupled switched reluctance motors based on net flux method," *IEEE Transactions on Industry Applications*, vol. 56, no. 3, pp. 2451–2461, 2020.

[26] P. Pramod, P. Nuli, R. Mitra, and S. Mehta, "Modeling and simulation of switched reluctance machines for control and estimation tasks," in *2019 IEEE International Electric Machines Drives Conference (IEMDC)*, May 2019, pp. 565–570.

[27] M. S. Islam, M. H. Chowdhury, A. Shrestha, M. Islam, and I. Husain, "Multi-Load point optimization of interior permanent magnet synchronous machines for high-performance variable speed drives," *IEEE Transactions on Industry Applications*, pp. 1–1.

[28] S. Mehta, P. Pramod, I. Husain, and M. A. Kabir, "Small-Signal modeling of mutually coupled switched reluctance motor," *IEEE Trans. Ind. Appl.*, pp. 1–1, 2020.

[29] M. A. J. Kondelaji and M. Mirsalim, "Segmented-Rotor modular switched reluctance motor with high torque and low torque ripple," *IEEE Transactions on Transportation Electrification*, vol. 6, no. 1, pp. 62–72, Mar. 2020.

Siddharth Mehta (Student Member, IEEE) received the Bachelor of Technology in Electrical and Electronics Engineering from Vellore Institute of Technology-Chennai, India in 2015 and his Master of Science in Electrical Engineering from North Carolina (NC) State University, Raleigh, NC, USA in 2017. He is currently working towards his Ph. D. in Electrical Engineering from NC State University. He worked with GE Healthcare on medical devices in Bangalore, India; and designed and tested switched reluctance motor drives for steering applications at Nexteer Automotive Corporation, Saginaw, MI, USA. He is currently a Graduate Research Assistant with the Future Renewable Electric Energy Delivery and Management Systems Center, Raleigh, NC. His research interests include modeling and control of electric motor drives and electric machine design.

Mehta is a member of IEEE, IEEE-IAS, and reviewer of IEEE journals on electric machines and drives.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.

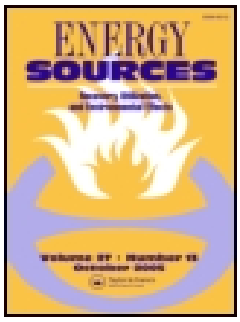
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIA.2021.3073384, IEEE Transactions on Industry Applications

Md Ashfanoor Kabir (Member, IEEE) received his Bachelor's and Master's degrees in Electrical and Electronic Engineering (EEE) from Bangladesh University of Engineering and Technology (BUET), Bangladesh in 2009 and 2012, respectively. From 2009 to 2012, he served the Department of EEE, BUET as a lecturer. In 2017, he received his Ph.D. degree in electrical engineering from the department of Electrical and Computer Engineering at North Carolina State University, Raleigh, NC, USA. From 2017 to 2020, he was with ABB Inc. working as a research scientist at their US Corporate Research Center in Raleigh, NC, USA. Currently he is working as a motor design engineer at TurnTide Technologies (formerly Software Motor Company) in Sunnyvale, California.
His research interests include design, modelling and control of electric machines using variable speed drives with mechanical stress, vibration and noise analysis. He focuses on high performance, high reliability designs for industrial and traction, ultra-high speed machines with WBG drives, and integrated motor drives. He has received the NCSU graduate fellowship, ABB fellowship, NCSU research symposium award and grand prize in Appalachian state energy conference. He is a member of IEEE, IEEE-IAS, and reviewer of IEEE and IET journals on electric machines and drives.

Prerit Pramod (Senior Member, IEEE) received his Bachelor of Engineering degree in Electrical Engineering from the Delhi College of Engineering, University of Delhi, India in 2010 and his Master of Science in Electrical Engineering Systems from the University of Michigan, Ann Arbor, Michigan, USA in 2012. He is currently an Engineering Manager in the Advanced Systems and Software Engineering group at Nexteer Automotive Corporation, Saginaw, Michigan, USA, where he is responsible for development of advanced motion control systems. His research interests include the design, modeling and sizing of electrical machines and power converters and the development and analysis of control, estimation and optimization algorithms for electric motor drive systems.

Iqbal Husain (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Texas AM University, College Station, TX, USA, in 1993.
He is the Director with the FREEDM NSF Engineering Center and the ABB Distinguished Professor with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA, where he also serves as the Director with the Advanced Transportation Energy Center. Prior to joining the North Carolina State University, he was with the University of Akron, where he built a successful power electronics and motor drives program. He was a Visiting Professor with Oregon State University, Corvallis, OR, USA, in 2001. His expertise is in the areas of power electronics, electric machines, motor drives, and system controls. In the FREEDM center, his research is focused on power electronics integration into power systems and transportation electrification. He has also developed innovative graduate and undergraduate courses on electric and hybrid vehicles and published a textbook on the subject.
Dr. Husain was the recipient of the 2006 SAE Vincent Bendix Automotive Electronics Engineering Award, the 2004 College of Engineering Outstanding Researcher Award, the 2000 IEEE Third Millennium Medal, and the 1998 IEEE-IAS Outstanding Young Member award. He was the Distinguished Lecturer of the IEEE Industry Applications Society for 2012-2013 and the Editor-in-Chief for the IEEE ELECTRIFICATION MAGAZINE.

0093-9994 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 09:44:56 UTC from IEEE Xplore. Restrictions apply.



Energy Sources, Part A: Recovery, Utilization, and Environmental Effects

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ueso20

Sensitivity analysis and multi-objective optimization of organic Rankine cycle integrated with vapor compression refrigeration system

Ashwni, Ahmad Faizan Sherwani, Deepak Tiwari & Anil Kumar

To cite this article: Ashwni, Ahmad Faizan Sherwani, Deepak Tiwari & Anil Kumar (2021): Sensitivity analysis and multi-objective optimization of organic Rankine cycle integrated with vapor compression refrigeration system, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, DOI: 10.1080/15567036.2021.1916132

To link to this article: https://doi.org/10.1080/15567036.2021.1916132

Published online: 15 Apr 2021.

Submit your article to this journal

Article views: 54

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
https://doi.org/10.1080/15567036.2021.1916132

Taylor & Francis
Taylor & Francis Group

ARTICLE

Check for updates

Sensitivity analysis and multi-objective optimization of organic Rankine cycle integrated with vapor compression refrigeration system

Ashwni [a], Ahmad Faizan Sherwani[a], Deepak Tiwari[b], and Anil Kumar[c,d]

[a]Department of Mechanical Engineering, Jamia Millia Islamia, Delhi, India; [b]Department of Mechanical Engineering, Bhagwan Parshuram Institute of Technology, Delhi, India; [c]Department of Mechanical Engineering, Delhi Technological University, Delhi, India; [d]Centre for Energy and Environment, Delhi Technological University, Delhi, India

ABSTRACT

In this paper, the organic Rankine cycle integrated vapor compression refrigeration system has been evaluated thermodynamically and optimized using Taguchi-integrated gray relation analysis. Exergetic efficiency and pressure ratios of compressor and expander are considered as the performance characteristics. The optimal setting of input parameters is 298 K temperature of condenser, R602 working fluid, 80% isentropic efficiency of the expander, 85% isentropic efficiency of the compressor, evaporator temperatures for ORC at 373 K, and VCR at 268 K. The condenser and the ORC evaporator have strongly affected the performance of the system (sensitivity analysis).

ARTICLE HISTORY
Received 9 November 2020
Revised 25 March 2021
Accepted 6 April 2021

KEYWORDS
ORC; Taguchi; gray relational analysis; VCR; sensitivity analysis

Introduction

The rise in income and population growth especially in developing countries like India has pushed the energy demand worldwide by more than one-fourth up to 2040 (Duffin 2019). Modern families are consuming a large part of the energy in the form of electricity (IEA 2018). Therefore, the share of electricity in overall global energy consumption is approaching 20% and around 15% of the electricity consumption is accounted for by the refrigeration sector (Coulomb 2008). Currently, fossil and nuclear fuel together account for 80% share in electricity production, and renewables are responsible for the rest 20% share. However, the share of renewable energy is expected to rise by 30% by 2040 (IEA 2018). Renewable energy encompasses solar, wind, biomass, hydro, tidal energy, etc. A rigorous effort is made to tap low-temperature energy sources like solar heat, geothermal heat, waste heat, etc. for electricity generation or refrigeration. Recently, ORC has gained worldwide attention to tap low-temperature energy sources to produce electricity. It uses low boiling point refrigerants as the working fluid, which makes ORC a perfect choice for low-temperature applications. Simplicity in structure and operation are also cited as their benefits (Goyal, Sherwani, and Tiwari 2019). Further, the researchers have focused on multipurpose ORC systems such as; ORC-based combined heating and power generation (CHP) systems, combined heating, power, and cooling (CHPC) systems, and combined power and refrigeration (CPR) systems. Multipurpose ORC systems exhibit better overall efficiencies and lower impact on the environment than single-purpose thermal systems (Aziz, Salim, and Kim 2019).

The integration of ORC and VCR systems to produce electricity and refrigeration throughout the year has been proved a promising solution. Many researchers have worked on a combined ORC-VCR system to study its thermodynamic performance in the context of first law and second law

CONTACT Deepak Tiwari deepakmathiya@gmail.com Department of Mechanical Engineering, Bhagwan Parshuram Institute of Technology, Delhi, 110 089, India

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2021 Taylor & Francis Group, LLC

efficiencies. The selection of working fluid is another major challenge along with performance enhancement of the ORC-VCR system thermodynamically. There is a direct link between the working fluid and the performance of the systems for low-temperature applications. According to (Li et al. 2013), R600 was a better fluid followed by R600a for the combined system. While (Bu, Wang, and Huashan 2014) proposed R600a as the best one based on several parameters like COP of the system, expander size, pressure ratio, and safety. However, (Kim and Perez-Blanco 2015) proposed butane and isobutane as suitable working fluids for the system at lower ORC expander inlet pressure and R134a and R152a for higher ORC expander inlet pressure. (Nasir and Kim 2016) used seven different working fluids like R123, R134a, R245fa, R1234yf, R1234ze (E), Butane, and isobutane for the combined system. R134a was found as the most suitable working fluid for ORC and isobutane for the VCR cycle. (Saleh 2016) used 10 different pure working fluids and proposed R600 as an optimum working fluid. In another study on the ORC-VCR system, (Saleh 2018) used 13 different pure working fluids and proposed R602 for the system based on overall thermal performance and environmental concerns.

Moreover, numerous research studies have been conducted to optimize the input conditions of design parameters affecting the performance of the ORC system. (Goyal, Sherwani, and Tiwari 2019) applied response surface methodology on low-temperature heat source driven ORC using zeotropic mixtures. The performance of the system is strongly dependent on heat sink temperature. (Bademlioglu et al. 2018) used the Taguchi and ANOVA methods on ORC to determine the impact weights of design parameters on the thermal efficiency of the system. The temperatures of evaporator and condenser and isentropic efficiency of the turbine were found the most significant factors affecting the thermal efficiency of the system by the impact ratio of 70%. In another study, (Bademlioglu, Canbolat, and Kaynakli 2020) carried out the optimization study of ORC using Taguchi-gray relational methods. Isentropic efficiency of the turbine, the effectiveness of the internal heat exchanger, and the temperature of the evaporator and condenser were taken as input parameters to determine their impact on the first law and second law efficiencies where the impact of the evaporator temperature was determined to be the maximum at 31.37%. (Salim and Kim 2019) used a non-dominated sorting genetic algorithm for optimal thermo-economic performance of the system. While (Patel, Desai, and Kachhwaha 2017) conducted thermo-economic optimization of ORC powered cascaded vapor compression – vapor absorption system. The overall size of the system and its cost were also optimized.

The Taguchi method has been proven a very powerful tool to determine the optimal combination of design parameters for single-objective problems. It makes use of an experimental design called orthogonal array to examine the quality of performance characteristics through the use of signal-noise ratios (Kackar 1985). However, single-objective optimization studies do not make real sense because only a single objective at a time is optimized. However, the overall performance of an actual thermal system depends on multiple indicators like thermal efficiency, exergetic efficiency, sizes of expanders, compressors, and heat exchangers. Therefore, the Taguchi method becomes ineffective for multi-objective problems. Therefore, It is imperative to convert such multi-objective problems into an equivalent single objective problem to utilize the benefits of the Taguchi method for finding an optimal combination of design parameters. Grey relational technique is being widely used to convert a multi-objective problem into an equivalent single objective through the determination of gray relational grade coefficient. Therefore, the Taguchi method has been integrated with gray relational analysis to determine the optimal combination of design parameters for multi-objective problems (Hussain, Khan, and Sarmah 2019).

Based on the exhausting review, the following conclusions can be drawn:

- It can be concluded that most of the optimization studies have been carried out on simple ORC systems (Bademlioglu et al. 2018; Goyal, Sherwani, and Tiwari 2019). Little work has been carried out on the optimization of the thermal performance ORC-VCR system except (Salim and Kim

2019). Moreover, there is no optimization study of the combined ORC-VCR system having multi-objectives using Taguchi integrated gray relational analysis.

- Previous multi-objective optimization studies were on simple ORCs or ORC systems where thermal efficiency, exergetic efficiency (Bademlioglu, Canbolat, and Kaynakli 2020; Bademlioglu et al. 2018), heat exchanger size parameter (based on heat exchanger area), and turbine size parameter (based on volume ratios) (Salim and Kim 2019) were considered in performance evaluation. Pressure ratios of expander (EPR) and compressor (CPR) are also very important parameters as performance characteristics as they also directly affect the sizes of the expander and compressor, respectively. They should be minimized as possible.
- No sensitivity analysis of the ORC-VCR system has been carried out to reduce exergetic destruction in components to increase the overall exergetic system performance.

This study not only evaluates but also further optimize thermal performance in terms of exergetic efficiency but also estimates the pressure ratios affecting the sizes of major costs affecting components of the system such as expander and compressor for ORC-VCR system.

Therefore, a multi-objective optimization study of the combined organic Rankine cycle-based vapor- compression refrigeration system has been carried out using Taguchi-gray relational analysis. In this study, EPR and CPR have been considered performance characteristics, which are to be minimized along with the exergetic efficiency that is to be maximized whereas the type of working fluid, the temperature of ORC and VCR evaporators, isentropic efficiencies of ORC expander, and VCR compressor, and condenser temperature are taken as operating parameters.

Materials and methods

Description of the system

The special characteristic of this system is the utilization of the same working fluid both in the ORC and VCR. The power produced by the expander is enough to power the VCR compressor and both are directly coupled to reduce energy losses. The system can be used to produce electricity when refrigeration is not needed in winter. Moreover, it can also be used to produce a refrigeration effect in summer. Therefore, owing to its operational flexibility makes this system more beneficial as per the climate requirement throughout the year.

The working of the system can be explained as follows (Figure 1). The ORC fluid enters the evaporator and takes heat from the low-temperature waste heat source (state 6 to state 7). It exits the ORC evaporator as a dry saturated vapor (state 7) and drives the expander (state 8) to produce enough work to run the compressor. At state 8, it enters the mixer where it mixes with the working fluid coming from the compressor. After mixing, the total mass flux of working fluid enters the condenser at state 4 where it rejects heat and exits the condenser at state 5. Flow-regulating valves divide the total mass flux into two streams, one of which circulates through the VCR system and the other one circulates through ORC. The ORC working fluid at state 5 enters the pump, which pumps it to the ORC evaporator.

The VCR system working fluid at state 5 passes through the expansion valve and experiences an isenthalpic process to state 1. In the VCR evaporator, it extracts the heat from the refrigerated space by evaporating in the evaporator and exits the evaporator at state 2. Then, it enters the compressor, which increases its pressure and temperature. In this way, the cycle repeats itself. Figure 2 depicts the processes on temperature-entropy property diagram of the system.

Selection of working fluid

The choice of working fluid for the system is critical as it directly affects the performance. A working fluid has desired thermo-physical properties (critical temperature, critical pressure, etc.), low ozone

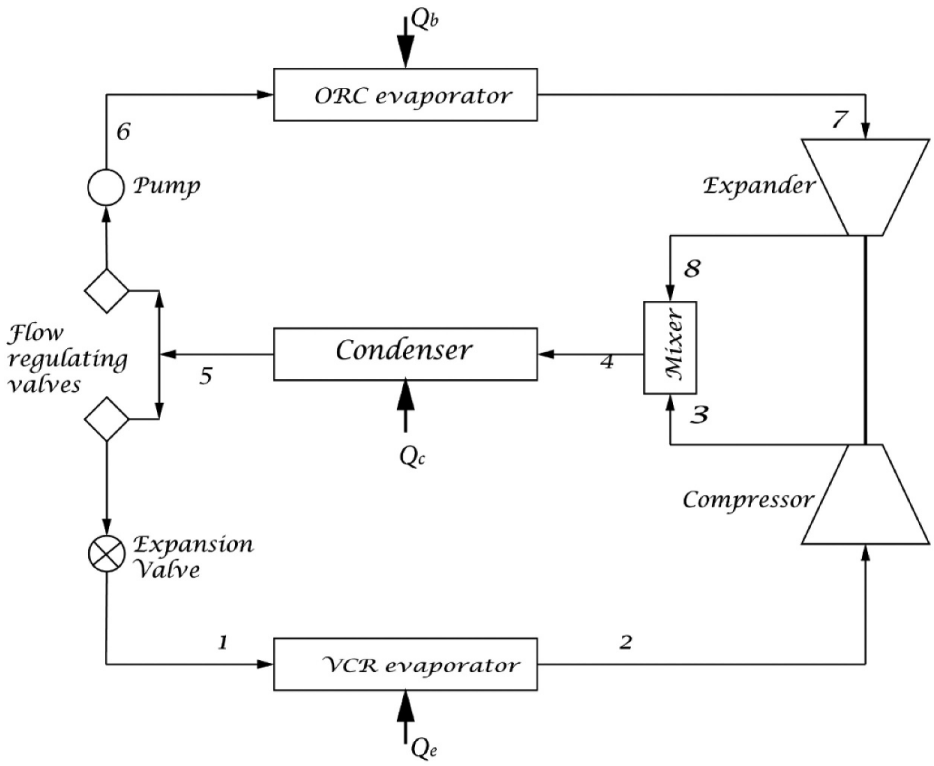


Figure 1. Schematic diagram of combined ORC–VCR system.

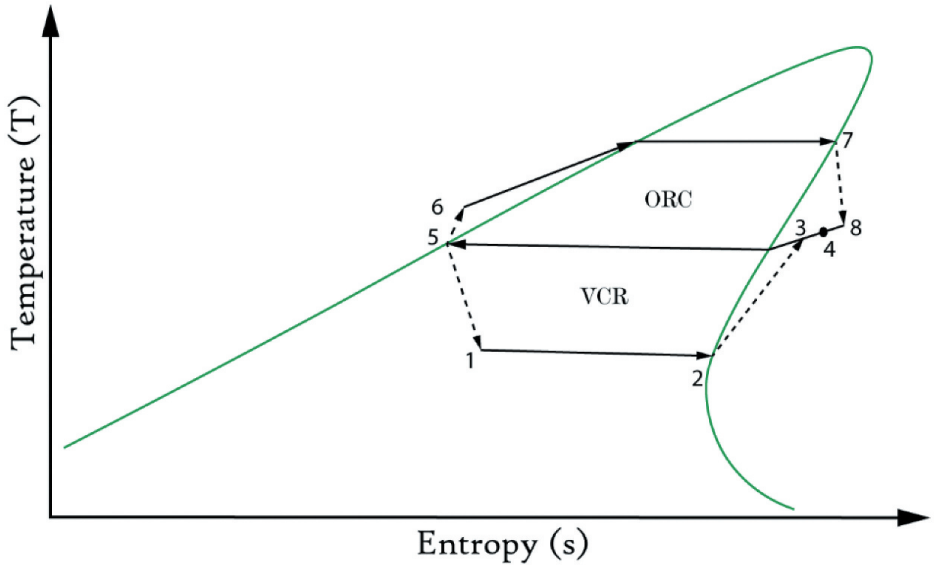


Figure 2. Temperature-entropy diagram of ORC- VCR system.

depletion potential (ODP), and global warming potential (GWP) for being selected (Wu et al. 2017). 3 working fluids have been chosen for the system namely; R600, R602, and R245fa, their thermo-physical properties are given in table 1. The selected working fluids for the present study have been chosen based on higher critical temperatures, zero ODP, and lower GWP [Table 1]. Moreover, they are dry working fluids for eliminating the chances of erosion for expander blades.

Thermodynamic modeling

Thermodynamic modeling of the proposed system is based on exergy balance relation and the law of mass and energy conservations. Following assumptions are considered in thermodynamic modeling of the system:

- Steady operating conditions, negligible potential, and kinetic energies.
- No pressure drop within piping or system components
- Negligible heat losses to the surroundings.

The thermodynamic equations of the system components are given in table 2. The exergetic efficiency (η_{ex}) of the ORC-VCR system, ORC expander pressure ratio (EPR), and VCR compressor pressure ratio (CPR) are performance characteristics (or parameters) that are intended to be optimized in this analysis [Table 2].

The state at the exit of the condenser is taken subcooled by 5 K to avoid the onset of cavitation at the inlet of the pump. Although (Mago et al. 2008) suggested for no superheating to avoid the irreversible losses yet (Muhammad et al. 2015) pointed out some degree of superheat to ensure stable operation of the system. Therefore, states of working fluid at the exit of ORC and VCR evaporators are taken superheated one with a degree of superheat of 5 K for stable operation of the system. The approach temperatures in the condenser, ORC, and VCR evaporators are 5 K, 15 K, and 5 K, respectively.

Taguchi integrated gray relational optimization of design parameters

The Taguchi method is the most popular statistical method that provides a robust experimental design. It provides the best combination of design input parameters at which the performance of the system optimizes (Kackar 1985). However, the Taguchi method is mostly used for single output characteristics optimization only. Therefore, problems start to materialize when more than one output characteristic to be optimized and the nature of those characteristics is of a conflicting type. For example, the one-output characteristic is to be maximized and the other one is to be minimized. On the other hand, Grey Relational Analysis (GRA) is gradually used for applications of multi-objective type. However, the robustness of the system is an issue if GRA is used stand-alone. Hence, it becomes necessary to integrate the GRA and Taguchi methods to get the advantages of both methods (Hussain, Khan, and Sarmah 2019; Xu et al. 2018).

In the present study, Taguchi-based GRA is used to determine the optimal combinations of design parameters that directly affect the performance of the system. The design parameters and their ranges [Table 3,4] have been decided based on previous studies (Bademlioglu et al. 2018 ; Saleh 2018).

It is necessary to choose an appropriate Orthogonal Array (OA) design for the application of Taguchi methodology. L_{27} [Table 5] is the appropriate orthogonal array model for six controlling parameters with three levels of each parameter, as reported in the literature (Bademlioglu et al. 2018).

Following steps are being taken to implement Taguchi integrated Grey relational analysis (Aslantas, Ekici, and Çiçek 2018):

Step 1: The first step involves the determination of the loss function called the Signal/Noise ratio (S/N). It is very useful in finding the deviation of the actual output response value to the desired one. Higher the better (HTB) S/N ratio [Equation 1] is used for beneficial output characteristics, while for non-beneficial output characteristics; lower-the-better (LTB) S/N ratio [Equation 2] is desired. Note that beneficial output characteristics are intended to maximize while non-beneficial output characteristics are desired to be minimized.

$$HTB = -10\log\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) \tag{1}$$

$$LTB = -10\log(\frac{1}{n}\sum_{i=1}^{n} 1/y_i^2) \tag{2}$$

The resulting decision matrix is given as:

$$D = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nm} \end{bmatrix}$$

where n is equal to several runs or different combinations of design parameters and m is the number of output characteristics, while y_i is the output response value of i^{th} run or combination.

Step 2: It involves the normalization of the S/N ratios of output characteristics data as output response values are of different units. Thus, normalization [Equation 3–4] is done to have a uniform set of data.

For beneficial output characteristics, normalization is done using Equation 3.

$$\mathrm{r}_{\mathrm{ij}}^{\mathrm{B}} = \frac{\mathrm{d_{ij}} - \min\{\mathrm{d_{ij}}, \mathrm{i} = 1 \ \ldots \ \ldots \ .\mathrm{n}\}}{\max\{\mathrm{d_{ij}}, \mathrm{i} = 1 \ \ldots \ \ldots \ .\mathrm{n}\} - \min\{\mathrm{d_{ij}}, \mathrm{i} = 1 \ \ldots \ \ldots \ .\mathrm{n}\}} \tag{3}$$

For non-beneficial output characteristics, normalization is done using Equation 4.

$$\mathrm{r}_{\mathrm{ij}}^{\mathrm{NB}} = \frac{\max\{\mathrm{d_{ij}}, \mathrm{i} = 1 \ \ldots \ \ldots \ .\mathrm{n}\} - \mathrm{d_{ij}}}{\max\{\mathrm{d_{ij}}, \mathrm{i} = 1 \ \ldots \ \ldots \ .\mathrm{n}\} - \min\{\mathrm{d_{ij}}, \mathrm{i} = 1 \ \ldots \ \ldots \ .\mathrm{n}\}} \tag{4}$$

The resulting normalized matrix is of the form given as:

$$\mathrm{R} = \begin{bmatrix} \mathrm{r_{11}} & \cdots & \mathrm{r_{1m}} \\ \vdots & \ddots & \vdots \\ \mathrm{r_{n1}} & \cdots & \mathrm{r_{nm}} \end{bmatrix}$$

Step 3: This step involves the determination of the Grey relation coefficient (γ_{ij}), a measure of the correlation between the normalized response value and the optimal one (γ_{ij}) given by Equation (5):

$$\gamma_{\mathrm{ij}} = \frac{\Delta_{\min} + \xi * \Delta_{\max}}{\Delta_{\mathrm{ij}} + \xi * \Delta_{\max}} \tag{5}$$

where Δ_{ij} is $= deviationsequenceofi^{th}runandj^{th}outputresponse$

$\Delta_{\min} = \min\{\Delta_{\mathrm{ij}}, \mathrm{i} = 1 \ldots\ldots\ldots\mathrm{n}, \mathrm{j} = 1 \ldots\ldots\mathrm{m}\}$ and $\Delta_{\max} = \max\{\Delta_{\mathrm{ij}}, \mathrm{i} = 1 \ldots\ldots\ldots\mathrm{n}, \mathrm{j} = 1 \ldots\ldots\mathrm{m}\}$

ξ is the distinguishing coefficient ($0 \leq \xi \leq 1$). Generally, the value of $\xi$is taken as 0.5.

Step 4: Grey relational grade (ε_i) is determined by taking mean of γ_{ij} corresponding to each output response characteristics [Equation 6]:

$$\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n} \gamma_{ij} \tag{6}$$

The highest value of ε_i is ranked one and the corresponding set of design parameters will be the one at which the output characteristics will be optimum.

Step 5: Grey relational analysis is a type of range analysis. A range analysis cannot detect fluctuations in output characteristics caused by changes in the levels of design parameters. In the last step, the analysis of variance test is done to overcome the above-said problem by investigating those input parameters that affect significantly the output of the system by calculating the *p*-values and F-values.

Sensitivity analysis

The sensitivity analysis helps in the analysis of how the change in the design parameter of an individual component affects its rate of exergy destruction, $EXD_{i,}$ and the rate of overall exergy destruction, EXDtot of the ORC-VCR system. It helps in the improvement in the performance of individual components on the overall performance of the system. This effect is quantified in terms of the coefficient of the structural bond (CSB) given by equation 7.

$$CSB = \left(\frac{d(EXD_i)}{d(EXD_{tot})}\right)_{P_i} \tag{7}$$

Where P_i is the design parameter relevant to the performance of the component. For example, the temperature of the ORC evaporator is relevant to its performance.

If CSB>1 then the reduction in the exergy rate of the component causes a decrease in the rate of overall exergy destruction of the ORC-VCR system otherwise not. It is required to design those components whose CSB is greater than 1.

Model validation

The thermodynamic model of the ORC-VCR system of this study is validated with that of reference (Saleh 2018) using the following input conditions: T_b = 373 K; T_c = 308 K; T_e = 273 K; $_{vcrs} = 1kg/s$; $\eta_p = 0.80$; $\eta_c = 0.75$; $\eta_{\exp} = 0.80$, and Hexane. Table 6 compares the results of the present study with that of reference (Saleh 2018) regarding η_{ex}, CPR and EPR. The results show that there is a marginal error of 3.42%, 1.5%, and 1.96% in the exergetic efficiency of the system, EPR, and CPR, respectively.

Results and discussion

The coding of thermodynamic analysis for the proposed system is done on MATLAB 2017b (Moler 2017) with the interfacing of Refprop 9.0 (Lemmon et al. 2010). The optimization has been carried out for a combined ORC-VCR system using Taguchi- gray relational analysis. The Signal/Noise ratios (Taguchi analysis) are being computed using Minitab 18.1 statistical software while gray relational analysis is done on Microsoft Excel version 2016. Moreover, the sensitivity analysis is done to evaluate the impact of the component performances on the overall performance of the ORC-VCR system. The design parameters are A: Type of working fluid (WF), B: ORC evaporator temperature (T_b), C: VCR evaporator temperature (T_e), D: Isentropic efficiency of ORC expander (η_{exp}), E: Isentropic efficiency of VCR Compressor (η_c) and F: condenser temperature (T_c). The output performance characteristics are exergetic efficiency of the combined ORC-VCR system (η_{ex}), expander pressure ratio (EPR), and compressor pressure ratio (CPR).

Signal/noise ratios

Table 5 summarizes the Signal/Noise ratios for each output performance characteristic for 27 combinations of design parameters. Each combination of design parameters consists of different levels of design parameters.

Thereafter, corresponding to each performance characteristic, the average S/N ratio for each design parameter corresponding to each level is computed.

Similarly, average S/N ratios are computed and given in table 7. Then, the contribution ratio (CR) of each design parameter to each output characteristic is computed.

The contribution ratio of working fluid type to the compressor pressure ratio is

$$CR_{wf,CPR} = \frac{Delta_{wf}}{Delta_{tot}}$$

where $Delta_{wf}$ is the difference between the maximum S/N ratio to the minimum S/N ratio of working fluid type while $Delta_{tot}$ is the summation of the delta of each design parameter for each performance characteristic. At the last rank of each system, the parameter is calculated depending upon the magnitude of its effect on output characteristics. All these types of data have been summarized in table 8.

Compressor pressure ratio (CPR) is affected by condenser temperature, F with a contribution ratio (CR) of 48.26%, followed by the type of working fluid, A (CR of 21.81%) and

VCR evaporator temperature, C (CR of 21.70%) while other design parameters have the least effect on CPR (Figure 3).

For the expander pressure ratio (EPR), the ORC evaporator, B is the most effective parameter (CR of 45.56%) followed by a condenser temperature, F (CR of 29.31%) (Figure 4). Type of working fluid, A comes at the distant third position with a CR of 15.70% while the rest of the parameters are least effective.

For exergetic efficiency (η_{exr}), condenser temperature, F is the most effective one (CR of 57.40%) followed by ORC evaporator temperature, B (CR of 18.98%), and isentropic efficiency of VCR compressor, E (CR of 12.32%) (Figure 5).

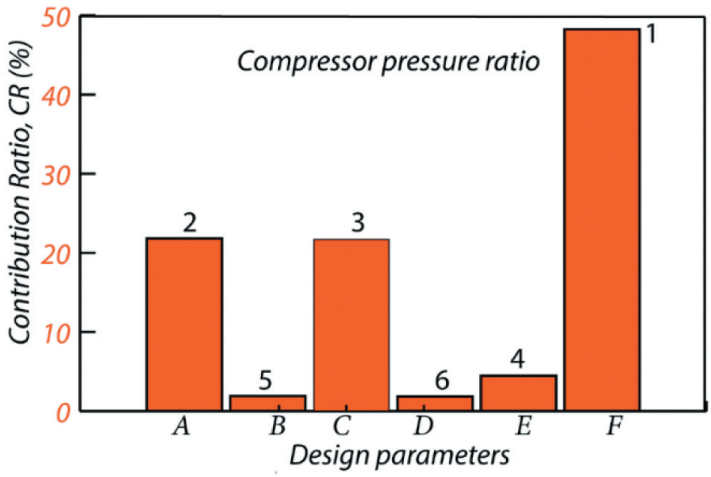


Figure 3. Contribution ratio v/s design parameters for compressor pressure ratio.

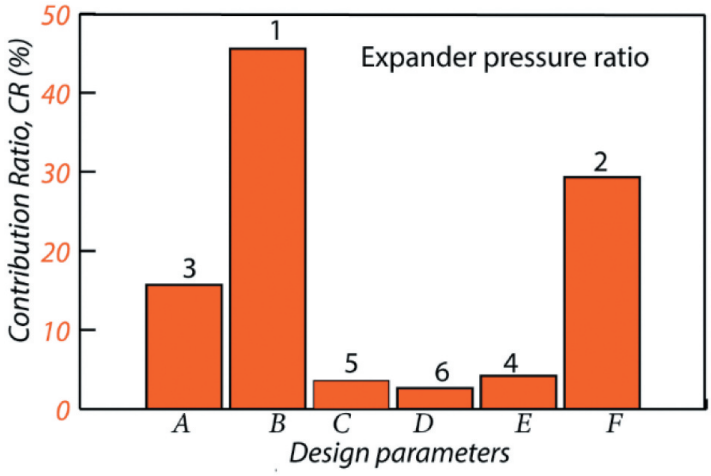


Figure 4. Contribution ratio v/s design parameters for expander pressure ratio.

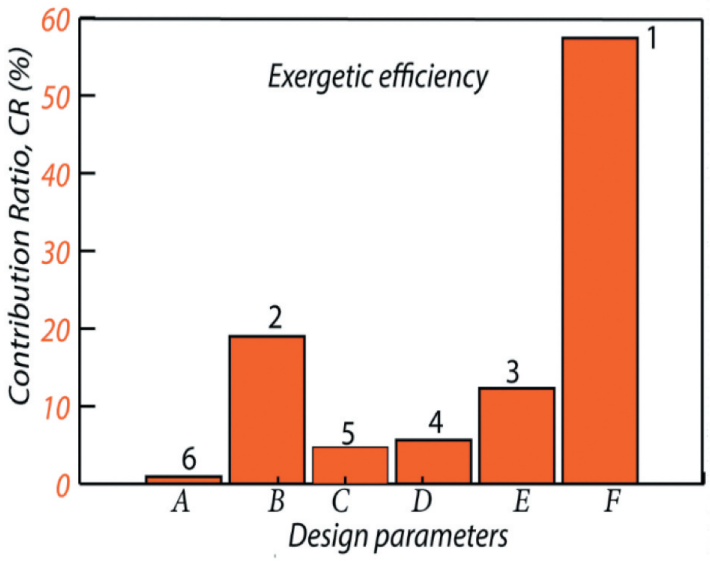


Figure 5. Contribution ratio v/s design parameters for exergetic efficiency.

Therefore, from the above analysis, one can gain insight as follows:

- Condenser temperature, F is the most important parameter for exergetic efficiency (CR of 57.40%) and CPR (CR of 48.26%).
- Similarly, the ORC evaporator temperature, B is for EPR with a CR of 45.56%. Moreover, its impact on exergetic efficiency is too considerable (CR of 18.98%)
- The type of working fluid, A contributes to the expander pressure ratio (CR of 21.81%) and compressor pressure ratio (CR of 21.70%). Thus, the type of working fluid is an important parameter in the sizing of the ORC expander and VCR compressor and ultimately on the cost of the system.

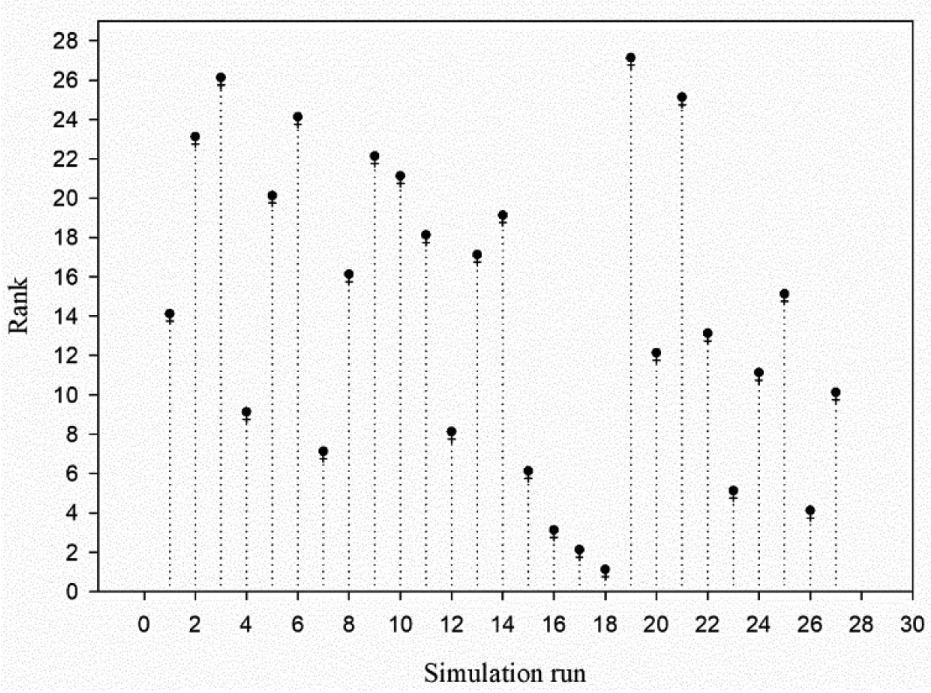


Figure 6. Rank v/s Simulation run.

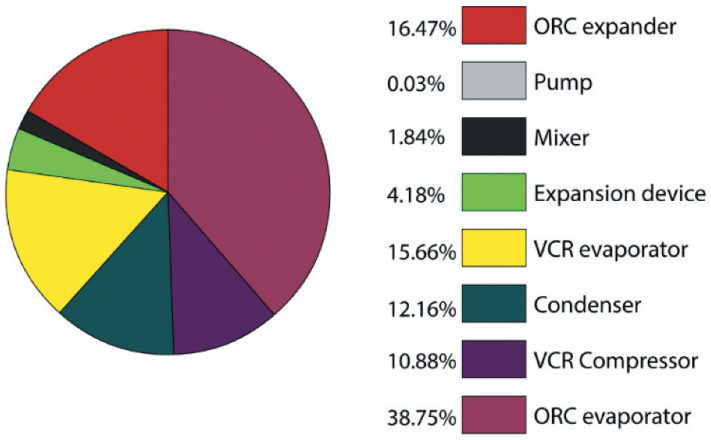


Figure 7. Contribution of components in the total exergy destruction rate.

Grey relational analysis

After computing Signal/Noise ratios of each performance characteristics [Table 5]. The next step is the determination of the gray relational coefficient, gray relational grade and then rank as per procedure stipulated in step 2 to step 4 in section 3. The corresponding data are given in table 8. Figure 6 shows the plot between simulation runs and their corresponding ranks.

The simulation run 18 is ranked at one. The corresponding combination of design parameters is $A_2B_3C_1D_2E_3F_1$. It means, the best combination of design parameters is working fluid of R602 type (A_2), ORC evaporator at a temperature of 373 K (B_3), VCR evaporator at a temperature of 268 K (C_1), isentropic efficiency of expander at 80% (D_2), isentropic efficiency of the compressor at 85% (E_3) and condenser temperature at 298 K (F_1). The system performance is optimum (EPR = 9.88, CPR = 4.38 and η_{ex} = 39.86%) at this combination of design parameters.

Thus, Taguchi integrated Grey analysis has provided the knowledge about most influencing design parameters and the combination of parameters at which EPR, CPR, and η_{ex} of the system is optimum.

Contribution of components to the overall exergy destruction rate, EXD_{tot}

The contribution of each component of the ORC-VCR system to the rate of overall exergy destruction at the optimal operating condition is shown in Figure 7. It is clear from the figure that major contribution to the overall exergy destruction rate is by 5 components such as ORC evaporator (38.8%) followed by ORC expander (16.5%), VCR evaporator (15.7%), condenser (12.16%) and, VCR compressor (10.88%). The significant finite temperature difference between the heat source and working fluid in the ORC evaporator is the main cause for its highest exergy destruction rate. The pump accounts for a negligible exergy destruction rate (0.03%) due to its handling of working fluid in a liquid state rather than a vapor or two-phase state.

Sensitivity analysis

The components that contribute to the overall exergy destruction rate significantly (analyzed in section 4.4) are considered for sensitivity analysis. For each component, its design parameter is varied (Table 3) while the other design parameters are kept at their optimal condition. For example, the isentropic efficiency is the design parameter for the ORC expander and it is varied from 0.75 to 0.85 for the curve plot in the sensitivity analysis. Such plots (Figure 8) help in the determination of slopes of the curves (or CSB values) by equation 7. Figure 9 depicts the variation of CSB of the components with the variation of design parameters. It can be concluded (Figure 9) that the CSB values of the ORC evaporator and the condenser are more than 3 indicating the strong impact of the

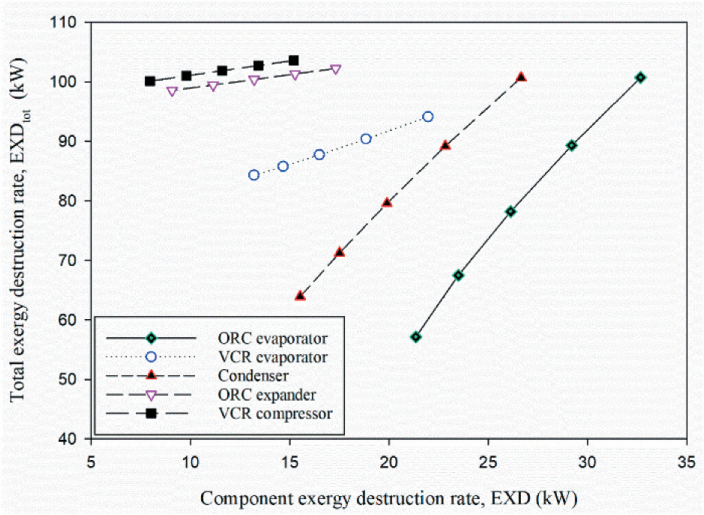


Figure 8. Total exergy destruction rate v/s exergy destruction rate of the component.

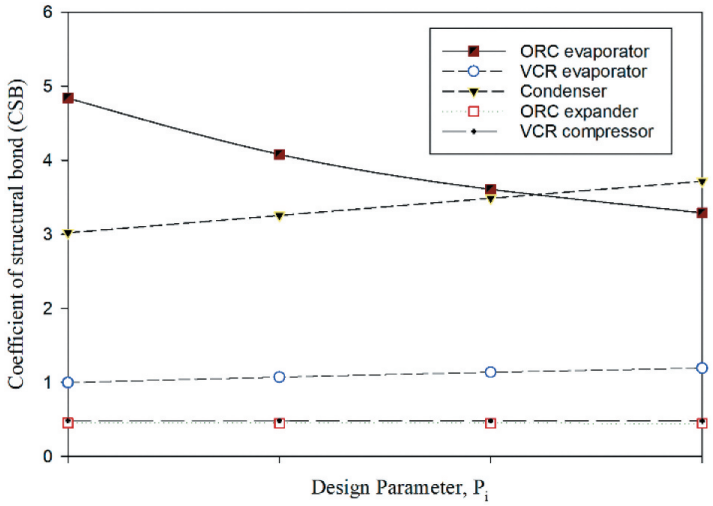


Figure 9. Coefficient of structural bond v/s design parameter of the component.

improvement in their exergy performances on that of the system. For the VCR evaporator, The CSB is close to unity and the reduction in its exergy destruction rate over the overall exergy destruction rate is small. However, the CSB values for the ORC expander and VCR compressor are less than one. It indicates that the reduction in their exergy destruction rates results in increased exergy destruction rates of other components that will ultimately increase the overall exergy destruction rate.

Conclusions

In this study, an ORC-VCR system is proposed that can utilize low-temperature heat sources to produce power and cooling. The conclusions drawn based on the results are as follows:

- The optimum design parameters are $A_2B_3C_1D_2E_3F_1$ i.e. working fluid of R602 type, ORC evaporator at a temperature of 373 K (level 3), VCR evaporator at a temperature of 268 K (level 1), isentropic efficiency of expander at 80% (level 2), isentropic efficiency of the compressor at 85% (level 3) and condenser temperature at 298 K (level 1).

- The optimal exergetic efficiency for the combined system is 39.86% whereas expander and compressor pressure ratios are 9.88 and 4.38, respectively.
- The compressor pressure ratio is mostly affected by condenser temperature (48.26%) followed by the working fluid (21.81%).
- For the expander pressure ratio, ORC evaporator temperature (45.56%) is the most effective parameter followed by the condenser temperature (29.31%).
- Condenser temperature (57.40%) is the most effective one for exergetic efficiency, followed by ORC evaporator temperature (18.98%).
- ORC evaporator contributes most, 38.8% to the overall exergy destruction rate at the optimal condition of the design parameters.
- The sensitivity analysis indicates that the performances of the evaporators of both ORC and VCR, and the condenser affect mostly the exergy performance of the system.

The outcomes of the study help design optimally ORC-VCR system powered by low-temperature heat sources and the development of the roadmap for ORC-based VCR systems in the future. In the future, a detailed thermodynamic analysis of the ORC-VCR system integrated with other renewable energy sources such as solar heat source is required to exploit the easily available solar energy.

Nomenclature		Subscripts	
$\dot{Q}$	Heat rate (KW)	comp	compressor
η	Efficiency	exp	expansion valve
W	Work (KW)	pump	pump
s	Specific entropy (kJ kg^{-1} K^{-1})	cond	condenser
.	Mass flux of working fluid (kgs^{-1})	mix	mixer
T_o	Ambient temperature (K)	turb	expander
h	Specific enthalpy (kJ kg^{-1})	evap	evaporator
EXD	Exergy destruction (KW)	tf	Total fluid
GRA	Grey relational analysis	b	ORC evaporator
CR	Contribution ratio		
GRG	Grey relational grade		

Notes on contributors

Ashwni pursuing Doctor of Philosophy in Department of Mechanical Engineering of Jamia Millia Islamia (A central university), Delhi, India. His Research area includes low- grade heat source power generation and refrigeration system.

Ahmad Faizan Sherwani is working as an Associate Professor in Department of Mechanical Engineering of Jamia Millia Islamia (A central university), Delhi, India. His research area includes solar energy, alternate refrigeration technologies, exergy analysis of thermal systems, and life cycle analysis. He has published plenty of research paper in reputed international and national journals. He has worked as reviewer in several reputed journals.

Deepak Tiwari working as Associate Professor in Mechanical Engineering Department of Bhagwan Parshuram Institute of Technology, Delhi, India. His research area includes low-grade heat source power generation.

Anil Kumar is Associate Professor and Associate Head in Department of Mechanical Engineering, with additional charge of Additional Coordinator-Centre for Energy and Environment, Delhi Technological University, Delhi, India. His research area includes; Energy Technology, Renewable Energy, Solar Energy Applications, Energy Economics, Heat Transfer, Natural Rubber Sheet Drying, and Environmental Issues.

References

Aslantas, K., E. Ekici, and A. Çiçek. 2018. Optimization of process parameters for micro milling of Ti-6Al-4V alloy using taguchi-based gray relational analysis. *Measurement: Journal of the International Measurement Confederation* 128:419–27. doi:10.1016/j.measurement.2018.06.066.

Aziz, F., M. S. Salim, and M.-H. Kim. 2019. Performance analysis of high temperature cascade organic rankine cycle coupled with water heating system. *Energy* 170:954–66. doi:10.1016/j.energy.2018.12.210.

Bademlioglu, A. H., A. S. Canbolat, N. Yamankaradeniz, and O. Kaynakli. 2018. Investigation of parameters affecting organic rankine cycle efficiency by using taguchi and ANOVA methods. *Applied Thermal Engineering* 145:221–28. doi:10.1016/j.applthermaleng.2018.09.032.

Bademlioglu, A. H., A. S. Canbolat, and O. Kaynakli. 2020. Multi-objective optimization of parameters affecting organic rankine cycle performance characteristics with taguchi-grey relational analysis. *Renewable and Sustainable Energy Reviews* 117:109483. doi:10.1016/j.rser.2019.109483.

Bu, X., L. Wang, and L. Huashan. 2014. Working fluids selection for fishing boats waste heat powered organic rankine-vapor compression ice maker. *Heat and Mass Transfer* 50 (10):1479–85. doi:10.1007/s00231-014-1350-0.

Coulomb, D. 2008. Refrigeration and cold chain serving the global food industry and creating a better future: Two key IIR challenges for improved health and environment. *Trends in Food Science and Technology* 19 (8):413–17. doi:10.1016/j.tifs.2008.03.006.

Duffin, E. 2019. India: Population by wealth 2019. *Statista*. https://www.statista.com/statistics/482579/india-population-by-average-wealth/ .

Goyal, A., A. F. Sherwani, and D. Tiwari. 2019. Optimization of cyclic parameters for ORC system using response surface methodology (RSM). *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*. doi:10.1080/15567036.2019.1633443.

Hussain, M. Z., S. Khan, and P. Sarmah. 2019. Optimization of powder metallurgy processing parameters of Al2O3/Cu composite through taguchi method with grey relational analysis. *Journal of King Saud University - Engineering Sciences*. doi:10.1016/j.jksues.2019.01.003.

IEA. 2018. World energy outlook 2018. Paris. https://www.iea.org/reports/world-energy-outlook-2018 .

Kackar, R. N. 1985. Off-Line Quality Control, Parameter Design, and the Taguchi Method. *Journal of Quality Technology* 17:176–88. doi:10.1007/978-1-4684-1472-1_4.

Kim, K. H., and H. Perez-Blanco. 2015. Performance Analysis of a Combined Organic Rankine Cycle and Vapor Compression Cycle for Power and Refrigeration Cogeneration. *Applied Thermal Engineering* 91:964–74. doi:10.1016/j.applthermaleng.2015.04.062.

Lemmon, E. W., I. H. Bell, M. L. Huber, and M. O. McLinden 2010. "Reference Fluid Thermodynamic and Transport Properties-REFPROP." Gaithersburg: National institute of standards and technology, standard reference data program,Gaithersburg.

Li, H., B. Xianbiao, L. Wang, Z. Long, and Y. Lian. 2013. Hydrocarbon working fluids for a rankine cycle powered vapor compression refrigeration system using low-grade thermal energy. *Energy and Buildings* 65:167–72. doi:10.1016/j.enbuild.2013.06.012.

Mago, P. J., L. M. Chamra, K. Srinivasan, and C. Somayaji. 2008. An examination of regenerative organic rankine cycles using dry fluids. *Applied Thermal Engineering* 28 (8–9):998–1007. doi:10.1016/j.applthermaleng.2007.06.025.

Moler, C. 2017. "MATLAB." The Mathworks, Inc.

.

Muhammad, U., M. Imran, D. Hyun, and B. Sik. 2015. Design and experimental investigation of a 1 KW organic rankine cycle system using R245fa as working fluid for low-grade waste heat recovery from steam. *Energy Conversion and Management* 103:1089–100. doi:10.1016/j.enconman.2015.07.045.

Nasir, M. T., and K. C. Kim. 2016. Working fluids selection and parametric optimization of an organic rankine cycle coupled vapor compression Cycle (ORC-VCC) for air conditioning using low grade heat. *Energy and Buildings* 129:378–95. doi:10.1016/j.enbuild.2016.07.068.

Patel, B., N. B. Desai, and S. S. Kachhwaha. 2017. Optimization of waste heat based organic rankine cycle powered cascaded vapor compression-absorption refrigeration system. *Energy Conversion and Management* 154:576–90. doi:10.1016/j.enconman.2017.11.045.

Saleh, B. 2016. Parametric and working fluid analysis of a combined organic rankine-vapor compression refrigeration system activated by low-grade thermal energy. *Journal of Advanced Research* 7 (5):651–60. doi:10.1016/j.jare.2016.06.006.

Saleh, B. 2018. Energy and exergy analysis of an integrated organic rankine cycle-vapor compression refrigeration system. *Applied Thermal Engineering* 141:697–710. doi:10.1016/j.applthermaleng.2018.06.018.

Salim, M. S., and M.-H. Kim. 2019. Multi-objective thermo-economic optimization of a combined organic rankine cycle and vapour compression refrigeration cycle. *Energy Conversion and Management* 199. doi:10.1016/j.enconman.2019.112054.

Wu, S. Y., H. Yang, L. Xiao, and L. Chun. 2017. Comparative investigation on thermo-economic performance between ORC and LiBr absorption refrigerating cycle in waste heat recovery. *Energy Procedia* 105:1446–53. doi:10.1016/j.egypro.2017.03.425.

Xu, F., S. Zhang, W. Kunying, and Z. Dong. 2018. Multi-response optimization design of tailor-welded blank (TWB) thin-walled structures using taguchi-based gray relational analysis. *Thin-Walled Structures* 131:286–96. doi:10.1016/j.tws.2018.07.007.

Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Simulation analysis of mixing quality in T-junction micromixer with bend mixing channel

Ranjan Prakash *, Mohammad Zunaid, Samsher

Department of Mechanical Engineering, Delhi Technological University, Delhi, India

ARTICLE INFO

Article history:
Received 2 January 2021
Received in revised form 26 February 2021
Accepted 7 March 2021
Available online xxxx

Keywords:
Passive micromixer
Chaotic advection
Mixing quality
Simulation

ABSTRACT

In this paper, an analysis of the mixing quality of micromixers with bend curvature is proposed by Computational fluid dynamics simulation. Species transport phenomena are described by a set of conservation equations that is the Continuity equation, Navier-Stokes equation, and Convection-Diffusion equation. Reynolds number is taken in the range of 30 to 500. Firstly the mixing quality of T-junction straight and T-junction with bend curvature are compared. At a higher Reynolds number, microchannel T junction with bend curvature gives higher mixing quality (approx. M_i =46%). The results show that as the Reynolds number increases, the mixing index value, as well as pressure drop value also increases. Further, the effect of bend's curvature is studied for better mixing quality. For this goal, three bend curvatures of the different radius of 1 mm, 2 mm, and 3 mm but with equal lengths were simulated. The results show that as the channel radius increases, mixing quality increases.
© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the 3rd International Conference on Computational and Experimental Methods in Mechanical Engineering.

1. Introduction

Microfluidics is used in the biochemical and biomedical sciences for laboratory research like sample preparation and analysis, DNA analysis, enzyme reactions, drug delivery, protein folding, etc. The main aim of microfluidics research is to develop two devices namely a micro-total analytical system and a lab on a chip. Micromixers are categories into two types: active and passive micromixers. Active micromixers use external powers such as Magnetic energy, Electrical energy, Thermal energy, Electrokinetic, Acoustic wave, etc. Passive micromixers use only fluid pumping energy. Passive micromixers depend on the design modification of the microchannel to achieve better mixing of the fluids. From the literature review, better mixing is obtained by either split and recombination mechanism or chaotic advection mechanism. Split and recombination mean split the mixing channel into two or more sub-streams and then rejoin them. SAR mechanism is used for small Reynolds numbers. A chaotic advection mechanism is generated by either fold/bend the flow or inserts obstacles in the way of the flow.

Mixing quality increases in chaotic advection based microchannel at high Reynolds numbers. In this context, Solehati et al. [1] proposed a T-type microchannel with a wavy form and develop a mathematical model for this micromixer. Wong et al. [2] proposed a microchannel T- type mixer for rapid mixing by experimental and CFD simulations. Wong et al. [3] analyzed the mixing process of micromixer in the cross-form having square grooves in the mixing channel. Afzal et al. [4] analyzed a convergent-divergent micro-channel by introducing a pulsatile flow and compared the mixing index of four other microchannel designs like Straight channel, Square wave, Zig-zag, Sinusoidal. Rahimi et al. [5] analyzed the mixing index of T-junction at various confluence angles and flow rate ratios. X. Chen et al. [6] study and analyzed the mixing quality of three micromixers square wave, multi-wave, and zig-zag by CFD simulation and experiment. Hossain et al. [7] designed Square wave, Zig-Zag, and Curve microchannel and compared mixing index calculation of them. X. Chen et al. [8] designed a square wave micro-channel with obstacles at a different plan. Vatankhah et al. [9] proposed spiral microchannel and compared with straight T-junction and also study of various parameters. Raza et al. [10] designed a short-length micromixer using split and recombination structures. Tokas et al. [11] designed and simulated T-junction with helical structure and compared the mixing performance with straight T- junction. Javad et al. [12] introduced nanofluids in spiral T-junction and compared mixing

* Corresponding author.
E-mail address: ranjanprakashmnnit@gmail.com (R. Prakash).

https://doi.org/10.1016/j.matpr.2021.03.172
2214-7853/© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the 3rd International Conference on Computational and Experimental Methods in Mechanical Engineering.

Please cite this article as: R. Prakash, M. Zunaid and Samsher, Simulation analysis of mixing quality in T-junction micromixer with bend mixing channel, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2021.03.172

Nomenclature

D	Diffusion coefficient	μ	Absolute viscosity of fluid
P	Pressure	ρ	Mass density of fluid
Mi	Mixing index	σ	Standard deviation
C	Concentration	Re	Reynolds number

index with straight T-junction. Tokas et al. [13] proposed a T-junction with a mixing channel in helical form and compared mixing quality results with a straight T-junction. Cetkin et al. [14] optimized Y- shape micromixer having one mother duct and two daughter ducts and also analyzed the effect of obstacles in the mother duct. Haghighinia et al. [15] studied T-junction microchannel in 2D and 3D and compared the simulation results of 2D and 3D geometry. Chen et al. [16] designed an E-shape microchannel using SAR and chaotic advection mechanism. Mondal et al. [17] designed raccoon and serpentine microchannel using experimental approaches. Ling et al. [18] designed a microreactor to synthesize nanoparticles using an experimental setup. Lobasov et al. [19] analyzed the mixing index in T-junction microchannel for different fluid properties like initial temperature, viscosity, density, and rheology of mixing fluids. Dauyeshova et al. [20] simulated a T-junction microchannel with obstacles using the LBM method. Quiroz et al. [21] designed a T-junction with inlets at different positions and studied the variation of mixing channel parameters on the mixing index. Rahmannezhad et al. [22] optimized grooved microchannel with obstacles. Chen et al. [23] optimized T-junction microchannel with cylindrical and square obstacles. Silva et al. [24] designed a T-junction micromixer with symmetrical and asymmetrical inlets and studied the effect of pulsed flow applied at inlets. Calado et al. [25] designed a T-junction microchannels of asymmetrical form at various heights and compared the mixing capacity of proposed micromixers. Balasubramaniam et al. [26] proposed a spiral micromixer and studied the effect of different cross-sections like trapezoidal, semi-circular, rectangular, and square. Dundi et al. [27] evaluated the swirl effect of mixing in a simple T junction microchannel. Shinde et al. [28] designed U shape pillar split and recombine micromixer and compared the mixing index results with other configurations. Ennen et al. [29] studied about the interaction of magnetic nanoparticles in microfluidic environments.

Based on the literature review, in this work T junction with bend curvature is proposed and simulated. The results of the mixing quality of T junction with bend curvature are compared with T junction straight. Then the effect of Reynolds numbers and different bend curvature radius (r = 1 mm, r = 2 mm, r = 3 mm) are also studied. Since the proposed geometry is based on a chaotic advection mechanism, it enhances the mixing quality. The importance of the present work is that as the curvature radius increases, the straight channel is converted into a curved or circular form. The simulation results show that a curved or circular form mixing channel is more efficient than a straight channel.

2. Micromixer design

T-junction straight and T-junction with bend curvature are shown in Fig. 1. Water flows into the microchannel from the right inlet, whereas water-dye flows from the left inlet. The inlet channels cross-sectional area is 100 μm × 100 μm, the mixing channel cross-sectional area is 100 μm × 200 μm, the total length of the inlets channel is 800 μm, and the mixing channel length is 3000 μm. For comparison between T straight and T bend, the length in the y-direction is taken equal for both cases.

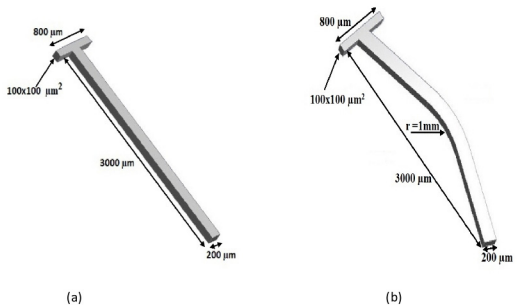


Fig. 1. Geometry of T-junction micromixers: (a) simple (b) bend curvature.

3. Simulation details

This paper contains, the calculation of the mixing index of micromixers using Fluent, rooted on the Finite Volume Method. This CFD package solves the continuity equation, Navier-Stokes equation, and Convection-Diffusion equation. In this case, flow is incompressible Newtonian but considering the condition of a no-slip boundary. The set of conservation equations are described in Balasubramaniam et al. [26] work and shown in equations (1) to (3) as:

$$\nabla.\mathrm{u} = 0 \tag{1}$$

$$\frac{\partial u}{\partial t} + \mathrm{u}.\nabla u = -\frac{1}{\rho}\nabla p + \frac{\mu}{\rho}\nabla^2 \mathrm{u} \tag{2}$$

$$\frac{\partial c}{\partial t} + (\mathrm{u}.\nabla)\mathrm{c} = \mathrm{D}\nabla^2 \mathrm{c} \tag{3}$$

Where, velocity vector u, viscosity μ, fluid density ρ, pressure p, concentration c, and diffusion coefficient D.

The mixing of two fluids water and water-dye having the same fluid properties were taken for flow simulation. This is the case of a single-phase mixture (liquid–liquid mixture). In this simulation laminar flow and species transport model was turned on. The absolute viscosity and mass density of water were taken as μ = 0.001 Ns/$m^2$and ρ = 1000 kg/m^3respectively, while the diffusion coefficient was taken as D = 10^{-10} m^2/s at 20 °C. For the continuity equation and momentum equation, the convergence criteria were taken as 10^{-8} and for the equation of species transport were taken as 10^{-10}. To solve the continuity and momentum equations, an upwind scheme of the second order was chosen and for the equation of species transport, a MUSCL scheme of the third order was chosen. Finally, for pressure–velocity coupling SIMPLE algorithm was applied. The same velocity was employed at the inlets and a zero-gauge pressure at the outlet. Mass fractions of 0 and 1 were fed at the inlets.

The standard deviation of the mixture was calculated at a particular plane of the micromixer in the direction perpendicular to the flow. The standard deviation of the mass fraction on a particular plane as described in Chen et al. [16] is defined as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(C_i - C_m)2} \quad (4)$$

In the above equation, the number of points taken on a particular plane N, the mass fraction at i sampling point c_i, and the mean value of the mass fraction c_{mean}. The mixing index calculation can be evaluated by the equation is given below:

$$\mathrm{Mi} = 1 - \frac{\sigma}{\sigma_{max}} \quad (5)$$

In the above equation, the standard deviation of the mixture at any cross-section σ and the maximum value of standard deviation σ_{max} (0.5 in our case). The value of the mixing index varies from 0 to 1 in which higher values of mixing index denotes higher mixing quality [9].

4. Results and discussions

The simulations were run in a T-type straight mixer and T-type mixer with bend curvature. The mixing quality of the T-type mixer with bend curvature is compared with the T-type straight mixer and also studied the impact of Reynolds numbers and impact of bend's curvature.

4.1. Grid independence analysis

For the accuracy of results of T junction microchannel, the effect of increasing computational cell has been well studied. Fig. 2 shows that the variation of the mixing index at the outlet by increasing the grid elements in million. As observed in Fig. 2, the mixing index value decreases by increasing the grid cells. The value of the mixing index becomes approx. same at higher grid cell 4.5 million and 5 million. There is only a 0.16% difference in mixing index value at grid cells 4.5 million and 5 million. The same tests were performed for T junction with bend curvature at radius 1 mm, 2 mm, and 3 mm.

4.2. Validation of computational work

In this section validation of computational work is performed. Results gained by the present simulation are compared to Dundi et al. [27]. Fig. 3 shows the comparison of results of the present work and Dundi et al. [27] results. There is a good agreement between the results of the present work and existing work.

4.3. Impact of geometry

Fig. 4 shows the mass fraction contours of microchannel T junction straight and T junction with bend curvature, in which bend structure is found to be better than a straight one. The mixing index at the outlet of the microchannel for the bend curvature is more (M_i = 0.459) than that of the straight one which value is (M_i = 0.280). The red and blue colors represent water solution with dye and pure water respectively. As the color is turned into green representing a mixture.

4.4. Impact of Reynolds number

The impact of Reynolds numbers on mixing index in both T junction straight and T junction with bend curvature was studied. For this purpose, the T junction straight and T junction with bend curvature having the equal length were simulated at various Reynolds numbers ranging from 30 to 500. Fig. 5 shows mixing index calculation versus Reynolds number for the T-mixer straight and T-mixer with bend curvature. As shown in this figure, mixing index in the T-mixer with bend curvature microchannel increases suddenly with increasing the Reynolds numbers.

As discussed above that the T-junction with bend curvature is more efficient than the T-junction straight. Since passive micromixers use only the fluid pumping energy, there is no need to use any external powers; the consumed energy in the device can be indicated by calculated pressure drop. Fig. 6 shows pressure drop in the T-mixer straight and T- mixer with bend curvature at different Reynolds number. As seen in this figure, the pressure drop in the T- mixer with bend structure is slightly higher. Chaotic advection based micromixers induce a secondary flow due to which causes a high-pressure drop in the micromixer. Further, the mixing index for T-mixer straight and T- mixer with bend curvature was calculated along the down-channel length. Fig. 7 shows the mixing index parameter along the down-channel length for the T-mixer straight and T- mixer with bend curvature. From the figure, as the down-channel length increases, the mixing index also increases.

4.5. Impact of bend's curvature

In this heading, the impact of bend's curvature is studied on the mixing quality of the micro-channel. For this goal, bend curvatures with equal lengths but with a different bend radius of 1 mm, 2 mm, and 3 mm were simulated. The calculated mixing index of the micromixers is shown in Fig. 8. This figure shows that as the bend radius increases, the mixing index of the micromixer increases.

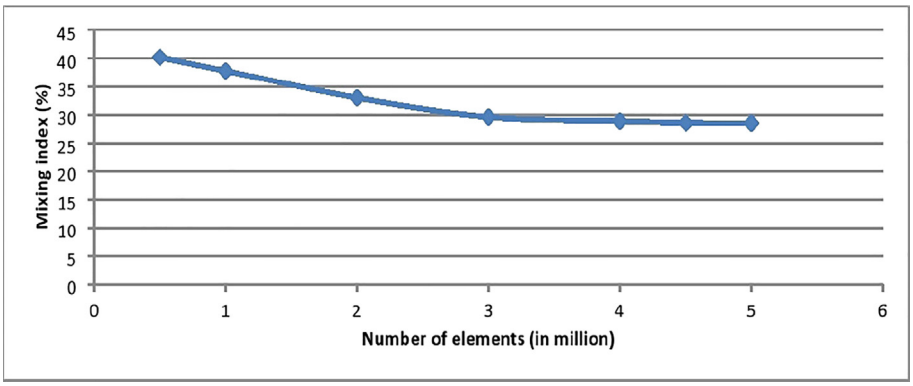


Fig. 2. Calculated mixing index with the number of the grid cells.

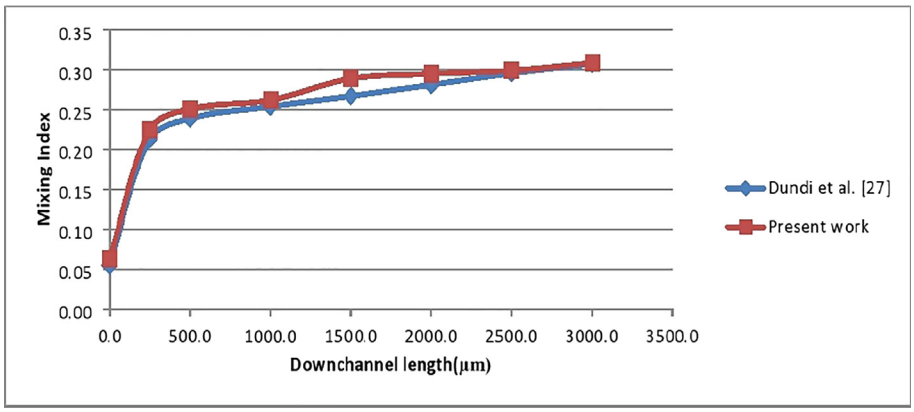


Fig. 3. Mixing index value comparison of present T junction and existing work with T junction.

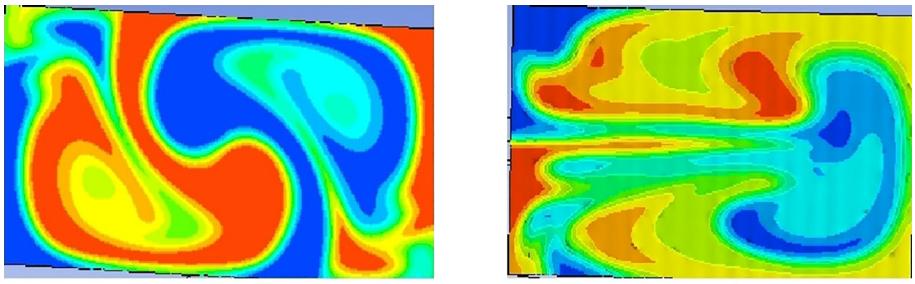


Fig. 4. Mass fraction contours of micro-channel T-junction at the outlet for (a) straight (b) bend structure at Re = 500.

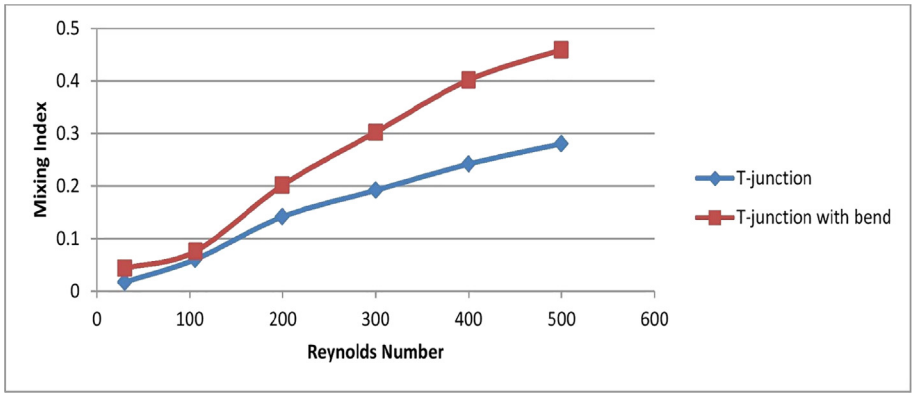


Fig. 5. Calculated mixing index for the T-type straight mixer and T-type mixer with bend curvature at different Reynolds number.

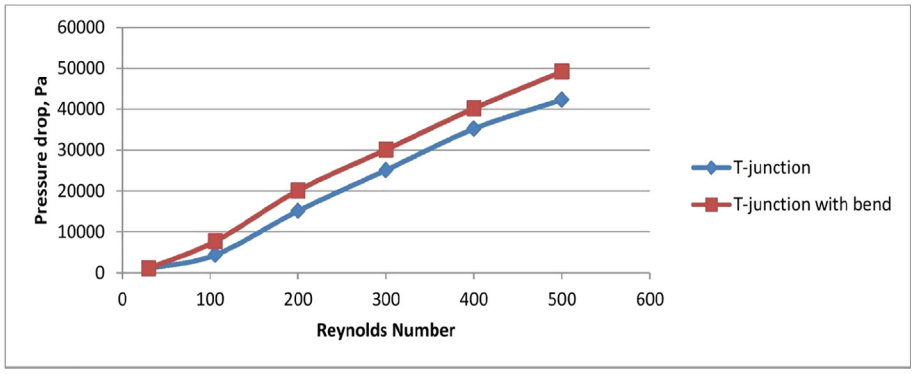


Fig. 6. Pressure drop in the T junction straight and T junction with bend curvature at different Re.

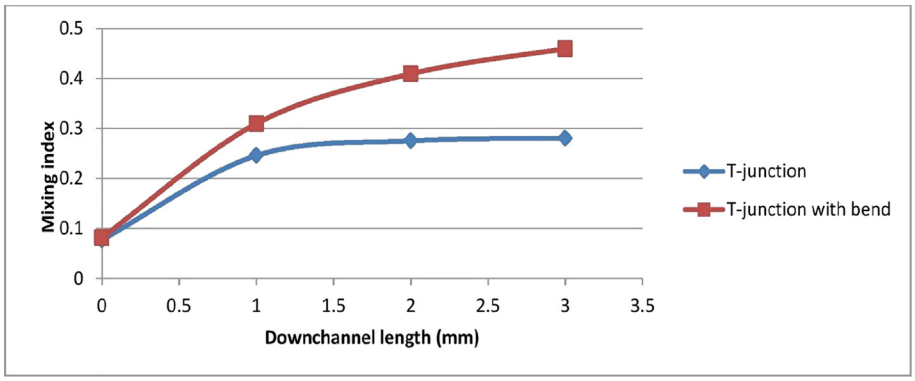


Fig. 7. Mixing index along the down-channel length for the fluids.

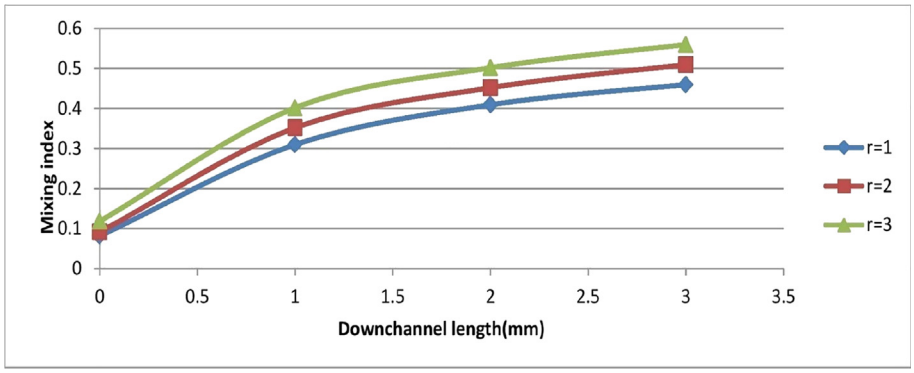


Fig. 8. Mixing index of three bend microchannel with different radius vs down-channel length.

5. Conclusions

In this work, the two micro-channel T junction with bend curvature and T junction straight have been simulated. The following points concluded that:

a) Microchannel with bend curvature gives a higher mixing index than that of a straight mixing channel due to chaotic advection mechanisms.
b) The pressure drop in T junction with bend curvature is slightly more than T junction straight.
c) As the mixing bend's curvature increases, mixing quality also increases.

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.

Acknowledgments

This work was supported by Delhi Technological University, Delhi.

References

[1] N. Solehati, J. Bae, A.P. Sasmito, Numerical investigation of mixing performance in microchannel T-junction with wavy structure, Computers Fluids 96 (2014) 10–19.
[2] S.H. Wong, M.C.L. Ward, C.W. Wharton, Micro T-mixer as a rapid mixing micromixer, Sens. Actuator B 100 (2004) 359–379.
[3] Seck Hoe Wong, Patrick Bryant, Michael Ward, Christopher Wharton, Investigation of mixing in a cross-shaped micromixer with static mixing elements for reaction kinetics studies, Sens. Actuator B 95 (2003) 414–424.
[4] A. Afzal, K.-Y. Kim, Convergent–divergent micromixer coupled with pulsatile flow, Sens. Actuators B 211 (2015) 198–205.
[5] M. Rahimi, M. Akbari, Mohammad Amin Parsamoghadam, CFD study on effect of channel confluence angle on fluid flow pattern in asymmetrical shaped microchannels, Comput. Chem. Eng. 73 (2014) 172–182.
[6] X. Chen, T. Li, H. Zeng, H.u. Zengliang, F.u. Baoding, Numerical and experimental investigation on micromixers with serpentine microchannels, Int. J. Heat Mass Transf. 98 (2016) 131–140.
[7] S. Hossain, M.A. Ansari, K-Y. Kim, Evaluation of the mixing performance of three passive micromixers, Chem. Eng. J. 150 (2009) 492–501.
[8] T. Li, X. Chen, Numerical investigation of 3D novel chaotic micromixer with obstacles, Int. J. Heat Mass Transf. 115 (2017) 278–282.
[9] P. Vatankhah, A. Shamloo, Parametric study on mixing process in an in-plane spiral micromixer utilizing chaotic advection, Anal. Chim. Acta 1022 (2018) 96–105.
[10] W. Raza, S. Hossain, K.Y. Kim, Effective mixing in a short serpentine split and recombination micromixer, Sens. Actuator B 258 (2017) 381–392.
[11] S. Tokas, M. Zunaid, Mubashshir Ahmad Ansari, Numerical investigation of the performance of 3D-helical passive micromixer with Newtonian fluid and non-Newtonian fluid blood, Asia-Pac. J. Chem. Eng. (2020).
[12] S.F. Javed, M. Zunaid, Numerical investigation inside spiral passive micromixer using nanofluid, Int. J. Mech. Prod. Eng. Res. Dev. (2020) 9825–9834.
[13] S. Tokas, M. Zunaid, Computational analysis of passive mixing in T-micromixer with Non-newtonian blood, Int. J. Mech. Prod. Eng. Res. Dev. (2020) 9889–9898.
[14] E. Cetkin, A.F. Miguel, Constructal branched micromixers with enhanced mixing efficiency: slender design, sphere mixing chamber and obstacles, Int. J. Heat Mass Transf. 131 (2019) 633–644.
[15] Arash haghighinia, Salman Movahedirad, Fluid micro-mixing in a passive microchannel: Comparison of 2D and 3D numerical simulations, Int. J. Heat Mass Transf. 139 (2019) 907-916.
[16] Xueye Chen, Jienan Shen, Numerical analysis of mixing behaviors of two types of E-shape micromixers, Int. J. Heat Mass Transf. 106 (2016) 593–600.
[17] Bappa Mondal, Sukumar Pati, Promod Kumar Patowari, Fabrication of wavy micromixer using soft lithography technique, Mater. Today Proc. 26 (2020) 1271–1278.
[18] Fiona W.M. Ling, Hayder A. Abdulbari, Chin Sim-Yee, Synthesis and characteristics of silica nano-particles using modified sol-gel method in microreactor, Mater. Today Proc.

[19] A.S. Lobasov, A.V. Minakov, Analyzing mixing quality in a T-shaped micromixer for different fluids properties through numerical simulation, Chem. Eng. Process. 124 (2018) 11–23.
[20] Bagdagul dauyeshova, Ernesto Monaco, Numerical simulation of diffusion process T- shape micromixer using Shan -Chen Lattice Boltzmann Method, Comput. Fluids, 167 (2018) 229-240.
[21] C.A. Cortes-Quiroz, A. Azarbadegan, M. Zangeneh, Effect of channel aspect ratio of 3-D T-mixer on flow patterns and convective mixing for a wide range of Reynolds number, Sens. Actuators B 239 (2016) 1153–1176.
[22] Javad Rahmannezhad, Seyed Ali Mirbozorgi, CFD analysis and RSM-based design optimization of novel grooved micromixers with obstructions, Int. J. Heat Mass Transfer 140 (2019) 483–497.
[23] Xueye Chen, Zhongyi Zhao, Numerical investigation on layout optimization of obstacles in a three- dimensional passive micromixer, Anal. Chim. Acta 964 (2017) 142–149.
[24] Jose Pedro Silva, Aires dos Santos, Viriato Semiao, Experimental characterization of pulsed Newtonian fluid flows inside T- shaped micromixers with variable inlets widths, Exp. Therm. Fluid Sci. 89 (2017) 249–258.
[25] Bernardo Calado, Aires dos Santos, Characterization of mixing regimes of Newtonian fluid flows in asymmetrical T- shaped micromixers, Exp. Therm. Fluid Sci. 72 (2015) 218–227.
[26] Lakshmi Balasubramaniam, Rerngchai Arayanarakool, Samuel David Marshall, Bing Li, Poh Seng Lee, Peter C.Y. Chen, Impact of cross-sectional geometry on mixing performance of spiral microfluidic channels characterized by swirling strength of Dean-vortices, J. Micromech. Microeng. 27 (2017) 095016.
[27] T. Manoj Dundi, V.R.K. Raju, V.P. Chandramohan, Numerical evaluation of swirl effect on liquid mixing in a passive T-micromixer, Aust. J. Mech. Eng. (2019).
[28] Anil B. Shinde, A.V. Patil, B.V. Patil, Enhance the mixing performance of water and ethanol at micro -level using geometrical modifications, Mater. Today Proc. (2020).
[29] Inga Ennen, Andreas Hutten, Magnetic nanoparticles meet microfluidics, Mater. Today Proc. 4 (2017) S160–S167.

2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)

Smart Robot for Collection and Segregation of Garbage

Akanksha
Electrical Engineering
Delhi TechnologicalUniversity
New Delhi, India
akankshaprasad8@gmail.com

Muskan Gupta
Electrical Engineering
Delhi TechnologicalUniversity
New Delhi, India
muskangupta2099@gmail.com

Madan Mohan Tripathi
Electrical Engineering
Delhi TechnologicalUniversity
New Delhi, India
mmtripathi@dce.ac.in

***Abstract*—This paper proposes a robotic bot to collect the litter thrown and segregate the garbage so as to reduce human effort and keep the environment clean. The proposed robotic bot autonomously identifies garbage and classifies it into biodegradable and non-biodegradable. The final collected garbage is, in a way, segregated automatically. The robot moves towards the garbage just like a rag picker does. When the robot's bins are full, a signalling system is created. This autonomous bot is capable of collecting, segregating and providing signals if the garbage bin is full. The yolov4 model was used to train a dataset of 10,000 images of both the categories and the model was able to detectthe test images with adequate efficiency of 99.66 percent, making it promising candidate for using as an alternative to human effort.**

***Index Terms*—machine learning, YOLOv4, object detection, garbage segregation, image classification, garbage collection, motors**

I. INTRODUCTION

With the advent of commercialization and modernization, the manufacturing and domestic processes are catching up the pace. They are developing faster than ever and the recent technologies have been a right hand in that. The more the production, the more the supply and the people would use more products. This means that the waste generated will also be more. The same means that the urban regions must adopt techniques which help in the best garbage disposal systems which allow the better utilization of resources and energy. Thankfully, the same can be achieved by the exact same technology we use to produce them. The amount of waste generated by each household, let alone a region, is increasing on a daily basis. The same can be accounted for by the increasing population and the increasing GDP. Just the proper disposal of waste would not ensure the complete process of garbage management [1].

The people in both urban and rural areas have a tendencyto litter on the roads and on other public spaces. Although, the various government initiatives have placed the bins in different areas, the people tend to make a more "convenient" choice that is, throwing the litter anywhere but in the bins. This problem creates several other issues regarding garbage collection and sorting. The workers responsible for picking up the garbage face usual problems like their hands getting cut by sharp objects, or other health problems. This problem can be solved by technology. At the same time, the same technology can be used to sort out the waste. Keeping all this in mind and also how important it is to distribute the waste according to biodegradable and non-biodegradable wastes, it is good to have an automated waste management tool or a robot which can collect the garbage and at the same time distinguish it as biodegradable and non-biodegradable waste and process it accordingly[2]. In a future perspective these types of robots can be used in a bigger picture so as to produce an automated smartcity.

The basic categorization of waste, which is widely accepted, is the two tier classification into biodegradable and non-biodegradable. The same has been done keeping in mind the environmental restrictions in eliminating the waste from the ground. The waste needs to be segregated in the aforementioned categories as it is easier to decide which waste has to be recycled. The 3 Rs (Reuse, Reduce and Recycle) play a major role in waste management. By segregating the waste into biodegradable and non-biodegradable categories, one can decide to make a wiser resource allocation [3].

This paper presents an innovative algorithm which is used to detect the garbage at a ground level and further segregate the same into two categories, biodegradable and non-biodegradable, with ten classes each. The algorithm focuseson the newly developed convolutional neural networks technique, You Only Look Once or YoloV4 which detects and classifies the garbage in just one frame. The model was further used to develop a bot which was able to perform garbage detection, segregation and collection. The results show that the model was able to detect and segregate the images captured with a mAP accuracy of 99.96 percent. The same was generated while training a model with 5000 images of each category and testing the same with the given dataset. The developed bot was able to perform the aforementioned actions with an accuracy of 80 percent. Furthermore, it was able to pick garbage weighing from 100-200 gramssuccessfully.

Rest of the paper is organized as follows. Section 2, presents the design of bot for garbage storage and segregation and section 3 discusses the motion control with wheels. The Image capture and centralization is discussed in section 4, whereas data preparation and processing is presented in section 5. Section 6 explains the working of the model and algorithm used for garbage detection and classification. Conclusion future scope is presented in section 7.

978-0-7381-3289-1/21/$31.00 ©2021 IEEE

2021 International Conference on Innovative Practices in Technology and Management (ICIPTM) | 978-1-6654-2530-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIPTM52218.2021.9388369

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:18 UTC from IEEE Xplore. Restrictions apply.

II. Design Of Bot For Garbage Storage And Segregation

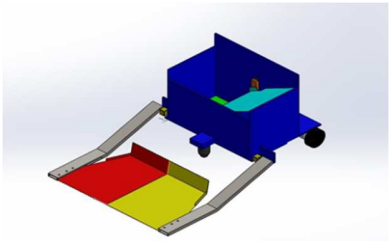


Fig. 1. The Bot designed using ANSYS tool for garbage collection

The bin is divided into two compartments using a wall in the middle such that one side is for biodegradable wasteand the other side is for non-biodegradable waste. A flap-type construction is made in the middle such that one side is hidden and inclination is provided when the garbage is being dropped in the other compartment. When the category of garbage is detected, it is thrown in the respective compartment. That is, if the garbage is biodegradable, the servo motor changes the direction of the flap such that the non-biodegradable side is covered and the garbage drops due to inclination in the compartment for biodegradable waste compulsorily. The same process is followed for non-biodegradable waste. The outer walls are constructed in a manner that their height enables the flaps to rest on them at an inclination. The motion of the flap is supported by a servo motor which is situated in the middle portion which gives it a turn based on the detected category of the garbage. The movement of arms comprises of four actions which are explainedbelow.

A. Inward Movement of Arms

When the bot reaches the garbage, the garbage is collected. This explains the movement of the arms of the bot. They move inward and coincide or overlap so that the garbage on the flap attached to the arms doesn't fall while picking it up. Even if the garbage falls, a steep slope catches the same. The steep slope is made behind the hand. The angle that the slope makes from the ground is greater than 90 degrees so that when lifting is done it is less than 90 degrees with the bin body. When garbage has to be picked, the arms move inward using two servo motors attached to the bottom of the body of the bin. The arms are attached in front of the body.

B. Upward Movement of Arms

This movement is against gravity; therefore, a high torque motor is used. It helps in efficient collection and picking of garbage. This action is completed by two servo motors. When once the garbage is lifted the arms move outwards so as to drop the garbage. This ensures that the garbage falls intothe bin. The arms move outwards by using the same servo motor which helps in collecting the garbage. Both the arms simultaneously move inward and outward to ensure that the garbage falls in thebin.

C. Downward Movement of Arms

After the arms move outwards and the garbage dumping process is completed, the arms are dropped down using the same servo motor which was earlier used for lifting. This hap- pens quickly so as to reduce torques that are being consumed due to gravity. This ensures an efficient operation.

III. Motion Control With Wheels

The bot has two arms which move with the help of two servo motors on each arm. A servo motor is used as an actuator which helps in precise movement of the device. In both the arms, one servo motor is used to give the horizontal movement and other for giving the vertical movement. When the bot reaches the garbage, first, the arms move inward by using a servo motor which provides horizontal movement such that garbage can be picked. When the garbage is picked, the same is lifted upusing another servo motor which is mounted above the former servo motor to give vertical inclination. This helps in a way that the garbage gets dropped in the appropriate bin. A servo motor is used as a component for motion of the wheels of the bot.

The L293D motor driver IC is used for controlling DC Motors with Raspberry Pi in the present work which works on the principle of Half-H Bridge. This allows the wheels (motors) to rotate in both clockwise and anticlockwise direction. The motor driver is very commonly used to drive two separate wheels (motors) with individual currents up to 600mA. The purpose of the wheels is to provide direction to the bot and move left, right or at any angle. The wheels are also capable of providing clockwise, anti-clockwise, straight and backward motion. When both the wheels (motors) move in clockwise direction, the bot moves forward where as when both the wheels move in anticlockwise direction, the bot moves backward. When the right wheel moves in clockwise direction and left is stationary, then bot moves such that it can take any angle in the left direction. However, when the left wheel moves in clockwise direction and right wheels are stationary, the bot can take any angle in the right direction. When the direction is being set based on input, the wheel's speed is reduced to 10 percent of its analog value so that it complies with the object position in the image which is found using yolov4 on raspberry pi. The garbage is detected using yolov4 trained custom model and bounding box is created. The coordinates of the bounding box are used to calculate the centroid of the image.
x coordinate of image = (top left x coordinate + bottom right x coordinate)/2 y coordinate of image = (topleft y coordinate + bottom right y coordinate)/2

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:18 UTC from IEEE Xplore. Restrictions apply.

The same is then compared with range set in image as to whether it lies in the particular center region or not so as to comply with the scope of the arms while picking. If it doesn't lie in that region, the bot eye checks whether it is in the right side or the left side of the center region of the image. The bot takes the turn based on its position of the object in the image such that the turn brings the image in center position w.r.t the bot's point of view.

That is, if the garbage detected has its centroid in the left with respect to the center region of the image, the bot takes a slight left turn, checks the centroid of the garbage detected in image again and if it is in the center region, it stops this process otherwise repeats the same process of turning and detection, moving towards the left until garbage is in the center. The bot takes a left turn for 1 microsecond and runs yolov4 custom detection model. It checks the coordinates, calculates the centroid and compares the values of centroids. If it is inthe center region, it stops turning left otherwise, it repeats the process. The same process is repeated if garbage's centroid lies on the right of the bot's centroid. Once the direction is set, the bot moves towards the garbage while performing real time object detection using custom yolov4 and calculating the percentage coverage after calculating the area:

$$area1 = I(x1 - x2) * (y1 - y2)I$$

Percentage coverage = area1*100/ (per known area of the image).

As soon as the coverage is greater than 70 percent the bot stops and it is ready to pick up the garbage. The coverage percentage was calculated keeping a track of the parameters including size, scope of hand, camera vision and distance from bot[4].

IV. Image Capture And Image Centralization

The raspberry pi cam is used to capture images.Then, on the captured image, objection detection using a custom trained yolov4 model is run [5]. It is located at a certain lower angle so that it captures the images of the object on the ground. When the image is captured, the Yolov4 model checks whether garbage is there or not. If the garbage is not there, the bot sleeps for 5 minutes. Then it resumes the same action. The range and the sleep function can be used to adjust the functional activity of the bot according to the preference which is taken as 5 minutes in this case. It identifies the type of object under the category of biodegradable or non-bio degradable and makes a check for finding the centroid of the coordinates of the object. If the object in the image does not lie in the center region of the image taken by the camera, it checks for the direction in the image taken to find out the side where the object lies and turns with 10 percent of its full speed [6].

During this process, the camera complies with real-time object detection using yolov4 until the garbage lies centrally [7]. This process has been referred to as centralization of image. This process of the bot has been inspired by the one a human being would do such that while picking the garbage, the same is in the reach of the hands. If in the image, captured by the camera, the object does not lie on the central region rather on the left, the object tries to move slightly towards left using the right front and the right back wheels. Whereas, if the object lies on the right side of the central position, the bot moves slightly right using the right front and back wheels unless the object aligns exactly with the central region. After the object is in the central region, the bot has to cover a certain distance so as to reach the object, pick it up and throw it in the bin. This work is done with the help of straight movement. The bot keeps moving towards the object until the object covers 70 percent of the image. Once the required coverage is found, the bot's arms reach the object and pick it up and then garbage is thrown in to the bin [8].

Two Infrared (IR) sensors are also installed on either side of the compartment, biodegradable and non-biodegradable, to indicate that the bin is full. When the garbage reaches a height and touches/in the proximity of the IR sensor, an LED blinks to indicate that the bin is full. The same is used to provide indication when a bin compartment is full. The bot stops performing whatever function is in motion, and only the LED isblinking.

V. Data Preparation And Processing

The dataset was prepared by clicking 300 images of all the 10 categories of both biodegradable and non-biodegradable types of wastes and then augmenting by cropping, contrast changing and applying filters (Gaussian, negative) to increase the dataset of images of each category to 500 images. One example is shown in fig. 2. The pictures were clicked with personal mobile devices and then clubbed together. The pictures were clicked from different angles and by placing the object in different environments.

Fig. 2. Images of biodegradable and non-biodegradable garbage

All the 500*10 + 500*10 = 10,000 images of the dataset were made of equal resolution (1280-960 pixel). All the images were renamed using a python program which

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:18 UTC from IEEE Xplore. Restrictions apply.

serialized the same. The images were annotated using the software, LabelImg which is an online tool that helps in easy annotation and dataset preparation. This software has a feature to upload the dataset of the images as a directory. Image annotation is a computer vision process which helps the machine to learn the various images and thereafter detect the images of the testing set once the algorithm is trained. LabelImg also helps the user to label the images and store the coordinates of the frames in a text file. The same folder, with all the text files, can be used as the custom dataset. Annotations are developed and saved as XML files in the PASCAL VOC format. Besides, it also supports YOLO format. The 10,000 annotated images along with the xml files are saved in a single folder and run in the YOLOv4 code to train the model on the Google colab platform.

VI. Algorithm For Garbage Detection And Classification And Working

Yolov 4 is completely based on the principle of CNN or Convolutional Neural Networks [9]. The principle isolates a particular image into different regions and envisioned a confined-edge box and probabilities of every region. The custom dataset was trained using Yolov4 on Google Colab platform for performing garbage identification and classification using the custom trained model on raspberrypi.

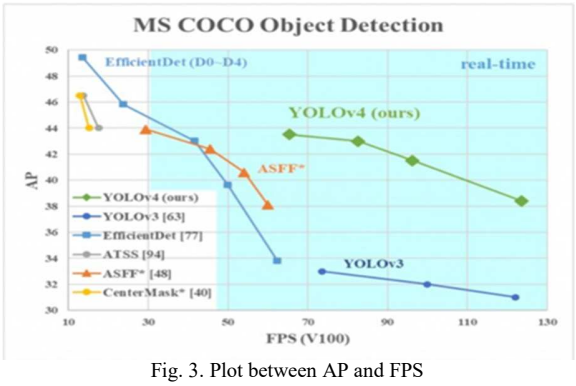


Fig. 3. Plot between AP and FPS

During the training of the dataset the MAP (Mean Average Precision) was used to obtain the accuracy of the algorithm. The Mean Average Precision is simply referred as Average Precision or (AP), which is used to measure the performance of the different PASCAL algorithms and models[10]. This gives us a single value known as mAP which is a quality measure for the proposed model. The YoloV4 model was used to train a dataset of 10,000 images of both the categories and the model was able to detect the test images with adequate efficiency. The fig. 3 shows that the proposed model was able to detect garbage and was able to classify the same on the basis of its type into biodegradable or non-bio degradable categories as 99. 66 percent accuracy was achieved.

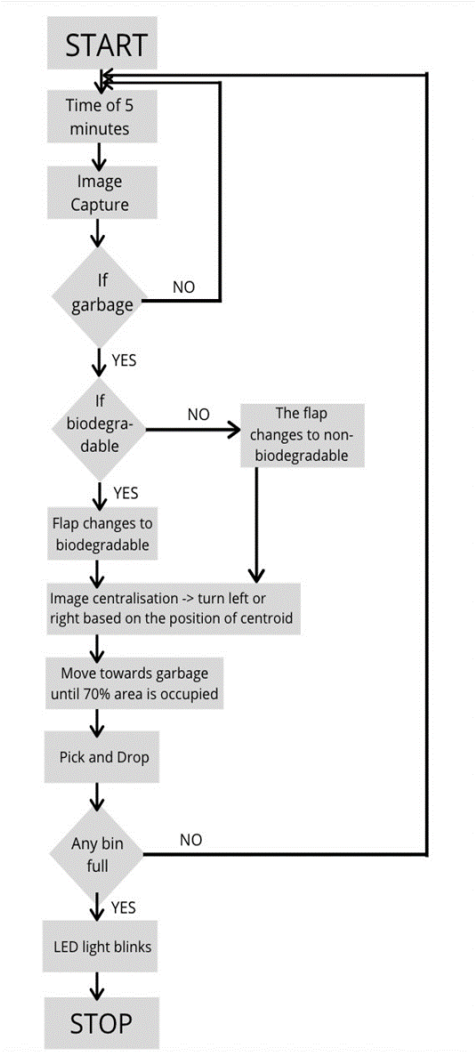


Fig. 4. Flow chart showing the proposed algorithm

The working of the proposed algorithm is shown in fig. 4 through a flow chart. The bot captures an image at a duration of every 5 minutes and runs object detection on a custom trained YoloV4 model to check if there is any garbage in the image capture dornot. If the garbage is detected, it is classified into respective categories. If the garbage collected is classified as biodegradable, the flap of the bin changes its direction with the help of a servo motor and exposes the bio degradable side of the bin. The inclination of the flap allows the collectedgarbage to fall into the appropriate compartment. The same process is repeated if the garbage is classified as non-biodegradable[11]. The next step is to do the image centralization such that the garbage is in the center of the image. The same is done using real time yolov4 garbage detection and taking the required left or right turn so that the image of garbage lies in the center region. The bin takes a soft left and right turn using the motor drivers and the dc motors attached in the wheels. The bot recognizes the coordinates and then

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:18 UTC from IEEE Xplore. Restrictions apply.

calculates the centroid of the object detected. If the object lies to the left side of the center region, the bot takes a slight left turn while running the real time custom object detection such that it brings the object in the middle region of the image. The same process is carried if garbage lies on the right of the center region. This process is repeated until the bot's centroid coincides with the centroid of theimage.

Now, the bot moves forward until the garbage covers 70 percent of the center region of the image. Then, it uses itsarms to pick and drop the garbage into the respective compartment and the hands are put back to their initial position. The bot then comes to the end of its garbage collection procedure and repeats the above cycle until any of its bin is full which is detected with the help of IR sensor installed at each side of thebins.

The proposed bot was able to recognize garbage of 10different types of non-biodegradable waste i.e. plastic bottles, wrappers of biscuits, chips and chocolates, milk packets, empty packs of frooti and appy, polythene bags and cans. It was also able to detect biodegradable waste like tissue, paper bags, paper plates and bowls, banana peels, pizza boxes, paper cups and newspaper. The bot detects the garbage and changes its alignment towards garbage. As the size of each object using which the dataset is created is almost the same, when 70 percent center region is overlapped by garbage, it performs pick and drop action. This process to identify the path and collect the garbage runs with an accuracy of 80 percent. This percentage is suitable for garbage with diameter between 7cm to 15cm. The construction of the arm is such that it is perpendicular while dropping the garbage in the bin. The process of collecting, picking and dropping the garbage ensures that the garbage falls in the correct container (biodegradable or non-biodegradable) of thebin.

The bot used real time custom object detection technique to align itself at an angle by the motion of the wheels based on the position of the object. The arms were efficiently able to pick garbage of weight 10-200 grams and drop the garbage in respective bins with falling in the right bin i.e.biodegradable or non-biodegradable. The flaps were designed in a way that the inclined angles provided support to the garbage while it gets into the appropriate bin. After the garbage is thrown into the bin, the bot stays there unless the next picture with the garbage is clicked and the garbage is detected.

VII. Conclusion And Future Scope

The custom trained YOLOv4 model was used to detect the garbage and classify it in the aforementioned categories with the mean average precision of 99.66 per cent. The collection of garbage including the process of direction setting i.e. image centralization, moving towards the garbage until the same covers the 70 per cent of the center portion was done with accuracy of 80 per cent. The bot was able to pick and drop the garbage dataset that is chosen i.e. garbage weight between 100 - 200 gm. The sensors were able to identify when the bin(s) were full. The bot stopped the operation, momentarily and resumed when the bins were emptied. The robot was able to perform the aforementioned tasks. The robot that has been designed can be further improved by increasing the dataset and hence the number of garbage items it could identify will increase. Also, in the future it can be designed to pick up garbageofupto4-5kgsweight.

VIII. Acknowledgement

This research was supported by IoT Lab of the Department of Electrical Engineering, Delhi Technological University. We thank Murli Manohar Pant, Delhi Technological University for helping in compiling the data set and elaborating on the process of augmentation and annotations. We also thank Bhavya Khatri, Delhi Technological University for helping us make the CAD model of the robot and guide us through the mechanics of the entire construction procedure.

References

[1] Gary Thung, Mingxiang Yang,"Classificationof Trash for Recyclability Status", https://www.semanticscholar.org/paper/Classification-of-Trash-for-Recyclability

[2] Kanchan Mahajan, Prof J.S. Chitode, "Waste bin monitoring system using integrated technology", International Journal of Innovative Research in Science, Engineering and Technology.

[3] S.S.Navghane, M.S. Killedar, Dr. V.M. Rohokale, "IoT based smart garbage and waste collection bin ",International journal of advance research in electronics and communication engineering, volume 5, Issue 5, May 2016

[4] Andrea, Zanella, Nicola But, Angelo Castellani, Lorenzo Vangelista, Michele Zorzi, "Internet of Things for Smart Cities", IEEE Internet of Things Journal (Volume: 1, Issue:1, Feb. 2014)

[5] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan MarkLiao,"YOLOv4: Optimal Speed and Accuracy of Object Detection", arXiv: 2009.07168v2

[6] YingWangandXuZhang,"Autonomous garbage detection for intelligent urban management",MATEC Web of Conferences232(4):0105 DOI:10.1051/matecconf/201823201056

[7] Satvilkar,Mandar(2018)ImageBasedTrashClassification using Machine Learning Algorithms for Recyclability Status. Masters thesis, Dublin, National College of Ireland.

[8] Siddhant Bansal, et al. AGDC: "Automatic garbage detection and collection",arXiv: 1908.05849

[9] M, Aghilan and M, Arun Kumar and TS, Mohammed Aafrid and A, Nirmal Kumar and S, Muthulakshmi, Garbage Waste Classification Using Supervised Deep Learning Techniques, IJETIE Vol. 6

[10] Yinghao Chu,Chen Huang,Xiaodan Xie, "Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling",November 2018 Computational Intelligence and Neuroscience 2018:1-9 ,DOI: 10.1155/2018/5060857

[11] 15 He K., Zhang X., Ren S., and Sun J., "Deep residual learning for image recognition," in Proceedings of Conference on Computer Vision and Pattern Recognition, Las Vegas Valley, NV, USA, June2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:55:18 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021).
IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4

2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) | 978-1-6654-1960-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICICV50876.2021.9388622

Software Defect Prediction using Dimensionality Reduction and Deep Learning

Massoud Massoudi
Software Engineering Department
Assistant Professor, Parwan University
Charikar, Afghanistan
Ph.D. candidate at DTU, Delhi, India
massoud.massoudi@hotmail.com

Nameet Kumar Jain
Department of Computer Engineering
Delhi Technological University, Delhi, India
nameetj3416@gmail.com

Pranay Bansal
Department of Computer Engineering
Delhi Technological University, Delhi, India
pranay.bansal200@outlook.com

Abstract: **Software Defect Prediction is a significant angle to guarantee programming quality. Deep Learning methods can likewise be utilized for the same. In this paper, we propose to extract a bunch of expressive features from an underlying arrangement of essential change measures using different dimensionality techniques namely PCA and Kernel PCA, and then train a classifier based on the extracted features using Decision tree and Artificial Neural Networks. We utilize five open source datasets from NASA Promise Data Repository to play out this comparative study. For evaluation, we have used three widely used metrics: Accuracy, F1 scores and Areas under Receiver Operating Characteristic curve. It is found that Kernel PCA along with Artificial Neural Network as a classifier outperformed all the other techniques.**

Keywords- Deep Learning, Artificial Neural Network, Principal Component Analysis, Kernel PCA, Decision Tree, Area under ROC curve.

I. INTRODUCTION

Software Defect Prediction is one of the most important activities of the Software Development Life Cycle. It detects the classes and modules which are error prone and require changes or multiple testing strategies. In order to build high quality softwares, defect prediction has become an important aspect as a lot of time and effort is put into software testing and its debugging otherwise. Defect prediction techniques are proposed to help prioritize software testing and debugging; they can recommend software components that are likely to be defective to developers. A lot of parameters are considered while predicting whether a software is buggy or not which include number of lines in the code, its complexity, the number of operators and operands used in the code and other factors.

We have considered a set of 22 initial features to predict whether the module is buggy. Our method of software defect prediction uses a newer classification based model which classifies any given instance into two classes: fault prone and non-fault prone. At first, already known examples, whose class we as of now perceive, are provided to the classifier. When the model is prepared, this is tested on the obscure occurrences and prediction performance of the technique is judged.

Deep learning is a new area and the most promising one in the machine learning literature, and has been adopted in a lot of research areas and has proven to be very effective, particularly in image processing and speech recognition[7].

ANN is a deep learning algorithm based on the working of biological neural networks. It has a number of layers and each layer has a number of nodes connected via weighted edges[8]. We here propose to use it as a binary classifier between the defective and the non-defective classes.

Decision Trees are a supervised learning method for classification and regression that works by applying classification or regression based on a particular function on a labeled training set. It works based on Greedy Algorithms. Here we propose to use Decision Trees as our second binary classifier.

PCA and Kernel PCA are ways used for dimensionality reduction[15][16]. PCA is an unsupervised linear transformation technique. On the other hand, KPCA is a non-linear form of PCA i.e. an extension to PCA that uses kernel methods[18].

The dataset consists of 22 initial features which can result in our model to overfit on our training dataset and perform poorly on other data. So, dimensionality reduction was carried out on our dataset before training the model on it using the classifiers namely the ANN and Decision Trees.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

On the basis on what classifier is used along with what dimensionality reduction method 4 different models were constructed which are as follows:

- PCA + Decision Tree
- PCA + ANN
- KPCA + Decision Tree
- KPCA + ANn

For evaluation, we have used three metrics: Accuracy, F1-scores and Areas under ROC curve. Classification accuracy alone can at times mislead and hence the other two metrics have also been considered. F1 scores are a weighted average of Precision and Recall taking into consideration both False Positives and False Negatives. Basically, it is the harmonic mean of the two. ROC curve analysis can be considered as a complete report of sensitivity and specificity[19]. The major analysis is done on the basis of the area under ROC curves. This is because, in software defect prediction kind of models, it is more important to predict the defective instances correctly than to predict the non-defective instances incorrectly.

DATASET DESCRIPTION:

Open source datasets are easily available online. The five datasets used for this project were taken from NASA Promise Dataset Repository namely kc1, kc2, pc1, cm1, jm1 each having no missing values and 22 attributes that come from McCabe and Halstead features extractors.

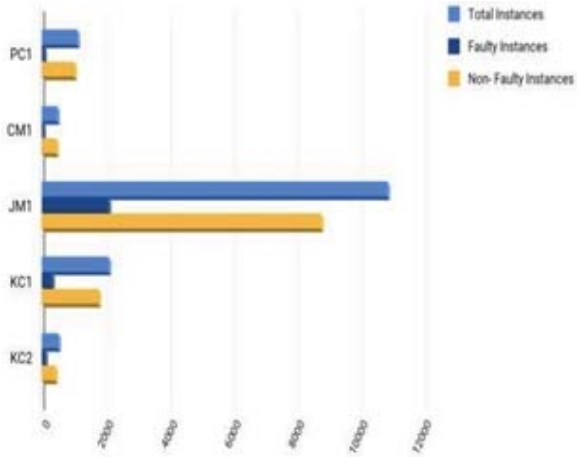


Fig. 1. Dataset Characteristics

Dataset	Language used	Total Instances	Defective Instances	Non-Defective Instances
PC1	C	1,109	77	1,032
CM1	C	498	49	449
JM1	C	10,885	2,106	8,779
KC1	C++	2,109	326	1,783
KC2	C++	522	105	415

Fig. 2. Dataset Characteristics

Feature Name	Description
McCabe's line count of code	It counts the lines of code in module.
McCabe "cyclomatic complexity"	It indicates complexity of the module on basis of number of linearly independent paths.
McCabe "essential complexity"	It indicates the extent to which a flowgraph can be reduced.
McCabe "design complexity"	It indicates cyclomatic complexity of its reduced flowgraph.
Halstead total operators + operands	It gives the count of operators and operands used in the module.
Halstead "volume"	It measures the product of length and log of vocabulary on base 2.
Halstead "program length"	It indicates the length of the program.
Halstead "difficulty"	It is related to the difficulty of the program to write or understand. Also computed as reciprocal of length.
Halstead "intelligence"	It determines amount of intelligence presented in the module. (Length * Volume)
Halstead "effort"	It translates into actual coding time.
Halstead	It is a base Halstead measure.
Halstead's time estimator	It evaluates the testing time of C/C++ codes.
Halstead's line count	It indicates the numbers of lines in the code.
Halstead's count of lines of comments	It indicates the number of lines of comments.
Halstead's count of blank	It gives the count of blank lines present in the module.
IOCodeAndComment	It gives the lines of code and comment in the module.
unique operators	It counts the total number of distinct operators in the module.
unique operands	It counts the total number of distinct operands in the module.
total operators	It counts the total number of operators in the module.
total operands	It counts the total number of operands in the module.
branchCount of the flow graph	It gives the count of branches in the flow graph.
defects	It indicates whether the module has any reported defect or not.

Fig.3. Description of the features present in the dataset

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

II. LITERATURE SURVEY

P. Deep Singh and A. Chug[20] examined the most famous and broadly utilized Machine Learning algorithms - ANN, PSO, DT, NB, LC and came to a conclusion that Linear Classifier performed better than the other algorithms used. Datasets utilized in this exploration were similar to what we used and were taken from an NASA Promise data repository. Classification is performed on the seven datasets and approved utilizing 10-fold cross validation.

Peng He et al. conducted a study on software defect prediction with a metric set extracting the top-k metrics[21]. Research was conducted using the data available on open source PROMISE repository. They came to a conclusion that the result of the top-k metric subset had better outputs compared to the initial dataset which had numerous metrics. In simple words, the simplified dataset helped the model train well.
Following the above idea, Instead of extracting the top-k metrics, we used dimensionality reduction to extract the top-6 features of the dataset.

Grishma BR et al. started by investigating the main reason for fault occurrence and their different types in various phases of SDLC. They did so by applying different clustering techniques like agglomerative clustering, COBWEB, k-means, density based scan, expectation maximization and farthest first. To the surprise finally they concluded that the k-means even after being a simpler algorithm outperformed all other techniques[22].

Anuradha Chug et al. focused more on defect prediction using various supervised-classification learning algorithms and unsupervised-clustering learning algorithms. The performance metrics that they used were similar to what we have used in our paper. Among all other classification algorithms Random Forest algorithm was highly accurate on one of the datasets and also yielded highest value in different performance metrics. Amongst all unsupervised algorithms k-means algorithm trained extremely fast and clustered the least incorrect instances[23].

Shamsul Huda et al. [24] researched that training the defect prediction model on the datasets with a large number of dimensions was a time-consuming process. In this research two novel hybrid SDP models are used for identifying the metrics. They used different datasets and classification algorithms such as SVM and ANN. The results were measured and compared by using AUC and MEWMA.

S. Choirunnisa, B. Meidyani and S. Rochimah[25] used the datasets from the NASA repository. They established that if different versions of the same software have different numbers of classes, then this will affect the correctness of predicting whether the software is defective or non-defective. They researched that some learning algorithms are used to identify possible decreases in the number of classes of program modules, thus affecting the optimum utilization of resources. To improve the correctness of the SDP model, they proposed a new framework to handle the class imbalance.

III. PROPOSED METHOD

The dataset used for Software Defect prediction in the project, is taken from NASA Promise Repository. All the 5 data sets have 22 attributes, though each having a different number of instances. ANN and Decision Tree classifier is used to make the model learn from the test set and then the model is tested on the training set and the performance measures are calculated. However, having so many attributes and instances can lead the model to overfit. Hence, we first reduced the dimensionality[9] of the data to a set of 6 cumulated features using 2 different techniques and then trained the model using ANN and Decision Tree classifier. A detailed comparison was then made based on the performance metrics that include Accuracy, F1-Scores and Area Under the Receiver Operating Characteristics (ROC)[10].

Following are the algorithms used for Dimensionality Reduction, along with a brief description:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

A. Principal Component Analysis (PCA)

The PCA algorithm is applied such that it extracts 6 new independent features that explain most the variance of the dataset, regardless of the dependent variable. Since the final class of each instance is not considered while turning the data into a low dimensional one, hence it is an Unsupervised Model[15].

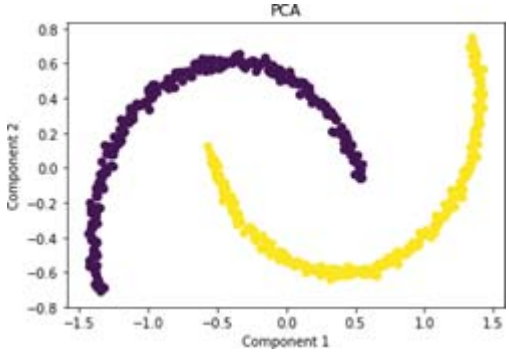


Fig. 4. Projection by PCA

B. Kernel Principal Component Analysis (KPCA)

Standard PCA works well only when the data is linear whereas if we have structures which are multi-dimensional, then it does not work well[6]. Gaussian kernel PCA reveals more complex hidden structures of the data than standard PCA. Gaussian RBF, Polynomial, Hyperbolic Tangent are some popular Kernel functions. We leveraged the Gaussian RBF kernel function to reduce the dimensionality of our data set[17].

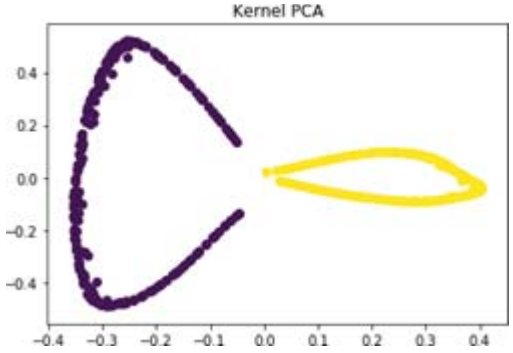


Fig.5. Projection by Gaussian KPCA

Following algorithms are used for Binary Classification:

A. Decision Tree Algorithm

The Decision Tree algorithm has been used as a classifier. Entropy of a dataset measures the impurity of the dataset. The most critical aspect of this algorithm is the attribute selection method employed at each node of the tree. The algorithm works on principle of greediness i.e., it looks for the solution that appears to be best at the moment without looking at the picture at large. The Decision Tree algorithm uses the Information Gain, which calculates the reduction in entropy or gain in information, to split the data set using a particular attribute. The algorithm is advantageous as it does not require much data cleaning and is not influenced by outliers and missing values to a fair extent[13].

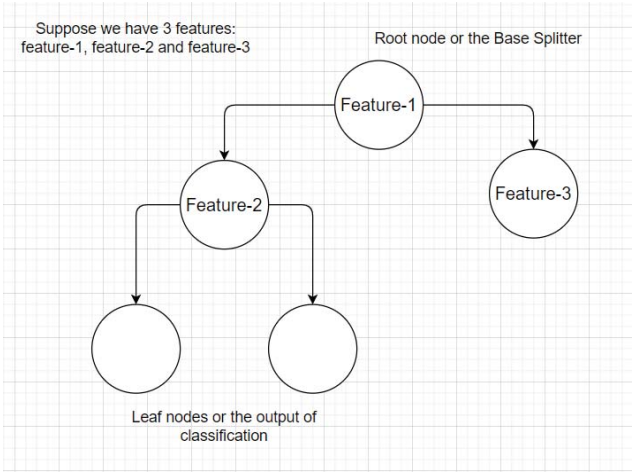


Fig.6. Decision Tree Skeleton

B. Artificial Neural Networks

ANN is a deep learning algorithm which resembles the working of human neural networks. It has a lot of layers and each layer has a lot of nodes connected via weighted edges[8]. We here propose to use it as a binary classifier between the defective and the non-defective classes. The Network here we used consisted of a single hidden layer consisting of 3 nodes which get their input from the first layer which consists of 6 nodes (the ones derived from the initial 22 features after dimensionality reduction) resulting into a final class.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

Following are the various Performance Measures used:

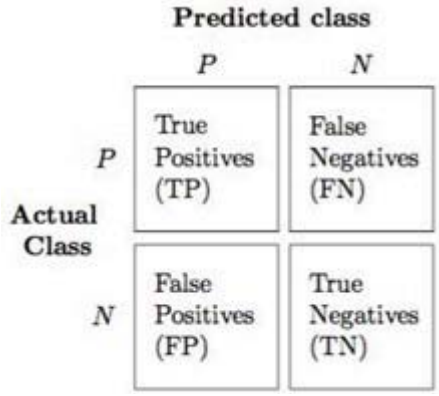


Fig.7. Confusion Matrix

A. Accuracy

Accuracy is calculated by dividing the correctly predicted instances of the test set with the total number of instances of the test set.

Accuracy = (TP + TN) / (TP + FN + FP + TN)

B. F1 Score

The F1 score is equivalent to a weighted average. A measurement of the precision and recall where the best possible score is at 1 and the worst possible score is at 0. The contributions of precision and recall to the F1 score are statistically equivalent. The formula for the F1 score is established.

F1 = 2* (Precision + Recall) / (Precision + Recall).

C. Area Under the Curve (AUC)

The performance of the predicted models was assessed by plotting the Receiver Operating Characteristics (ROC) curve and evaluating the area under the curve. ROC curve is a successful strategy for assessing the quality or execution of predicted models.

For validation, we have utilized the hold out cross validation method to validate the informational collection. Since all the informational indexes utilized had a significant enormous number of occasions, the training set and test set were isolated in the proportion 3:1. The preparation set was utilized to prepare the classifier and afterward the model was approved on the test set[12].

IV. EXPERIMENTAL ANALYSIS

Confusion matrix was obtained by applying various dimensional reduction techniques and training the dataset by Decision Tree classifier. Accuracy and F1 score thus obtained are following:

	PCA+DT	PCA+ANN	KPCA+DT	KPCA+ANN
Cm1	0.76	0.76	0.76	0.84
Jm1	0.72	0.74	0.74	0.72
Kc1	0.75	0.755	0.76	0.77
Kc2	0.87	0.827	0.85	0.84
Pc1	0.80	0.86	0.80	0.78

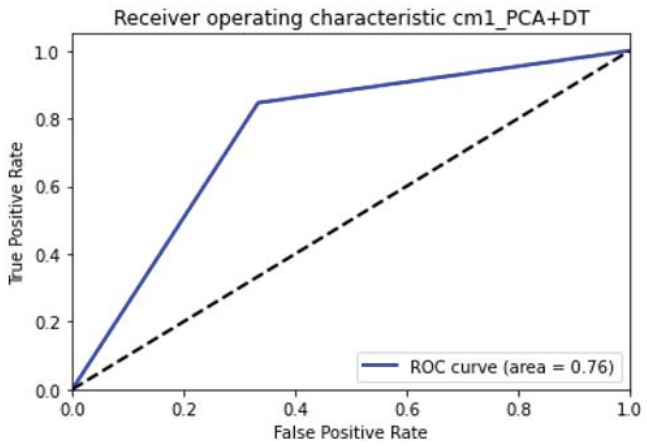
TABLE 1: ACCURACY FOR EACH TECHNIQUE

	PCA+DT	PCA+ANN	KPCA+DT	KPCA+ANN
Cm1	0.76	0.76	0.76	0.83
Jm1	0.72	0.74	0.75	0.72
Kc1	0.75	0.755	0.76	0.78
Kc2	0.87	0.827	0.85	0.83
Pc1	0.80	0.85	0.82	0.79

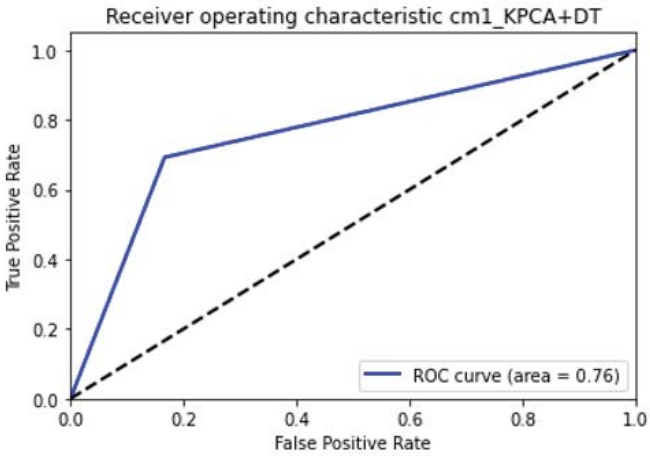
TABLE 2: F1 SCORES FOR EACH TECHNIQUE

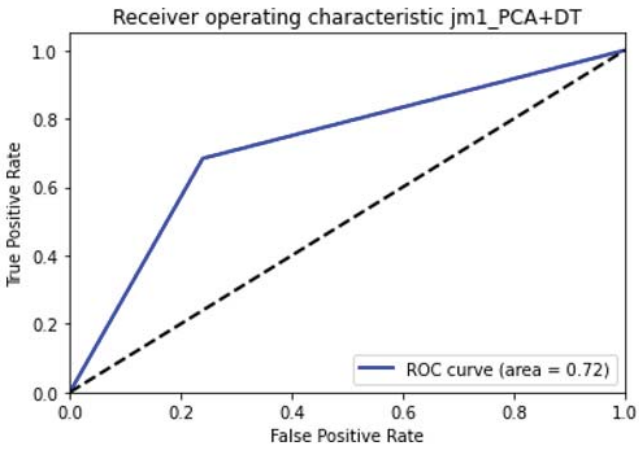
Next, we plotted the ROC Curves and calculated the AUC to check the validity of our model. Following are the ROC curves for the respective confusion matrices.

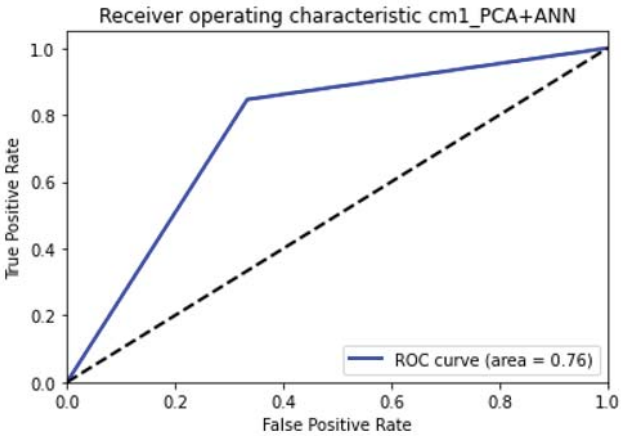
Note: The headings of the graphs symbolize the algorithm and the dataset. For eg: pc1_KPCA+DT symbolizes KPCA and Decision Tree on the pc1 dataset.

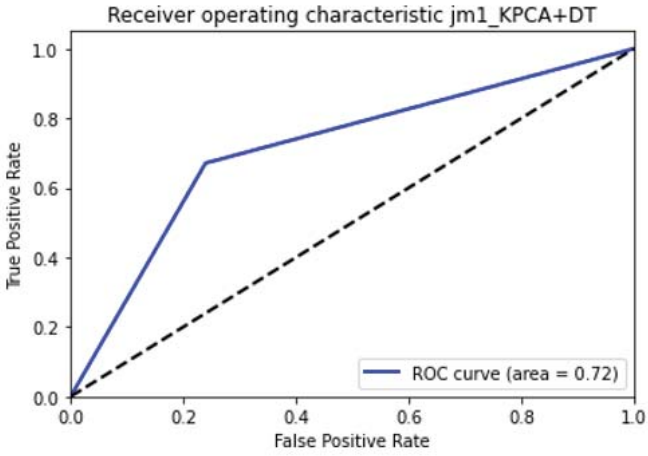


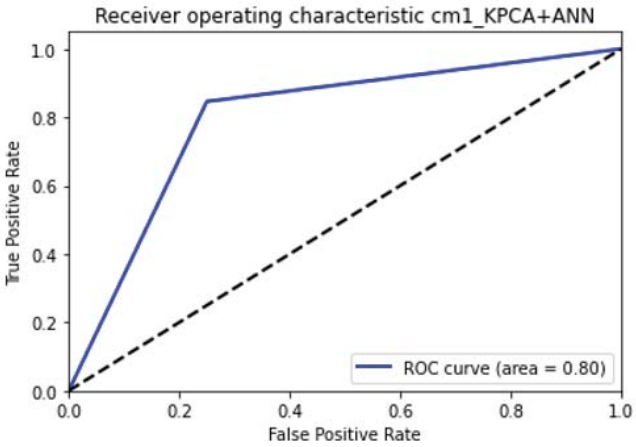
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

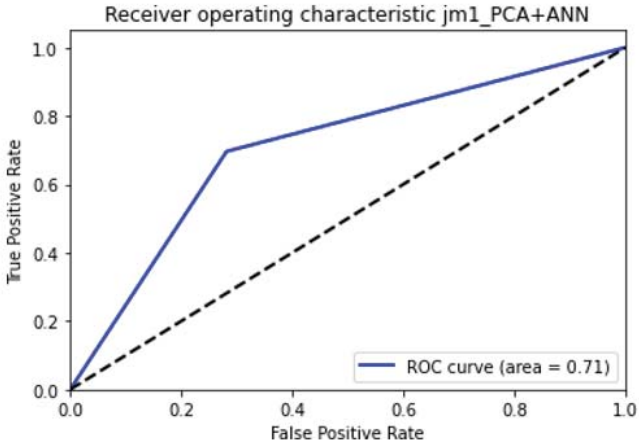




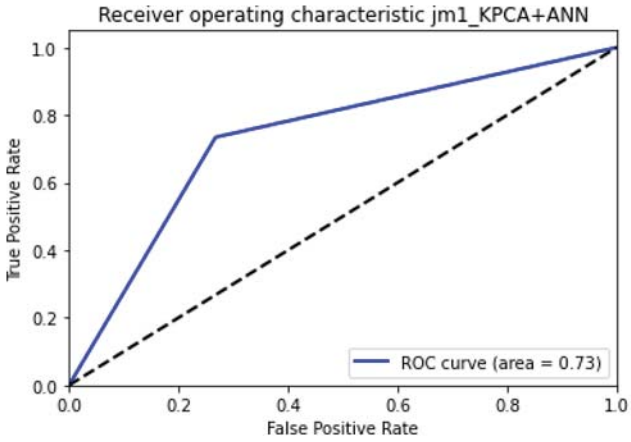


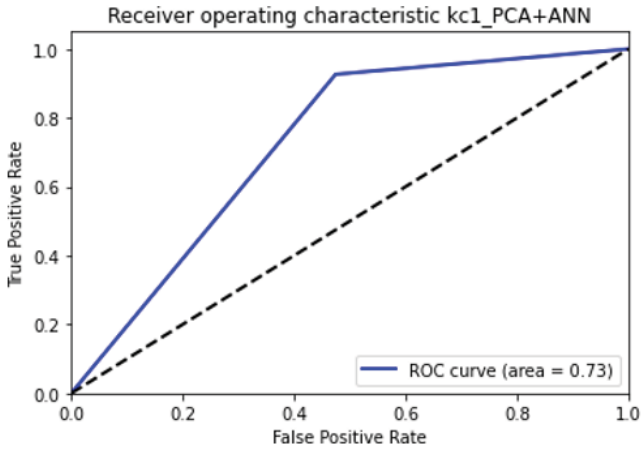


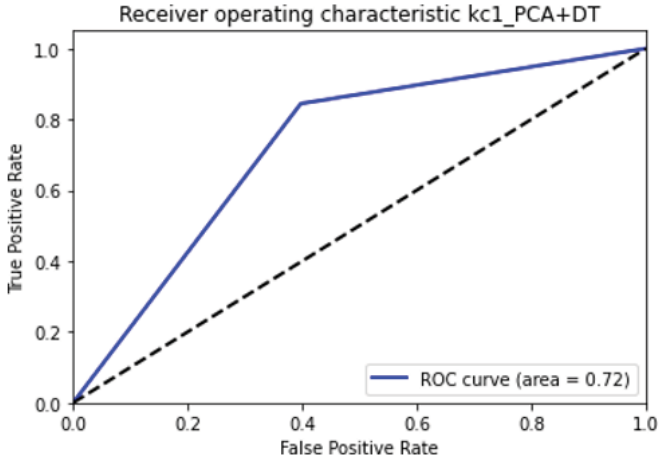


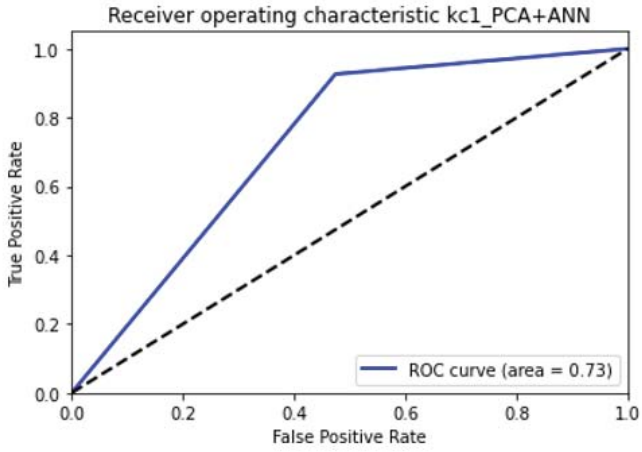


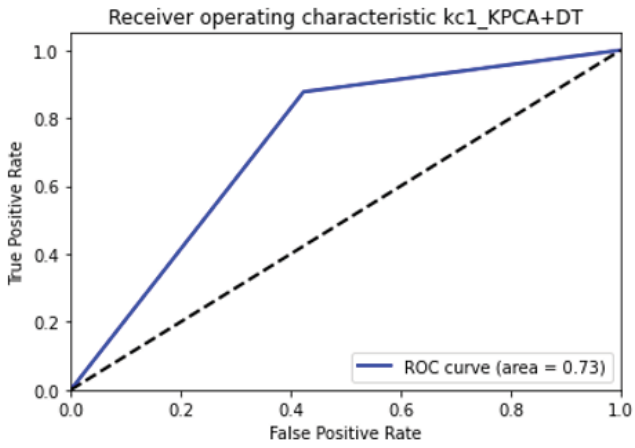
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

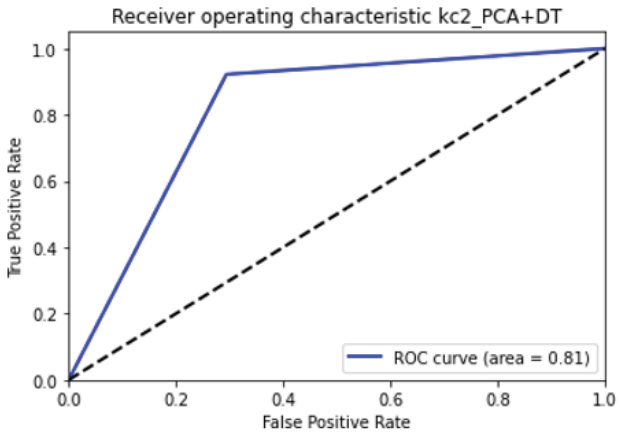




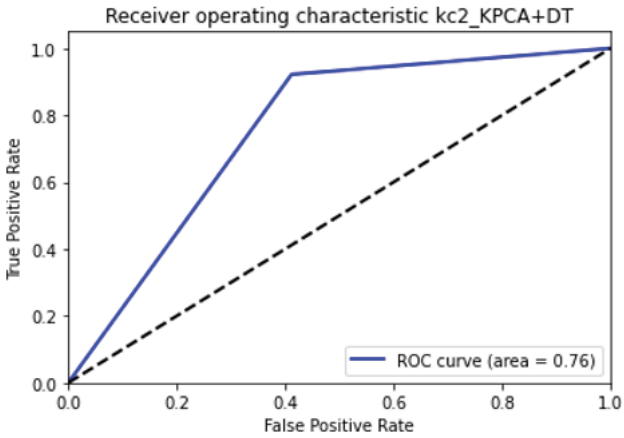


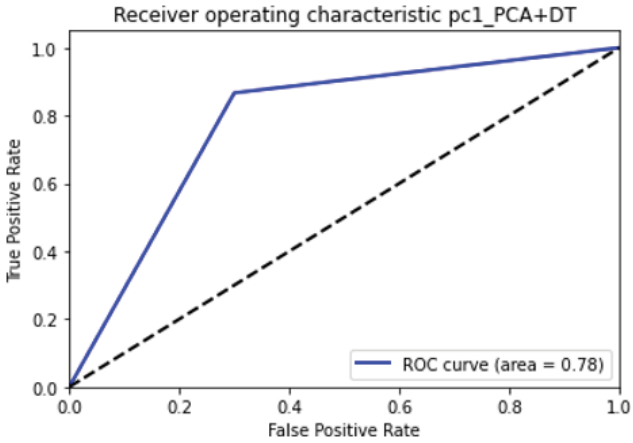


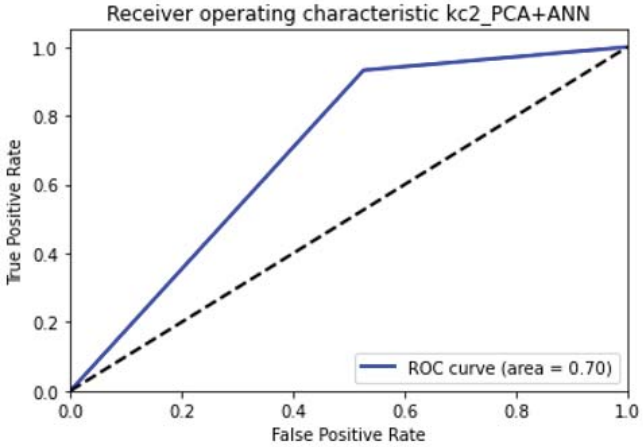


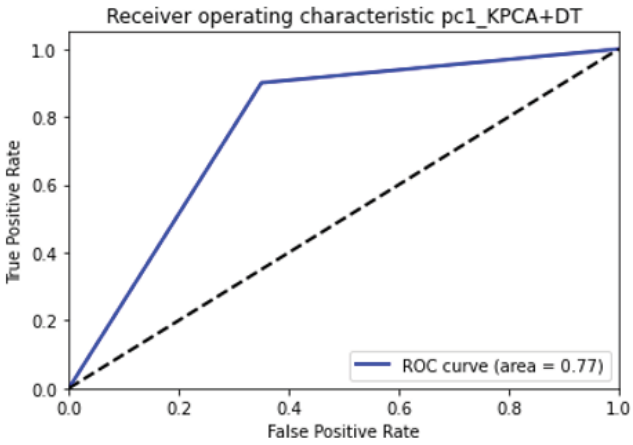


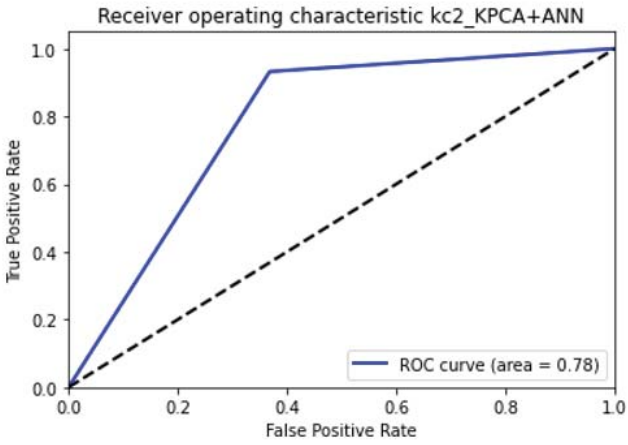
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

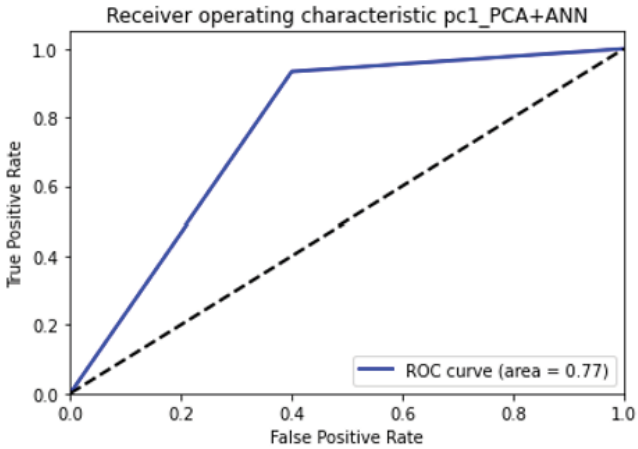




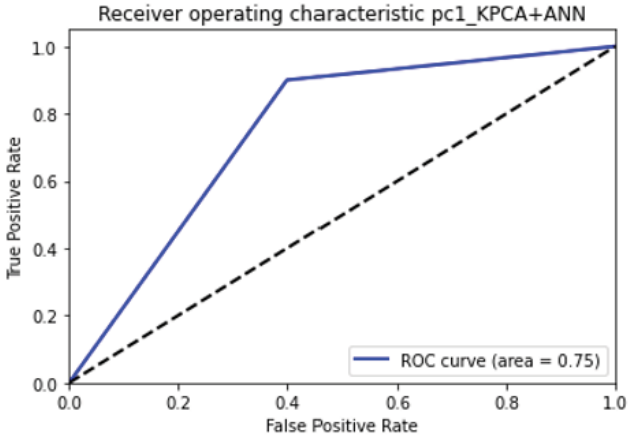








Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.



	PCA+DT	PCA+ANN	KPCA+DT	KPCA+ANN
Cm1	0.76	0.76	0.76	0.80
Jm1	0.72	0.71	0.72	0.73
Kc1	0.72	0.72	0.72	0.73
Kc2	0.81	0.70	0.76	0.78
Pc1	0.78	0.77	0.77	0.75

TABLE 3: AUC SCORES FOR DIFFERENT TECHNIQUES

V. CONCLUSION AND FUTURE WORKS

After closing analyzing the different scores we conclude that using ANN as a classifier along with Kernel PCA yielded the best results. Although the results from that of PCA + Decision Tree were also very close. Thus, it can be concluded that dimensionality Reduction and Deep Learning can prove to be beneficial when the data under consideration has quite a large number of attributes. In the future, we plan to improve the neural network model by changing various parameters which include the number of hidden layers, neurons in each layer, optimizers and the cost function.

Various other techniques need to be experimented for reducing the dimensionality of the parameters and as per the current techniques[13], changing the number of components taken can also result in certain improvement. We also plan to test other classifiers such as Naive Bayes, K-Nearest Neighbors, Kernel Support Vector Machines and ensemble techniques such as Random Forest and compare the results to those found by using Decision Tree classifier.

Threats to Validity:

Following is a brief discussion on the various threats to validity of our comparative study[14]:

A. Construct Validity

We gauge the exhibition of models utilizing hold out cross validation method. Preparing and test sets are developed haphazardly so they may over fit the information. Utilizing other execution assessment strategies, they may give various outcomes. Aside from the thought about characteristics, there may be different elements influencing the presence of deformities.

B. Internal Validity

The data set used contains information regarding features determined by McCabe and Halstead feature extractors, which are known to have certain limitations.

C. External Validity

We used only 5 open source datasets taken from NASA Promise Repository and so our results may not generalize to all software. Replication of this comparative study taking into account other datasets may produce more generalized results.

D. Conclusion Validity

The datasets being used have class imbalance problems. So, we used AUC to evaluate the performance of our model but it can still be partial for non-buggy instances.

VI. REFERENCES

[1] X. Yang, D. Lo, X. Xia, Y. Zhang, and J. Sun, "Deep learning for just-in-time defect prediction," in QRS'15: Proc. of the International Conference on Software Quality, Reliability and Security , 2015.
[2] Y. Kamei, E. Shihab, B. Adams, A. E. Hassan, A. Mockus, A. Sinha, and N.Ubayashi, "A large-scale empirical study of just-in-time quality assurance," TSE, vol. 39, no. 6, pp. 757–773, 2013.
[3] T. Jiang, L. Tan, and S. Kim, "Personalized defect prediction," in ASE, 2013, pp. 279–28.
[4] M. Tan, L. Tan, S. Dara, and C. Mayeux, "Online defect prediction for imbalanced data," in ICSE'15, pages 99–108.
[5] S. Wang, T. Liu, and L. Tan, "Automatically learning semantic features for defect prediction," in ICSE'16: Proc. of the International Conference on Software Engineering, 2016.
[6] Y. Kamei, S. Matsumoto, A. Monden, K.-i. Matsumoto, B. Adams, and A. E. Hassan, "Revisiting common bug prediction findings using effort-aware models," in ICSM, 2010, pp. 1–10.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

[7] Jian Li, Pinjia He, Jieming Zhu, and Michael R. Lyu, "Software Defect Prediction via Convolutional Neural Network".
[8] P. D. Singh and A. Chugh, "Software Defect Prediction Analysis Using Machine Learning Algorithms," in International Conference on Cloud Computing, Data Science & Engineering – Confluence, 2017.
[9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.
[10] G. E. Hinton, "Learning multiple layers of representation," Trends in cognitive sciences , vol. 11, no. 10, pp. 428–434, 2007.
[11] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J.Williams et al.,"Recent advances in deep learning for speech research at microsoft," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on , 2013, pp. 8604–8608.
[12] F. Rahman and P. Devanbu, "How, and why, process metrics are better," in ICSE, 2013, pp. 432–441.
[13] J. Ali, R. Khan, N. Ahmad, I. Maqsood, "Random Forests and Decision Trees," in International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.
[14] R. Jindal, R. Malhotra, A. Jain, "Software Defect Prediction Using Neural Networks," IEEE Conference 2014
[15] V. A. Kumar, N. Elavarasan, "A Survey on Dimensionality Reduction Technique," in International Journal of Emerging Trends & Technology in Computer Science, Volume 3 Issue 6, November-December 2014
[16] S. Mosci, L. Rosasco, A. Verri, "Dimensionality Reduction and Generalization," in the 24th International Conference on Machine Learning, Corvallis, OR, 2007.
[17] N. Varghese, V. Verghese, Gayathri. P and Dr. N. Jaisankar, "A Survey of Dimensionality Reduction And Classification Methods," in International Journal of Computer Science & Engineering Survey, Vol.3, No.3, June 2012
[18] Q. Wang, "Kernel Principal Component Analysis and its Applications in Face Recognition and Active Shape Models," Rpi, Troy, Ny, Usa, 2011. Copyright 2011
[19] J. Hanley, BJ. McNeil, "The meaning and use of the area under a Receiver Operating Characteristic ROC curve", Radiology, 143, 1982, pp. 29-36.
[20] P. Deep Singh and A. Chug, "Software defect prediction analysis using machine learning algorithms," 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, Noida, 2017, pp. 775-781, doi: 10.1109/CONFLUENCE.2017.7943255.
[21] He, Peng, et al. "An empirical study on software defect prediction with a simplified metric set." Information and Software Technology 59 (2015): 170-190.
[22] Grishma, B. R., and C. Anjali. "Software root cause prediction using clustering techniques: A review." Communication Technologies (GCCT), 2015 Global Conference on. IEEE, 2015.
[23] Chug, Anuradha, and Shafali Dhall. "Software defect prediction using supervised learning algorithm and unsupervised learning algorithm." (2013): 5-01
[24] Huda, Shamsul, et al. "A Framework for Software Defect Prediction and Metric Selection." IEEE Access (2017)
[25] S. Choirunnisa, B. Meidyani and S. Rochimah, "Software Defect Prediction using Oversampling Algorithm: A-SUWO," 2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS), Batu, East Java, Indonesia, 2018, pp. 337-341, doi: 10.1109/EECCIS.2018.8692874

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:05:24 UTC from IEEE Xplore. Restrictions apply.

International Journal of Logistics Research and Applications

A Leading Journal of Supply Chain Management

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjol20

Strategic framework for developing resilience in Agri-Food Supply Chains during COVID 19 pandemic

Pravin Kumar & Rajesh Kumar Singh

To cite this article: Pravin Kumar & Rajesh Kumar Singh (2021): Strategic framework for developing resilience in Agri-Food Supply Chains during COVID 19 pandemic, International Journal of Logistics Research and Applications, DOI: 10.1080/13675567.2021.1908524

To link to this article: https://doi.org/10.1080/13675567.2021.1908524

Published online: 28 Mar 2021.

Submit your article to this journal

Article views: 808

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=cjol20

INTERNATIONAL JOURNAL OF LOGISTICS: RESEARCH AND APPLICATIONS
https://doi.org/10.1080/13675567.2021.1908524

Taylor & Francis
Taylor & Francis Group

Check for updates

Strategic framework for developing resilience in Agri-Food Supply Chains during COVID 19 pandemic

Pravin Kumar[a] and Rajesh Kumar Singh[b]

[a]Department of Mechanical Engineering, Delhi Technological University, Delhi, India; [b]Management Development Institute, Gurgaon, India

ABSTRACT

The Agri-food Supply Chains (AFSC) dealing with perishable items have suffered a lot due to the COVID 19 pandemic. During this uncertain time, developing resiliency has become a priority for management. Therefore, the main purpose of the study is to explore the impact of COVID-19 on AFSC and the possible strategies for improving the resilience of the AFSC. Impacts on AFSC and possible strategies are explored through the literature review and experts' opinions. The importance rating of impacts of COVID-19 on AFSC is determined using the Best-Worst Method (BWM). These impacts are further correlated with the strategies for improving the resilience of AFSC using Quality Function Deployment (QFD). The findings of the study will help farmers, food processors, distributors, and retailers to maintain the uninterrupted flow of food items from farmers to end-users during the uncertain business environment. It will also contribute to formulating effective policies by the government and other stakeholders for improving the resilience of AFSCs.

ARTICLE HISTORY
Received 22 December 2020
Accepted 21 March 2021

KEYWORDS
Agri-Food Supply Chains; COVID-19; Resilience; Best-Worst Method; Quality Function Deployment; Supply chain strategies

1. Introduction

In December 2019, an unknown pneumonia case was detected in Wuhan, China. Initially, any possibility of influenza and other coronaviruses was denied by laboratory testing but, later on, the Chinese authority announced it as a novel coronavirus (nCOV). This virus was named COVID-19 by the Word Health Organization (WHO) (Sahin et al. 2020). An increasing number of positive cases have been caused by cross-infection, i.e. human-to-human transmission (Munster et al. 2020). Due to lockdown for a long period, most of the business activities completely disrupted during that time. Food is a common need of everyone. Food production, processing, and distribution are also disrupted due to the impact of COVID-19.

In general, Agri-Food Supply Chain (AFSC) consists of five major components: farming, post-harvest activities, food processing, distribution and retailing, and consumption. Tsolakis et al. (2014) has also defined AFSC as a set of activities comprising of farming, processing/production, testing, packaging, warehousing, transportation, distribution, and marketing. All activities of AFSC were disrupted during the period of outbreak of COVID-19 and subsequent lockdown happened. The food supply chain covers a wide variation in food items including the staple to highly perishable items (Hueston and McLeod 2012). The supply chains of high-value products (perishable products) are most affected by the fast spread of novel coronavirus and subsequent imposition of lockdown. The staple product production and harvesting are also affected up to some extent due

CONTACT Rajesh Kumar Singh rajesh.singh@mdi.ac.in
© 2021 Informa UK Limited, trading as Taylor & Francis Group

to the migration of workers. Some of the crops could not be harvested due to the shortage of workers, maintaining social distancing, transport restriction, etc.

Both the demand and supply-side shocks have been observed in AFSC due to the changing behaviour of consumers (Changed patterns of Demand) and transport restriction, import disruption, shortage of labourers, etc. (Hobbs 2020). A major shift in the demand pattern has also been observed, with a significant increase in the use of online platforms and direct delivery of food items. The demand for food in restaurants, hotels decreased rapidly. Labour shortage due to migration is given more importance in the agriculture sector (Larue 2020). The labour shortage also influenced the operations of food processing firms and the downstream supply chain. The main reasons for the labour shortage were self-isolation, food security, movement restrictions, migration of workers to their native places etc.

In this study, the Resource-Based Theory (RBT) is considered to analyse the various resources used to mitigate the risks of supply chain disruptions due to COVID-19. The various resources and remedial approaches are used to normalise the disrupted AFSC at the earliest. Stakeholder theory is highly integrated to RBT and Institutional theory concerning sustainable practices (Yuen et al. 2019). Saint Ville, Hickey, and Phillip (2017) studied the implications of the interrelationships between the different stakeholders concerning the associated policies of AFSCs. Dania, Xing, and Amer (2018) also used the same theory to study the collaboration behavioural factors in AFSC. They called it the Resource Dependent Theory (RDT). Pfeffer and Salancik (2003) introduced the RDT as an organisational theory by observing the dependency of the organisation on the different factors to minimise environmental uncertainty (Ketchen, Jr, and Hult 2007). This study mainly focuses on the supply chain of the food grains.

The COVID-19 has a significant impact on operations of AFSC. Disruptions have impacted the routine operations of AFSC. Sharma et al. (2020) divided all risks in AFSC due to disruptions in various categories. These are supply risks, demand risks, financial risks, logistics and infrastructure risks, management and operational, policy and regulation, and biological and environmental risks. Although many studies have been conducted in context to manufacturing sectors post COVID 19 but limited research has been done to analyse impacts and possible strategies to develop resilience in AFSC during this pandemic. But these studies lack in mapping the various strategies with respect to the impacts of COVID-19 on AFSC. Considering the complexity of AFSC and perishability of food items, the present study has answered the following research questions

(1). What are the major impacts of COVID-19 on AFSC and how can these impacts be quantified?
(2). How are these impacts correlated with different remedial approaches/strategies?
(3). How to prioritise the remedial approaches/strategies to improve the resilience of the AFSC?

In this study, a mapping between the impacts of COVID-19 and various strategies is demonstrated using Quality Function Deployment (QFD) approach. The mapping and the importance rating of the different impacts are based on the opinion of the farmers, experts, and the end-users. The Best-Worst method (Rezaei 2015) is used to find the importance rating of COVID-19 impacts on AFSC and Quality Function Deployment (Akao 2004) is used to establish the correlation between the impacts and remedial approaches. These impacts are treated as the voice of customers (VOC) in QFD. The various remedial approaches/strategies are considered as Design Characteristics (DC) in QFD. Finally, the different approaches are prioritised using Quality Function Deployment (QFD). The findings of the study will contribute significantly to bridging the existing research gaps. The remaining part of the paper is organised as follows: The second section represents the literature review. The third section describes the research methodology. The fourth section represents the result of the discussion and implications. The fifth section concludes the study and presents the limitations and future scope of the research.

2. Literature review

The term AFSC can be described as the distribution of agricultural products from farmers to the end consumers (Aramyan et al. 2007). AFSC is formed by the farmers, processing units, distribution, retailing, and end consumers. The AFSC is very similar to the other commercial supply chains but it can be differentiated only in terms of the perishability of the products, weather dependency of the crops, limited shelf-life, price variability, etc. (Ahumada and Villalobos 2009). In recent years, AFSC has received more attention from researchers. During the lockdown, after the outbreak of the COVID-19 the entire transportation activities and import-export were disrupted. The effective distribution of food items has become a major challenge. The demand pattern of the food items and consumer behaviour were changed. To avoid panic in the public, the government had to announce that there is no shortage of food items and the transport restriction on the necessary items was relaxed. Besides, many impacts have been observed in AFSC as discussed in the following subsections.

2.1 Impacts of COVID-19 on AFSC

The AFSC is the most vulnerable with the uncertainties related to weather, demand, and business environment. All the major impacts have been divided into four groups i.e. operational impacts, logistics impacts, financial impacts, and social and environmental impacts for the review purpose. Impacts will be considered as Voice of Customers (VOC) for applying quality function deployment. VOC also represents 'WHAT' in QFD approach.

High Production and Distribution cost (O11): Outbreak of the COVID-19 and subsequent lockdown affected most supply chain activities. Due to transport restriction, migration of workers, and maintaining the social distancing, the production and distribution of the agri-food products became very difficult and resulted in increased production and distribution cost. Due to movement restriction and labour shortage, the price gap between wholesale and retail was increased during the first phase of lockdown (Narayanan and Saha 2020). A significant increase in the food price was observed due to the higher transport cost, limited availability of labour, and logistics uncertainties.

Lack of trust in International trade (O12): In contrast to the increased cost of production and distribution due to movement restriction and labour shortage, a price drop of food products was also observed due to export restriction imposed by many governments to ensure the availability of the food products in the domestic markets for the short- term. This resulted in a major loss to the farmers. Due to export restriction, the traders lost their reputation/trust in the international markets. It also affected the global business opportunity in the future (Espitia, Rocha, and Ruta 2020).

*Food Quality and Safety Issue (O13)***:** It is very important to prevent the contamination of food items from infectious diseases. However, no report has been received which shows food contamination due to the nCOV virus (Aday and Aday 2020). During the lockdown, the physical verification and certification of food items become very difficult due to movement restrictions and labour shortages (OECD, April 29 2020a). The quality of food items may be influenced by the remaining shelf-life (Li and Wang 2017). The safety measures cover controlled food processing and handling, employee's health, use of PPE (Personal Protection Equipment) kits, sanitisation, social distancing, etc. The safety measures are more critical towards the end of supply chains due to the increased handling and involvement of more people (Rizou et al. 2020). Maintaining effective coordination in the AFSC is essential for ensuring the quality of food items (Singh 2014).

*Lack of response (O14)***:** During the first phase of the lockdown, it has been observed that there was a lack of response at various points in the supply chain. The lack of response was also observed at the governance level. Many state governments were reluctant to take the initiatives such as the prevention of pandemic spread, the arrangement of food for poor people, a shelter for the labourers, and arrangement of transport facilities for the migrating workers. Due to the lack of response and

poor information, people started to store the food items unnecessarily. A long queue in all the supermarkets and a crowd of people at grocery stores were observed during the lockdown (Hobbs 2020). This type of situation could be avoided by the proper response, product fill rate, effective handling of consumer complaints, etc. (Persson and Olhager 2002).

Poor accessibility and availability of key intermediates (L11): The COVID-19 pandemic also affected the supply of key intermediate inputs to the farmers such as fertilisers, seeds, pesticides, etc. For the moment, there were no shortages of these items but the farmers were facing extra difficulties in sourcing (Farm Policy News, April 6 2020). The entire AFSC is dependent on the production and supply of the crops from the farmers. Any delay in the availability of the key intermediates may result in less production and hence affects the entire food processing industry. The accessibility and availability of the information during the crisis is also stressed in the literature because it leads to transparency in the food supply chains (Forslund and Jonsson 2009).

Transport restriction (L12): Transport restriction and quarantine are found to be the important measures to contain the spread of COVID-19 (Amjath-Babu et al. 2020) which may result in delays and disruptions of the entire AFSC. Border closures and extra precautions to check the CORONA positive cases and food safety measures led to congestion and extra delays. The fruit and vegetable sector was found to be most affected by the border inspection (OECD, May 11 2020c). The social distancing led to the reduced number of inspectors at borders and increased lead time for inspection and customs clearance. The closure of airline services caused a major problem for the export of perishable foods like seafood, fruit, and vegetables, etc. By contrast, the effect of the disruption on supplies of staple food items (cereal) was not significant because of bulk transportation and minimal labour requirement for the handling.

Import Disruption (L13): Port closures posed a serious threat to fulfilling the demand of the seeds required for the next crop and pesticides as it was sourced from China. Many shipping containers were stuck at the Chinese ports due to the restriction on imports (OECD, May 11 2020c). However, the seed needed for the March, April, and May sowing period (such as maize, soybean, and spring wheat) had arrived before the import disruption. But it was the major problem for ensuring the timely arrival of the seed for the next growing seasons (OECD, June 2 2020b). The seed is also often transported by air which was also completely disrupted. China is a major supplier of pesticides, thus, the fulfilment of the demand for pesticides was also challenging.

Supply and Distribution uncertainties (L14): The production of the crops and their distribution for the food processing industry are the most vulnerable. The disruption of transport and other supply chain activities are the major sources of vulnerability. The dependency on the weather for crop production is also responsible for the vulnerability. Borodin et al. (2016) have found the need to develop a robust AFSC so that it can reduce the impacts of existing sources of uncertainty and high perishability of food products. The uncertainty concerning the continuation of restriction on transportation, import-export, and returning of labour on work-place are the major issues in the agriculture sector.

Liquidity and capital shortage (F11): Financial liquidity and capital shortage are crucial for farmers. The productivity of crop production and capital-intensive systems is largely affected by credit availability (Mitra, Khan, and Nielsen 2019). Some of the barriers like revenue problems due to lower sales value, reduced production, harvest losses due to migration of workers, difficulties in accessing the output markets, eroded liquidity, and led the capital shortage for the next cycle of crop production. A lower sales revenue and capital shortage influenced the purchasing decisions of an individual farmer, for the 2020 monsoon Kharif cropping season (July – November) and hence, the working of the entire value chain (Hassouneh et al. 2012). The reduced sales of agricultural products and reduced cash flows resulted in a liquidity crunch and the traders' ability to maintain stocks of critical inputs (iDE 2020).

Price volatility (F12): Here, price volatility means variation in the price of agricultural output over time. The smooth variation in the price with a certain trend is not price volatility. Price volatility is concerned with unexpected price changes that are not anticipated. It leads to inaccurate

decision-making by the supply chain partners. A significant increase in the food price is more problematic for low-income people in the least developed countries as they use more than half of their earnings on food items. In spite of safety net mechanisms, even in the developed countries, the increased food prices can create a problem for the poor, (FAO 2011a).

Loss or wastage of high-value (perishable) products (F13): Disruption downstream of AFSC leads to accumulation of food items at different points of the supply chains and puts a strain on the storage facilities. Also, it causes a loss for highly perishable products. The agri-food items need to control environmental factors like temperature and humidity (Zambrano et al. 2019; Schanes, Dobernig, and Gözet 2018). These items also need a controlled environment with changing weathers (Kumar, Singh, and Modgil 2020). The supply-side disruptions cause food loss. The farmers waste their crops due to the labour shortage and inaccessibility of the markets.

Migration of workers (S11): The seasonal employment of the labour, especially for planting, harvesting, processing, or transporting crops to markets is common in many countries. The directives of maintaining the social distancing, using one-third of the workers in the plant, food security, and safety led to the migration of workers and anxiety among the workers to reach their native place at the earliest. The food supply chain got significantly affected by labour shortages due to sickness and migration. It created an issue of food safety, especially in the COVID-19 affected areas (Stephens et al. 2020)

Delays in farming (S12): Due to import disruption and transport restriction, the COVID-19 also affected the availability of key intermediate inputs for farmers for the sowing of the next crops (Farm Policy News, April 6 2020) and the entire supply chain from farm to fork was disrupted (Ramakumar 2020; Rawal et al. 2020). Not only procurement was disrupted, but harvesting and selling of the crop produced were also disrupted due to a lack of buyers in the market and transport restriction (Lai 2020). Delay in procurement and selling activities affected the yield of the next crop because the agricultural activities are dependent on season and weather.

Shift in consumer demands/behaviour (S13): A drastic change in the pattern of consumer demand was observed during the period of lockdown. The demand for food from restaurants and hotels was shifted to food at home. Also, the demand for healthy and immunity-boosting food was increased in comparison to junk food (OECD, June 2 2020b). The demand for frozen and packaged foods increased dramatically (IRI, May 15 2020). The demand for food items varied with the price and the income level of consumers, consumption, and shopping preferences, and time constraints. The frequency of visits to food stores was also changed with the amount of money on food per visit changed (Bakalis et al. 2020; Cranfield 2020).

*Lack of food security and nutritious food among the workers (S14)***:** The lack of food security and availability of nutritious food for the people is a major issue for underdeveloped and developing countries. Developing countries are facing the challenge of ensuring the availability of nutritious and safe food to poor people (Pingali, Mittra, and Rahman 2017). The availability of food grains is crucial for countries like India because a large section of the population is unable to get the food (Medialdea et al. 2018; Clapp 2017). The governments (central and state) share responsibilities to ensure the availability of food grains to the targeted population (Reardon et al. 2019; Kumar, Singh, and Modgil 2020).

2.2 Supply chain strategies to reduce the impact of COVID-19 on AFSC (HOWS)

Many strategies have been discussed in the literature to meet the challenges in the AFSC. Based on the frequency analysis, Dania, Xing, and Amer (2018) outlined some major strategies. In this study, we have used the frequency analysis and opinion of the executives working in the food industries to select some important strategies and found the relationships with the impacts of COVID 19 in AFSC. These strategies will be considered as Design Characteristics (DC). These characteristics represent 'HOW' while applying quality function deployment. These strategies are discussed in the following subsections.

Supply Chain responsiveness (H1): Supply chain responsiveness means the ability to quickly respond to any change that occurs in the supply chain. The prompt response is a basic requirement of the resilient supply chain which improves the agility, accuracy, and flexibility of the supply chain (Alicke, Azcue, and Barriball 2020). It is concerned with the provision of the right products and services at the right time and the right place. Agility is one of the critical aspects of responsive supply chains. The main measure of responsiveness is the evaluation of the effectiveness of product management by adapting the traceability system (Persson and Olhager 2002). It can be measured in terms of fill rate, product lateness, lead-time, customers' complaints, and shipping errors (Stranieri et al. 2021). Parreno-Merchante et al. (2014) and Mai et al. (2010) used the traceability systems to improve the responsiveness in aquaculture and seafood supply chain respectively.

Coordination between stakeholders (H2): An increased coordination between the stakeholders of AFSC is required to mitigate risks and create a more resilient food system (Amjath-Babu et al. 2020). Coordination may be defined as the perspectives of interactions among the supply chain partners to achieve the goals by accomplishing the tasks jointly (Gulati, Wohlgezogen, and Zhelyazkov 2012). It helps in minimising the potential ambiguity and future conflicts by structuring effective problem-solving processes (Chounta et al. 2014). Centralised coordination may help in knowledge sharing, capacity enhancement, and information dissemination (Heimeriks and Schreiner 2002). Badraoui, Van der Vorst, and Boulaksil (2020) also highlighted the importance of horizontal collaboration in AFSC. Singh et al. (2019) have also observed that the application of IT tools may help in improving coordination in the supply chains.

Trust Development (H3): Trust is an act of one party relying on another party. All the parties have a high level of confidence and belief that everyone will act for achieving the common goal rather than taking benefits individually (Anbanandam, Banwet, and Shankar 2011). Trust also influences organisational collaboration behaviour to grab new business opportunities and partnerships (Qu and Yang 2015). In AFSC, trust is more important because it is a collaboration of many parties including farmers, processors, distributors, and retailers. Also, it is more vulnerable to different kinds of uncertainties and external factors as observed during the lockdown.

Information Sharing (H4): Information sharing is crucial for maintaining coordination among the stakeholders in AFSC (Dwaikat et al. 2018). The different types and approaches of information sharing vary with the types of relationships among the stakeholders of the supply chain (Zhang and Chen 2013). Nowadays, organisations are using technology-based traceable systems to track the information related to food processing, purchasing, inventory, marketing, etc. the different mechanisms of information sharing helps in resource management and coordination within the AFSC (Prince, Barrett, and Oborn 2014; Patel and Cassou 2015).

Supply Chain Collaboration (H5): This is the inherent characteristics of a supply chain that enhance co-creation activities. It develops a harmonious relationship among partners by reducing conflicts. Its features include the no-blame culture, the fluidity of collaboration, organisational task orientation, mutual understanding, and relationship. Burkhardt et al. (2009) highlighted the importance of verbal communications which reflects the quality of collaboration. To achieve the benefits of collaboration across the entire supply chain, it is necessary to suppress individual interests. Joint efforts and commitments are required to converge the focus of all organisations to align their interests and achieve the overall objectives (Meng 2012). In AFSC, there is a requirement of strong collaboration between the supply chain partners to get the collaborative benefits.

Resource Sharing (H6): This is a process of exchange of the various kinds of resources such as skills, assets, technology, etc. among supply chain partners (Cao and Zhang 2011). It is one of the major components of successful collaboration (Samaddar and Kadiyala 2006). Heimeriks and Schreiner (2002) also supported the importance of complementary resources in the success of collaboration (Dania, Xing, and Amer 2018). Resource sharing during the pandemic is more crucial because Governments are facing financial burdens due to the resources required for many social assistance programmes. Organisations sharing specific assets in a collaboration earn trust from others that they put efforts into alleviating uncertainties and extending new opportunities

(Fynes, Voss, and de Búrca 2005). Inadequate funding may lead to a reduction in demand for agricultural production over the medium term. The reduced demand may harm the many emerging private sectors (FAO 2020).

*Safety measures and addressing labour constraints (H7)***:** The major concern of the food companies is to maintaining the employee's health and employing an alternate workforce in place of the workers who do not want to work because of sickness or coronavirus fear (FAO and WHO 2020). The consumers' confidence regarding food safety and security is crucial for AFSC. At this time of crisis, accessibility of food is very important for the end consumers (OECD, June 2 2020b) especially for those working in the unorganised sectors.

One of the important safety measures is to monitor the health of the workers and all other visitors, before entering the facility. Use of wear personal protection equipment, face mask, gloves, and sanitisers, etc, must be mandatory. Other precaution may be taken in terms of employing the less number of workers in the smaller groups in an extended number of shifts.

Digitisation of the process (H8): Digitisation may enable the farmers to reach more profitable markets directly and sell their products at a higher price. Many e-commerce platforms have already been started to help farmers. Some platforms are also monitored by government agencies. Nowadays, many advanced technologies such as blockchain technology, Industry 4.0 technologies are in use to digitise the entire supply chain processes. Blockchain technology is used in the supply chain for data security and correct information flow. It uses the encrypted data flow and eliminates intermediaries. It may decrease the data administration cost and complexity, transaction duration, data duplication, incompatibility, paperwork, communication overhead, etc. (Katsikouli et al. 2020). One of the major limitations of blockchain technology is the lack of compatibility and interoperability with different systems.

The robotic systems of materials handling may be employed to ensure food safety and hygiene by minimising the involvement of humans. Industry 4.0 technologies now play an important role in supply chain management. Cyber-Physical System (CPS) can monitor unsafe or low-quality products in the food supply chain (Iqbal, Khan, and Khalid 2017; Bowler, Bakalis, and Watson 2020). For the past few years, some leading technologies such as the Internet of Things (IoT), cloud computing, cyber-physical systems, etc. have been in use to integrate the entire supply chains.

Commitment (H9): Commitment refers to the dedication of an organisation to maintaining and improving relationships with partners for value co-creation. Trust is one of the most important behavioural factors for commitment and hence for successful collaboration (Fynes, Voss, and de Búrca 2005). Commitment also influences information sharing positively in supply chain collaboration (Wu, Chuang, and Hsu 2014). For strong commitment, individual desire is to be sacrificed for long-term achievement.

Joint Planning (H10): Joint planning represents the collective decision-making, execution, and assessment at the strategic, tactical, and operational levels. It consists of the determination of common objectives and linking them with the limited resources of individual organisations toward the achievement of the goals. Joint planning leads to the strategies to solve the problems related to demand management, production planning, scheduling, procurement, delivery, and distribution systems (Cao and Zhang 2011). All of these activities are relevant for continuous improvement as a joint effort (Burkhardt et al. 2009).

Risk and reward sharing (H11): It is very crucial for the resilience of the supply chain (Tse, Zhang, and Jia 2018). No business can compete as stand-alone entities; rather it needs to compete as a supply chain. Risk and reward sharing have become essential due to the increasing complexity of a supply chain and tough competition from the global markets (Kleindorfer and Van Wassenhove 2004). It also boosts the relationship, coordination, and collaboration among the supply chain partners (Hall 1999). All the risks and rewards need to be shared and balanced among upstream and downstream supply chain partners for successful collaborative partnerships (Anbanandam, Banwet, and Shankar 2011; Meng 2012). In AFSC, risk and reward sharing may be established between the farmers, processing industry, distributors, and the retailers so that the

collaboration among these partners can be improved and losses due to the supply chain disruption can be minimised.

Adaptation (H12): Adaptation may be defined as the ability of an organisation to tune its goals, strategies, processes, resources in line with the needs and capabilities of its partners (Boeck and Wamba 2008; Fynes, Voss, and de Búrca 2005). AFSC is the most vulnerable due to lack of integration and the self-oriented interest of the stakeholders, uncertainties associated with market conditions, and customer demands. Adaptation is very close to tolerance to diversity and successful collaboration. Thus, adaptation is one of the important requirements to handle unplanned disruptions (Ivanov and Sokolov 2012).

Operational Flexibility (H13): Restriction on the movement and complete lockdown requires operational flexibility at a higher scale. This may include flexibility in the working hours and work from home. Extended flexible hours of working with some incentives to employees may be provided to fulfil the production requirements in AFSC. However, the people working in the food industry may not have the opportunity to work from home (Nicola et al. 2020). Jie, Parton, and Mustafid (2016) have worked on the improvement of performance flexibility in the Australian beef industry by supply chain practices. Kumar, Shankar, and Yadav (2008) have discussed many dimensions of flexibility in the supply chain including the sourcing and distribution flexibility so that the procurement and the distribution of the product can be continued without any hurdle. To maintain flexibility in the inventory system, First Expiry First Out (FEFO) and product rotation are suggested to minimise the risk of product expiry (Chowdhury et al. 2020).

Multiple Sourcing (H14): Multiple sourcing is concerned with enhancement in the flexibility from different sources (Stone and Rahimifard 2018). Sourcing flexibility is useful in common product platform modular product, multiple pathways, supply contract flexibility, and alternate suppliers (Tomlin 2006). In AFSC, multiple sourcing may be used for delayed commitment/production, alternate distribution channels, and fast re-routing of requirements (Carvalho et al. 2012). It is required to fulfil the demand of the farmers related to the seed, pesticides, and fertilisers. Similarly, the demand of the consumers may be fulfilled by the multiple sourcing of the food items from different food processing companies and suppliers.

*Local*isa*tion/decentral*isa*tion of operations (H15)*: Localisation of the operations and sourcing is one of the alternative solutions for the supply chain disruption due to transport restriction. Sourcing of fresh food products from the local supplier will be less vulnerable to cross-border disruptions. Consumers show an interest in local foods and the motivations for this interest are economic, social, environmental, and health benefits (Cranfield, Henson, and Blandon 2012). During the lockdown, the customers have already seen the empty store shelves and long queues outside the grocery stores. The consumers were worried about the availability of the food items for the future if the same situation is continued for a long –time. This crisis may give a boost to the local food movement in the short to medium term. However, the local food supply may be less cost-efficient having less variety of food items and it may not have competitive advantages over the mainstream supply chains (Hobbs 2019).

Centralised food manufacturing is another important factor for the disruption of food chains because of the rigid and lengthy supply chain. The small number of very large production facilities may face supply chain disruption due to transport restrictions and labour crises (Almena et al. 2019). Decentralisation is aimed at fostering innovative new start-ups in certain geographically clustered areas where the ecosystem is more conducive (Lerner 2020) and there is no single point of failure as in the case of a very centralised approach (Rossello and Dewitte 2020).

Waste Management (H16): The Food and Agriculture Organization of the United Nations (FAO) differentiated the food loss and food waste. Food loss is concerned with the decrease in quality and quantity of the edible food mass that occurs in the primary stages of the supply chain while food waste is concerned with the food losses occurring at the end of the food (FAO 2011b). Although, both terms are considered together as FLW for quantification (Corrado and Sala 2018; Wunderlich and Martinez 2018). The changes like food production and consumption habits

may lead to the generation of FLW along the supply chain (Jribi et al. 2020) and other aspects of sustainability (Song et al. 2019). Some of the perishable products like vegetables and fruits could not be supplied due to transport restrictions and labour crises. Farmers have to throw away the waste. The waste management, temperature, and humidity-controlled storage functioning of local food markets, and value chain systems can be facilitated by the government.

Based on the above discussions and literature analysis, the following research gaps have been identified:

(i) In the past, most of the pandemics affected the certain regions and zone of the world, but the impact of novel coronavirus is completely different. It caused the complete shutdown and had the global impact. Therefore studies in context to different regions and countries are required.
(ii) There is a lack of sufficient literature that leads to an in-depth analysis of the impact of COVID-19 on AFSCs.
(iii) Only few studies are found on strategies to minimise the impact of COVID-19 on operations of AFSCs.

3. Research methodology

In this study, some of the major challenges in AFSC due to the impact of COVID-19 are extracted through the literature review and the opinion of the farmers and the end consumers. During the lockdown period, the farmers and the end consumers are the great sufferers. Therefore, their opinion helped in selecting the major impacts on AFSC. For the importance rating, Best-Worst Method (Rezaei 2015, 2016a) has been used. To reduce the impact of COVID-19 on AFSC, some major supply chain strategies are explored through the literature review using frequency analysis and selected with the opinion of the experts from the food processing industry. A correlation matrix of Quality Function Deployment (QFD) is prepared to establish the correlation between the Impacts on AFSC (VOC) and Supply Chain Strategies (DC) with the help of experts' opinions. The correlation is categorised into three categories: Strong (9), Medium (6), and weak (3). Based on the QFD analysis, the priority ranking of the supply chain strategies to mitigate the impacts on AFSC is determined. Figure 1 shows the steps of research methodologies used in the study.

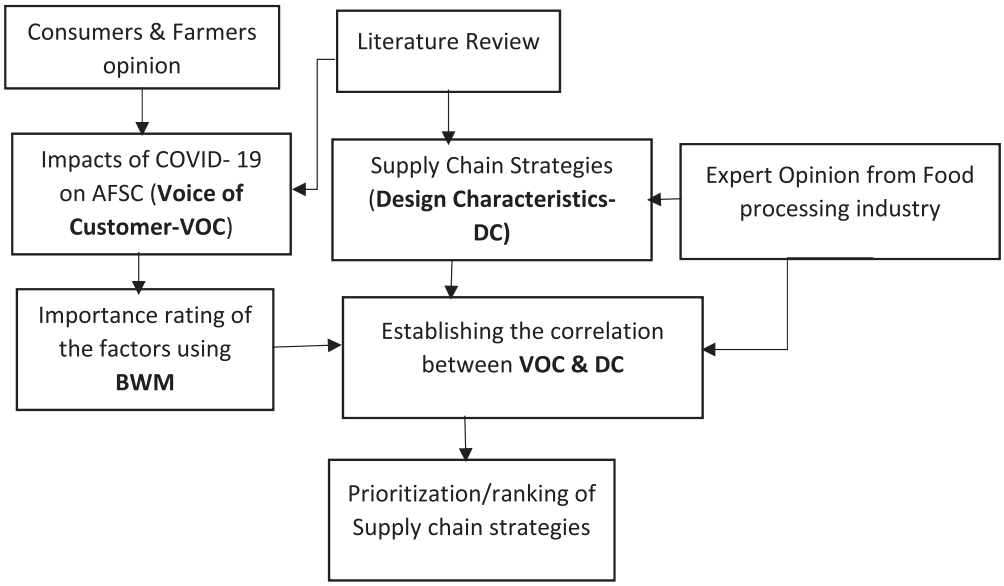


Figure 1. A framework of the research methodology.

Before applying the mathematical tools to analyse the impacts of COVID-19 on AFSC and the strategies to meet these challenges. Some farmers, industry experts, and end consumers were consulted for their opinion. They were asked to give some important impacts on AFSC due to the lockdown, logistics disruptions, migration of workers, etc. Most of the concerns were very close to the observations extracted from the literature review. All the impacts were divided into four groups: Operational, Logistical, Financial, and Socio-economic impacts. They were asked to find the most important impact and least important impact in each group. After that for implementing BWM, we asked experts to rate the most important impact with respect to all the other impacts and all the other impacts with respect to the least important impact using Saaty's 9-point rating scale. Similarly, the mapping of various strategies to meet these challenges is completely based on the opinion of these experts through QFD. Experts were asked about all other strategies that how a specific strategy will help to find the remedy for a particular impact on AFSC.

3.1 Application of the Best-Worst method for ranking of COVID 19 impacts

The BMW method is applied to determine the importance rating of the impacts of COVID-19 on AFSC. It helps in reducing the complexity and efforts applied in a pairwise comparison of each impact. Only the best (most important) impact is compared with other impacts and other impacts are compared with the worst (Least important) impact. It gives more reliable results in comparison to other multi-criteria decision-making tools such as AHP (Analytic Hierarchy Process), SMART (Simple Multi-Attribute Rating Technique), and SAW (Simple Additive Weighting). Some of the advantages of BWM over the other multi-criteria decision-making tools are given below as (Rezaei 2015; 2016a; Rezaei et al. 2016b; Mi et al. 2019):

- It is more flexible and can accommodate the opinions of many decision-makers.
- It is consistent compared to the other multi-criteria tools.
- It involves a limited number of pairwise comparisons in comparison with other methods like AHP.

Many authors have reported the application of BWM in the different areas of management including supply chain sustainability (Ahmed et al. 2018; Ahmadi, Kusi-Sarpong, and Rezaei 2017); logistics performance index (Rezaei, van Roekel, and Tavasszy 2018), enablers for technological innovation in Indian micro, small, and medium enterprises (Gupta and Barua 2016), transportation, and supplier selection (Moktadir et al. 2020; Malek and Desai 2019), etc. The Mathematical steps of BWM is given below:

Step 1: Identify the set of impacts in AFSC through literature review and opinion of the farmers and the end consumers.

Step 2: Identify the best impact (the most significant) and the worst impact (the least significant impact) as per the opinion of the farmers and the end consumers.

Step 3: Determine the preference of the best impact over (BO) the other impacts on a 9-point rating scale and construct X_B (best-to-other) vector as shown in Equation (1).

$$X_B = \{x_{B1}, x_{B2}, \ldots, x_{Bn}\} \tag{1}$$

where x_{Bj} indicates the preference of the best impact B over impact j. It is clear that $x_{BB} = 1$

The preference of best (most important) impact over the other impacts and sub impacts in each category (Operational, Logistic, Financial, and Socio-Economic impacts) are mentioned in Tables 1–5. These preferences are given based on discussion with the farmers, industry experts, and consumers. A total of eight respondents has been considered for taking inputs to rank different factors through BWM. Each respondent has selected the best or worst factor based on his

Table 1. BO Vectors for four clusters of the COVID-19 impact on AFSC.

Respondent No.	Best (Most Important) Impact ↓	Operational Impact (OI)	Logistical Impact (LI)	Financial Impact (FI)	Socio-economic Impact (SI)
1	Logistical Impact (LI)	2	1	6	3
2	Logistical Impact (LI)	3	1	8	5
3	Operational Impact (OI)	1	2	7	4
4	Operational Impact (OI)	1	3	7	4
5	Operational Impact (OI)	1	2	8	4
6	Logistical Impact (LI)	2	1	5	3
7	Socio-economic Impact (SI)	6	2	4	1
8	Logistical Impact (LI)	2	1	4	6

Table 2. BO Vectors for Operational Impacts of COVID-19 on AFSC.

Respondent No.	Best (Most Important) Impact ↓	High Production and Distribution Cost (O11)	Lack of Trust in International Trade (O12)	Food quality/ safety issue (O13)	Lack of Response (O14)
1	O11	1	3	8	5
2	O11	1	2	8	4
3	O13	2	4	1	6
4	O14	3	6	8	1
5	O11	1	7	3	3
6	O12	3	1	7	5
7	O14	5	3	8	1
8	O12	2	1	6	4

Table 3. BO vector for the logistical impact of COVID-19 on AFSC.

Respondent No.	Best (Most Important) Impact ↓	Poor accessibility and availability (L11)	Transport restriction (L12)	Import disruption (L13)	Supply & Distribution Uncertainties (L14)
1	**L11**	1	6	3	2
2	**L13**	2	5	1	2
3	**L13**	2	6	1	3
4	**L14**	2	3	5	1
5	**L11**	1	3	8	5
6	**L14**	2	3	5	1
7	**L11**	1	7	5	2
8	**L12**	3	1	2	5

Table 4. BO vector for financial impacts on AFSC.

Respondent No.	Best (Most Important) Impact ↓	Liquidity and capital shortage (F11)	Price volatility (F12)	Loss/wastage of Perishable (High Value) products (F13)
1	Price volatility **(F12)**	3	1	5
2	Price volatility **(F12)**	2	1	4
3	Liquidity and capital shortage **(F11)**	1	2	4
4	Loss/wastage of Perishable (High Value) products **(F13)**	5	3	1
5	Price volatility **(F12)**	3	1	5
6	Liquidity and capital shortage **(F11)**	1	3	4
7	Liquidity and capital shortage **(F11)**	1	3	6
8	Price volatility **(F12)**	3	1	6

Table 5. BO vector for Socio-economic impacts of COVID-19 on AFSC.

Respondent No.	Best (Most Important) Impact ↓	Migration of Workers (S11)	Delays in farming (S12)	Shift in Consumer demand/behaviours (S13)	Lack of food security and nutritious food among the workers (S14)
1	Lack of food security and nutritious food among the workers **(S14)**	2	5	3	1
2	Lack of food security and nutritious food among the workers **(S14)**	3	5	7	1
3	Migration of Workers **(S11)**	1	8	3	2
4	Migration of Workers **(S11)**	1	7	4	3
5	Delays in farming **(S12)**	2	1	5	3
6	Lack of food security and nutritious food among the workers **(S14)**	2	3	5	1
7	Migration of Workers **(S11)**	1	7	4	2
8	Shift in Consumer demand/behaviours **(S13)**	2	6	1	3

perception. Notations for all clusters and subfactors have been already mentioned in the literature review section.

Step 4: Determine the preference of the impact *j* over the worst impact (OW) on a 9-point rating scale and construct X_W (other-to-worst) vector as shown in Equation (2).

$$X_W = \{x_{1W}, x_{2W}, \ldots, x_{nW}\}^T \tag{2}$$

where x_{jW} indicates the preference of the impact *j* over the worst impact *W*. It is clear that $x_{WW} = 1$

The preference of other impacts over the Worst (Least important) impact in each category (Operational, Logistic, Financial, and Socio-Economic impacts) are mentioned in Tables 6–10 respectively. These preferences are given based on discussion with farmers and consumers.

Step 5: Determination of optimal weights of each impact ($W_1^*, W_2^*, \ldots, W_n^*$).

The main aim is to determine the optimal weights of the impacts, such that the maximum absolute differences $\left|\frac{W_B}{W_j} - x_{Bj}\right|$ and $\left|\frac{W_j}{W_W} - x_{jW}\right|$ for all j is minimised, which may be translated to the following minimax model:

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

Subject to

$W_j \geq 0$, for all *j*

The model (3) may be represented in the following form

$$\min \xi \tag{4}$$

Table 6. OW Vectors for four clusters of the COVID-19 impact on AFSC.

	Respondents							
Worst (least Important) Impact →	1 Financial Impact (FI)	2 Financial Impact (FI)	3 Financial Impact (FI)	4 Financial Impact (FI)	5 Financial Impact (FI)	6 Financial Impact (FI)	7 Operational Impact (OI)	8 Socio-economic Impact (SI)
Operational Impact (OI)	3	3	7	7	8	3	1	3
Logistical Impact (LI)	6	8	3	2	4	5	3	5
Financial Impact (FI)	1	1	1	1	1	1	2	2
Socio-economic Impact (SI)	2	2	2	2	2	2	6	1

Table 7. OW Vectors for Operational Impacts of COVID-19 on AFSC.

	Respondents							
Worst (Least Important) Impact →	1 O13	2 O13	3 O14	4 O13	5 O13	6 O13	7 O13	8 O13
High Production and Distribution Cost **(O11)**	8	8	3	3	7	3	2	3
Lack of Trust in International Trade **(O12)**	3	4	2	2	1	7	3	6
Food quality/safety issue **(O13)**	1	1	6	1	2	1	1	1
Lack of Response **(O14)**	2	2	1	8	2	2	8	2

Table 8. OW vector for the logistical impact of COVID-19 on AFSC.

	Respondents							
Worst (Least Important) Impact →	1 L12	2 L12	3 L12	4 L13	5 L12	6 L13	7 L12	8 L14
Poor accessibility and availability **(L11)**	3	3	3	3	8	3	7	2
Transport restriction **(L12)**	1	1	1	2	3	2	1	5
Import disruption **(L13)**	5	5	6	1	1	1	2	3
Supply & Distribution Uncertainties **(L14)**	3	3	2	5	2	5	4	1

Table 9. OW vector for financial impacts on AFSC.

	Respondents							
Worst (Least Important) Impact →	1 F13	2 F13	3 F13	4 F11	5 F13	6 F13	7 F13	8 F13
Liquidity and capital shortage **(F11)**	2	2	4	1	2	4	6	2
Price volatility **(F12)**	5	4	2	2	5	2	2	6
Loss/wastage of Perishable (High Value) products **(F13)**	1	1	1	5	1	1	1	1

Table 10. OW vector for Socio-economic impacts of COVID-19 on AFSC.

	Respondents							
Worst (Least Important) Impact →	1 S12	2 S13	3 S12	4 S12	5 S13	6 S13	7 S12	8 S12
Migration of Workers **(S11)**	3	2	8	7	3	3	7	3
Delays in farming **(S12)**	1	2	1	1	5	2	1	1
Shift in Consumer demand/behaviours **(S13)**	2	1	3	2	1	1	2	6
Lack of food security and nutritious food among the workers **(S14)**	5	7	4	2	2	5	3	2

Table 11. 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

Subject to

$$\left|\frac{W_B}{W_j} - X_{Bj}\right| \leq \xi, \text{ for all } j$$

$$\left|\frac{W_j}{W_B} - X_{jB}\right| \leq \xi, \text{ for all } j$$

$$\sum_j W_j = 1$$

$W_j \geq 0$, for all j.

The optimal weights (W_1^*, W_2^*, …, W_n^*) and ξ^* are found by solving the linear programming model as shown in Equation (4) and using the consistency index (CI) as given in Table 3 (Rezaei 2015).

The consistency ratio can be obtained using $\xi^*$and the corresponding CI as given below

$$CR = \frac{\xi^*}{CI} \tag{5}$$

The value of CR very close to zero is considered as higher consistency, however, no threshold value for the CR in BWM is mentioned (Rezaei 2015; 2016a) (Table 11).

The weights of all the different impacts and sub impacts as per the calculation are shown in Tables 12–16.

Table 12. Optimal Weights of the four clusters of COVID-19 impacts on AFSC.

	Optimal Weight of Impacts					
Respondent No.	Operational Impact (OI)	Logistical Impact (LI)	Financial Impact (FI)	Socio-economic Impact (SI)	Ksi*	CR
1	0.250	0.500	0.083	0.167	0.000	0.000
2	0.206	0.598	0.072	0.124	0.020	0.012
3	0.529	0.255	0.078	0.137	0.020	0.012
4	0.579	0.187	0.084	0.150	0.019	0.011
5	0.533	0.267	0.067	0.133	0.000	0.000
6	0.254	0.485	0.092	0.169	0.023	0.014
7	0.081	0.270	0.135	0.514	0.027	0.017
8	0.273	0.500	0.136	0.091	0.045	0.028
Average	**0.338**	**0.383**	**0.094**	**0.186**	**0.019**	**0.012**

Table 13. Optimal Weights of the Operational impacts of COVID-19 impacts on AFSC.

	Optimal Weight OF Operational Impacts					
Respondent No.	High Production and Distribution Cost **(O11)**	Lack of Trust in International Trade **(O12)**	Food quality/ safety issue **(O13)**	Lack of Response **(O14)**	Ksi*	CR
1	0.598	0.206	0.072	0.124	0.021	0.013
2	0.533	0.267	0.067	0.133	0.000	0.000
3	0.270	0.135	0.514	0.081	0.027	0.017
4	0.214	0.107	0.071	0.607	0.036	0.022
5	0.557	0.082	0.180	0.180	0.016	0.010
6	0.208	0.589	0.079	0.125	0.034	0.021
7	0.124	0.206	0.072	0.598	0.021	0.013
8	0.270	0.514	0.081	0.135	0.027	0.017
Average	**0.347**	**0.263**	**0.142**	**0.248**	**0.023**	**0.014**

Table 14. Optimal weight of logistical impacts on AFSC.

Respondent No.	Optimal Weight: Poor accessibility and availability **(L11)**	Transport restriction **(L12)**	Import disruption **(L13)**	Supply & Distribution Uncertainties **(L14)**	Ksi*	CR
1	0.234	0.085	0.447	0.234	0.021	0.013
2	0.234	0.085	0.447	0.234	0.021	0.013
3	0.250	0.083	0.500	0.167	0.000	0.000
4	0.254	0.169	0.092	0.485	0.023	0.014
5	0.598	0.206	0.072	0.124	0.021	0.013
6	0.254	0.169	0.092	0.485	0.023	0.014
7	0.533	0.072	0.113	0.282	0.031	0.019
8	0.169	0.485	0.254	0.092	0.023	0.014
Average	**0.316**	**0.169**	**0.252**	**0.263**	**0.020**	**0.013**

Table 15. Optimal weight of Financial impacts on AFSC.

Respondent No.	Liquidity and capital shortage (F11)	Price volatility (F12)	Loss/wastage of Perishable (High Value) products (F13)	Ksi*	CR
1	0.225	0.650	0.125	0.025	0.025
2	0.286	0.571	0.143	0.000	0.000
3	0.571	0.286	0.143	0.000	0.000
4	0.125	0.225	0.650	0.025	0.025
5	0.225	0.650	0.125	0.025	0.025
6	0.629	0.229	0.143	0.057	0.057
7	0.667	0.222	0.111	0.000	0.000
8	0.222	0.667	0.111	0.000	0.000
Average	**0.369**	**0.437**	**0.194**	**0.017**	**0.017**

Table 16. Optimal weight of Socio-economic impacts on AFSC.

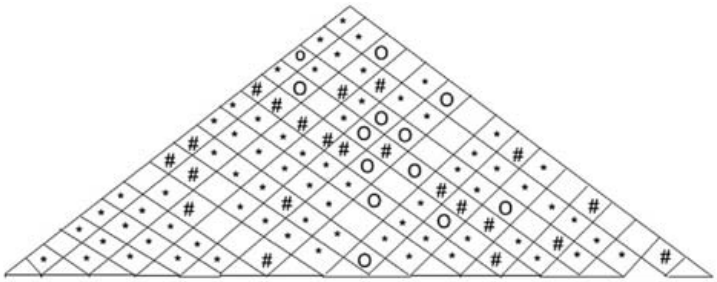
Respondent No.	Optimal weight: Migration of Workers (S11)	Delays in farming (S12)	Shift in Consumer demand/behaviours (S13)	Lack of food security and nutritious food among the workers (S14)	Ksi*	CR
1	0.254	0.092	0.169	0.485	0.023	0.014
2	0.195	0.126	0.080	0.598	0.034	0.021
3	0.506	0.062	0.173	0.259	0.012	0.008
4	0.579	0.084	0.150	0.187	0.019	0.011
5	0.254	0.485	0.092	0.169	0.023	0.014
6	0.254	0.169	0.092	0.485	0.023	0.014
7	0.529	0.078	0.137	0.255	0.020	0.012
8	0.250	0.083	0.500	0.167	0.000	0.000
Average	**0.353**	**0.148**	**0.174**	**0.325**	**0.019**	**0.012**

3.2 Application of Quality Function Deployment (QFD) for developing relationships

QFD, as a cross-functional product design and development approach, was developed at the site of Kobe Shipyard, a Japanese shipbuilding company, in 1972 (Mizuno and Akao 1978). It was then adopted by the Japanese auto industry (Toyota group) in the early 1980s and followed by the U.S. auto industry by the mid-1980s (Hauser and Clausing 1988). Bottani and Rizzi (2006) have used QFD for modelling logistics services. Kumar, Shankar, and Yadav (2012) used the fuzzy extension of the QFD for supplier development. In this study, QFD is used to mitigate the impacts of COVID-19 on AFSC due to the impact of COVID-19. QFD is a tool that is used to correlate or map the voice of customers with the design characteristics of the product.

The mathematical steps of QFD are given below

Step 1: After finding the importance rating of the impacts (also known as Voice of Customers (VOC) in QFD), these impacts are correlated with the supply chain strategies (Also known as



VOC (Impact of COVID-19 on AFSC)		Importance Rating	DC (Remedies of Impact of COVID-19 on AFSC)															
			Supply Chain Responsiveness (H1)	Coordination between Stakeholders (H2)	Trust Development (H3)	Information Sharing (H4)	Supply Chain Collaboration (H5)	Resource Sharing (H6)	Safety Measures and addressing Labour Constraints (H7)	Digitization of the Processes (H8)	Commitment (H9)	Joint Planning (H10)	Risk and reward Sharing (H11)	Adaptation (H12)	Operational Flexibility (H13)	Multiple Sourcing (H14)	Localization/regionalization of Operations (H15)	Waste Management (H16)
Operational impact	High Production and Distribution Cost **(O11)**	0.116	*	#		O	*	*		*	O	O	#	O	*	*	#	*
	Lack of Trust in International Trade **(O12)**	0.089		*	*	*	*	#			*	*	#	#				
	Food quality/safety issue **(O13)**	0.048	*	*		O	*	*		*	*	#						O
	Lack of Response **(O14)**	0.084		*	O	*	*			*	*	O	*	O		O	O	
Logistical Impact	Poor accessibility and availability **(L11)**	0.121		*	#	*	*	*		*				#		#	#	
	Transport restriction **(L12)**	0.065		*		#	*	#		O	O	#	#	#	#	*	*	
	Import disruption **(L13)**	0.097		*	*	#	*			*		#	#	#		*	*	
	Supply & Distribution Uncertainties **(L14)**	0.101	*	*	#	*	*	*	#	*	*	#	*	*	*	*	*	
Financial Impact	Liquidity and capital shortage **(F11)**	0.035	*		*		*	*			*	#	#	O				
	Price volatility **(F12)**	0.041	*		*	*	*			#	O			*	*	*	*	
	Loss/wastage of Perishable (High Value) products **(F13)**	0.018	*	*		#	*	*		#		O	O	#	#			*
Socio-economic Impact	Migration of Workers **(S11)**	0.066			*	*			*		*		#	O	O			
	Delays in farming **(S12)**	0.028					O	#				O	#	#	#			*
	Shift in Consumer demand/behaviours **(S13)**	0.032			*	#			O		*				*			
	Lack of food security and nutritious food among the workers **(S14)**	0.06	#	*		#	*	#	*		*	*		#		#	#	
Relative Ranking			13	2	9	3	1	5	15	4	6	12	11	8	14	7	10	16

Weights of * = 9; # = 6, O = 3

Figure 2. Application of House of Quality to determine the relative weights of supply chain strategies.

Design characteristics (DC) in QFD) using the three-point scale i.e. strong, medium and weak correlation. The correlation weight for strong, medium and weak correlations are used as 9, 6, 3 respectively. The correlation between the impacts of COVID-19 and the supply chain strategies to make it resilient is established based on the opinion of the experts from the food processing industry as shown in the House of Quality (Figure 2).

Table 17. Local and Global Ranking of the Impacts in Agri-Food Supply Chain.

Clusters of Impacts of COVID-19 on AFSC	Weight	Impacts of COVID-19 on AFSC	Local Weight	Local Ranking	Global Weight	Global Ranking
Operational Impacts	**0.338**	High Production and Distribution Cost **(O11)**	0.343	1	0.116	**2**
		Lack of Trust in International Trade **(O12)**	0.263	2	0.089	**5**
		Food quality/safety issue **(O13)**	0.142	4	0.048	**10**
		Lack of Response **(O14)**	0.248	3	0.084	**6**
Logistical Impacts	**0.383**	Poor accessibility and availability **(L11)**	0.316	1	0.121	**1**
		Transport restriction **(L12)**	0.169	4	0.065	**8**
		Import disruption **(L13)**	0.252	3	0.097	**4**
		Supply & Distribution Uncertainties **(L14)**	0.263	2	0.101	**3**
Financial Impacts	**0.094**	Liquidity and capital shortage **(F11)**	0.369	2	0.035	**12**
		Price volatility **(F12)**	0.437	1	0.041	**11**
		Loss/wastage of Perishable (High Value) products **(F13)**	0.194	3	0.018	**15**
Socio-economic Impacts	**0.186**	Migration of Workers **(S11)**	0.353	1	0.066	**7**
		Delays in farming **(S12)**	0.148	4	0.028	**14**
		Shift in Consumer demand/ behaviours **(S13)**	0.174	3	0.032	**13**
		Lack of food security and nutritious food among the workers **(S14)**	0.325	2	0.060	**9**

Step 2: Establish the interrelationship between the supply chain strategies so the impact of one over another can be easily observed. These interrelationships are also established based on the opinion of the experts from the food processing industry and are shown on the roof of the House of Quality (Figure 2).

Step 3: Multiply the importance rating of the impacts with the correlation of the ith Supply chain strategy and find the weighted sum as shown in Equation (6). This is known as the absolute weightage of the ith supply chain strategy.

$$W_{ai} = \sum_{j=1}^{n} W_j \times r_{ij} \tag{6}$$

where $W_j$is the importance rating of the jth impact and $r_{ij}$is the correlation between jth impact and ith supply chain strategy. $i=1, 2, \ldots, m$ represents the supply chain strategies, and $j=1,2, \ldots, n$ represents the impacts in the supply chain.

The importance rating of the COVID-19 impacts is shown in the third column of the House of quality (Figure 2).

Step 3: After finding the absolute weightage of all the supply chain strategies, find the relative weights of the ranking of all the supply chain strategies using Equation (7).

$$W_{ri} = \frac{W_{ai}}{\text{Max. value of } W_{ai}};\ \forall i \tag{7}$$

The relative weights or ranking of the supply chain strategies are shown in the last row of the House of Quality (Figure 2).

4. Results and discussion

We have found a local and global ranking of COVID 19 impacts by applying the Best-worst method as explained in the previous section. The ranking of the impacts of COVID-19 on AFSC is given in Table 1. It has been observed that poor accessibility and availability are the most important impact

of COVID-19 on AFSC as it has the highest global weight based on BWM calculation. Due to the restriction of movement, transportation, complete lockdown, it becomes very tough to access the required products (food items) and services. There was poor traceability, the empty shelf of the supermarket, and later complete lockdown of the supermarkets and mall facilities. Due to the unavailability of the transport services and migration of workers, production was breakdown and distribution of the product became very difficult. It resulted in high production and distribution costs. It is the second most important impact of COVID-19 on AFSC as shown in Table 17.

Due to the infectious nature of the pandemic and unavailability of reliable treatment/medicine for the disease and the lack of personal protection equipment (PPE) kit, everything was uncertain. It was very difficult to say that when the crisis will end and all the supply chain activities will become normal. Lockdown was extended and all the people were suggested to maintain a social distancing and avoid coming out of the house unnecessarily. All the supplies and distributions became uncertain. The supply and distribution uncertainties are found to be the third most important impact of COVID-19. The geographical borders were closed and all the air services were stopped. Import was completely disrupted. The import disruption is observed as the fourth important impact of COVID-19. Due to disruption of export-import, the reputation of the exporters and traders also got impacted in the international markets. It caused a lack of trust in international trade. Therefore, lack of trust is observed as the fifth most important impact of the COVID-19.

Due to lack of the logistics facilities, migration of workers, restriction of the movement of people post COVID 19, response rate to consumers' requirements reduced drastically. It is found as the sixth important impact. Due to disruptions in business, many workers lost their jobs, and food security became an issue for them. They started to migrate from the workplace to their native place. Due to the migration of workers, not only the food processing and distribution activities are affected but the on-time agricultural activities are also affected. It is found as the seventh important impact of COVID-19 on AFSC. However, most of the impacts are interrelated with each other. Any impact cannot be treated in standalone manner. For example, the eighth important impact i.e. restrictions on transportation may lead to a lack of responsiveness. A long queue was scene in front of the grocery stores due to the poor responsiveness of the shopkeeper (Hobbs 2020). The ninth impact is lack of food security. It lead to migration of workers. The tenth important impact was food quality and safety issues. It lead to the adverse health of the common people. Transport restriction and decrease in demand for many food items lead to price volatility which is the eleventh important impact. The importance of maintaining the food quality during the lockdown was also highlighted by OECD (April 29 2020a). Similarly, the ranking of all the other impacts of COVID-19 such as liquidity and capital shortages, a shift in consumer demand/behaviour, delays in farming, loss of perishable items are ranked as twelfth, thirteenth, fourteenth, and fifteenth importance ranks.

As a next part of the study, we tried to develop a relationship between impacts of COVID-19 on AFSC and probable supply chain strategies to neutralise their effect by using QFD. The correlations between the impacts and supply chain strategies are defined on basis of three scales (Strong, medium, and weak) as per the norms of QFD.

The relative weights obtained from the matrix House of Quality (as shown in Figure 2) show that supply chain collaboration is the most important strategy and it is advocated by many authors as discussed in the literature review. If we observe the roof of the House of Quality, supply chain collaboration is strongly correlated with the other strategies such as supply chain responsiveness, coordination between the stakeholders, trust development, information sharing, resource sharing, digitisation of the processes, joint planning, commitment, risk and reward sharing, etc.

The second most important strategy to mitigate the risks of the pandemic is coordination between the stakeholders. Effective coordination can minimise the impact of supply chain disruptions. It is strongly correlated with the other strategies such as supply chain responsiveness, trust development, information sharing, supply chain collaboration, resource sharing, commitment,

joint planning, multiple sourcing, and waste management. Thus, for successful collaboration in a supply chain, coordination between all the stakeholders is required.

Similarly, the third important strategy is information sharing. It is also strongly correlated with the digitisation of the process, commitment, and joint planning, etc. To minimise the communication gap and impact of a chaotic environment, proper information flow throughout the supply chain is required. The fourth important strategy is the digitisation of the process. It may help in quality management, reducing lead times, information sharing, and controlling the entire processes from the remote location using various technologies such as the Internet of Things, Cloud computing, Cyber-physical systems, automation, etc.

The fifth important supply chain strategy is resource sharing. It may help in the effective management of available resources like finance, human resources, technology, etc. In this way, we can see that all the supply chain strategies are prioritised as per their importance. But, most of the strategies are interrelated to each other. If we improve any one strategy, it will improve the other too. Other strategies are also ranked using the QFD, these rankings are the commitment of stakeholders (Sixth), priority to multiple sourcing in place of single sourcing (seventh rank), Adaptation (ninth), option for localisation of supply chain activities other than the globalisation (tenth), risk and reward sharing among the supply chain partners (eleventh), joint planning (twelfth), supply chain responsiveness (thirteenth), operational flexibility (fourteenth), safety measures and addressing labour constraints (fifteenth), and waste management (sixteenth). Some of the other strategies are ranked low but they are interlinked with the high important strategies. For example, supply chain responsiveness and flexibility depend on many other factors which are shown in the roof of the house of quality. Thus, one can understand the relationship among different strategies through QFD.

Managerial Implications: This study may help the managers of the food processing industries and other stakeholders of AFSC in developing resiliency in supply chains. Effective strategy formulation may improve the accessibility and availability of the products and services with proper management of information flow, coordination, a collaboration between the different supply chain partners. Organisations will be motivated to use various advanced technologies to digitise the processes. Findings may be also helpful in enhancing the safety measures and reducing the labour constraints during uncertain business environments caused by COVID 19 outbreak. Findings of this study may help manage other types of disruptions in AFSC and minimising the loss to the farmers, consumers, and other stakeholders.

Theoretical Implications: This study may help the researchers and the academia to explore other possible disruptions and strategies to overcome emerging business problems. This study also illustrates a mechanism to correlate COVID 19 impacts with possible strategies. A similar approach can be also applied in a different context for handling disruptions in the supply chains. Researchers may also think to validate the findings with more robust analysis.

5. Conclusions

After the COVID-19 outbreak, managing the continuous supply of perishable products has become a major challenge for managers. In this study, we have explored the impacts of the COVID-19 on the AFSC and possible strategies to minimise the effect. Each impact is correlated with appropriate supply chain strategy. All the impacts of the COVID-19 on AFSC are explored through the literature review and finalised after discussion with the end consumers and other stakeholders. To make the supply chain more resilient, various strategies are also identified through the literature review and discussion with experts from the food processing industry. The importance rating of the COVID 19 impacts is determined based on the feedback of the end-users using BWM. Then these impacts are correlated with the supply chain strategies by using QFD. The relationship has been developed based on our discussion with the industry experts.

Poor accessibility and availability, high production and distribution costs, supply and distribution uncertainties, import disruption, the lack of trust in international trade are the first five

important impacts of COVID 19 on AFSC. Similarly, supply chain collaboration, coordination between the stakeholders, information sharing, digitisation of the processes, resource sharing are found to be the important supply chain strategies to control the effect of the pandemic.

Based on the analysis of the study, the following recommendations are made.

(i) Effective coordination, proper information sharing. and commitments are required to make the supply chain more resilient.
(ii) Resource collaboration may help in resuming the supply chain quickly.
(iii) To avoid the stock-out and lack of responsiveness, localisation of supply chain activities is a better option including the globalisation of supply chains.
(iv) The dependency of a single supplier may not be an effective strategy when supply-side disruption occurs. Thus, multiple sourcing may be one of the options. Maintaining the relationship with multiple vendors may be beneficial during the supply chain disruptions.
(v) Workers are an important part of the supply chain, the welfare, food security, and health issues must be part of strategies to avoid the migration of workers.

Some recommendations may be more specific to the type of disruptions in supply chains, but the above-mentioned recommendations can be considered in general to control the impact of disruptions in a supply chain.

Limitations and Future scope of the work: In this study, we have taken the opinion of a limited number of consumers and businessmen for prioritising the impacts of COVID 19. Similarly, for establishing the correlation between the impacts and supply chain strategies, we have taken the opinion of the limited number of experts from the food processing industry to develop the House of Quality. The opinion may change with other group of experts. Therefore, findings may not be generalised. Proposed framework may be validated by empirical study and few case studies can be carried out for detailed strategic analysis of key emerging issues post pandemic as future scope of this study. Researchers may also analyse the impact of COVID 19 on the sustainability of AFSC. Study may be further extended to analyse impact of COVID 19 on sustainability of operations. Decision making post disruption had been a challenge for managers. Therefore, researchers may also propose an appropriate decision making framework during pandemic time. How to measure performance of supply chains during pandemic is also highly unexplored area. Therefore, it may be also explored as a part of future studies.

Disclosure statement

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

References

Aday, S., and M. S. Aday. 2020. "Impacts of COVID-19 on Food Supply Chain." *Food Quality and Safety*, doi:10.1093/fqsafe/fyaa024.

Ahmadi, H. B., S. Kusi-Sarpong, and J. Rezaei. 2017. "Assessing the Social Sustainability of Supply Chains Using Best Worst Method." *Resources, Conservation and Recycling* 126: 99–106.

Ahmed, F., M. Naeem, W. Ejaz, M. Iqbal, and A. Anpalagan. 2018. "Resource Management in Cellular Base Stations Powered by Renewable Energy Sources." *Journal of Network and Computer Applications* 112: 1–17.

Ahumada, O., and J. R. Villalobos. 2009. "Application of Planning Models in the Agri-Food Supply Chain: A Review." *European Journal of Operational Research* 196 (1): 1–20.

Akao, Y. 2004. *Quality Function Deployment: Integrating Customer Requirements Into Product Design*. New York: Productivity Press.

Alicke, K., X. Azcue, and E. Barriball. 2020. "Supply-Chain Recovery in Corona Virus Times—Plan for Now and the Future." https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Operations/Our%20Insights/Supply%20chain%20recovery%20in%20coronavirus%20times%20plan%20for%20now%20and%20the%20future/Supply-chain-recovery-in-coronavirus-times-plan-for-now-and-the-future.pdf.

Almena, A., P. J. Fryer, S. Bakalis, and E. Lopez-Quiroga. 2019. "Centralized and Distributed Food Manufacture: A Modeling Platform for Technological, Environmental and Economic Assessment at Different Production Scales." *Sustainable Production and Consumption* 19: 181–193.

Amjath-Babu, T. S., T. J. Krupnik, S. H. Thilsted, and A. J. McDonald. 2020. "Key Indicators for Monitoring Food System Disruptions Caused by the COVID-19 Pandemic: Insights from Bangladesh Towards Effective Response." *Food Security* 12 (4): 761–768.

Anbanandam, R., D. K. Banwet, and R. Shankar. 2011. "Evaluation of Supply Chain Collaboration: A Case of Apparel Retail Industry in India." *International Journal of Productivity Performance and Management* 60: 82–98.

Aramyan, L. H., A. G. Oude Lansink, J. G. Van Der Vorst, and O. Van Kooten. 2007. "Performance Measurement in Agri-Food Supply Chains: A Case Study." *Supply Chain Management: International Journal* 12 (4): 304–315.

Badraoui, I., J. G. Van der Vorst, and Y. Boulaksil. 2020. "Horizontal Logistics Collaboration: An Exploratory Study in Morocco's Agri-Food Supply Chains." *International Journal of Logistics Research and Applications* 23 (1): 85–102.

Bakalis, S., V. P. Valdramidis, Argyropoulos, L. Ahrne, J. Chen, P. J. Cullen, E. Cummins et al. 2020. "Perspectives from CO + RE: How COVID-19 Changed our Food Systems and Food Security Paradigms." *Current Research in Food Science* 3: 166–172.

Boeck, H., and S. F. Wamba. 2008. "RFID and Buyer-Seller Relationships in the Retail Supply Chain." *International Journal of Retail & Distribution Management* 36 (6): 433–460.

Borodin, V., J. Bourtembourg, F. Hnaien, and N. Labadie. 2016. "Handling Uncertainty in Agricultural Supply Chain Management: A State of the art." *European Journal of Operational Research* 254 (2): 348–359.

Bottani, E., and A. Rizzi. 2006. "Strategic Management of Logistics Service: A Fuzzy QFD Approach." *International Journal of Production Economics* 103 (2): 585–599.

Bowler, A. L., S. Bakalis, and N. J. Watson. 2020. "A Review of in-Line and on-Line Measurement Techniques to Monitor Industrial Mixing Processes." *Chemical Engineering Research and Design* 153: 463–495.

Burkhardt, J. M., F. Détienne, A. M. Hébert, L. Perron, S. Safin, and P. Leclercq. 2009. "*An Approach to Assess the Quality of Collaboration in Technology-Mediated Design Situations*." In *Proceedings of ECCE 2009: European Conference on Cognitive Ergonomics*.

Cao, M., and Q. Zhang. 2011. "Supply Chain Collaboration: Impact on Collaborative Advantage and Firm Performance." *Journal of Operations Management* 29 (3): 163–180.

Carvalho, H., A. P. Barroso, S. Azevedo, and V. H. Cruz-Machado. 2012. "Supply Chain Redesign for Resilience Using Simulation." *Computers & Industrial Engineering* 62 (1): 329–341.

Chounta, I. A., T. Hecking, H. U. Hoppe, and N. Avouris. 2014, September. "Two Make a Network: Using Graphs to Assess the Quality of Collaboration of Dyads." In *CYTED-RITOS International Workshop on Groupware*, edited by Nelson Baloian, Frada Burstein, Hiroaki Ogata, Flavia Santoro, and Gustavo Zurita, 53–66. Cham: Springer.

Chowdhury, M. T., A. Sarkar, S. K. Paul, and M. A. Moktadir. 2020. "A Case Study on Strategies to Deal with the Impacts of COVID-19 Pandemic in the Food and Beverage Industry." *Operations Management Research*, 1–13. doi:10.1007/s12063-020-00166-9.

Clapp, J. 2017. "Food Self-Sufficiency: Making Sense of it, and When it Makes Sense." *Food Policy* 66: 88–96.

Corrado, S., and S. Sala. 2018. "Food Waste Accounting Along Global and European Food Supply Chains: State of the art and Outlook." *Waste Management* 79: 120–131.

Cranfield, J. A. 2020. "Framing Consumer Food Demand Responses in a Viral Pandemic." *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, doi:10.1111/cjag.12234.

Cranfield, J., S. Henson, and J. Blandon. 2012. "The Effect of Attitudinal and Sociodemographic Factors on the Likelihood of Buying Locally Produced Food." *Agribusiness* 28 (2): 205–221.

Dania, W. A. P., K. Xing, and Y. Amer. 2018. "Collaboration Behavioural Factors for Sustainable Agri-Food Supply Chains: A Systematic Review." *Journal of Cleaner Production* 186: 851–864.

Dwaikat, N. Y., A. H. Money, H. M. Behashti, and E. Salehi-Sangari. 2018. "How Does Information Sharing Affect First-Tier Suppliers' Flexibility? Evidence from the Automotive Industry in Sweden." *Production Planning & Control* 29 (4): 289–300.

Espitia, A., N. Rocha, and M. Ruta. 2020. "Covid-19 and Food Protectionism: The impact of the pandemic and Export Restrictions on World Food Markets." *Policy Research Working Paper;No. 9253. World Bank, Washington, DC.*

FAO. 2011a. "Price Volatility in Food and Agricultural Markets: Policy Responses." http://www.fao.org/fileadmin/templates/est/Volatility/Interagency_Report_to_the_G20_on_Food_Price_Volatility.pdf.

FAO (Food and Agriculture Organization of the United Nations). 2011b. "Lessons from the World Food Crisis of 2006–08, Food and Agriculture Organization [Online]." http://www.fao.org/3/i2330e/i2330e04.pdf. Accessed on Apr. 02, 2020.

FAO (Food and Agriculture Organization of the United Nations). 2020. "Extension and Advisory Services: At the Frontline of the Response to COVID-19 to Ensure Food Security [Online]." http://www.fao.org/3/ca8710en/CA8710EN.pdf. Accessed on Apr. 24, 2020.

FAO (Food and Agriculture Organization of the United Nations) and WHO (World Health Organization). 2020. "COVID-19 and Food Safety: Guidance for Food Businesses: Interim Guidance [Online]." Accessed April 17, 2020. http://www.fao.org/3/ca8660en/CA8660EN.pdf.

Farm Policy News. April 6 2020. https://farmpolicynews.illinois.edu/2020/04/covid-19-impacting-food-purchasing-dynamics-as-ag-labor-concerns-persist/.

Forslund, H., and P. Jonsson. 2009. "Obstacles to Supply Chain Integration of the Performance Management Process in Buyer-Supplier Dyads." *International Journal of Operations & Production Management* 29 (1): 77–95.

Fynes, B., C. Voss, and S. de Búrca. 2005. "The Impact of Supply Chain Relationship Quality on Quality Performance." *International Journal of Production Economics* 96 (3): 339–354.

Gulati, R., F. Wohlgezogen, and P. Zhelyazkov. 2012. "The two Facets of Collaboration: Cooperation and Coordination in Strategic Alliances." *Academy of Management Annals* 6 (1): 531–583.

Gupta, H., and M. K. Barua. 2016. "Identifying Enablers of Technological Innovation for Indian MSMEs Using Best–Worst Multi Criteria Decision Making Method." *Technological Forecasting and Social Change* 107: 69–79.

Hall, R. 1999. "Rearranging Risks and Rewards in a Supply Chain." *Journal of General Management* 24 (3): 22–32.

Hassouneh, I., A. Radwan, T. Serra, and J. M. Gil. 2012. "Food Scare Crises and Developing Countries: The Impact of Avian Influenza on Vertical Price Transmission in the Egyptian Poultry Sector." *Food Policy* 37 (3): 264–274.

Hauser, J. R. and D. Clausing. 1988. "The House of Quality." *Harvard Business Review* 66 (3): 63–73.

Heimeriks, K., and M. Schreiner. 2002. "Alliance Capability, Collaboration Quality, and Alliance Performance: An Integrated Framework." *Eindhoven Center for Innovation Studies, Eindhoven*, Netherlands Working Paper 02.05, 31–49. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.896.9348&rep=rep1&type=pdf.

Hobbs, J. E. 2019. "Heterogeneous Consumers and Differentiated Food Markets: Implications for Quality Signaling in Food Supply Chains." *Canadian Journal of Agricultural Economics/Revue Canadienne D'agroeconomie* 67 (3): 237–249.

Hobbs, J. E. 2020. "Food Supply Chains During the COVID-19 Pandemic." *Canadian Journal of Agricultural Economics/Revue Canadienne D'agroeconomie*. doi: 10.1111/cjag.12237.

Hueston, W., and A. McLeod. 2012. "Overview of the Global Food System: Changes Over Time/Space and Lessons for Future Food Safety." In *Institute of Medicine (US). Improving Food Safety Through a One Health Approach: Workshop Summary*, 189. Washington, DC: National Academies Press (US).

iDE. 2020. *COVID-19 Rapid Market Analysis Report for CSISA-MEA Project, Internal Report*. CSIA-MEA, CIMMYT.

Iqbal, J., Z. H. Khan, and A. Khalid. 2017. "Prospects of Robotics in Food Industry." *Food Science and Technology* 37: 159–165.

IRI. May 15, 2020. *COVID-19 Impact: Consumer Spending Tracker For Measured Channels*. https://www.iriworldwide.com/IRI/media/Library/2020-05-15-IRI-BCG-COVID-Global-Consumer-Spend-Tracker.pdf.

Ivanov, D., and B. Sokolov. 2012. "Structure Dynamics Control Approach to Supply Chain Planning and Adaptation." *International Journal of Production Research* 50 (21): 6133–6149.

Jie, F., K. A. Parton, and G. Mustafid. 2016. "Supply Chain Performance Flexibility in the Australian Beef Industry." *International Journal of Logistics Research and Applications* 19 (4): 300-–3317.

Jribi, S., H. Ben Ismail, D. Doggui, and H. Debbabi. 2020. "COVID-19 Virus Outbreak Lockdown: What Impacts on Household Food Wastage?" *Environment, Development and Sustainability* 22: 3939–3955.

Katsikouli, P., A. S. Wilde, N. Dragoni, and H. Høgh-Jensen. 2020. "On the Benefits and Challenges of Blockchains for Managing Food Supply Chains." *Journal of the Science of Food and Agriculture*. doi:10.1002/jsfa.10883.

Ketchen, D. J. Jr, and G. T. M. Hult. 2007. "Bridging Organization Theory and Supply Chain Management: The Case of Best Value Supply Chains." *Journal of Operations Management* 25 (2): 573–580.

Kleindorfer, P. R., and L. N. Van Wassenhove. 2004. "Managing Risk in Global Supply Chains." In *The INSEAD-Wharton Alliance on Globalizing*, edited by H. Gatignon and J. R. Kimberley, 288–305. Cambridge: Cambridge University Press.

Kumar, P., R. Shankar, and S. S. Yadav. 2008. "Flexibility in Global Supply Chain: Modeling the Enablers." *Journal of Modelling in Management* 3 (3): 277–297.

Kumar, P., R. Shankar, and S. S. Yadav. 2012. "An Analysis of Supplier Development Issues in Global Context: an Approach of Fuzzy Based Modelling." *International Journal of Logistics Systems and Management* 11 (3): 407–428.

Kumar, A., R. K. Singh, and S. Modgil. 2020. "Exploring the Relationship Between ICT, SCM Practices and Organizational Performance in Agri-Food Supply Chain." *Benchmarking: An International Journal* 27 (3): 1003–1041.

Lai, L. 2020. "COVID-19: India's Harvests also Locked Down." Inter Press Service News Agency April 24 2020. http://www.ipsnews.net/2020/04/covid-19-indiasharvests-also-locked/.

Larue, B. 2020. "Labour Issues and COVID-19." *Canadian Journal of Agricultural Economics/Revue Canadienne D'agroeconomie*. doi: 10.1002/cjag.12233.

Lerner, J. 2020. "Government Incentives for Entrepreneurship." NBER Working Paper No. 26884, National Bureau of Economic Research.

Li, D., and X. Wang. 2017. "Dynamic Supply Chain Decisions Based on Networked Sensor Data: An Application in the Chilled Food Retail Chain." *International Journal of Production Research* 55 (17): 5127–5141.

Mai, N., S. G. Bogason, S. Arason, S. V. Arnason, and T. G. Matthiasson. 2010. "Benefits of Traceability in Fish Supply Chains Case-Studies." *British Food Journal* 112 (9): 976–1002.
Malek, J., and T. N. Desai. 2019. "Prioritization of Sustainable Manufacturing Barriers Using Best Worst Method." *Journal of Cleaner Production* 226: 589–600.
Medialdea, J. T., J. A. P. Ruiz, C. F. García, A. C. Capilla, J. C. Martorell, and J. B. Rodenas. 2018. "Potential of Science to Address the Hunger Issue: Ecology, Biotechnology, Cattle Breeding and the Large Pantry of the sea." *Journal of Innovation & Knowledge* 3 (2): 82–89.
Meng, X. 2012. "The Effect of Relationship Management on Project Performance in Construction." *International Journal of Project Management* 30 (2): 188–198.
Mi, X., M. Tang, H. Liao, W. Shen, and B. Lev. 2019. "The State-of-the-art Survey on Integrations and Applications of the Best Worst Method in Decision Making: Why, What, What for and What's Next?" *Omega* 87: 205–225.
Mitra, S., M. A. Khan, and R. Nielsen. 2019. "Credit Constraints and Aquaculture Productivity." *Aquaculture Economics & Management* 23 (4): 410–427.
Mizuno, S., and Y. Akao. 1978. *Quality Function Deployment: A Companywide Quality Approach*. Tokyo: JUSE Press.
Moktadir, M. A., A. Kumar, S. M. Ali, S. K. Paul, R. Sultana, and J. Rezaei. 2020. "Critical Success Factors for a Circular Economy: Implications for Business Strategy and the Environment." *Business Strategy and the Environment*. doi:10.1002/bse.2600.
Munster, V. J., M. Koopmans, N. van Doremalen, D. van Riel, and E. de Wit. 2020. "A Novel Coronavirus Emerging in China—key Questions for Impact Assessment." *New England Journal of Medicine* 382 (8): 692–694.
Narayanan, S., S. Saha, and Urban Food Markets and the Lockdown in India. May 12, 2020. Available at SSRN: https://ssrn.com/abstract=3599102 or http://dx.doi.org/10.2139/ssrn.3599102.
Nicola, M., Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, and R Agha. 2020. "The Socio-Economic Implications of the Coronavirus Pandemic (COVID-19): A Review." *International Journal of Surgery* 78: 185–193.
OECD. April 29, 2020a. "COVID-19 and the Food and Agriculture Sector: Issues and Policy Responses." https://www.oecd.org/coronavirus/policy-responses/covid-19-and-the-food-and-agriculture-sector-issues-and-policy-responses-a23f764b/#back-endnotea0z2.
OECD. June 2, 2020b. "Food Supply Chains and COVID-19: Impacts and Policy Lessons." http://www.oecd.org/coronavirus/policy-responses/food-supply-chains-and-covid-19-impacts-and-policy-lessons-71b57aea/.
OECD. May 11, 2020c. "OECD Scheme for the Application of International Standards for Fruit and Vegetables." https://www.oecd.org/agriculture/fruit-vegetables/oecd-covid-19-impact-on-fruit-and-vegetables-trade.pdf.
Parreno-Marchante, A., A. Alvarez-Melcon, M. Trebar, and P. Filippin. 2014. "Advanced Traceability System in Aquaculture Supply Chain." *Journal of Food Engineering* 122: 99–109.
Patel, P., and D. Cassou. 2015. "Enabling High-Level Application Development for the Internet of Things." *Journal of Systems and Software* 103: 62–84.
Persson, F., and J. Olhager. 2002. "Performance Simulation of Supply Chain Designs." *International Journal of Production Economics* 77 (3): 231–245.
Pfeffer, J., and G. R. Salancik. 2003. *The External Control of Organizations: A Resource Dependence Perspective*. Stanford, CA: Stanford University Press.
Pingali, P., B. Mittra, and A. Rahman. 2017. "The Bumpy Road from Food to Nutrition Security–Slow Evolution of India's Food Policy." *Global Food Security* 15: 77–84.
Prince, K., M. Barrett, and E. Oborn. 2014. "Dialogical Strategies for Orchestrating Strategic Innovation Networks: The Case of the Internet of Things." *Information and Organization* 24 (2): 106–127.
Qu, W. G., and Z. Yang. 2015. "The Effect of Uncertainty Avoidance and Social Trust on Supply Chain Collaboration." *Journal of Business Research* 68 (5): 911–918.
Ramakumar, R. 2020. "Agriculture and the Covid-19 Pandemic: An Analysis with Special Reference to India." *The Journal* 10 (1): 72–110.
Rawal, V., M. Kumar, A. Verma, and J. Pais. 2020. *COVID-19 Lockdown: Impact on Agriculture and Rural Economy*. New Delhi: Society for Social and Economic Research.
Reardon, T., R. Echeverria, J. Berdegué, B. Minten, S. Liverpool-Tasie, D. Tschirley, and D. Zilberman. 2019. "Rapid Transformation of Food Systems in Developing Regions: Highlighting the Role of Agricultural Research & Innovations." *Agricultural Systems* 172: 47–59.
Rezaei, J. 2015. "Best-worst Multi-Criteria Decision-Making Method." *Omega* 53: 49–57.
Rezaei, J. 2016a. "Best-worst Multi-Criteria Decision-Making Method: Some Properties and a Linear Model." *Omega* 64: 126–130.
Rezaei, J., T. Nispeling, J. Sarkis, and L. Tavasszy. 2016b. "A Supplier Selection Life Cycle Approach Integrating Traditional and Environmental Criteria Using the Best Worst Method." *Journal of Cleaner Production* 135: 577–588.
Rezaei, J., W. S. van Roekel, and L. Tavasszy. 2018. "Measuring the Relative Importance of the Logistics Performance Index Indicators Using Best Worst Method." *Transport Pol* 68: 158–169.

Rizou, M., I. M. Galanakis, T. M. S. Aldawoud, and C. M. Galanakis. 2020. "Safety of Foods, Food Supply Chain and Environment Within the COVID-19 Pandemic." *Trends in Food Science & Technology* 102: 293–299.
Rossello, S., and P. Dewitte. 2020. "Anonymization by Decentralization? The Case of COVID-19 Contact Tracing apps." https://europeanlawblog.eu/2020/05/25/anonymization-by-decentralization-the-case-of-covid-19-contact-tracing-apps/.
Sahin, A. R., A. Erdogan, P. M. Agaoglu, Y. Dineri, A. Y. Cakirci, M. E. Senel, and A. M. Tasdogan. 2020. "2019 Novel Coronavirus (COVID-19) Outbreak: A Review of the Current Literature." *EJMO* 4 (1): 1–7.
Saint Ville, A. S., G. M. Hickey, and L. E. Phillip. 2017. "How do Stakeholder Interactions Influence National Food Security Policy in the Caribbean? The Case of Saint Lucia." *Food Policy* 68: 53–64.
Samaddar, S., and S. S. Kadiyala. 2006. "An Analysis of Interorganizational Resource Sharing Decisions in Collaborative Knowledge Creation." *European Journal of Operational Research* 170 (1): 192–210.
Schanes, K., K. Dobernig, and B. Gözet. 2018. "Food Waste Matters-A Systematic Review of Household Food Waste Practices and Their Policy Implications." *Journal of Cleaner Production* 182: 978–991.
Sharma, R., A. Shishodia, S. Kamble, A. Gunasekaran, and A. Belhadi. 2020. "Agriculture Supply Chain Risks and COVID-19: Mitigation Strategies and Implications for the Practitioners." *International Journal of Logistics Research and Applications*, 1–27. doi:10.1080/13675567.2020.1830049.
Singh, R. K. 2014. "Assessing Effectiveness of Coordination in Food Supply Chain: A Framework." *International Journal of Information Systems and Supply Chain Management* 7 (3): 104–117.
Singh, R. K., S. Luthra, S. K. Mangla, and S. Uniyal. 2019. "Applications of Information and Communication Technology for Sustainable Growth of SMEs in India Food Industry." *Resources, Conservation and Recycling* 147: 10–18.
Song, G., X. Gao, P. Fullana-i-Palmer, D. Lv, Z. Zhu, Y. Wang, and L. B. Bayer. 2019. "Shift from Feeding to Sustainably Nourishing Urban China: A Crossing-Disciplinary Methodology for Global Environment-Food-Health Nexus." *Science of The Total Environment* 647: 716–724.
Stephens, E. C., G. Martin, M. van Wijk, J. Timisina, and V. Snow. 2020. "Editorial: Impacts of COVID-19 on Agricultural and Food Systems Worldwide and on Progress to the Sustainable Development Goals." *Agricultural Systems* 183: 102873.
Stone, J., and S. Rahimifard. 2018. "Resilience in Agri-Food Supply Chains: A Critical Analysis of the Literature and Synthesis of a Novel Framework." *Supply Chain Management: An International Journal* 23 (3): 207–238.
Stranieri, S., F. Riccardi, M. P. Meuwissen, and C. Soregaroli. 2021. "Exploring the Impact of Blockchain on the Performance of Agri-Food Supply Chains." *Food Control* 119: 107495.
Tomlin, B. 2006. "On the Value of Mitigation and Contingency Strategies for Managing Supply Chain Disruption Risks." *Management Science* 52 (5): 639–657.
Tse, Y. K., M. Zhang, and F. Jia. 2018. "The Effects of Risk and Reward Sharing on Quality Performance." *International Journal of Operations & Production Management* 38 (123): 2367–2388.
Tsolakis, N. K., C. A. Keramydas, A. K. Toka, D. A. Aidonis, and E. T. Iakovou. 2014. "Agrifood Supply Chain Management: A Comprehensive Hierarchical Decision-Making Framework and a Critical Taxonomy." *Biosystems Engineering* 120: 47–64.
Wu, L., C. H. Chuang, and C. H. Hsu. 2014. "Information Sharing and Collaborative Behaviors in Enabling Supply Chain Performance: A Social Exchange Perspective." *International Journal of Production Economics* 148: 122–132.
Wunderlich, S. M., and N. M. Martinez. 2018. "Conserving Natural Resources Through Food Loss Reduction: Production and Consumption Stages of the Food Supply Chain." *International Soil and Water Conservation Research* 6 (4): 331–339.
Yuen, K. F., X. Wang, F. Ma, G. Lee, and X. Li. 2019. "Critical Success Factors of Supply Chain Integration in Container Shipping: an Application of Resource-Based View Theory." *Maritime Policy & Management* 46 (6): 653–668.
Zambrano, M. V., B. Dutta, D. G. Mercer, H. L. MacLean, and M. F. Touchie. 2019. "Assessment of Moisture Content Measurement Methods of Dried Food Products in Small-Scale Operations in Developing Countries: A Review." *Trends in Food Science & Technology* 88: 484–496.
Zhang, J., and J. Chen. 2013. "Coordination of Information Sharing in a Supply Chain." *International Journal of Production Economics* 143 (1): 178–187.

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1754-2731.htm

Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda

Received 18 February 2021
Revised 3 March 2021
Accepted 10 March 2021

Rajesh Kumar Singh
Management Development Institute, Gurgaon, India
Saurabh Agrawal
Delhi School of Management, Delhi Technological University, Delhi, India
Abhishek Sahu
Department of Mechanical, Production and Industrial Engineering, Delhi Technological University, Delhi, India, and
Yigit Kazancoglu
Logistics Management, Yasar University, Izmir, Turkey

Abstract

Purpose – The proposed article is aimed at exploring the opportunities, challenges and possible outcomes of incorporating big data analytics (BDA) into health-care sector. The purpose of this study is to find the research gaps in the literature and to investigate the scope of incorporating new strategies in the health-care sector for increasing the efficiency of the system.

Design/methodology/approach – Fora state-of-the-art literature review, a systematic literature review has been carried out to find out research gaps in the field of healthcare using big data (BD) applications. A detailed research methodology including material collection, descriptive analysis and categorization is utilized to carry out the literature review.

Findings – BD analysis is rapidly being adopted in health-care sector for utilizing precious information available in terms of BD. However, it puts forth certain challenges that need to be focused upon. The article identifies and explains the challenges thoroughly.

Research limitations/implications – The proposed study will provide useful guidance to the health-care sector professionals for managing health-care system. It will help academicians and physicians for evaluating, improving and benchmarking the health-care strategies through BDA in the health-care sector. One of the limitations of the study is that it is based on literature review and more in-depth studies may be carried out for the generalization of results.

Originality/value – There are certain effective tools available in the market today that are currently being used by both small and large businesses and corporations. One of them is BD, which may be very useful for health-care sector. A comprehensive literature review is carried out for research papers published between 1974 and 2021.

Keywords Big data, Big data analytics, Health-care analytics, Key challenges

Paper type Research paper

1. Introduction

Healthcare is the aggregation and convergence of industries within the economic structure that provides curative, preventive, rehabilitative and palliative care products and services for the treatment of patients. It consists of organizations that offer medical services, manufacture medical equipment or remedies, offer medical insurance or otherwise make it easier for patients to access healthcare. The health-care sector is one of the fastest-growing sectors generating huge and complex data (Davies, 2018). With an increase in health-care records,

The TQM Journal
© Emerald Publishing Limited
1754-2731
DOI 10.1108/TQM-02-2021-0051

compliances and regulatory requirements, the health-care industry is growing drastically (Raghupathi and Raghupathi, 2014). Organizations should manage their data and resources very carefully for sustainable competitive advantage (Singh *et al.*, 2008). Big data (BD) can play a very crucial role in the evolution of health-care practices. According to Economist Intelligence Unit (2019), World industry outlook (2019), over 2019–2023, the global healthcare spending is likely to increase at a compound annual growth rate (CAGR) of 5%, an increment of 2.7% from 2014 to 2018. This has grabbed the attention of many researchers, industry and the government toward "Big Data in Healthcare Sector" (Soleimani-Roozbahani *et al.*, 2019).

Global spending of healthcare, contributing to gross domestic product (GDP) is likely to stay at about 10.2% by 2023, equivalent to the 2018 ratio (Deloitte Insights – 2020 Global Health Care Outlook). BD in healthcare contains data including numerous medical prescriptions, insurance records, patients' feedbacks and responses, medication details, imaging, multiple test reports, demographic details and pharmaceutical prescriptions with various outcome details (Miller, 2012; Wamba *et al.*, 2013). Since we live in a digital era, BD has become an integral part of our lives. The role of data in business, sociology and engineering has been highlighted by the literature (Duan and Xiong, 2015). Utilizing useful data from this huge pool of data may be beneficial for the stakeholders. Previously, there have been technical constraints that have undermined the ability of data scientists to utilize these data. With recent technological advancements, generating and analyzing huge volumes of data has become faster than ever (Wamba *et al.*, 2015). Challenges related to information technology (IT) like lack of incorporation of IT system into the health-care systems is a severe hindrance of the value addition of the US health-care sector (Foidl and Felderer, 2015). Big data analytics (BDA) may be the one of the solutions. BDA enables the analysis of a huge volume of varied data and velocity of data across whole of the health-care system decisions (Raghupathi and Raghupathi, 2014). Real-time data, decision support systems (DSSs) and business intelligence (BI) tools can assist end users in gaining an insight into business processes and may be helpful to the business organizations (Bag, 2017). Wamba *et al.* (2017) mentioned BDA as the "changer that improves the efficiency and effectiveness of the business, thanks to its high operational and strategic potential." BD consists of 3V's: volume, velocity and variety (Russom, 2011; Duan and Xiong, 2015). Further, it has been stated by Wamba *et al.* (2015) that BD consists of 5V's: volume, velocity, variety, veracity and value. Hadoop is one of the prevalent systems used in managing and analyzing unstructured BD (Dwivedi and Dubey, 2014; Gandomi and Haider, 2015; Manwal and Gupta, 2017).

Humans are getting familiar with the rapid technological advancements, computing processes data with greater variety, volume and velocity (Lee, 2017). The significance of 3V's is elaborated by Zikopoulos and Eaton (2011) and Chen *et al.* (2014). Volume is concerned with volume of BD, which is used to handle a huge amount of data, while variety means that BD consists of data from multiple sources that are processed in various formats such as structured, unstructured, multifactor, probabilistic. Velocity implies data processing speed, i.e. data handled by BD require quick data processing to assist in decision-making processes (Mcafee *et al.*, 2012b). Later on, two more characteristics of BD are added with 3V's, i.e. "veracity" and "value." Veracity refers to how appropriate the data are, for the user and its credibility (Mcafee *et al.*, 2012b). Value refers to the savings made by using BD. It is the "economic value" that benefits health-care organizations (Wan and Alagar, 2016). It further refers to knowledge discovery in data analytics (Wixom *et al.*, 2013; Alotaibi *et al.*, 2020).

BDA deals with numerous data patterns and gives the meaningful insights from them (Anuradha, 2015). BD provides a range of health-care industry opportunities, such as increased health-care productivity and quality, health threat identification, early disease diagnosis and quicker decision-making for better decisions (European Public Health Alliance, Health and care in the single digital market: refection paper, 2017; Leite *et al.*, 2020) (Kruse *et al.*, 2016). BDA helps to understand, assess and improve the effectiveness of costly

treatments (European Public Health Alliance, Health and care in the digital single market: refection paper, 2017). However, storing, standardizing and transferring data, privacy and data protection are some of the major challenges of the industry (Bhadani and Jothimani, 2016; Vinodh *et al.*, 2020).

The application of BDA is continuously growing in health-care industry with an increase in the amount of data generated (Galetsi and Katsaliaki, 2020; Kamble *et al.*, 2018). The application of BDA includes adequate resource utilization, prediction of any outbreak of infectious disease and monitoring and prevention of disease (Sahoo *et al.*, 2016). They help in identifying the patient with rare diseases and to monitor the quality of services provided to patients by private and government hospitals (Chen and Zhang, 2014; Archenaa and Anita, 2015; Sahoo *et al.*, 2016). Mehta and Pandit (2018) explained the data collected in healthcare as structured data (e.g. genomics data) or unstructured data (e.g. medical imaging). In healthcare, the heterogeneous data mainly collect the demographic details of the patients (Malik *et al.*, 2018), the patient's treatment past records (Ozminkowski *et al.*, 2015) and the investigative reports (Amirian *et al.*, 2017).

Kolasa *et al.* (2020a) explained the shortage of financial resources that badly impacted the health-care sector. Because of fluctuating behavior of BD, the health-care organizations are also changing and updating themselves by deploying new technologies in their organizations (Harerimana *et al.*, 2018; Zhang *et al.*, 2015). Kamal *et al.* (2020) presented a new algorithm, NextSTMove where the future movement of vehicles is predicted with or without Apache Spark. This study is useful in predicting the future location of emergency vehicles. The algorithm has successfully reduced the processing time up to 300% and is accurate in 75–85% of typical cases. Soleimani-Roozbahani *et al.* (2019) have done a meta-analysis on published research work during the period 2008–2018. In their study, they evaluated 256 papers from seven databases and found each work had an average of four writers. Most of the studies conducted in this period have used "Meta-analysis and evidence" as a methodology for the study. Daghistani *et al.* (2020) proposed a framework to handle demand uncertainty and recompensate patient no-show appointment. This study has analyzed a rich data set consisting of 2,011,813 visits, collected from patients and utilized machine learning techniques to formulate prediction. According to Stanford Medicine (2020), the democratization of healthcare (2018) explored the growth of artificial intelligence (AI) in healthcare. They conducted a survey and explained that 34% said that healthcare would be more accessible to people, 31% responded the diagnosis would be more accurate and faster, 27% responded that AI would make recommendations better and another 27% responded people would have health-care specialist with them any time anywhere on their devices. Therefore, it is important to answer the following research question:

RQ1. What are the current issues/challenges of BD in health-care sectors?

RQ2. How can we deal with these issues/challenges?

To answer the abovementioned research questions, the proposed study and their detailed discussion will have the potential to answer all these questions.

Remaining proposed article is organized as follows: Section 2 describes the research methodology followed for the literature review. Key opportunities and benefits of BDA in health-care sector and issues and challenges are discussed in section 3 and section 4, respectively. Strategies are explained in section 5. In section 6, application of BD is described. Synthesis of findings and research gap analysis are discussed in Section 7. At last, article is concluded along with its limitations and future scope of research in the Conclusion section.

2. Research methodology

The state-of-the-art literature review approach suggested by Seuring and Gold (2012) and Mayring and Fenzl (2014) is followed as the research methodology for the proposed research. A similar literature review approach has been used by authors such as Agrawal *et al.* (2015), Singh *et al.* (2016), Kumar and Singh (2017), Senthilkumar *et al.* (2018) and Khanra *et al.* (2020) in the area of healthcare pertaining to BD. The step-by-step approach is explained in the following section.

2.1 Material collection

Material collection is the process of collecting relevant materials through intuitive and innovative search from the relevant literature. It begins with the primary attention such as definitions and keywords. The unit of analysis is also defined for the search and collection. The unit of analysis for the proposed research is defined as a single research article/book/report. The articles were collected potentially by using the database of Scopus, Web of Science and Google Scholar. The keywords "big data," "big data analytics," "healthcare," "challenges," "issues" and "healthcare analytics" and their all-possible combinations were used for searching and collecting the research articles. Using above criteria, the initial search resulted into 1,600 articles for further search. Further, screening criteria were applied after going through abstract and 350 articles were selected for further analysis. Finally, 215 articles comprised of 200 articles from international journals of repute were selected for the study.

2.2 Descriptive analysis

Based on Kamal and Irani (2014), a review protocol is developed to guide the researchers for conducting a literature review. It helps to make research process clear and consistent. The review procedure provides detailed guidelines for defining and screening relevant review articles, along with detailing the methods of review for the entire process. This helps to decrease biasness in the evaluation process and minimize duplication with existing reviews.

By this analysis, a solid background is developed upon which the theoretical analysis is conducted to assess the different formal aspects of publications. A number of articles from different journals were sorted out and illustrated in Figure 1. Major number of the articles are

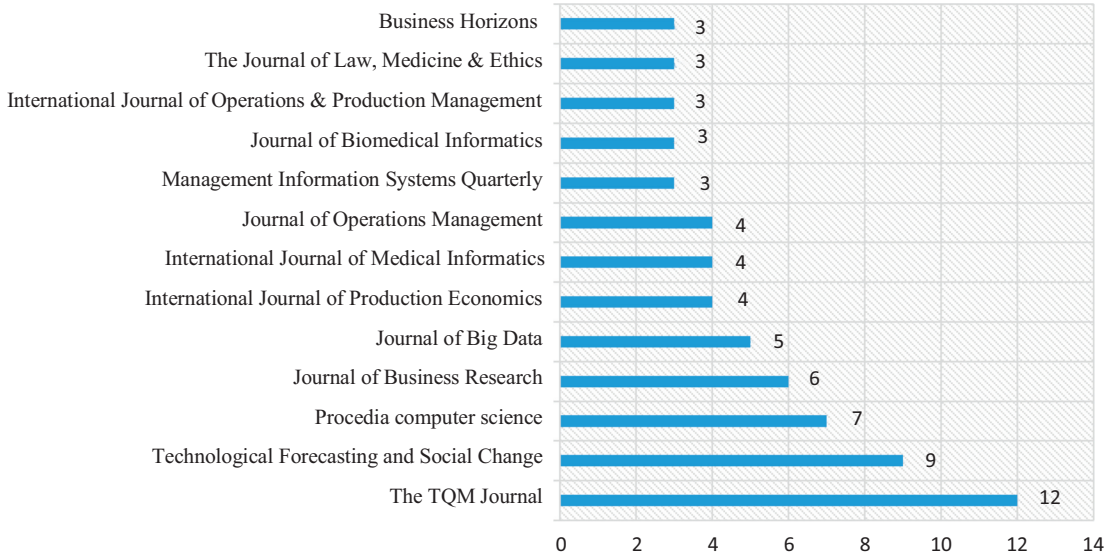


Figure 1.
Journal-wise distribution of reviewed articles

published in prominent journals such as *The TQM Journal*, *Technological Forecasting and Social Change*, *Procedia Computer Science*, *Journal of Business Research* and *Journal of Big Data*. The articles are also sorted out based on annual publications from the year 1974 to 2021. It is evident from Figure 2 that most of the articles are published in the last seven years (138 articles) and recently numbers have increased considerably because of growing interest in this area of research.

2.3 Category selection

The selected 215 articles were categorized into five research categories, as shown in Figure 3. Further, both the inductive and deductive approaches were used to select the research articles. Some key challenges and issues were selected through deductive approach and others were selected after the completion of the literature review through inductive approach.

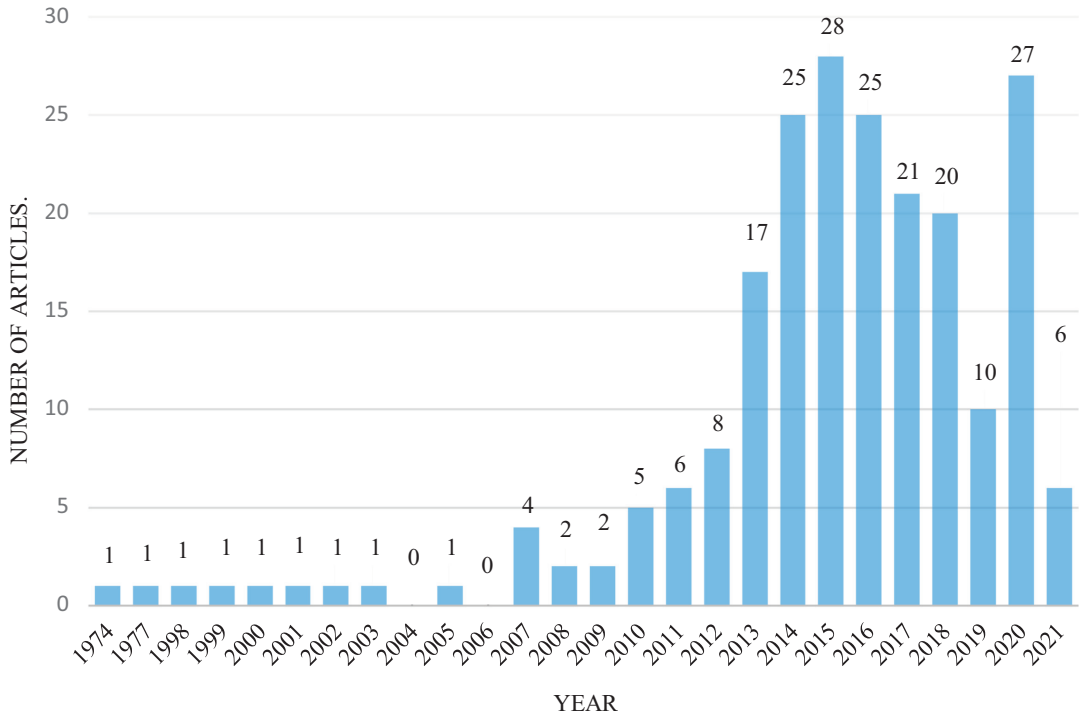


Figure 2.
Annual distributions of the published articles (1974–2021)

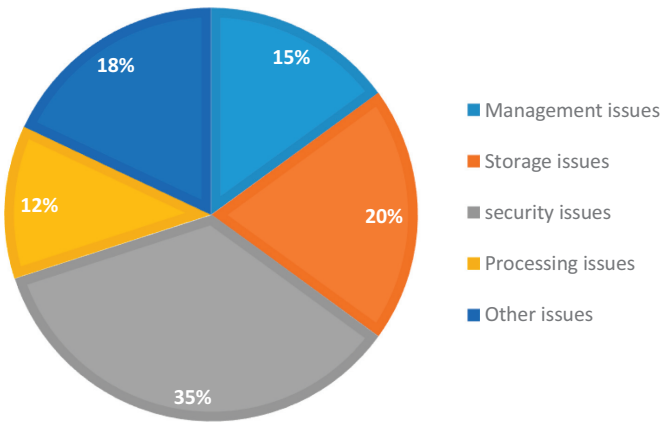


Figure 3.
Issue-wise distribution of published articles on BDA in healthcare

The issues can be categorized into management issues, storage issues, processing issues, security issues and other issues. These categories are further subcategorized, which is represented through a pie chart, i.e. shown in Figure 3.

2.4 Material evaluation

The material evaluation is the process of identifying the relevancy of the selected research articles. A total of four researchers were selected to ensure the validity of the selected articles that used inductive approaches for increasing the reliability of research. These articles are further cross-examined through the number of citations, site reliability, the number of volumes and online feedback features.

In order to find out the opportunities, benefits, the challenges faced in using BDA in the health-care sector and formulate the different strategies for the smooth functioning of BD in the health-care sector, various papers have been reviewed mainly from Emerald Insight (www.emeraldinsight.com) and ScienceDirect (www.sciencedirect.com). Year-wise selection of the articles has been shown in the above bar graph. The keywords for the search were "Big data in healthcare," "Hadoop" and "Healthcare."

3. Key opportunities and benefits of BDA in health-care sector

It may not appear at first glance that the worlds of BD and healthcare may have something in common or go well together, but there are a lot of benefits in store. BD has found widespread use in the health-care sector to forecast epidemics, cure illnesses, improve the quality of life, avoid avertible deaths (Manyika *et al.*, 2011; Stadler *et al.*, 2016; Abouelmehdi *et al.*, 2018; Majnarić *et al.*, 2021). The individual's genetic properties are used to help in selecting most appropriate drugs according to the genetic history and may help in avoiding any side effects (Luo *et al.*, 2016). Decision-making culture driven by data is needed, where decision makers make a conclusion based on data rather than taking decisions by intuition only (McAfee *et al.*, 2012a). Currently, machine learning mechanism is being explored to track pigmented lesions over time and identify those at a higher risk of malignancy (Li *et al.*, 2016). Machine learning has recently been used to come up with a skin cancer diagnosis algorithm (Esteva *et al.*, 2017). As machine learning includes both AI and the use of huge databases, there is a lot of scope in research for applying machine learning in order to resolve the task (Chin, 1998; Van der Velde, 2010; Elsebakhi *et al.*, 2015). Current challenges show that it is going to play a very important role in BDA (Hoang and Pham, 2016).

Cloud computing includes many information and computing resources like operating systems, storage, networks, hardware, databases and software applications which are used in BDA (Chiregi and Navimipour, 2016; Mourtzis and Vlachou, 2016). It provides important financial advantages to many business organizations. The attention now rests entirely on thoroughly knowing the patient, as early as possible in their life, and ideally picking up any symptoms of a serious illness early, making treatment easier (Azad and Navimipour, 2017). BD is used and analyzed by companies to increase accountability and decision-making and also to improve collaboration (Hazen *et al.*, 2014). It was observed by Vachon and Klassen (2008), Pagell and Wu (2009), Bag *et al.* (2020a), Trakulsunti *et al.* (2020) that collaboration among supply chain partners is used to meet sustainability goals and address environmental issues. The use of BDA is aiding in this and helping organizations to form collaborations with more ease. BDA capability is the ability to handle a large amount of volume data in order to allow users to conduct data analysis and respond (Hurwitz *et al.*, 2013). The different data analytics capabilities as stated by Wang *et al.* (2016) are as follows:

(1) Analytical skills for patterns of care: It involves analytical procedures usually used in BDA to handle enormous quantities of data that differ in velocity through the

implementation of data storage, management, analytics and visualization techniques (Chen *et al.*, 2012).

(2) Unstructured data analytical capability: Data are collected from sources related to health-care sectors, stored in disseminated database systems. They are filtered and then analyzed in order to get meaningful outcomes (Wang *et al.*, 2016).

(3) Decision support and predictive capacity: Predictive analytics in healthcare refers to the use of unstructured historical data to forecast the future, which is commonly used to decrease the degree of uncertainty and helps managers to make informed decisions more quickly (Russom, 2011; Suresh, 2016).

(4) Traceability: It is the capability to monitor output data from all the components of the system of entire organization (Groves *et al.*, 2013).

Duan and Xiong (2015) suggested that research is currently at an embryonic stage based on the benefits of big data and predictive analytics (BDPA). Currently, the studies on BD are more inclined toward device setup: data collection, storage, networking and computation of distributed systems. It is time for the system infrastructure to concentrate on critical and important resources that will build advance BDPA capabilities in the coming years (McAfee *et al.*, 2012a; Bag *et al.*, 2021). Shang and Seddon (2002) developed a framework describing the various advantages of using BDA in healthcare. The advantages are as follows:

(1) The benefits of IT infrastructure

(2) Operational advantages

(3) Organizational advantages

(4) Managerial advantages

(5) Strategic advantages

In the sense of smart cities, smart health is a new implementation of BD that offers an extraordinarily smart, user-centric climate. Smart health is an upcoming but highly hopeful area of research for medical informatics, public health and business, using intellectual capabilities via the Internet of Things (IoT) (Pramanik *et al.*, 2017; Saqlain *et al.*, 2020; Au-Yong-Oliveira *et al.*, 2021). Timely monitoring of patients using wearables and embedded systems (such as activity tracker, pacemakers, etc.) helps to minimize daily activities (Kolasa *et al.*, 2020b). BD has contributed to the modern age of electronic and mobile health in smart health-care systems, which increase efficiency and reduce the cost and time involved (Varshney, 2014). Due to the ever-increasing use of social media and electronic devices, smart health has been gaining a lot of popularity these days (Raghupathi and Raghupathi, 2014). Mobile-based health applications of BD and AI in healthcare have numerous benefits (Cavallone and Palumbo, 2020). These m-health-based mobile applications help to remotely monitor the health (physical and mental) condition of the patients (Reinertsen and Clifford, 2018). The data collected from healthcare can be used for early fraud detection and prevention, improvement of public health surveillance, development of new and effective devices and drugs for patients, to speed up the service time and to reduce cost (Abouelmehdi *et al.*, 2018; Raghupathi and Raghupathi, 2014). BDA helps in real-time monitoring of hospitals and helps them to ensure, improve and optimize the quality of the services (Gowsalya *et al.*, 2014). It helps in caring of patients (Salomi and Balamurugan, 2016) and reduces the readmission rates of them to hospitals (Gowsalya *et al.*, 2014). It helps in safe monitoring and provides proactive care (Sendra *et al.*, 2018). Dash *et al.* (2019) explain that organizations can save over 25% annually with the implementation of BD because organizations will predict the disease early,

with which the diagnosis becomes easy and the readmission rate decreases. BDA improves the pharmaceutical supply chain, forecasts demand, optimize staffing and streamline patient care, which ultimately helps in cost reduction. Bakker *et al.* (2020) conducted a study of 71 papers and followed preferred reporting items for systematic reviews and meta-analyses extension for scoping reviews (PRISMA-scr) checklist (Tricco *et al.*, 2018). They found that the analytics may generate savings, improve the result and performance of the organization. Adoption of BDA helps in effective communication with the patients (Dobrzykowski and Tarafdar, 2015; Bag *et al.*, 2021). BDA helps organizations to develop a cost-effective health-care system which primarily focuses on patient's care (Zhang *et al.*, 2018). Yu *et al.* (2020) explained the importance of information sharing and coordination in health-care supply chain. Information sharing in health-care system helps the organization in generating new business quickly (Galbraith, 1974). Information processing increases the operational flexibility of healthcare (Yu *et al.*, 2020). Adopting information systems (e.g. real-time locating systems, electronic health record systems, etc.) helps the organizations to process and analyze the data in real time. This further helps to improve the data-driven decision-making process of the organizations (Wang and Byrd, 2017; Wang *et al.*, 2019; Ward *et al.*, 2014), improves patient's experience and delivery service time (Dobrzykowski and Tarafdar, 2015; Mandal, 2017; Zepeda *et al.*, 2016). Mehta *et al.* (2019) examined that BD tools should be explored to improve the quality of health-care service. Harlow (2018) explained that BDA helps in the improvement of knowledge sharing. Integration of supply chain in healthcare helps in free flow of information, goods and patients in hospitals (Chen *et al.*, 2013; Drupsteen *et al.*, 2013; Bag *et al.*, 2020b). Lack of this integration leads to inaccurate and distorted supply of information among the system, which leads to providing poor facility to the patients (Wong *et al.*, 2011). Application of BDA helps in strategic collaboration among the health-care network (Dobrzykowski and Tarafdar, 2015; Srinivasan and Swink, 2018; Wang *et al.*, 2019). BDA also helps in revenue generation and helps to reduce a significant amount when used effectively (Kwon *et al.*, 2016). Table 1 provides detailed descriptions of key opportunities and benefits of BDA in health-care sector.

4. Issues and challenges

While organizations are adopting BD, there are a number of issues and challenges. These issues can be categorized into management issues, storage issues, processing issues (Ji, 2012), security issues and other issues.

(1) Management issues

The data involved are a collection of large volumes of structured, semi-structured and unstructured data from an organization, government sector, private and public administration (Russom, 2013). The main goal of BD management is to ensure the good quality of data, data ownership, responsibilities, standardization and documentation of data. It also helps in making the data accessible. Gartner (2014) gave his views on this issue stating that "Big data" challenge is more than just managing huge volumes of data.

(2) Storage issues

The storage of BD is dealt with the method of visualization, where it holds a large set of sensor information, media, videos, e-business records and cell phone signal coordinators. Most of the companies which use BD storage like EMC, IBM, Netapp and Amazon handle large volumes of data by using tools like NoSQL, Apache Drill and Horton Works (Saraladevi *et al.*, 2015). Storing the real-time information, updating it after few moments and filtering the redundant facts are a difficult task (Sharma *et al.*, 2020; Suciu *et al.*, 2015).

(3) Processing issues

Opportunity/benefit	References
I. Complex algorithms like machine learning and IoT are used to set up a predictive model	Raghupathi and Raghupathi (2014), Li *et al.* (2016), Russom (2011), Pramnik *et al.* (2017), Hoang and Pham (2016), Kaur *et al.* (2018), Wang *et al.* (2018), Yadav *et al.* (2020)
II. Big data in smart health-care systems has led to the new era of electronic and mobile health, which has increased efficiency and reduced the cost and time involved	Sahoo *et al.* (2016), Jiang *et al.* (2014), Varshney (2014), Raghupathi and Raghupathi (2014), Palanisarmy and Thirunavukarasu (2019)
III. Data-driven decision-making	Abouelmehdi *et al.* (2017), Wang and Byrd (2017), Ward *et al.* (2014), McAfee *et al.* (2012a), Russom (2011), Hurwitz *et al.* (2013), Waller and Fawcett (2013), Hazen *et al.* (2014), Wang *et al.* (2016), Kache and Seuring (2017), Gold *et al.* (2010), Sahu *et al.* (2020)
IV. Improved transparency within the organization	Waller and Fawcett (2013), Schoenherr and Speier-Pero (2015), Hazen *et al.* (2014), Wang *et al.* (2016), Kache and Seuring (2017), Gold *et al.* (2010)
V. Helps to form easy collaborations, social and environmental sustainability	Vachon and Klassen (2008), Pagell and Wu (2009), Keeso (2014)
VI. Transition from reactive to proactive and patient centered	Chen *et al.* (2012), Wang *et al.* (2018)
VII. Traceability	Groves *et al.* (2013), Hurwitz *et al.* (2013), Palanisarmy and Thirunavukarasu (2019)
VIII. Analytical capability and conversion of unstructured to structured data	Chen *et al.* (2012), Wang *et al.* (2015), Hurwitz *et al.* (2013), Palanisarmy and Thirunavukarasu (2019)
IX. Increased use of mobile phones and the Internet for healthcare, leading to an increase in the profits and producibility	Chen *et al.* (2014), Burrows and Savage (2014), Chiregi and Navimipour (2017), Palanisarmy and Thirunavukarasu (2019)
X. Helps in improvement of the infrastructure, strategy making and operational tasks	Dobrzykowski and Tarafdar (2015), Srinivasan and Swink (2018), Wang *et al.* (2019), Gupta and George (2016), Duan and Xiong (2015), McAfee *et al.* (2012a), Shang and Seddon (2002), Wang *et al.* (2018)
XI. Helps in identifying fraudulent activities (like fraudulent health insurance claims or any terrorism activity)	Chen (2018), Archenaa and Anita (2015), Sahoo *et al.* (2016), Strang and Alamieyeseigha (2015), Terry (2015), Srinivasan and Arunasalam (2013)
XII. Helps in early detection of diseases	Abouelmehdi *et al.* (2018), Wang and Hajli (2017), European Public Health Alliance, Health and care in the digital single market: refection paper (2017), Chen and Zhang (2014), Archenaa and Anita (2015), Sahoo *et al.* (2016), Gu *et al.* (2017), Raghupathi and Raghupathi (2014)
XIII. Helps in hospital monitoring	Archenaa and Anita (2015), Srivathsan and Arjun (2015), Sultanow and Chircu (2015)
XIV. Helps in patient's care	Salomi and Balamurugan (2016), Gowsalya *et al.* (2014)
XV. BDA helps in cost cutting and increase revenue	Bakker *et al.* (2020), Dash *et al.* (2019), Abouelmehdi *et al.* (2018), Kwon *et al.* (2016), Raghupathi and Raghupathi (2014)
XVI. Improves the pharmaceutical supply chain, forecasts demand, optimize staffing and streamline patient care	Dash *et al.* (2019), Abouelmehdi *et al.* (2018), Sedevich-Fons *et al.* (2014)

Table 1.
Key opportunities and benefits of BDA in health-care sector

The processing of BD analyzes the size of BD in petabyte, exabyte or even in zettabyte, either in batch processing or stream processing.

(4) Security issues

As BD consists of large volumes of data, there are challenges faced in managing this set of data in a secure manner. There are more threats involved in the public and private database, leakage of data is also common. It is very easy for the hackers to collect private resources because of the flawed private and public policy. Researchers have warned about the privacy and security concern of patient's data during intersystem exchange (Weng and Kahn, 2016; Mohammed *et al.*, 2014). Data and system hackers often store the data from public databases in their own Universal Serial Bus (USB) or other personal devices. They attack the data storage by sending messages like denial of service (Chen, 2007) spoofing attack and brute force attack (Bernstein, 2005).

Kolasa *et al.* (2020a, b) explain data protection issues which are very common to digital services platforms. Hence, data security and data privacy should be given more importance in the coming years. BD incurs huge initial investments (Sultanow and Chircu, 2015; Szlezak *et al.*, 2014; Wu *et al.*, 2016). Still, many organizations are investing heavily in BD to reap its benefits. But organizations are struggling a lot and lacks to unfold the potential benefits of BD in healthcare (Wang and Hajli, 2017; Wang *et al.*, 2019). Batarseh *et al.* (2020) expressed concern about the growth of fraudulent activities in healthcare. They explained that there is a lack of scientific proof and approvals from government authorities for any recommendations made to the patients with regard to clinics or pharmaceutical companies. However, Sahoo *et al.* (2016), Archenaa and Anita (2015), Strang and Alamieyeseigha (2015), Terry (2015), Chen and Zhang (2014) and Hamilton (2020) explained that BD helps in early detection of fraudulent activities in healthcare. These activities can be any fraudulent health claims or any terrorist activities in the field of healthcare.

(5) Other issues

The other issues with BD in healthcare include the cost and difficulty in the storage of data. The processing of data is another limitation. The data need advanced techniques for the analysis and formatting, but the main concerns are about accuracy, reliability and security (Mackenzie *et al.*, 2017; Wehner *et al.*, 2017; Brunetti *et al.*, 2020). While the tools like ERM or IRA-NRM are used for the integration into clinical work flows, they have their own drawbacks such as privacy of the patient, the unreadiness of the entire industry, the lack of information system infrastructure and structure and the data ownership (Mounia and Habiba, 2015; Feldman *et al.*, 2015). Due to lack of technical support and a lag in the use of analytical techniques in the BDA, it is not completely implemented (Sakr and Elgammal, 2016; Ferranti *et al.*, 2010; Fihn *et al.*, 2014). In the health-care industry, the BD platforms require a great deal of technical and programming skills, the end user may not possess. The lack of technical skills and the resistance to change, both affect the quality of the data collected from the organization (Zhang *et al.*, 2015; Wang *et al.*, 2018; Brunetti *et al.*, 2020). This poor quality of data further affects the quality of insights from that data, leading to insufficient insights and misleading suggestions (Sáez and García-Gómez, 2018; McNutt *et al.*, 2016). In a survey, it was found that 60% of surveyed health-care organizations struggle to develop a clear, integrated business strategy and vision for the implementation of analytics across a wide variety of functions (Taylor, 2015).

A lack of understanding on economic potential of BDA is also among one of the reasons in the failure (Groves *et al.*, 2013; Murdoch and Detsky, 2013). BD comes with inherent challenges that include volume, velocity, variety and veracity. The various challenges faced due to the use of BD in health-care sector can be classified as follows.

4.1 Data challenges

Volume of data is an important factor. The IoT is creating exponential growth in data. Large scale and the high volume of data is a big challenge presented in front of us. The latter

argument is also supported by Barnaghi *et al.* (2013) and Sivarajah *et al.* (2017). These data collected in healthcare include personal information, biometric sensor readings, radiology images, 3D imaging and patient's personal medical records (Rghioui *et al.*, 2020). Velocity is equally important. BD velocity deals with the pace at which data flow in. The flow of data is massive and continuous. Sampling data can help to deal with issues like volume and velocity. Big healthcare data are devastating not because of their heterogeneous nature (Ullah *et al.*, 2017: Frost, 2015). One of the biggest challenges of BD is defined in variety. It can be unstructured and so many different data types can be used. Variability is distinct from diversity. Six different coffee blends can be sold by a coffee shop but if you get the same blend every day and it tastes different every day that is variability. The same is true for data, which can have a major effect on your data homogenization if the context is constantly changing. Veracity is all about ensuring that the information is reliable, which includes procedures to prevent the systems from accumulating bad data. In today's world, visualization is important. It is much more powerful to express context using charts and graphs to visualize vast volumes of complex data than spreadsheets and reports chock-full of numbers and formulas. Value is the end game. You want to make sure that your company gets value from the data after addressing volume, speed, variety, variability, veracity and visualization, which takes a lot of time, effort and money. Another big issue is heterogeneity of data. It is a major challenge in BDA (Dash *et al.*, 2019). The huge amount of data having variety of structured, unstructured and semi-structured data increases the complexity. Structured data are the clinical data, unstructured and semi-structured data are the medical records, images or doctor notes. The major challenges here are combining the traditional data with the new data to get a piece of insightful information about the patient (Rghioui *et al.*, 2020). In healthcare sector, the velocity of data is equally important. For example, the life and death of patient can depend on real time basis. Hence, the diseases should be detected as early as possible to save lives (Moran *et al.*, 2013; Rghioui *et al.*, 2020). Lastly, data storage is one of the primary challenges which many organizations are facing.

4.2 Process challenges

Process challenges are the collection of challenges that occur as the data are analyzed and understood in-depth, from gathering the data at the beginning to evaluating and presenting the final results. Since large groups of data used are typically unstructured, the processing of such semi-structured data sets presents a much greater challenge than the data problems themselves.

4.3 Management challenges

(1) Health information ownership, use and security

The lack of advanced technology that guarantees data protection such as honesty, confidentiality, accessibility and transparency, the problems of data security are magnified as data sources become pervasive (Demchenko *et al.*, 2013). In the debate about how to handle BD in healthcare, critical questions are also posed regarding who controls those types of health information, how and by whom that information can be used and for what purposes.

(2) Personally, identifiable information

Among the biggest privacy risks, there is the loss of personally identifiable details such as birth dates, driver's license numbers and security numbers. Although external threats dominate top-of-mind discussions, data breaches are increasing, posing the potential for substantial consumer damages, high claims for compensation, litigation and permanent reputation damage.

4.3.1 Clinical data. A broad variety of patient-specific information is found in electronic health records including medication data, care history and other data. A hacker may use it to access unwanted medical treatment or bill for services never received along with policy number. The leakage and/or corruption of such data may also lead to irrevocable harm to the personal and professional life of the individual. The use of BD in healthcare can contribute at different levels such as early detection, effectiveness and the quality of treatments by the identification of early signals and disease intervention, decreased probability of adverse reactions, etc. (Pastorino *et al.*, 2019).

4.3.2 Financial data. The medical industry is becoming an easy target for hackers because other industry like banks and individuals are becoming more vigilant about securing their financial information. Outsourcing billing operations and increasing Internet and mobile healthcare participation can further create opportunities for possible data theft, which may result in legal implications.

4.3.3 Behavioral data. Thanks to surveillance devices, Global Positioning System (GPS) tracking, Internet site visits, social media, buying patterns, workout activity and self-reporting, behavioral data are the newest and probably fastest-growing in healthcare. Increasingly, behavioral data are becoming the "hot favorite" for cyber criminals as they help to create shockingly accurate depictions of human actions that are of great demand for marketing firms and often with criminal intentions. This information is becoming more vulnerable to fraud with the increasing use of tablets, smartphones and other mobile devices.

4.3.4 Data governance. Another complicated twist is that privacy becomes more difficult with the vast amount of data available, well beyond the safety of protected health information (PHI) or credit card numbers. It is easier to reveal intellectual property or potentially harmful information that may have been contained in unstructured data sources. Inferences (e.g. using Semantic Web technologies) can be made with a large number of data points to reveal information that has historically been much harder to piece together. In addition, effective BD governance is important to ensure the quality of data collected and analyzed from a broad data set pool (Hashem *et al.*, 2015). Several businesses are comfortable with holding information at their own sites. This helps them to have access and control over data anytime. It also helps in securing the data but storing data at their own infrastructure and maintaining them on their own incur huge cost. Hence, it is suggested to adopt cloud-based storage using IT infrastructure (Milani and Navimipour, 2016). Cloud-based storage offers easy recovery of the data. Organizations should choose their providers wisely, who understands the compliances and security issues in this sector (Dash *et al.*, 2019; Confetto and Covucci, 2021). Kolasa *et al.* (2020a) explained 80% of data present today are unstructured data and are difficult to process. Security and privacy are yet other issues to be addressed in today's era. Table 2 provides detailed information about the various issues/challenges while adopting BDA in health-care sector.

5. Strategies

The findings of a systematic review on the current state of BD science state that technical issues have been emphasized during most of the research (Dubey *et al.*, 2019). However, organizations will not be able to exercise most of the benefits of using BDA unless they are able to manage the managerial challenges effectively and efficiently (Mcafee *et al.*, 2012b; Sony *et al.*, 2021), make correct strategic choices and proper configuration of resource (Xu *et al.*, 2016) and properly understand the managerial, economic and strategic impact that BDA will have on this sector (Raghupathi and Raghupathi, 2014; Ward *et al.*, 2014). A single system does not exist that involves all the issues including reliability, scalability, security and ease of access in BD of healthcare, and thus, it is thoroughly disregarded (Hayes and Krippendorff, 2007).

Issue/challenge	References
I. Maintaining high quality of data	Daghistani *et al.* (2020), Gartner (2014) Lukoianova and Rubin (2014), Hashem *et al.* (2015), Chen *et al.* (2016)
II. Storage of a variety of data together	Dash *et al.* (2019), Frost (2015), White (2014), Barnaghi *et al.* (2013)
III. Processing and understanding the data due to lack of infrastructure	Demchenko *et al.* (2013), Chen (2014), Feldman *et al.* (2015)
IV. Storing large volumes of data	Sharma *et al.* (2020), Dash *et al.* (2019), Daghistani *et al.* (2020), Barnaghi *et al.* (2013), Chen *et al.* (2016), Palanisarmy and Thirunavukarasu (2019)
V. Leakage of private information	Sharma *et al.* (2020), Kolasa *et al.* (2020b), Thorpe and Gray (2015), Hoffman and Podgurski (2013), Kshetri (2014), Filkins *et al.* (2016), Rothstein (2015) Feldman *et al.* (2015), Chen (2007), Palanisarmy and Thirunavukarasu (2019)
VI. Authenticity of the data is hampered as some unwanted alterations happen because of lack of integrity	Demchenko *et al.* (2013), Palanisarmy and Thirunavukarasu (2019)
VII. Increase in the number of hackers who are finding new ways to collect private data	Filkins *et al.* (2016), Demchenko *et al.* (2013), Chen (2007)
VIII. Research focused more on the technical issues rather than managerial issues	Sakr and Elgammal (2016), Ferranti *et al.* (2010), Fihn *et al.* (2014), Wamba *et al.* (2015)
IX. Most of the health-care organizations do not have a suitable strategy and vision for analytics development	Sakr and Elgammal (2016), Ferranti *et al.* (2010), Fihn *et al.* (2014), Brynjolfsson and McAfee (2011), Chen *et al.* (2016)
X. A lack of understanding on the economic potential	Groves *et al.* (2013), Murdoch and Detsky (2013)
XI. Initial investment, deployment cost	Morse *et al.* (2020), Szlezak *et al.* (2014), Wu *et al.* (2016), Sultanow and Chircu (2015)

Table 2.
Issues/challenges of adopting big data in health-care sector

A two-way method was introduced by Mounia and Habiba (2015) to ensure patients' privacy. The first procedure focused on controlling unauthorized access to medical data during production, processing, sharing and storing data. This is done by the application of a variety of cryptographic algorithms and data backup strategies. The second half procedure consists of applying some privacy constraints such as one-time passwords (OTPs). The second procedure was focused on the generation and protection of the patients' medical records. A recent survey by the Stevenson and Davis (2013) found that some of the training methods include mentoring, cross-functional team-based training and self-study which might be able to help employees. Some other ways which can help are

(1) Implementing data governance

(2) Information-sharing culture

(3) Employees should be trained to use BDA

(4) Cloud computing techniques should be used.

Wang *et al.* (2015) proposed a framework that comprised of five layers: (1) data, (2) data aggregation, (3) analytics, (4) information exploration and (5) data governance. Each of these layers has specific functions and will help in transforming the clinical data into clinical information, which helps in the decision-making.

Important-performance analysis (IPA) model was proposed by Martilla and James (1977) in which two axes, one for importance and the other for performance is used to divide it into

four grids: keep up the good job, emphasis here, low priority and potential overkill. Analysis and strategic decision-making with the help of this method becomes very easy. However, according to some scholars, IPA led to errors in the decision-making process (Mikulić and Prebežac, 2008). According to Hansen and Bush (1999) and Lin (2015), IPA is a clear and effective model of strategic evaluation. Smart city is described by IBM as the intelligent use of advanced technology to identify, analyze, process and incorporate large volumes of useful information about core systems in running cities (Stahl, 2011). It can provide smart answers to different daily needs (Pramanik *et al.*, 2016; Qin *et al.*, 2010).

The structure of a smart city consists of three important layers: perception layer, network layer and application layer (Su *et al.*, 2011). Each of the layers processes and analyzes a large volume of sets of data. Ubiquitous computing and environmental intelligence concepts are used to build health-care systems that are predictive, personalized, preventive and participatory (Röcker *et al.*, 2014). In health-care systems, nowadays modeling of data and processing is almost real time so that the patients can be updated instantaneously. According to Middleton *et al.* (2013), by 2020, 26 bn IoT devices will be functional. Real-time responsiveness should be considered as the highest priority for health-care applications like clinical text mining, predictive analytics, longevity analysis, patient similarity, genetic data analysis and intelligent services (Sakr and Gaber, 2014; Groves *et al.*, 2013; Pashazadeh and Navimipour, 2018). Proper governance and proper tools can harness large volumes of heterogeneous data (Bardhan *et al.*, 2015; Bates *et al.*, 2014). Thus, in healthcare, good analytical tools and techniques along with proper business and analytics strategies are very important for decision-making. Strongly committed sponsorship and people working to their full potential are also of utmost importance (Wang *et al.*, 2018).

Another model consisting of the resource-based theory (RBT) and capability building view was proposed. RBT is basically the theoretical foundation on how transformation of resources into ones competitive advantage is possible (Barney, 2001). Capability building view basically compensates the drawbacks of RBT (Bharadwaj, 2000; Doherty and Terry, 2009; Karimi *et al.*, 2007; Santhanam and Hartono, 2003; Saraf *et al.*, 2007; Wang *et al.*, 2012). According to Karimi *et al.* (2007), capability building is "the ability of companies to develop unique skills that can leverage their resources." Healthcare often requires real-time data (Tang *et al.*, 2019). A useful decision can be drawn from the data collected and can be used by decision makers of this field (Wang and Byrd, 2017; Prasser *et al.*, 2019). IoT devices having radio-frequency identification (RFID) tags and sensors help in providing the real-time information and help the devices to communicate. This device helps in monitoring the health of the patient (Gubbi *et al.*, 2013). Kolasa *et al.* (2020b) explains how BD is reshaping the decision-making process in health-care sector. They stated that decision makers should be sure about usage of data. Data should be limited, transparent, easy to handle and systematic. Significant amount of effort is needed when the data security and privacy issues are addressed. Information exchange and interoperability are crucial; hence, the medical information should be decentralized rather than storing all the information at one place. Table 3 provides detailed analytical strategies of the BD.

6. Application of big data

Some major companies that offer BD services in the health-care sector are mentioned in this section. Table 4 provides detailed application and benefits of BD in health-care sector.

7. Synthesis of findings and research gaps analysis

BDA can be applied in almost every area of the health-care sector. The potential areas of application are forecast epidemics, fraud detection, avoid avertible deaths and improved

Strategy	Reference
I. Data governance	Bakker *et al.* (2020), Stevenson and Davis (2013), Bardhan *et al.* (2015), Bates *et al.* (2014), Hashem *et al.* (2015)
II. Improved data analytics	Wang *et al.* (2018), Chen (2018), Palanisarmy and Thirunavukarasu (2019)
III. Focus on the managerial strategies also	Wamba *et al.* (2015)
IV. Real-time responsiveness	Rghioui *et al.* (2020), Tang *et al.* (2019), Sakr and Gaber (2014), Groves *et al.* (2013), Chen *et al.* (2016), Palanisarmy and Thirunavukarasu (2019)
V. Committed employees and sponsorships	Wang *et al.* (2018)
VI. Proper decision-making	Middleton *et al.* (2013), Prasser *et al.* (2019), Wang *et al.* (2018), Palanisarmy and Thirunavukarasu (2019), Bakker *et al.* (2020), Kolasa *et al.* (2020b), Rghioui *et al.* (2020)
VII. Security of data	Mounia and Habiba (2015)
VIII. Good technical algorithms with the help of IoT to be used for proper predictive analysis	Chen *et al.* (2016), Dash *et al.* (2019), Wang *et al.* (2018), Palanisarmy and Thirunavukarasu (2019), Cunha *et al.* (2020)
IX. Important-performance analysis model (IPA)	Martilla and James (1977), Hansen and Bush (1999), Lin (2015)
X. Resource-based theory model (RBT)	Barney (2001), Bharadwaj (2000), Santhanam and Hartono (2003), Doherty and Terry (2009), Karimi *et al.* (2007), Saraf *et al.* (2007), Wang *et al.* (2012)

Table 3.
Analytical strategies of the big data

quality of life. Further, the analytical capability is used in a BD system to increase accountability and decision-making support that improves collaboration among supply chain partners (Bag *et al.*, 2020b). The key opportunities of BDA in the health-care sector are used to set up predictive models that support smart health-care systems. Consequently, it will be helpful in the improvement of the infrastructure, strategy making and identifying fraudulent activities. Sometimes, BDA helps in cost-cutting and improves the pharmaceutical supply chain, forecasts demand, optimizes staffing and streamlines patient care.

The successful implementation of BDA in the health-care sector has been explored in many developed economies. The understanding of key issues and challenges are essential for the decision-making of the BDA in the health-care sector. The previous studies were focused on identification of key issues and challenges while adopting BDA in the health-care sector as shown in Table 2. These issues can be categorized into management issues, storage issues, processing issues, security issues and other issues. Further, data protection issues are very common to digital services platforms. Hence, data security and data privacy should be given more importance in the coming years. Further, these issues may differ from sector to sector and transform over the period. Previous studies were mainly focused on an individual aspect of these issues; however, in this study, we considered all these issues and challenges simultaneously in the adoption of BDA in the health-care sector.

The research gaps are identified from the systematic literature review. It leads to the development of future research directions, which are specified as follows:

(1) The majority of the articles are focused on the case study and qualitative analysis rather than quantitative analysis. Therefore, statistical modeling can also be performed to evaluate the impact of each BD strategy on health-care practices.

(2) Managerial, storage, processing and security issues on adopting BDA in the health-care sector have been considered in the very few studies in context of developing economies.

Company name	Applications	Benefits	Web link
IBM Watson (in partnership with GSK)	GSK launched 16 virtual assistants in 10 months Provides online messaging and chat channels (1) Emails (2) Online database (3) CRM systems (4) In-house custom solutions	In this tenure, through Watson Assistant 82,000 questions were taken with 7,000 unique answered	https://www.ibm.com/watson https://www.ibm.com/blogs/watson/2020/12/how-glaxosmithkline-launched-16-virtual-assistants-in-10-months-with-watson-assistant/
Digital reasoning systems	Used for cancer diagnosis and treatment process (1) It captures the data from the front end of cancer diagnosis and treatment process and uses natural language processing (NLP) and machine learning technology to read it (2) The piece of software then works on real-time basis and discovers, classifies, prioritizes the follow-up (3) The application makes sure the doctors get the right information at the right time	It improves the customer experience Better outcome and quality service (1) Streamline the operations	https://digitalreasoning.com/solutions-patient-intelligence/
Linguamatics	Provides text mining platform for unstructured data to various companies for numerous applications	(1) Makes a significant contribution to early detection of heart disease (2) Helps in timely treatment of disease (3) Cost saving	https://www.linguamatics.com/solutions/big-data-analytics-healthcare
Apixio	An AI-based platform that converts unstructured data into actionable data	(1) It helps to take better decision (2) Improves the quality (3) It provides affordable care	www.apixio.com/
Lumiata	Provides service for analytics and risk management	(1) Early risk prediction (2) Helps in efficient outcome	https://www.lumiata.com/
OptumHealth	Provides health-care facility, pharmacy services, health-care financial services	(1) Acts as a one stop for all the services which helps in cost saving	https://www.optum.com/

Table 4.
Application and benefits of big data services

(3) The study analyzes that data processing and data protection are the major challenges of the digital services platform. Hence, data security and data privacy should be given more importance in the coming years.

(4) The study identifies that the integration of BDA in the health sector helps in the early detection of diseases and real-time monitoring of hospitals. This ensures that academicians and physicians can have more industrial control over the health-care system, improving and optimizing the health-care strategies, quality of service, resilience and agility.

(5) The study analyzes various strategic factors in the devolvement of BDA. However, none of them consider the strategic factors such as data governance, information-sharing culture, real-time responsiveness and IPA model that can make a very diverse environment. Further, no study was found to analyze these factors for BDA in the health-care sector.

(6) The study identifies various applications and benefits of BD services in the health-care sector, such as in GSK launcher, to provide in-house customer solutions, software for real-time basis, discovery and text mining platform for handling unstructured data. However, none of the studies considered these applications in health-care services.

(7) The study analyzes various challenges in the development of BDA in the health-care sector. However, none of them prioritize these challenges using a heuristic (Tabu search, simulated annealing and genetic algorithm) and multicriteria decision-making (MCDM) (analytic network process [ANP], analytic hierarchy process [AHP], interpretive structural modeling [ISM] and decision-making trail and evaluation laboratory [DEMATEL]) approach for decision-making.

(8) Limited studies are available on empirical validation of BDA applications in the context of developing countries. So, comparative study can be done for developing and developed countries.

(9) Very few studies have identified contribution of BDA in the context of sustainable operations, so researchers may explore it further.

(10) Very few studies are available on barriers and critical success factors for implementing BDA. There is scope for developing scales in this context.

(11) Very few studies on developing human resource for implementing BDA offer a lot of potential for future research related to human resource development (HRM) for BDA capabilities.

8. Conclusion

In future, the healthcare will be more patient centric. The virtual care will be promoted and the tracking of daily activities will be given more importance. The coordination will be there on real-time basis. The payment services and financing will be promoted digitally. This system is based on data collected in real time and can be applied to any predictive case using sensors and mobile data generated. In addition, the difference between structured and unstructured data sources is leveraged through BDA. A well-known obstacle to overcome is the transition to an integrated data setting. The study of such data will provide additional insights into procedural, technological, medical and other forms of health-care changes. Indeed, organizations providing health-care analytics and clinical transformation services are leading to better and more successful results. Common objectives of these organizations include the reduction of analytical costs, the development of efficient clinical decision support (CDS) systems, the provision of platforms for enhanced treatment strategies and the detection and prevention of BD fraud.

The integrated data pool of health-care organizations and biomedical scientists has resulted in a stronger perspective, commitment and treatment of different diseases. It has also helped to create a better and healthier framework for personalized healthcare. Interestingly, the BD philosophy focuses heavily on the concept that the more knowledge can be derived from the data. It can make accurate predictions for future events. Data scientists have been overwhelmed by high volumes of medical data collected through heterogeneous systems for careful integration and implementation. Therefore, it is proposed that a health-care revolution is needed to put bioinformatics, health informatics and analytics together to facilitate customized and more efficient treatments.

In total, two research questions are asked in this study. In order to answer the first research question, i.e. what are the current challenges of BDA application in health-care sectors (RQ1)?, study has explored the 11 current issues/challenges of implementing BDA in the health-care sector. These are data protection, data leakage, data security, efficient processing of large quantities of medical imaging data, confidentiality and security of information, misuse of health. For the second question (RQ2), i.e. how can we deal with these issues/challenges?, the present study has made some suggestions to make the functioning of BDA application smooth in this sector. Analysis of these challenges and issues will enable business entities to formulate effective strategies from the beginning and find ways to counter them. Findings of the study would help health-care academicians, practitioners and researchers in exploring new BDA applications for efficient and cost-effective health-care management.

8.1 Managerial implications

BDA can be applied in almost every area of the health-care sector. This integrated data pool of health-care organizations and biomedical scientists has resulted in a stronger perspective, commitment and treatment of different diseases. The potential areas of application are forecasting epidemics, fraud detection, avoiding avertible deaths and improving the quality of life. This implies that academicians and researchers can have better industrial control on the supply chain processes. The main focus of this literature review is to utilize both the inductive and deductive research methodology approaches.

References

Abouelmehdi, K., Beni-Hessane, A. and Khaloufi, H. (2018), "Big healthcare data: preserving security and privacy", *Journal of Big Data*, Vol. 5 No. 1, pp. 1-18.

Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H. and Saadi, M. (2017), "Big data security and privacy in healthcare: a review", *Procedia Computer Science*, Vol. 113, pp. 73-80.

Agrawal, S., Singh, R.K. and Murtaza, Q. (2015), "A literature review and perspectives in reverse logistics", *Resources, Conservation and Recycling*, Vol. 97, pp. 76-92.

Alotaibi, S., Mehmood, R. and Katib, I. (2020), "The role of big data and Twitter data analytics in healthcare supply chain management", *Smart Infrastructure and Applications*, Springer, Cham, pp. 267-279, doi: 10.1007/978-3-030-13705-2_11.

Amirian, P., van Loggerenberg, F., Lang, T., Thomas, A., Peeling, R., Basiri, A. and Goodman, S.N. (2017), "Using big data analytics to extract disease surveillance information from point of care diagnostic machines", *Pervasive and Mobile Computing*, Vol. 42, pp. 470-486, doi: 10.1016/j.pmcj.2017.06.013.

Anuradha, J. (2015), "A brief introduction on big data 5Vs characteristics and Hadoop technology", *Procedia Computer Science*, Vol. 48, pp. 319-324.

Archenaa, J. and Anita, E.A.M. (2015), "A survey of big data analytics in health care and government", *Procedia Computer Science*, Vol. 50, pp. 408-413.

Au-Yong-Oliveira, M., Pesqueira, A., Sousa, M.J., Dal Mas, F. and Soliman, M. (2021), "The potential of big data research in HealthCare for medical doctors' learning", *Journal of Medical Systems*, Vol. 45 No. 1, pp. 1-14.

Azad, P. and Navimipour, N.J. (2017), "An energy-aware task scheduling in the cloud computing using a hybrid cultural and ant colony optimization algorithm", *International Journal of Cloud Applications and Computing*, Vol. 7 No. 4, pp. 20-40.

Bag, S. (2017), "Big data and predictive analysis is key to superior supply chain performance: a South African experience", *International Journal of Information Systems and Supply Chain Management*, Vol. 10 No. 2, pp. 66-84.

Bag, S., Gupta, S. and Wood, L. (2020a), "Big data analytics in sustainable humanitarian supply chain: barriers and their interactions", *Annals of Operations Research*, pp. 1-40, doi: 10.1007/s10479-020-03790-7.

Bag, S., Pretorius, J.H.C., Gupta, S. and Dwivedi, Y.K. (2021), "Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities", *Technological Forecasting and Social Change*, Vol. 163, 120420.

Bag, S., Wood, L.C., Xu, L., Dhamija, P. and Kayikci, Y. (2020b), "Big data analytics as an operational excellence approach to enhance sustainable supply chain performance", *Resources, Conservation and Recycling*, Vol. 153, 104559.

Bakker, L., Aarts, J., Uyl-de Groot, C. and Redekop, W. (2020), "Economic evaluations of big data analytics for clinical decision-making: a scoping review", *Journal of the American Medical Informatics Association*, Vol. 27 No. 9, pp. 1466-1475.

Bardhan, I., Oh, J.H., Zheng, Z. and Kirksey, K. (2015), "Predictive analytics for readmission of patients with congestive heart failure", *Information Systems Research*, Vol. 26 No. 1, pp. 19-39.

Barnaghi, P., Sheth, A. and Henson, C. (2013), "From data to actionable knowledge: big data challenges in the web of things [guest editors' introduction]", *IEEE Intelligent Systems*, Vol. 28 No. 6, pp. 6-11.

Barney, J.B. (2001), "Resource-based theories of competitive advantage: a ten-year retrospective on the resource-based view", *Journal of Management*, Vol. 27 No. 6, pp. 643-650.

Batarseh, F.A., Ghassib, I., Chong, D.S. and Su, P.H. (2020), "Preventive healthcare policies in the US: solutions for disease management using big data analytics", *Journal of Big Data*, Vol. 7 No. 1, pp. 1-25.

Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A. and Escobar, G. (2014), "Big data in health care: using analytics to identify and manage high-risk and high-cost patients", *Health Affairs*, Vol. 33 No. 7, pp. 1123-1131.

Bernstein, D.J. (2005), "Understanding brute force", Workshop Record of ECRYPT STVL Workshop on Symmetric Key Encryption, eSTREAM report, Vol. 36.

Bhadani, A.K. and Jothimani, D. (2016), "Big data: challenges, opportunities, and realities", in *Effective Big Data Management and Opportunities for Implementation*, IGI Global, pp. 1-24.

Bharadwaj, A.S. (2000), "A resource-based perspective on information technology capability and firm performance: an empirical investigation", *Management Information Systems Quarterly*, Vol. 24 No. 1, pp. 169-196.

Brunetti, F., Matt, D.T., Bonfanti, A., De Longhi, A., Pedrini, G. and Orzes, G. (2020), "Digital transformation challenges: strategies emerging from a multi-stakeholder approach", *The TQM Journal*, Vol. 32 No. 4, pp. 697-724.

Brynjolfsson, E. and McAfee, A. (2011), "The big data boom is the innovation story of our time", *The Atlantic*, Vol. 21, p. 30.

Burrows, R. and Savage, M. (2014), "After the crisis? Big data and the methodological challenges of empirical sociology", *Big Data and Society*, Vol. 1 No. 1, 2053951714540280.

Cavallone, M. and Palumbo, R. (2020), "Debunking the myth of industry 4.0 in health care: insights from a systematic literature review", *The TQM Journal*, Vol. 32 No. 4, pp. 849-868.

Chen, C.P. and Zhang, C.Y. (2014), "Data-intensive applications, challenges, techniques and technologies: a survey on big data", *Information Sciences*, Vol. 275, pp. 314-347.

Chen, D.Q., Preston, D.S. and Xia, W. (2013), "Enhancing hospital supply chain performance: a relational view and empirical test", *Journal of Operations Management*, Vol. 31 No. 6, pp. 391-408.

Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from big data to big impact", *Management Information Systems Quarterly*, Vol. 36 No. 4, pp. 1165-1188.

Chen, M., Mao, S. and Liu, Y. (2014), "Big data: a survey", *Mobile Networks and Applications*, Springer, Vol. 19 No. 2, pp. 171-209.

Chen, S. (2014), "Architecting the future of big data", *Horton Technical Preview for Apache Spark*, Horton Works, SAMOA, IKANOW, Hadoop, Map Reduce and Grid Gain.

Chen, P.T. (2018), "Medical big data applications: intertwined effects and effective resource allocation strategies identified through IRA-NRM analysis", *Technological Forecasting and Social Change*, Vol. 130, pp. 150-164.

Chen, S. (2007), *Application Denial of Service*, Hack Tics.

Chin, W.W. (1998), "The partial least squares approach to structural equation modeling", *modern Methods for Business Research*, Vol. 295 No. 2, pp. 295-336.

Chen, Y., Argentinis, J.E. and Weber, G. (2016), "IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research", *Clinical Therapeutics*, Vol. 38 No. 4, pp. 688-701.

Chiregi, M. and Navimipour, N.J. (2016), "A new method for trust and reputation evaluation in the cloud environments using the recommendations of opinion leaders' entities and removing the effect of troll entities", *Computers in Human Behavior*, Vol. 60, pp. 280-292.

Chiregi, M. and Navimipour, N.J. (2017), "A comprehensive study of the trust evaluation mechanisms in the cloud computing", *Journal of Service Science Research*, Vol. 9 No. 1, pp. 1-30.

Confetto, M.G. and Covucci, C. (2021), "A taxonomy of sustainability topics: a guide to set the corporate sustainability content on the web", *The TQM Journal*.

Cunha, T.P., Méxas, M.P., da Silva, A.C. and Quelhas, O.L.G. (2020), "Proposal guidelines to implement the concepts of industry 4.0 into information technology companies", *The TQM Journal*, Vol. 32 No. 4, pp. 741-759.

Daghistani, T., AlGhamdi, H., Alshammari, R. and AlHazme, R.H. (2020), "Predictors of outpatients' no-show: big data analytics using Apache spark", *Journal of Big Data*, Vol. 7 No. 1, pp. 1-15.

Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S. (2019), "Big data in healthcare: management, analysis and future prospects", *Journal of Big Data*, Vol. 6 No. 1, p. 54.

Davies, S.C. (2018), "Annual report of the Chief Medical Officer", *Health 2040 – Better Health within Reach*, Department of Health and Social Care.

Deloitte Insights (2020), "Global health care outlook, laying a foundation for the future", available at: https://documents.deloitte.com/insights/2020globalhealthcareoutlook.

Demchenko, Y., Grosso, P., De Laat, C. and Membrey, P. (2013), "Addressing big data issues in scientific data infrastructure", *IEEE International Conference on Collaboration Technologies and Systems (CTS)*, pp. 48-55.

Dobrzykowski, D.D. and Tarafdar, M. (2015), "Understanding information exchange in healthcare operations: evidence from hospitals and patients", *Journal of Operations Management*, Vol. 36, pp. 201-214.

Doherty, N.F. and Terry, M. (2009), "The role of IS capability in delivering sustainable improvements to competitive positioning", *The Journal of Strategic Information Systems*, Vol. 18 No. 2, pp. 100-116.

Drupsteen, J., van der Vaart, T. and van Donk, D.P. (2013), "Integrative practices in hospitals and their impact on patient flow", *International Journal of Operations and Production Management*, Vol. 33 No. 7, pp. 912-933.

Duan, L. and Xiong, Y. (2015), "Big data analytics and business analytics", *Journal of Management Analytics*, Vol. 2 No. 1, pp. 1-21.

Dubey, R., Gunasekaran, A., Childe, S.J., Papadopoulos, T., Luo, Z., Wamba, S.F. and Roubaud, D. (2019), "Can big data and predictive analytics improve social and environmental sustainability?", *Technological Forecasting and Social Change*, Vol. 144, pp. 534-545.

Dwivedi, K. and Dubey, S.K. (2014), "Analytical review on Hadoop distributed file system", *2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence)*, IEEE, pp. 174-181.

Economist Intelligence Unit (EIU) (2019), "World industry outlook: healthcare and pharmaceuticals", September 29.

Elsebakhi, E., Lee, F., Schendel, E., Haque, A., Kathireason, N., Pathare, T. and Al-Ali, R. (2015), "Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms", *Journal of Computational Science*, Vol. 11, pp. 69-81.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017), "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, Vol. 542 No. 7639, pp. 115-118.

European Public Health Alliance (2017), "Health and care in the digital single market: refection paper", available at: https://epha.org/wp-content/uploads/2018/02/Health-and-care-indigital-single-market-position-paper.pdf (accessed 9 March).

Feldman, K., Davis, D. and Chawla, N.V. (2015), "Scaling and contextualizing personalized healthcare: a case study of disease prediction algorithm integration", *Journal of Biomedical Informatics*, Vol. 57, pp. 377-385.

Ferranti, J.M., Langman, M.K., Tanaka, D., McCall, J. and Ahmad, A. (2010), "Bridging the gap: leveraging business intelligence tools in support of patient safety and financial effectiveness", *Journal of the American Medical Informatics Association*, Vol. 17 No. 2, pp. 136-143.

Fihn, S.D., Francis, J., Clancy, C., Nielson, C., Nelson, K., Rumsfeld, J. and Graham, G.L. (2014), "Insights from advanced analytics at the veterans health administration", *Health Affairs*, Vol. 33 No. 7, pp. 1203-1211.

Filkins, B.L., Kim, J.Y., Roberts, B., Armstrong, W., Miller, M.A., Hultner, M.L. and Steinhubl, S.R. (2016), "Privacy and security in the era of digital health: what should translational researchers know and do about it?", *American Journal of Tourism Research*, Vol. 8 No. 3, p. 1560.

Foidl, H. and Felderer, M. (2015), "Research challenges of industry 4.0 for quality management", *International Conference on Enterprise Resource Planning Systems*, Springer, pp. 121-137.

Frost, S. (2015), "Drowning in big data? Reducing information technology complexities and costs for healthcare organizations".

Galbraith, J.R. (1974), "Organization design: an information processing view", *Interfaces*, Vol. 4 No. 3, pp. 28-36.

Galetsi, P. and Katsaliaki, K. (2020), "A review of the literature on big data analytics in healthcare", *Journal of the Operational Research Society*, Vol. 71 No. 10, pp. 1511-1529, doi: 10.1080/01605682.2019.1630328.

Gandomi, A. and Haider, M. (2015), "Beyond the hype: big data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

Gartner (2014), *STAMFORD, Conference*, September 17, available at: http://www.gartner.com/newsroom/id/2848718.

Gold, S., Seuring, S. and Beske, P. (2010), "Sustainable supply chain management and inter-organizational resources: a literature review", *Corporate Social Responsibility and Environmental Management*, Vol. 17 No. 4, pp. 230-245.

Gowsalya, M., Krushitha, K. and Valliyammai, C. (2014), "Predicting the risk of readmission of diabetic patients using MapReduce", *Sixth International Conference on Advanced Computing (ICoAC)*, December, IEEE, pp. 297-301.

Groves, P., Kayyali, B., Knott, D. and Van Kuiken, S. (2013), "The 'big data' revolution in healthcare: accelerating value and innovation", McKinsey, *McKinsey Quarterly*, Vol. 2, pp. 10-14.

Gu, D., Li, J., Li, X. and Liang, C. (2017), "Visualizing the knowledge structure and evolution of big data research in healthcare informatics", *International Journal of Medical Informatics*, Vol. 98, pp. 22-32, doi: 10.1016/j.ijmedinf.2016.11.006.

Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, M. (2013), "Internet of things (IoT): a vision, architectural elements, and future directions", *Future Generation Computer Systems*, Vol. 29 No. 7, pp. 1645-1660.

Gupta, M. and George, J.F. (2016), "Toward the development of a big data analytics capability", *Information and Management*, Vol. 53 No. 8, pp. 1049-1064.

Hamilton, B. (2020), "Big data is the future of healthcare", available at: https://www.cognizant.com/industries-resources/healthcare/Big-Data-is-the-Future-of-Healthcare.pdf.

Hansen, E. and Bush, R.J. (1999), "Understanding customer quality requirements: model and application", *Industrial Marketing Management*, Vol. 28 No. 2, pp. 119-130.

Harerimana, G., Jang, B., Kim, J.W. and Park, H.K. (2018), "Health big data analytics: a technology survey", *IEEE Access*, Vol. 6, pp. 65661-65678, doi: 10.1109/ACCESS.2018.2878254.

Harlow, H.D. (2018), "Developing a knowledge management strategy for data analytics and intellectual capital", *Meditari Accountancy Research*, Vol. 26 No. 3, pp. 400-419.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. and Khan, S.U. (2015), "The rise of big data on cloud computing: review and open research issues", *Information Systems*, Vol. 47, pp. 98-115.

Hayes, A.F. and Krippendorff, K. (2007), "Answering the call for a standard reliability measure for coding data", *Communication Methods and Measures*, Vol. 1 No. 1, pp. 77-89.

Hazen, B.T., Boone, C.A., Ezell, J.D. and Jones-Farmer, L.A. (2014), "Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications", *International Journal of Production Economics*, Vol. 154, pp. 72-80.

Hoang, N.D. and Pham, A.D. (2016), "Hybrid artificial intelligence approach based on metaheuristic and machine learning for slope stability assessment: a multinational data analysis", *Expert Systems with Applications*, Vol. 46, pp. 60-68.

Hoffman, S. and Podgurski, A. (2013), "Big bad data: law, public health, and biomedical databases", *Journal of Law Medicine and Ethics*, Vol. 41, pp. 56-60.

Hurwitz, J., Nugent, A., Hapler, F. and Kaufman, M. (2013), *Big Data for Dummies*, John Wiley and Sons, Hoboken, New Jersey.

Ji, C., Li, Y., Qiu, W., Awada, U. and Li, K. (2012), "Big data processing in cloud computing environments", *12th International Symposium on Pervasive Systems, Algorithms and Networks*, December, IEEE, pp. 17-23.

Jiang, P., Winkley, J., Zhao, C., Munnoch, R., Min, G. and Yang, L.T. (2014), "An intelligent information forwarder for healthcare big data systems with distributed wearable sensors", *IEEE Systems Journal*, Vol. 10 No. 3, pp. 1147-1159.

Kache, F. and Seuring, S. (2017), "Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management", *International Journal of Operations and Production Management*, Vol. 37 No. 1, pp. 10-36.

Kamal, M.D., Tahir, A., Kamal, M.B. and Naeem, M.A. (2020), "Future location prediction for emergency vehicles using big data: a case study of healthcare engineering", *Journal of Healthcare Engineering*, Vol. 2020.

Kamal, M.M. and Irani, Z. (2014), "Analysing supply chain integration through a systematic literature review: a normative perspective", *Supply Chain Management: An International Journal*, Vol. 19 Nos 5/6, pp. 523-557.

Kamble, S.S., Gunasekaran, A., Goswami, M. and Manda, J. (2018), "A systematic perspective on the applications of big data analytics in healthcare management", *International Journal of Healthcare Management*, Vol. 12 No. 3, pp. 226-240, doi: 10.1080/20479700.2018.1531606.

Karimi, J., Somers, T.M. and Bhattacherjee, A. (2007), "The role of information systems resources in ERP capability building and business process outcomes", *Journal of Management Information Systems*, Vol. 24 No. 2, pp. 221-260.

Kaur, P., Sharma, M. and Mittal, M. (2018), "Big data and machine learning based secure healthcare framework", *Procedia Computer Science*, Vol. 132, pp. 1049-1059.

Keeso, A. (2014), "Big data and environmental sustainability: a conversation starter", Smith School of Enterprise and the Environment, Working Paper Series, pp. 14-04.

Khanra, S., Dhir, A. and Mäntymäki, M. (2020), "Big data analytics and enterprises: a bibliometric synthesis of the literature", *Enterprise Information Systems*, Vol. 14 No. 6, pp. 737-768.

Kolasa, K., Goettsch, W., Petrova, G. and Berler, A. (2020a), "Without data, you're just another person with an opinion", *Expert Review of Pharmacoeconomics and Outcomes Research*, Vol. 20 No. 2, pp. 147-154.

Kolasa, K., Redekop, W.K., Berler, A., Zah, V. and Asche, C.V. (2020b), "Future of data analytics in the era of the general data protection regulation in Europe", *PharmacoEconomics*, Vol. 38 No. 10, pp. 1021-1029.

Kruse, C.S., Goswamy, R., Raval, Y.J. and Marawi, S. (2016), "Challenges and opportunities of big data in health care: a systematic review", *JMIR Medical Informatics*, Vol. 4 No. 4, p. e38.

Kshetri, N. (2014), "Big data's impact on privacy, security and consumer welfare", *Telecommunications Policy*, Vol. 38 No. 11, pp. 1134-1145.

Kumar, R. and Singh, R.K. (2017), "Coordination and responsiveness issues in SMEs supply chains: a review", *Benchmarking: An International Journal*, Vol. 24 No. 3, pp. 635-650.

Kwon, I.W.G., Kim, S.H. and Martin, D.G. (2016), "Healthcare supply chain management; strategic areas for quality and financial improvement", *Technological Forecasting and Social Change*, Vol. 113, pp. 422-428.

Lee, I. (2017), "Big data: dimensions, evolution, impact, and challenges", *Business Horizons*, Vol. 60 No. 3, pp. 293-303.

Leite, H., Lindsay, C. and Kumar, M. (2020), "COVID-19 outbreak: implications on healthcare operations", *The TQM Journal*, Vol. 33 No. 1, pp. 247-256.

Li, Y., Esteva, A., Kuprel, B., Novoa, R., Ko, J. and Thrun, S. (2016), "Skin cancer detection and tracking using data synthesis and deep learning", arXiv preprint arXiv:1612.01074.

Lin, C.L. (2015), "A novel hybrid decision-making model for determining product position under consideration of dependence and feedback", *Applied Mathematical Modelling*, Vol. 39 No. 8, pp. 2194-2216.

Lukoianova, T. and Rubin, V.L. (2014), "Veracity roadmap: is big data objective, truthful and credible?", *Advances in Classification Research Online*, Vol. 24 No. 1, pp. 4-15.

Luo, J., Wu, M., Gopukumar, D. and Zhao, Y. (2016), "Big data application in biomedical research and health care: a literature review", *Biomedical Informatics Insights*, Vol. 8, pp. BII-S31559.

Mackenzie, R.W., Levandoski, K.A., Martin, K. and Asgari, M.M. (2017), "Research techniques made easy: an intro to big data used in derma".

Majnarić, L.T., Babič, F., O'Sullivan, S. and Holzinger, A. (2021), "AI and big data in healthcare: towards a more comprehensive research framework for multimorbidity", *Journal of Clinical Medicine*, Vol. 10 No. 4, p. 766.

Malik, M.M., Abdallah, S. and Ala'raj, M. (2018), "Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review", *Annals of Operations Research*, Vol. 270 No. 1, pp. 287-312, doi: 10.1007/s10479-016-2393-z.

Mandal, S. (2017), "The influence of dynamic capabilities on hospital-supplier collaboration and hospital supply chain performance", *International Journal of Operations and Production Management*, Vol. 37 No. 5, pp. 664-684.

Manwal, M. and Gupta, A. (2017), "Big data and hadoop – a technological survey", *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*, IEEE, pp. 1-6.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011), "Big data: the next frontier for innovation, competition, and productivity", *McKinsey Global Institute*, Vol. 112 No. 119, pp. 1-6.

Martilla, J.A. and James, J.C. (1977), "Importance-performance analysis", *Journal of Marketing*, Vol. 41 No. 1, pp. 77-79.

Mayring, P. and Fenzl, T. (2014), "Qualitative inhaltsanalyse", *Handbuch Methoden der empirischen Sozialforschung*, Springer VS, Wiesbaden, pp. 543-556.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. and Barton, D. (2012a), "Big data: the management revolution", *Harvard Business Review*, Vol. 90 No. 10, pp. 60-68.

McAfee, P.C., Reah, C., Gilder, K., Eisermann, L. and Cunningham, B. (2012b), "A meta-analysis of comparative outcomes following cervical arthroplasty or anterior cervical fusion: results from 4 prospective multicenter randomized clinical trials and up to 1226 patients", *Spine*, Vol. 37 No. 11, pp. 943-952.

McNutt, T.R., Moore, K.L. and Quon, H. (2016), "Needs and challenges for big data in radiation oncology", *International Journal of Radiation Oncology, Biology, Physics*, Vol. 95 No. 3, pp. 909-915, doi: 10.1016/j.ijrobp.2015.11.032.

Mehta, N. and Pandit, A. (2018), "Concurrence of big data analytics and healthcare: a systematic review", *International Journal of Medical Informatics*, Vol. 114, pp. 57-65, doi: 10.1016/j.ijmedinf.2018.03.013.

Mehta, N., Pandit, A. and Shukla, S. (2019), "Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study", *Journal of Biomedical Informatics*, Vol. 100, p. 103311.

Middleton, P., Kjeldsen, P. and Tully, J. (2013), "Forecast: the internet of things, worldwide", *Gartneryrket*, available at: http://www.gartner.com/document/2625419.

Mikulić, J. and Prebežac, D. (2008), "Prioritizing improvement of service attributes using impact range-performance analysis and impact-asymmetry analysis", *Managing Service Quality: An International Journal*, Vol. 18 No. 6, pp. 559-576.

Milani, B.A. and Navimipour, N.J. (2016), "A comprehensive review of the data replication techniques in the cloud environments: major trends and future directions", *Journal of Network and Computer Applications*, Vol. 64, pp. 229-238.

Miller, K. (2012), "Big data analytics in biomedical research", *Biomedical Computation Review*, Vol. 2, pp. 14-21.

Mohammed, E.A., Far, B.H. and Naugler, C. (2014), "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends", *BioData Mining*, Vol. 7 No. 1, pp. 1-23, doi: 10.1186/1756-0381-7-22.

Moran, G.J., Rothman, R.E. and Volturo, G.A. (2013), "Emergency management of community-acquired bacterial pneumonia: what is new since the 2007 Infectious Diseases Society of America/

American Thoracic Society guidelines", *The American Journal of Emergency Medicine*, Vol. 31 No. 3, pp. 602-612.

Morse, K.E., Bagely, S.C. and Shah, N.H. (2020), "Estimate the hidden deployment cost of predictive models to improve patient care", *Nature Medicine*, Vol. 26 No. 1, pp. 18-19.

Mounia, B. and Habiba, C. (2015), "Big data privacy in healthcare Moroccan context", *Procedia Computer Science*, Vol. 63, pp. 575-580.

Mourtzis, D. and Vlachou, E. (2016), "Cloud-based cyber-physical systems and quality of services", *The TQM Journal*, Vol. 28 No. 5, pp. 704-733.

Murdoch, T.B. and Detsky, A.S. (2013), "The inevitable application of big data to health care", *Journal of the American Medical Association*, Vol. 309 No. 13, pp. 1351-1352.

Ozminkowski, R.J., Wells, T.S., Hawkins, K., Bhattarai, G.R., Martel, C.W. and Yeh, C.S. (2015), "Big data, little data, and care coordination for medicare beneficiaries with medigap coverage", *Big Data*, Vol. 3 No. 2, pp. 114-125, doi: 10.1089/big.2014.0034.

Pagell, M. and Wu, Z. (2009), "Building a more complete theory of sustainable supply chain management using case studies of 10 exemplars", *Journal of Supply Chain Management*, Vol. 45 No. 2, pp. 37-56.

Palanisamy, V. and Thirunavukarasu, R. (2019), "Implications of big data analytics in developing healthcare frameworks – a review", *Journal of King Saud University-Computer and Information Sciences*, Vol. 31 No. 4, pp. 415-425.

Pashazadeh, A. and Navimipour, N.J. (2018), "Big data handling mechanisms in the healthcare applications: a comprehensive and systematic literature review", *Journal of Biomedical Informatics*, Vol. 82, pp. 47-62.

Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W. and Boccia, S. (2019), "Benefits and challenges of big data in healthcare: an overview of the European initiatives", *The European Journal of Public Health*, Vol. 29, pp. 23-27.

Pramanik, M.I., Zhang, W., Lau, R.Y. and Li, C. (2016a), "A framework for criminal network analysis using big data", *2016 IEEE 13th International Conference on E-Business Engineering (ICEBE)*, November, IEEE, pp. 17-23.

Pramanik, M.I., Lau, R.Y., Demirkan, H. and Azad, M.A.K. (2017), "Smart health: big data enabled health paradigm within smart cities", *Expert Systems with Applications*, Vol. 87, pp. 370-383.

Prasser, F., Spengler, H., Bild, R., Eicher, J. and Kuhn, K.A. (2019), "Privacy-enhancing ETL-processes for biomedical data", *International Journal of Medical Informatics*, Vol. 126, pp. 72-81, doi: 10.1016/j.ijmedinf.2019.03.006.

Qin, H., Li, H. and Zhao, X. (2010), "Development status of domestic and foreign smart city", *Global Presence*, Vol. 9, pp. 50-52.

Raghupathi, W. and Raghupathi, V. (2014), "Big data analytics in healthcare: promise and potential", *Health Information Science and Systems*, Vol. 2 No. 1, p. 1.

Reinertsen, E. and Clifford, G.D. (2018), "A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses", *Physiological Measurement*, Vol. 39 No. 5, p. 05TR01.

Rghioui, A., Lloret, J. and Oumnad, A. (2020), "Big data classification and internet of things in healthcare", *International Journal of E-Health and Medical Communications*, Vol. 11 No. 2, pp. 20-37.

Röcker, C., Ziefle, M. and Holzinger, A. (2014), "From computer innovation to human integration: current trends and challenges for pervasive health technologies", *Pervasive Health*, Springer, London, pp. 1-17.

Rothstein, M.A. (2015), "Ethical issues in big data health research: currents in contemporary bioethics", *Journal of Law Medicine and Ethics*, Vol. 43 No. 2, pp. 425-429.

Russom, P. (2011), "Big data analytics", *TDWI Best Practices Report*, fourth quarter, Vol. 19 No. 4, pp. 1-34.

Russom, P. (2013), "Managing big data", *TDWI Best Practices Report*, TDWI Research, pp. 1-40.

Sáez, C. and García-Gómez, J.M. (2018), "Kinematics of big biomedical data to characterize temporal variability and seasonality of data repositories: functional data analysis of data temporal evolution over non-parametric statistical manifolds", *International Journal of Medical Informatics*, Vol. 119, pp. 109-124, doi: 10.1016/j.ijmedinf.2018.09.015.

Sahoo, P.K., Mohapatra, S.K. and Wu, S.L. (2016), "Analyzing health care big data with prediction for future health condition", *IEEE Access*, Vol. 4, pp. 9786-9798.

Sahu, A.K., Kumar, A., Sahu, A.K. and Sahu, N.K. (2020), "Evaluation of machine tool substitute under data-driven quality management system: a hybrid decision-making approach", *The TQM Journal*.

Sakr, S. and Elgammal, A. (2016), "Towards a comprehensive data analytics framework for smart healthcare services", *Big Data Research*, Vol. 4, pp. 44-58.

Sakr, S. and Gaber, M. (Eds) (2014), *Large Scale and Big Data: Processing and Management*, 25 June, Auerbach Publications.

Salomi, M. and Balamurugan, S.A.A. (2016), "Need, application and characteristics of big data analytics in healthcare – a survey", *Indian Journal of Science and Technology*, Vol. 9 No. 16, pp. 1-5, doi: 10.17485/ijst/2016/v9i16/87960.

Santhanam, R. and Hartono, E. (2003), "Issues in linking information technology capability to firm performance", *Management Information Systems Quarterly*, Vol. 27 No. 1, pp. 125-153.

Saqlain, M., Munir, M.M., Rehman, S.U., Gulzar, A., Naz, S., Ahmed, Z. and Mashhood, M. (2020), "Knowledge, attitude, practice and perceived barriers among healthcare workers regarding COVID-19: a cross-sectional survey from Pakistan", *Journal of Hospital Infection*, Vol. 105 No. 3, pp. 419-423.

Saraf, N., Langdon, C.S. and Gosain, S. (2007), "IS application capabilities and relational value in interfirm partnerships", *Information Systems Research*, Vol. 18 No. 3, pp. 320-339.

Saraladevi, B., Pazhaniraja, N., Paul, P.V., Basha, M.S. and Dhavachelvan, P. (2015), "Big data and Hadoop – a study in security perspective", *Procedia Computer Science*, Vol. 50, pp. 596-601.

Schoenherr, T. and Speier-Pero, C. (2015), "Data science, predictive analytics, and big data in supply chain management: current state and future potential", *Journal of Business Logistics*, Vol. 36 No. 1, pp. 120-132.

Sedevich-Fons, L. (2014), "Financial indicators in healthcare quality management systems", *The TQM Journal*, Vol. 26 No. 4, pp. 312-328.

Sendra, S., Parra, L., Lloret, J. and Tomás, J. (2018), "Smart system for children's chronic illness monitoring", *Information Fusion*, Vol. 40, pp. 76-86.

Senthilkumar, S.A., Rai, B.K., Meshram, A.A., Gunasekaran, A. and Chandrakumarmangalam, S. (2018), "Big data in healthcare management: a review of literature", *American Journal of Theoretical and Applied Business*, Vol. 4 No. 2, pp. 57-69.

Seuring, S. and Gold, S. (2012), "Conducting content-analysis based literature reviews in supply chain management", *Supply Chain Management: International Journal*.

Shang, S. and Seddon, P.B. (2002), "Assessing and managing the benefits of enterprise systems: the business manager's perspective", *Information Systems Journal*, Vol. 12 No. 4, pp. 271-299.

Sharma, R., Agarwal, P. and Mahapatra, R.P. (2020), "Evolution in big data analytics on internet of things: applications and future plan", *Multimedia Big Data Computing for IoT Applications*, Springer, Singapore, pp. 453-477.

Singh, R.K., Garg, S.K. and Deshmukh, S.G. (2008), "Strategy development by SMEs for competitiveness: a review", *Benchmarking: An International Journal*, Vol. 15 No. 5, pp. 525-547.

Singh, R.K., Kumar, R. and Kumar, P. (2016), "Strategic issues in pharmaceutical supply chains: a review", *International Journal of Pharmaceutical and Healthcare Marketing*, Vol. 10 No. 3, pp. 234-257.

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017), "Critical analysis of big data challenges and analytical methods", *Journal of Business Research*, Vol. 70, pp. 263-286.

Soleimani-Roozbahani, F., Ghatari, A.R. and Radfar, R. (2019), "Knowledge discovery from a more than a decade studies on healthcare big data systems: a scientometrics study", *Journal of Big Data*, Vol. 6 No. 1, pp. 1-15.

Sony, M., Antony, J. and Douglas, J.A. (2021), "Motivations, barriers and readiness factors for quality 4.0 implementation: an exploratory study", *The TQM Journal*.

Srinivasan, U. and Arunasalam, B. (2013), "Leveraging big data analytics to reduce healthcare costs", *IT Professional*, Vol. 15 No. 6, pp. 21-28.

Srinivasan, R. and Swink, M. (2018), "An investigation of visibility and flexibility as complements to supply chain analytics: an organizational information processing theory perspective", *Production and Operations Management*, Vol. 27 No. 10, pp. 1849-1867.

Srivathsan, M. and Arjun, K.Y. (2015), "Health monitoring system by prognotive computing using big data analytics", *Procedia Computer Science*, Vol. 50, pp. 602-609.

Stadler, J.G., Donlon, K., Siewert, J.D., Franken, T. and Lewis, N.E. (2016), "Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards", *Big Data*, Vol. 4 No. 2, pp. 129-135.

Stahl, E. (2011), "Performance and capacity implications for a smarter planet", available at: www.ibm.com/smarterplanet/us/en/ smarter_cities/overview.

Stanford Medicine (2020), "The democratization of healthcare", *Stanford Medicine Health Trends Report*, Accessed June.

Stevenson, S. and Davis, E. (2013), *Conquering Big Data: Building Analytical Skills in Your Organization*, American Management Association (AMA).

Strang, K.D. and Alamieyeseigha, S. (2015), "What and where are the risks of international terrorist attacks: a descriptive study of the evidence", *International Journal of Risk and Contingency Management*, Vol. 4 No. 1, pp. 1-20.

Su, K., Li, J. and Fu, H. (2011), "Smart city and the applications", *Proceedings of 2011 International Conference on Electronics, Communications and Control (ICECC)*, September, IEEE, pp. 1028-1031.

Suciu, G., Suciu, V., Martian, A., Craciunescu, R., Vulpe, A., Marcu, I. and Fratu, O. (2015), "Big data, internet of things and cloud convergence–an architecture for secure e-health applications", *Journal of Medical Systems*, Vol. 39 No. 11, p. 141.

Sultanow, E. and Chircu, A.M. (2015), "Improving healthcare with data-driven track-and-trace systems", *Strategic Data-Based Wisdom in the Big Data Era*, IGI Global, pp. 65-82.

Suresh, S. (2016), "Big data and predictive analytics: applications in the care of children", *Pediatric Clinics*, Vol. 63 No. 2, pp. 357-366.

Szlezak, N., Evers, M., Wang, J. and Pérez, L. (2014), "The role of big data and advanced analytics in drug discovery, development, and commercialization", *Clinical Pharmacology and Therapeutics*, Vol. 95 No. 5, pp. 492-495, doi: 10.1038/clpt.2014.29.

Tang, V., Choy, K.L., Ho, G.T., Lam, H.Y. and Tsang, Y.P. (2019), "An IoMT-based geriatric care management system for achieving smart health in nursing homes", *Industrial Management and Data Systems*, Vol. 119 No. 8, pp. 1819-1840, doi: 10.1108/IMDS-01-2019-0024.

Taylor, K. (2015), *Connected Health: How Digital Technology is Transforming Health and Social Care*, Deloitte Centre for Health Solutions, London.

Terry, N. (2015), "Navigating the incoherence of big data reform proposals", *Journal of Law Medicine and Ethics*, Vol. 43 No. 1, pp. 44-47.

Thorpe, J.H. and Gray, E.A. (2015), "Big data and public health: navigating privacy laws to maximize potential", *Public Health Reports*, Vol. 130 No. 2, pp. 171-175.

Trakulsunti, Y., Antony, J. and Douglas, J.A. (2020), "Lean Six Sigma implementation and sustainability roadmap for reducing medication errors in hospitals", *The TQM Journal*, Vol. 33 No. 1, pp. 33-55.

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D. and Straus, S.E. (2018), "PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation", *Annals of Internal Medicine*, Vol. 169 No. 7, pp. 467-473.

Ullah, F., Habib, M.A., Farhan, M., Khalid, S., Durrani, M.Y. and Jabbar, S. (2017), "Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare", *Sustainable Cities and Society*, Vol. 34, pp. 90-96.

Vachon, S. and Klassen, R.D. (2008), "Environmental management and manufacturing performance: the role of collaboration in the supply chain", *International Journal of Production Economics*, Vol. 111 No. 2, pp. 299-315.

Van der Velde, F. (2010), "Where artificial intelligence and neuroscience meet: the search for grounded architectures of cognition", *Advances in Artificial Intelligence*, 16877470.

Varshney, U. (2014), "Mobile health: four emerging themes of research", *Decision Support Systems*, Vol. 66, pp. 20-35.

Vinodh, S., Antony, J., Agrawal, R. and Douglas, J.A. (2020), "Integration of continuous improvement strategies with industry 4.0: a systematic review and agenda for further research", *The TQM Journal*, Vol. 33 No. 2, pp. 441-472.

Waller, M.A. and Fawcett, S.E. (2013), "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management", *Journal of Business Logistics*, Vol. 34 No. 2, pp. 77-84.

Wamba, S.F., Anand, A. and Carter, L. (2013), "A literature review of RFID-enabled healthcare applications and issues", *International Journal of Information Management*, Vol. 33 No. 5, pp. 875-891.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015), "How 'big data' can make big impact: findings from a systematic review and a longitudinal case study", *International Journal of Production Economics*, Vol. 165, pp. 234-246.

Wamba, S.F., Gunasekaran, A., Akter, S., Ren, S.J.F., Dubey, R. and Childe, S.J. (2017), "Big data analytics and firm performance: effects of dynamic capabilities", *Journal of Business Research*, Vol. 70, pp. 356-365.

Wan, K. and Alagar, V. (2016), "Characteristics and classification of big data in health care sector", *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, August, IEEE, pp. 1439-1446.

Wang, G., Gunasekaran, A., Ngai, E.W. and Papadopoulos, T. (2016), "Big data analytics in logistics and supply chain management: certain investigations for research and applications", *International Journal of Production Economics*, Vol. 176, pp. 98-110.

Wang, N., Liang, H., Zhong, W., Xue, Y. and Xiao, J. (2012), "Resource structuring or capability building? An empirical study of the business value of information technology", *Journal of Management Information Systems*, Vol. 29 No. 2, pp. 325-367.

Wang, Y. and Byrd, T.A. (2017), "Business analytics-enabled decision-making effectiveness through knowledge absorptive capacity in health care", *Journal of Knowledge Management*, Vol. 21, pp. 517-539.

Wang, Y. and Hajli, N. (2017), "Exploring the path to big data analytics success in healthcare", *Journal of Business Research*, Vol. 70, pp. 287-299.

Wang, Y., Kung, L.A. and Byrd, T.A. (2015), "Big data analytics: understanding its capabilities and potential benefits in healthcare organizations", *Technological Forecasting and Social Change*, Vol. 126, pp. 3-13.

Wang, Y., Kung, L. and Byrd, T.A. (2018), "Big data analytics: understanding its capabilities and potential benefits for healthcare organizations", *Technological Forecasting and Social Change*, Vol. 126, pp. 3-13.

Wang, Y., Kung, L., Gupta, S. and Ozdemir, S. (2019), "Leveraging big data analytics to improve quality of care in healthcare organizations: a configurational perspective", *British Journal of Management*, Vol. 30 No. 2, pp. 362-388.

Ward, M.J., Marsolo, K.A. and Froehle, C.M. (2014), "Applications of business analytics in healthcare", *Business Horizons*, Vol. 57 No. 5, pp. 571-582.

Wehner, M.R., Levandoski, K.A., Kulldorff, M. and Asgari, M.M. (2017), "Research techniques made simple: an introduction to use and analysis of big data in dermatology", *Journal of Investigative Dermatology*, Vol. 137 No. 8, pp. e153-e158, doi: 10.1016/j.jid.2017.04.019.

Weng, C. and Kahn, M.G. (2016), "Clinical research informatics for big data and precision medicine", *Yearbook of Medical Informatics*, Vol. 25 No. 1, pp. 211-218, doi: 10.15265/IY-2016-019.

White, S.E. (2014), "A review of big data in healthcare: challenges and opportunities", *Bioinformatics*, Vol. 4 No. 6, pp. 13-18.

Wixom, B.H., Yen, B. and Relich, M. (2013), "Maximizing value from business analytics", *MIS Quarterly Executive*, Vol. 12 No. 2, pp. 111-123.

Wong, C.Y., Boon-Itt, S. and Wong, C.W. (2011), "The contingency effects of environmental uncertainty on the relationship between supply chain integration and operational performance", *Journal of Operations Management*, Vol. 29 No. 6, pp. 604-615.

Wu, J., Li, H., Cheng, S. and Lin, Z. (2016), "The promising future of healthcare services: when big data analytics meets wearable technology", *Information and Management*, Vol. 53 No. 8, pp. 1020-1033, doi: 10.1016/j.im.2016.07.003.

Xu, Z., Frankwick, G.L. and Ramirez, E. (2016), "Effects of big data analytics and traditional marketing analytics on new product success: a knowledge fusion perspective", *Journal of Business Research*, Vol. 69 No. 5, pp. 1562-1566.

Yadav, N., Shankar, R. and Singh, S.P. (2020), "Impact of Industry 4.0/ICTs, Lean Six Sigma and quality management systems on organisational performance", *The TQM Journal*, Vol. 32 No. 4, pp. 815-835.

Yu, W., Zhao, G., Liu, Q. and Song, Y. (2020), "Role of big data analytics capability in developing integrated hospital supply chains and operational flexibility: an organizational information processing theory perspective", *Technological Forecasting and Social Change*, Vol. 163, p. 120417.

Zepeda, E.D., Nyaga, G.N. and Young, G.J. (2016), "Supply chain risk management and hospital inventory: effects of system affiliation", *Journal of Operations Management*, Vol. 44, pp. 30-47.

Zhang, F., Cao, J., Khan, S.U., Li, K. and Hwang, K. (2015), "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications", *Future Generation Computer Systems*, Vol. 43, pp. 149-160, doi: 10.1016/j.future.2014.06.009.

Zhang, L., Wang, H., Li, Q., Zhao, M.H. and Zhan, Q.M. (2018), "Big data and medical research in China", *British Medical Journal*, p. 360.

Zikopoulos, P. and Eaton, C. (2011), *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media.

Further reading

Sentinelinitiative (2021), available at: https://www.sentinelinitiative.org.

About the authors

Dr Rajesh Kumar Singh is professor in operations management area at Management Development Institute (MDI), Gurgaon, India. His areas of interest include supply chain management, Industry 4.0,

circular economy. Currently he is serving as associate editor of *International Journal of Consumer Studies* (Wiley) and *International Journal of Global Business and Competitiveness* (Springer). He is also in editorial board of *Journal of Supply Chain Management Systems* and *Vision-The Journal of Business Perspective*. He has also worked as guest editor for special issues of *Resources, Conservation and Recycling*, *Sustainable Production and Consumption*, *Competitiveness Review*, *Journal of Modelling in Management*, *Administrative Sciences*, *Global Journal of Operations and Strategic Sourcing,* etc. He has published research papers in reputed international/national journals and conferences. He has published papers in journals such as *Resources, Conservation and Recycling*, *Journal of Cleaner Production*, *Production planning and Control*, *Annals of Operations Research*, *International Journal of Production Research*, *Applied Soft Computing*, *Clean Technologies and Environmental Policy*, *Industrial Management and Data Systems*, *Singapore Management Review*, *International Journal of Productivity and Performance Management*, *Benchmarking: An International Journal*, *Journal of Modelling in Management*, *Competitiveness Review: An International Business*. Rajesh Kumar Singh is the corresponding author and can be contacted at: rajesh.singh@mdi.ac.in

Dr Saurabh Agrawal works as associate professor in Delhi School of Management, Delhi Technological University (DTU), Delhi, India. He has vast experience of academics, research and the industry both in India and the USA. He has completed his PhD in reverse supply chain from DTU. His research focus is in the areas of supply chain management, reverse logistics, sustainability and e-waste management. He has completed his undergraduate degree from the Indian Institute of Technology, Roorkee, India. He has received his master's degree in business administration from Oregon State University, the USA, and master's degree in industrial engineering from the Indian Institute of Technology, Delhi, India. He has published research papers in international journals of repute including *Resources, Conservation and Recycling*, *Journal of Cleaner Production*, *Competitiveness Review* and *Journal of Industrial Engineering: International.*

Abhishek Sahu is research scholar in Mechanical Engineering Department, Delhi Technological University (DTU), Delhi, India.

Dr Yigit Kazancoglu is professor at Yaşar University şirketinde Professor, Turkey. His research areas are operations management, sustainability, circular economy, total quality management and statistical quality control, operations research, multicriteria decision-making (MCDM) methods and green logistics.

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**

Subthreshold Analytical Model of Asymmetric Gate Stack Triple Metal Gate All Around MOSFET (AGSTMGAAFET) for Improved Analog Applications

Arvind Ganesh, Kshitij Goel, Jaskeerat Singh Mayall, Sonam Rewari

Department of Electronics and Communication Engineering, Delhi Technological University,
New Delhi, India
chikki.ag@hotmail.com, kshitijg9899@gmail.com, jaskeeratsinghmayall@gmail.com, rewarisonam@gmail.com

***Abstract*— In this paper, we have proposed a 2D analytical model for Asymmetric gate stack triple metal gate MOSFET(AGSTMGAAFET) and performed a comparative analysis with the simulation results obtained using the SILVACO 3D simulation software. Existing devices such as gate all around single metal (SMGAAFET), gate all around triple metal (TMGAAFET), gate stack single metal (GSSMGAAFET), gate stack triple metal (GSTMGAAFET) and asymmetric gate stack single metal (AGSTMGAAFET) have been compared with our proposed structure AGSTMGAAFET. Our device provides excellent performance in terms of drain current, transconductance, output conductance, current gain, maximum transducer power gain which shows our device's suitability for various analog applications moreover the potential and electric field plots obtained have twostep profile and extremely low electric field near the drain region which ordains our device with the ability to suppress various SCE's like DIBL and hot-carrier effect. The analytical model and simulation results show good convergence in values which validate the correctness of the proposed model.**

Keywords— Single Metal Gate All Around (SMGAAFET) MOSFET, Triple Metal Gate All Around (TMGAAFET) MOSFET, Gate Stack Single Metal Gate All Around (GSSMGAAFET) MOSFET, Gate Stack Triple Metal Gate All Around (GSTMGAAFET) MOSFET, Asymmetric Gate Stack Single Metal Gate All Around (AGSSMGAAFET) MOSFET, Asymmetric Gate Stack Triple Metal Gate All Around (AGSTMGAAFET) MOSFET, Short Channel Effects (SCE).

I. INTRODUCTION

In modern day world where there is need for continuous miniaturization of MOSFET used in variety of integrated circuits, device scaling has been one of the prime tools to achieve this goal. However, device scaling is restricted up to a limit mainly due to short channel effects (SCE) [1]. The continued downscaling of CMOS has led to the need for ultra-thin gate dielectrics as well. However, the thinning of oxide layer is limited by direct tunneling which increases the leakage current. Thus, combination of these effects leads to major reliability issues and causes performance degradation over a period of time.

Thus, a number of device structures like Schottky barrier MOSFET (SBCGAA MOSFET), Schottky barrier graded stack MOSFET (SBGS-CGAA MOSFET) and Dual metal Schottky barrier graded stack MOSFET (DMSBGSCGAA) have been proposed earlier to overcome short channel effects. These gate all around MOSFET have greater control over the channel as they surround the substrate from all sides, provide increased packing efficiency and avoid corner effects [2]. However, they suffer from DIBL effects, threshold voltage fluctuations and difficulties in fabrication.
In this paper we have proposed a triple material gate FET that helps in overcoming existing DIBL, impact ionization and eliminates oxide breakdown effects to large extent by improved transport efficiency. This when used with gate stack provides high transconductance. For better carrier generation and improved performance, a high-k dielectric material like HfO_2 has been used along with SiO_2 in the proposed structure which reduces leakage current [3]. However, high-k materials result in fringing fields from the source/drain regions which tends to decrease the control of the gate over the channel. Thus, an ultra-thin SiO2 interlayer between the high-K layer and silicon substrate has been introduced which improves the quality of interface by decreasing trap density and provides stability [2]. To overcome shallow source/drain at junction and for easy fabrication metallic material instead of doped material is used. Hence overall proposed structure consists of asymmetric gate stack and triple material gate which collectively improve drain current characteristics and reduce SCE's by providing better gate control.
Using graphical simulation for the proposed Asymmetric Gate Stack Triple Gate (AGSTMGAAFET) the device showed best performance results in terms of drain current, conductance, intrinsic gain, transconductance, maximum transducer power gain (MTPG), current gain and output conductance (g_d) as compared to Asymmetric gate stack single metal FET(AGSSMGAAFET) ,Gate all around single metal FET (SMGAAFET) ,Gate all around triple metal FET(TMGAAFET), Gate stack single metal FET(GSSMGAAFET) and Gate stack triple metal FET(TMGSGAAFET) which have been proposed earlier to overcome SCE's to a great extent. Section 2 describes the

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.

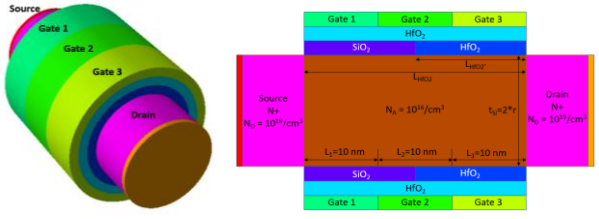
device structure of the proposed and the models used to simulate the various devices for comparison. Section 3 presents the analytical model of the device. Section 4 displays the comparative analysis of our device against those previously proposed and the comparison is made on the basis of various parameters such as electric field, potential, subthreshold current for different lengths and radius, MTPG, current gain, transconductance, output conductance, I_{DS} vs V_{GS} and I_{DS} vs V_{DS} graphs and also the potential contour plots of all the devices has been studied. In section 5 we provide the conclusion, in section 6 we acknowledge the inputs of all those who have guided us throughout the course of our research and section 7 presents the calculated coefficients of the analytical model.

II. DEVICE STRUCTURE

The three-dimensional view of Asymmetric Gate Stack Triple Metal Gate All Around (AGSTMGAAFET) MOSFET is presented in the figure 1(a). The gate material is composed of three metals each having different work functions. The highest work function metal (4.8 eV) is present at the source end and between the two the metal with work function 4.62 eV is present while the with the least work function metal (4.4 eV) is present at the drain end. Source and drain are doped with a concentration of 10^{19} cm^{-3} (n-type) and the substrate is doped with concentration of 10^{16} cm^{-3} (p-type). The thickness of the silicon substrate is t_{Si}=10 nm. Hafnium Oxide (HfO_2) is the gate stack material used in the device (t_{Hfo2}=2 nm, $\varepsilon_{r,Hfo2}=22\varepsilon_0$). Two methods namely Newton & Gummel are used for solving using ATLAS-3D.

The decreasing gradient of work functions from source towards the drain helps in reducing the value of electric field at the drain end thereby regulating it throughout the channel. Due to this, electron overshooting is prevented which is basically the hot-carrier effect (another type of SCE).

The 2-D cross-sectional view of AGSTMGAAFET is shown in figure 1(b). This topology divides the device into 4 regions. First region consisting of the gate material M_1 (first). Second region consisting of half of gate material M_2 (second), with SiO_2 as the layer between HfO_2 and the substrate. Third region consisting of the other half of gate material M_2 (second), with HfO_2 as the layer between HfO_2 and the substrate and fourth region consisting of the gate material M_3 (third).



(a) (b)

Fig 1.(a) 3-D Structural View of Asymmetric Gate Stack Triple Metal Gate All Around MOSFET (b) 2-D Cross sectional View of Asymmetric Gate Stack Triple Metal Gate All Around MOSFET

The list of various parameters used for all the single metal and all the triple metal device structures are tabulated in table 1 and 2 respectively. The list and description of models used for the simulation of the device using ATLAS-3D are tabulated in table 3 [4].

TABLE 1. Single Metal Devices Structural parameters

Parameters	SMGAA MOSFET	GSSMGAA MOSFET	AGSSMGAA MOSFET
L (nm)	30	30	30
N_A (cm^{-3})	1 x 10^{16}	1 x 10^{16}	1 x 10^{16}
N_D (cm^{-3})	1 x 10^{19}	1 x 10^{19}	1 x 10^{19}
t_{ox} (nm)	2	2	2
t_{hfo2} (nm)	-	2	2
$L_{hfo2'}$ (nm)	-	-	15
t_{Si} (nm)	20	20	20
ϕ_m (eV)	4.8	4.8	4.8
$L_{S/D}$ (nm)	10	10	10

TABLE 2. Triple Metal Devices Structural parameters

Parameters	TMGAA MOSFET	GSTMGAA MOSFET	AGSTMGAA MOSFET
L (nm)	30	30	30
N_A (cm^{-3})	1 x 10^{16}	1 x 10^{16}	1 x 10^{16}
N_D (cm^{-3})	1 x 10^{19}	1 x 10^{19}	1 x 10^{19}
t_{ox} (nm)	2	2	2
t_{hfo2} (nm)	-	2	2
$L_{hfo2'}$ (nm)	-	-	15
t_{Si} (nm)	20	20	20
ϕ_{m1} (eV)	4.8	4.8	4.8
ϕ_{m2} (eV)	4.62	4.62	4.62
ϕ_{m3} (eV)	4.4	4.4	4.4
$L_{S/D}$ (nm)	10	10	10

TABLE 3. List of Models used

S. No.	Model Used	Description
1.	SRH	Shockley-Read-Hall which is a generation and recombination model is used with fixed lifetimes.
2.	CONMOB	Concentration dependent mobility model is used which basically is a doping-mobility table (valid at 300K).
3.	FLDMOB	This specifies the transverse field degradation for electrons. Standard saturation models are used for both electrons and holes (specified from the values of EVSATMOD and HVSATMOD which indicates which parallel field dependent mobility model to be used).
4.	BOLTZMANN	Boltzmann is the carrier statistics model which is used for preventing impact ionization and overshooting.

III. ANALYTICAL MODEL

The Asymmetric gate stack triple metal exhibits a symmetry which is cylindrical in nature therefore to take advantage of this it is a good choice to resort to the use of cylindrical coordinates instead of cartesian coordinates because it greatly simplifies the solution procedure that is the solution is gotten using the 2D Poisson's equation rather than 3D Poisson's equation this luxury is afforded to us as a result of the fact that the potential and electric field distribution in our device do not depend on the variation of the azimuthal angle . The 2D Poisson's equation for a AGSTMGAA MOSFET is given by-

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \Phi_i(r,z)}{\partial r}\right) + \frac{\partial^2 \Phi_i(r,z)}{\partial z^2} = \frac{qN_A}{\varepsilon_{si}} \quad (1)$$

$0 \le r \le \frac{t_{si}}{2}$ and $L_{i-1} \le z \le L_i$ where i =1,2,3,4

Here 'r' represents the channel radius and 'z' represents the axis along which we measure the length of the channel and Φ_i (r,z) represents the potential as a function of r and z.

The form of the solution is as shown below were the use of boundary condition gives us the required coefficients by using the parabolic approximation method.

$$\phi(r,z) = a_0 + a_1 r + a_2 r^2 \quad (2)$$

Where a_0, a_1, a_2 represent the coefficients which are a function of 'z'. The calculation of these coefficients is done using the following boundary conditions: -

1. The electric field vanishes along the axis of symmetry

$$\frac{\partial \Phi_i(r,z)}{\partial r}\Big|_{r=0} = 0 \quad (3)$$

2. The electric field should be invariant crossing over from the dielectric to the silicon body

$$\frac{\partial \Phi_i(r,z)}{\partial r}\Big|_{r=\frac{t_{si}}{2}} = \frac{\varepsilon_j}{\varepsilon_{si}}\frac{(\phi_{gsi} - \phi_{s_i})}{t_k} \quad (4)$$

Where $\phi_{gsi} = V_{gs} - V_{fbi}$, $\phi_{s_i} = \Phi_i(r,z)|_{r=0}$ and i=1,2,3,4 ,j=1,2,k=1,2

$$\varepsilon_1 = \varepsilon_{ox} \text{ for } i = 1,2 \text{ and } \varepsilon_2 = \varepsilon_{hf} \text{ for } i = 3,4$$

$$t_1 = R\ln(1 + \frac{t_{oxeff1}}{R}) \text{ and } t_2 = R\ln(1 + \frac{t_{oxeff2}}{R}),\ R = \frac{t_{si}}{2} \quad (5)$$

$$t_{oxeff1} := t_{ox} + \frac{\varepsilon_{ox}}{\varepsilon_{hf}} \cdot t_{hf},\ t_{oxeff2} := t_{ox} + t_{hf} \quad (6)$$

3. The potential should not vary as we move from a region under one metal to another having different work functions.

$$\Phi_i(r, L_i) = \Phi_i(r, L_{i+1}),\ i=1,2,3 \quad (5)$$

4. The electric field should not vary as we move from a region under one metal to another having different work functions.

$$\frac{\partial \Phi_i(r,z)}{\partial z}\Big|_{z=L_i} = \frac{\partial \Phi_i(r,z)}{\partial z}\Big|_{z=L_{i+1}},\ i=1,2,3 \quad (6)$$

5. Potential measured at the source should be V_{bi}

$$\Phi_1(r,0) = V_{bi} \quad (7)$$

Where $V_{bi} = V_T \ln(\frac{N_A N_D}{n_i^2})$, $V_T = \frac{kT}{q}$

6. Potential measured at the drain should be $V_{bi} + V_{DS}$

$$\Phi_4(r, L_4) = V_{bi} + V_{DS} \quad (8)$$

We apply the first two boundary conditions and rearrange the equation to solve for the surface potential.

$$\frac{\partial^2 \phi_{s_1}(z)}{\partial z^2} - K_1^2 \phi_{s_1}(z) = Y_1,\ 0 \le z \le L_1 \quad (9)$$

$$\frac{\partial^2 \phi_{s_2}(z)}{\partial z^2} - K_1^2 \phi_{s_2}(z) = Y_2,\ L_1 \le z \le L_2 \quad (10)$$

$$\frac{\partial^2 \phi_{s_3}(z)}{\partial z^2} - K_2^2 \phi_{s_3}(z) = Y_3,\ L_2 \le z \le L_3 \quad (11)$$

$$\frac{\partial^2 \phi_{s_4}(z)}{\partial z^2} - K_2^2 \phi_{s_4}(z) = Y_4,\ L_3 \le z \le L_4 \quad (12)$$

Where,

$$K_1 := \left(\frac{2\cdot\varepsilon_{ox}}{R^2\cdot\varepsilon_{Si}\cdot\ln\left(1+\frac{t_{oxeff1}}{R}\right)}\right)^{0.5} \quad K_2 := \left(\frac{2\cdot\varepsilon_{hf}}{R^2\cdot\varepsilon_{Si}\cdot\ln\left(1+\frac{t_{oxeff2}}{R}\right)}\right)^{0.5}$$

$$Y_1 := \frac{q\cdot N_A}{\varepsilon_{Si}} - K_1^2\cdot(V_{gs} - V_{FB1}) \quad Y_2 := \frac{q\cdot N_A}{\varepsilon_{Si}} - K_1^2\cdot(V_{gs} - V_{FB2})$$

$$Y_3 := \frac{q\cdot N_A}{\varepsilon_{Si}} - K_2^2\cdot(V_{gs} - V_{FB3}) \quad Y_4 := \frac{q\cdot N_A}{\varepsilon_{Si}} - K_2^2\cdot(V_{gs} - V_{FB4})$$

$$V_{FBi} = \phi_{Mi} - \phi_{si},\ i = 1,2,3$$

Here ϕ_{Mi} represents the work function of the gate material and ϕ_{si} represents the work function of the silicon substrate.

$\phi_{si} = \chi_{si} + \frac{E_g}{2q} + \phi_f$, were $\phi_f = V_T \ln(\frac{N_A}{n_i})$

The solution for this equation is obtained in two parts that is in terms of the forced and natural solutions.

$$\phi_S(z) := \begin{cases} A_1 \cdot e^{K_1 \cdot z} + B_1 \cdot e^{-K_1 \cdot z} - \frac{Y_1}{K_1^2} & \text{if } 0 \le z \le L_1 \\ A_2 \cdot e^{K_1 \cdot z} + B_2 \cdot e^{-K_1 \cdot z} - \frac{Y_2}{K_1^2} & \text{if } L_1 \le z \le L_2 \\ A_3 \cdot e^{K_2 \cdot z} + B_3 \cdot e^{-K_2 \cdot z} - \frac{Y_3}{K_2^2} & \text{if } L_2 \le z \le L_3 \\ A_4 \cdot e^{K_2 \cdot z} + B_4 \cdot e^{-K_2 \cdot z} - \frac{Y_4}{K_2^2} & \text{if } L_3 \le z \le L_4 \end{cases} \quad (13)$$

To obtain the necessary coefficients we employ the boundary condition discussed above. The coefficients are presented in the appendix at the end. The electric field which is the negative gradient of the potential is obtained simply by taking the first derivative of the potential.

$$EF(z) = -\frac{\partial \Phi_i(r,z)}{\partial z} \quad (14)$$

$$EF(z) := \begin{cases} K_1 \cdot A_1 \cdot e^{K_1 \cdot z} - K_1 \cdot B_1 \cdot e^{-K_1 \cdot z} & \text{if } 0 \le z \le L_1 \\ K_1 \cdot A_2 \cdot e^{K_1 \cdot z} - K_1 \cdot B_2 \cdot e^{-K_1 \cdot z} & \text{if } L_1 \le z \le L_2 \\ K_2 \cdot A_3 \cdot e^{K_2 \cdot z} - K_2 \cdot B_3 \cdot e^{-K_2 \cdot z} & \text{if } L_2 \le z \le L_3 \\ K_2 \cdot A_4 \cdot e^{K_2 \cdot z} - K_2 \cdot B_4 \cdot e^{-K_2 \cdot z} & \text{if } L_3 \le z \le L_4 \end{cases} \quad (15)$$

The subthreshold current can be calculated using the formula [5].

$$I_{sub} := 2 \cdot R \cdot \pi \cdot \mu \cdot k \cdot T \cdot \eta_i \cdot \frac{\left(1 - e^{-q \cdot \frac{V_{ds}}{k \cdot T}}\right)}{\int_0^{L_1+L_2+L_3+L_4} \frac{1}{\int_0^R e^{\frac{\phi_i(r,z)}{k \cdot T}} dr} dz} \quad (16)$$

Where k=1.38 x 10^{23} J/K represents the boltzmann's constant, T represents temperature and μ =1300 cm^2/V_s represents the electron mobility.

And the corresponding sub threshold slope is given by [5].

$$SS = \frac{1}{\frac{d \log I_{sub}}{dV_{gs}}} \quad (17)$$

IV. RESULTS AND DISCUSSION

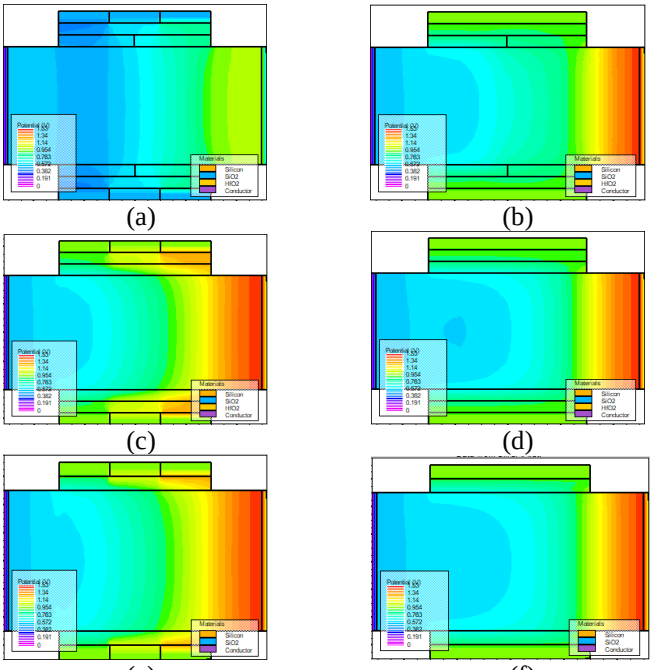


Figure 2: Contour potential plots for the various devices under study where (a) AGSTMGAAFET (b) AGSGSMGAAFET (c) GSTMGAAFET (d) GSSMGAAFET (e) TMGAAFET (f) SMGAAFET

Potential is an important parameter for understanding device physics as it explains the different FET's behavior under the same biasing condition as potential would instigate the charge distribution across the channel. On this charge distribution across the channel, complete device characteristics are dependent. We can clearly observe from the contour plots in Fig 2. that the increase in potential in triple metal FET's is much more as compared to the single metal FET's this basically leads to a reduction in short channel effects like hot carrier effect and impact ionization. Also, the use of asymmetric gate stack with high dielectric leads to a large capacitance and greater control over the channel. This in turn increases the electric field throughout the channel and also reduces any possibility of gate tunneling.

Figure 3 gives us an idea of the dependence of potential on position in the channel for all the devices under study. We are analyzing devices with a single and triple gate material; the latter case leads us to observe two steps in the potential moving from source to drain, this ordains our device with the ability to overcome DIBL to a great extent also this leads to an improvement in the efficiency of carrier transport and also the carrier speed leading to improved drain current [2]. This can be attributed to the change in work function moving from source to drain. Another point of interest is that devices with gate stack exhibit relatively lower potential which can be attributed to the use of triple metal gate with different work function [5].

Figure 4 analyses the dependence of electric field on the position the channel. Electric field is the first order derivative of potential. This graph exemplifies the benefit of

using an asymmetric gate stack which as shown leads to lower electric field values. In nano-scale device architectures reduction of SCE's is of paramount importance and tis lower electric field especially visible at the drain end goes a long way in achieving that [2].

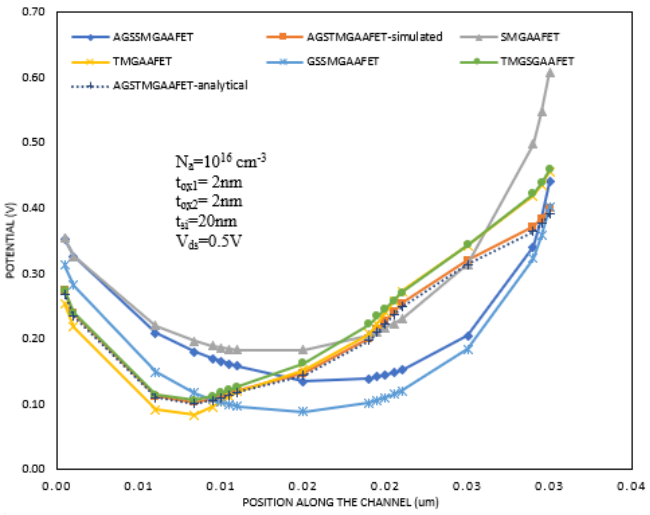


Figure 3: Surface Potential for all 6 models and analytical model for AGSTMGAAFET

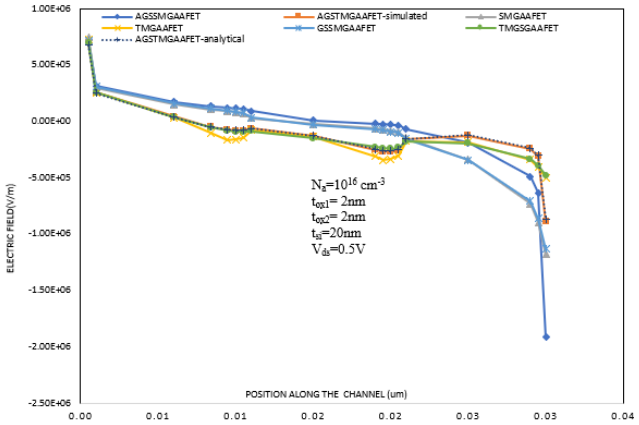
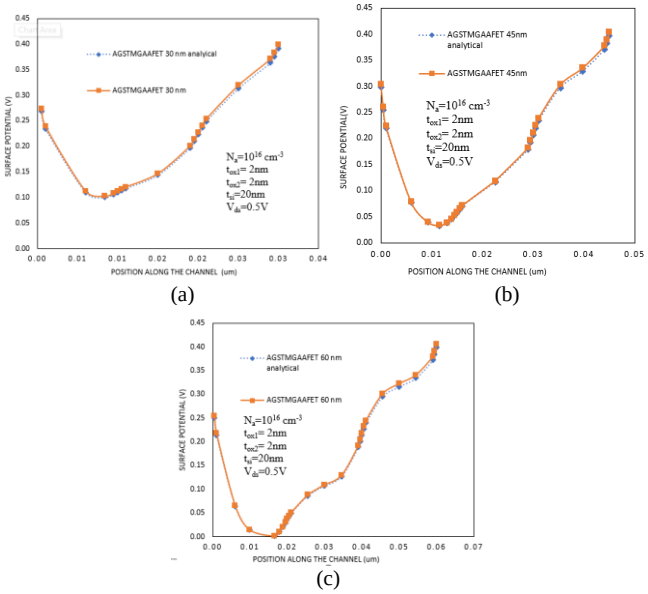


Figure 4: Electric Field for all 6 models and analytical model for AGSTMGAAFET

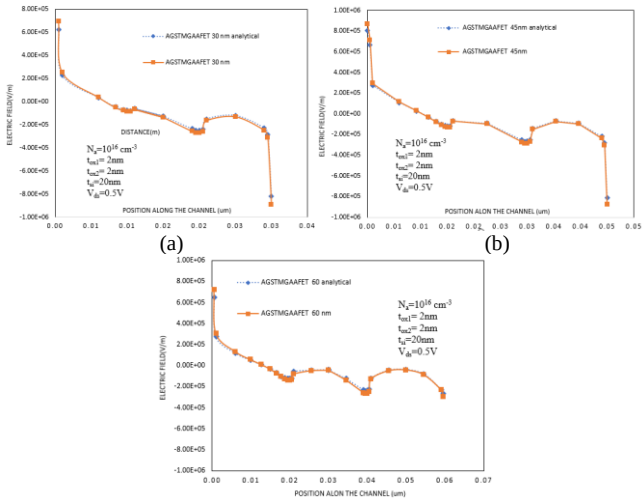


(a) (b)

(c)

Figure 5: Surface Potential vs Channel Length for AGSTMGAAFET (a) L = 30 nm (b) L = 45 nm (c) L = 60 nm

Figure 5 gives us the distribution of surface potential as a function of channel length ,now the first thing to observe here is the fact that the potential has a minima which progressively decreases as we increase the channel length this can be attributed to the decrease in charge control linear region which leads to the entire potential to slide in the direction of source [5].The analytical results are much in co-ordinance with the simulated results.



(a) (b)

(c)

Figure 6: Electric Field vs Channel Length for AGSTMGAAFET (a) L = 30 nm (b) L = 45 nm (c) L = 60 nm

Figure 6 presents the electric field distribution at the surface as a function of position in the channel for three values of channel length (L=30 nm,45nm,60nm). The triple metal asymmetric gate stack MOSFET gives us a twostep profile for the electric field where the steps occur at the interface of two gate metals [6]. This is extremely beneficial because it leads to the electric field being averaged out throughout the length of the channel moreover, we observe a marked reduction in electric field near the drain, both of these help in reducing SCEs [2].

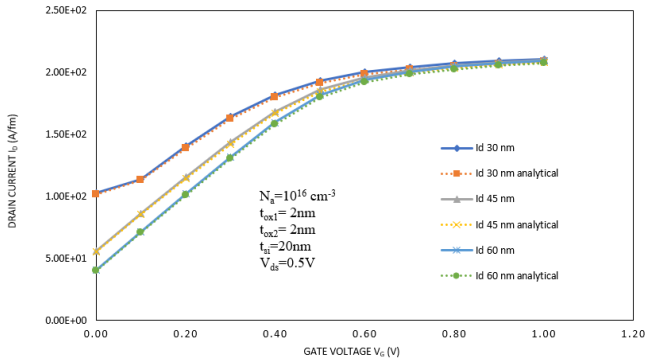
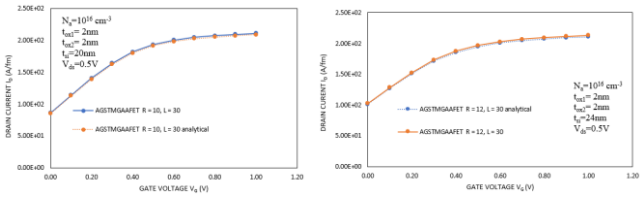


Figure 7: Subthreshold current (I_D) vs gate voltage (V_{GS}) for AGSTMGAAFET for L = 30, 45 and 60 nm

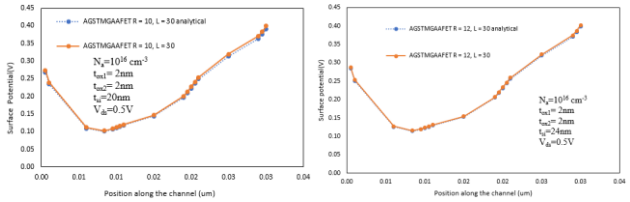
Figure 7 gives the graphical relationship between subthreshold current and the applied gate input voltage. Now the thing to observe here is that as we decrease the length of the channel correspondingly the value of subthreshold current increases leading to increased power consumption and other related SCEs [5].



(a) (b)

Figure 8: Subthreshold current (I_D) vs gate voltage (V_{GS}) for AGSTMGAAFET with L = 30 nm and different radius (a) R = 10 nm (b) R = 12 nm

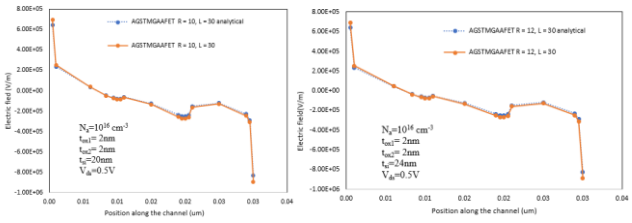
In figure 8, we have shown a comparison between source to drain and gate input voltage for different radii of the channel region and for a particular channel length(30nm). The observation to make here is that smaller the channel radius greater the subthreshold slope leading to lower off current and hence lower power dissipation [7].



(a) (b)

Figure 9: Surface Potential for AGSTMGAAFET with L = 30 nm and different radius (a) R = 10 nm (b) R = 12 nm

In figure 9, we show the variation of surface potential with different values of channel radii. The observation to be made here is that as the length of the channel goes up the position of potential minima also rises up as a result of the fact that controllability of the gate over the channel goes down with increasing channel thickness, this causes V_T to decrease as we increase channel radius so a smaller channel radius is desirable to keep off current in check [8].



(a) (b)

Figure 10: Electric Field for AGSTMGAAFET with L = 30 nm and different radius (a) R = 10 nm (b) R = 12 nm

In figure 10, the clear decline in electric field near the drain region because of the high -k dielectric (HfO_2) and triple metal gate this gives us the benefit of reduced hot-carrier effect and therefore additional reduction in power dissipation [9].

Figure 11 shows the variation in the drain current (I_{DS}) with the gate voltage (V_{GS}) for SMGAAFET, TMGAAFET, GSSMGAAFET, GSTMGAAFET, AGSSMGAAFET and AGSTMGAAFET MOSFET devices. It can be seen that the AGSTMGAAFET generates the highest drain current. Because of the amorphous nature, HfO_2 has high kinetic stability. This ensures less leakages in the corresponding MOSFET [10]. Also, in its presence the concentration of electrons increases in the channel because of higher capacitance which then results in higher I_{DS} [11]. The strength of the fringing electric fields also increases thereby increasing I_{DS} [12]. Gate stack devices give higher current than non-gate stack devices because of their presence of high dielectric (HfO_2) graded channel. Triple metal devices give higher current than single metal devices as they reduce the overshooting of the electrons. The general order of drain current is asymmetric gate stack greater than gate stack greater than non-gate stack FETs.

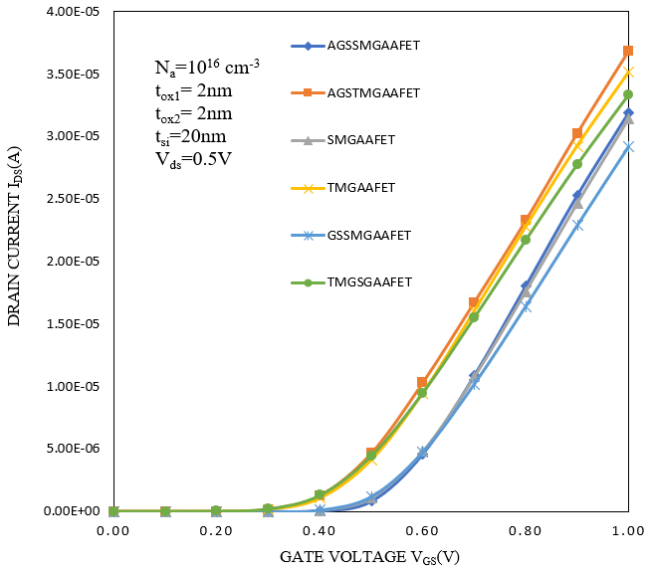


Figure 11: Plot of the change in drain current (I_{DS}) relative to change in gate voltage (V_{GS})

Figure 12 shows the variation in the transconductance (g_m) with the gate voltage (V_{gs}) for SMGAAFET, TMGAAFET, GSSMGAAFET, GSTMGAAFET, AGSSMGAAFET and AGSTMGAAFET devices. Transconductance is calculated by derivation of drain current relative to the gate voltage at constant V_{ds} [13].

$$g_m = \frac{\partial I_{ds}}{\partial v_{gs}}\bigg|_{v_{ds}}$$

It can be seen that initially the transconductance increases since the device is in linear region and then it tends to decrease when device enters the saturation region. In the presence of triple metal surrounding gates and HfO_2 the carrier mobility increases and the gate control over the channel region becomes more enhanced. This, in turn, increases the drain current and hence increases the transconductance (since g_m is proportional to I_{DS}). As g_m increases, it also implies a higher cut off frequency (f_T) [13].

Figure 13 shows the variation of the current flowing from source to drain with respect to the applied voltage V_{DS}. It can be seen clearly from the figure that our proposed model AGSTMGAAFET gives superior performance in terms of current as compared to all other devices under inspection. Owing to the amorphous nature of HfO_2 which gives us the benefit of high kinetic stability, leakage current is greatly reduced further the asymmetric gate stack provides a much higher drain current owing to its high relative permittivity, also the triple metal gate leads to a large reduction in short channel effects such as hot carrier effect, impact ionization, etc.

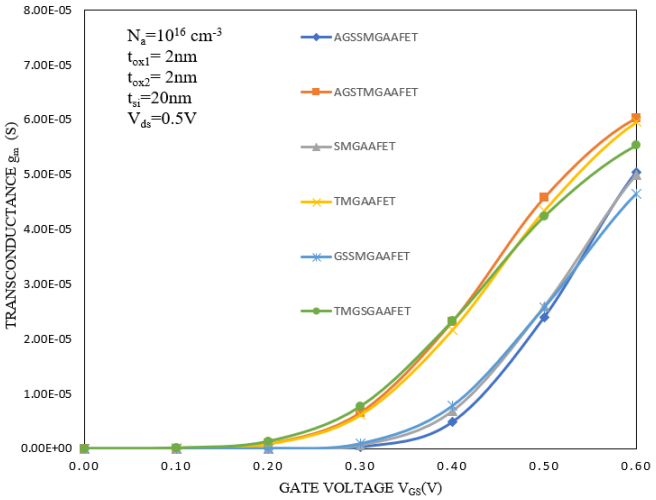


Figure 12: Plot of the change in transconductance (g_m) relative to change in gate voltage (V_{GS})

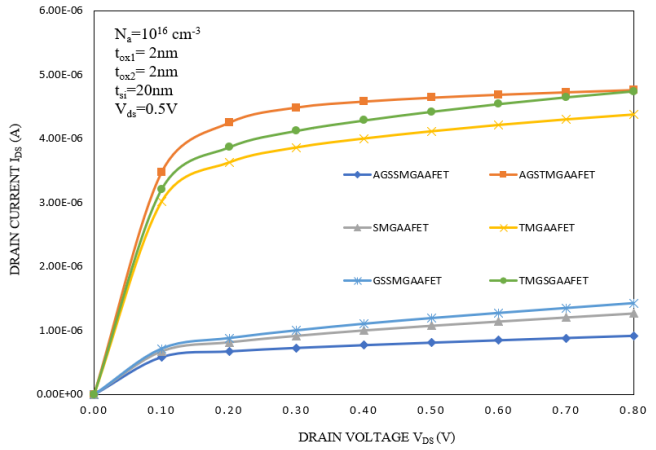


Figure 13: Plot of the change in drain current (I_{DS}) relative to change in drain voltage (V_{DS})

Figure 14 gives the variation of output conductance (g_d) with respect to the drain to source voltage (V_{DS}). The mathematical expression for output conductance is given as [13]:

$$g_d = \frac{\partial I_{DS}}{\partial V_{DS}}|_{V_{GS}}$$

It is observed that increased drain current obtained by the AGSTMGAAFET because of the excellent control over the channel which is provided by the asymmetric gate stack. The high relative permittivity of HfO_2 increases the capacitance of the device which in turn enhances the drain current and subsequently g_d as well.

Figure 15 clearly gives us the idea that the current gain offered by the proposed device AGSTMGAAFET is superior to the other devices under study, this result can be attributed to the fact that our device simply provides excellent performance in terms of drain current. The use of high-K dielectric (HfO_2) increases the gate capacitance, this in turn enhances the electric field as a by-product of which electron velocity increases leading to better current gain. Moreover, the decreasing work function of the gate material as we move from the source to the drain reduces unwanted SCE's.

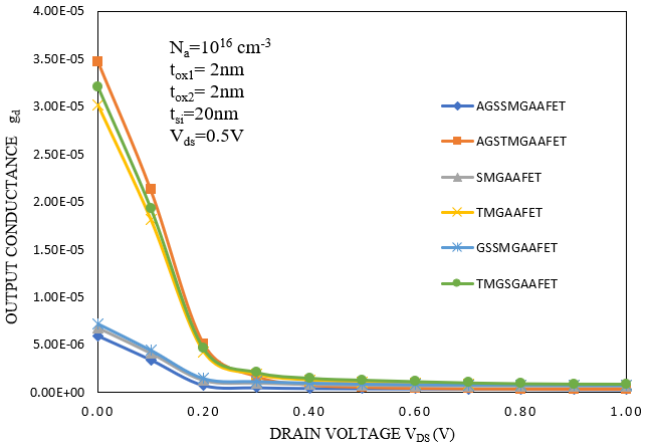


Figure 14: Plot of the change in output conductance (g_d) relative to the change in drain voltage (V_{DS})

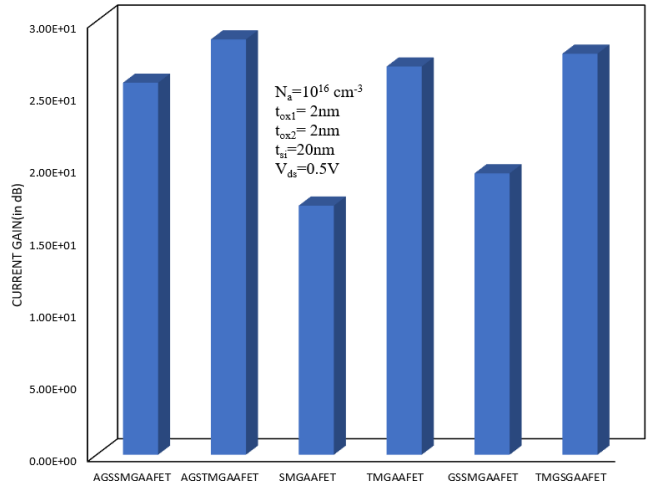


Figure 15: Current gain for different devices

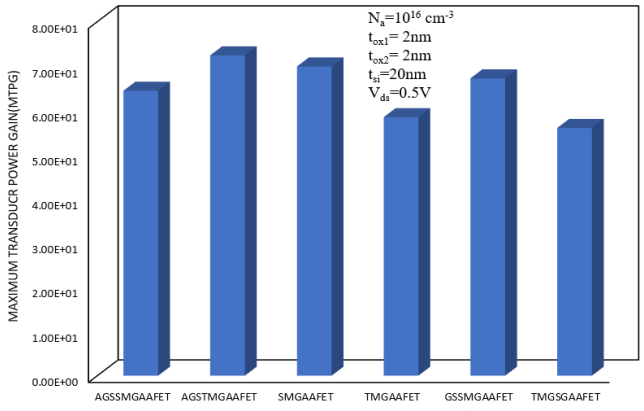


Figure 16: Maximum transducer power gain (MTPG for various devices)

Figure 16 gives the comparison of the Maximum transducer power gain (MTPG) of the various devices under study. This parameter basically gives the highest amount of power delivered to the load as a ratio of the power available from the source. This situation occurs when the load (Z_L) is matched to the devices output impedance (Z_O). The proposed device gives excellent performance in this respect as well.

$$MTPG = \frac{P_{load}}{P_{source}}|Z_L = Z_O{}^*$$

Now the combination of asymmetric gate stack with high-K dielectric and triple metal gate leads to the device having a large capacitance which in turn leads to a large electric field which causes a high saturation velocity for the charge carriers which leads to a large current value going from source to drain leading to a large MTPG.

V. CONCLUSION

In this paper AGSTMGAAFET is studied in terms of various device parameters. Excellent performance in terms of drain current, transconductance, output conductance, current gain, maximum transducer power gain which shows our device's suitability for various analog applications. Our device provides the advantages of triple work function gate and asymmetric gate stack which leads to increased carrier velocity, improvement of efficiency in carrier transport and alleviation of DIBL, hot-carrier effect and other SCE's moreover it leads to reduced power dissipation. Overall, the proposed device is suitable for various analog application requiring high gain and low power dissipation. The proposed analytical model for AGSTMGAAFET has been analyzed for different channel length (L= 30,45,60 nm) and radius of substrate (R = 10,12 nm). It is so found that the analytical results for potential, electric field and subthreshold current are in close accordance with the simulated results.

VI. ACKNOWLEDMENT

We would like to thank Professor R.S. Gupta (Department of Electronics and Communication Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India) for his constant guidance and support throughout the course of our research especially for his help with SILVACO ATLAS 3D simulation software.

VII. FUNDING STATEMENT

Not Applicable

VIII. CONFLICT OF INTEREST

There are no conflicts of interest amongst the authors

IX. AUTHOR CONTRIBUTIONS

The authors have contributed mutually regarding this paper.

X. AVAILABILITY OF DATA AND MATERIAL

Not Applicable

XI. COMPLIANCE WITH ETHICAL STANDARDS

Not Applicable

XII. CONSENT TO PARTICIPATE

All the authors have complete consent to participate.

XIII. CONSENT FOR PUBLICATION

All the authors have complete consent for publication.

XIV. APPENDIX

$$P := -\frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot \left[V_{bi} + X_1 - \frac{(X_1 - X_2)}{2 \cdot d_1^{-1}} \right] \cdot d_2 + \frac{(K_2 + K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot \frac{(X_1 - X_2)}{2 \cdot d_1^{-1}} \cdot d_2^{-1} + \frac{(X_3 - X_4)}{2 \cdot d_4} \cdot d_5^{2} + \frac{(X_3 - X_4)}{2 \cdot d_4^{-1}}$$

$$B_1 := \frac{(V_{bi} + V_{ds} + X_4) \cdot d_5 + \frac{(X_2 - X_3)}{2 \cdot d_3} \cdot d_5^{2} + \frac{(X_2 - X_3)}{2 \cdot d_3^{-1}} - \frac{(K_2 + K_1)}{2 \cdot K_2 \cdot d_3} \cdot \left[V_{bi} + X_1 - \frac{(X_1 - X_2)}{2 \cdot d_1} \right] \cdot d_2 \cdot d_5^{2} + \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3} \cdot \frac{(X_1 - X_2)}{2 \cdot d_1^{-1}} \cdot d_2^{-1} \cdot d_5^{2} + P}{\frac{(K_2 + K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot d_2^{-1} - \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot d_2 + \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3} \cdot d_2^{-1} \cdot d_5^{2} - \frac{(K_2 + K_1)}{2 \cdot K_2 \cdot d_3} \cdot d_2 \cdot d_5^{2}}$$

$$B_2 := B_1 + \frac{(X_2 - X_1)}{2 \cdot d_1^{-1}} \qquad B_3 := \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot A_2 \cdot d_2 + \frac{(K_1 + K_2)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot B_2 \cdot d_2^{-1} - \frac{K_2 \cdot (X_2 - X_3)}{2 \cdot K_2 \cdot d_3^{-1}}$$

$$B_4 := \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot \left[V_{bi} - B_1 + X_1 - \frac{(X_1 - X_2)}{2 \cdot d_1} \right] d_2 + \frac{(K_2 + K_1)}{2 \cdot K_2 \cdot d_3^{-1}} \cdot \left[B_1 - \frac{(X_1 - X_2)}{2 \cdot d_1^{-1}} \right] \cdot d_2^{-1} - \frac{(X_2 - X_3)}{2 \cdot d_3^{-1}} - \frac{(X_3 - X_4)}{2 \cdot d_4^{-1}}$$

$$A_1 := V_{bi} - B_1 + X_1 \qquad A_2 := A_1 - \frac{(X_1 - X_2)}{2 \cdot d_1} \qquad A_3 := \frac{(K_1 + K_2)}{2 \cdot K_2 \cdot d_3} \cdot A_2 \cdot d_2 + \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3} \cdot B_2 \cdot d_2^{-1} - \frac{K_2 \cdot (X_2 - X_3)}{2 \cdot K_2 \cdot d_3}$$

$$A_4 := \frac{(K_2 + K_1)}{2 \cdot K_2 \cdot d_3} \cdot \left[V_{bi} - B_1 + X_1 - \frac{(X_1 - X_2)}{2 \cdot d_1} \right] \cdot d_2 + \frac{(K_2 - K_1)}{2 \cdot K_2 \cdot d_3} \cdot \left[B_1 - \frac{(X_1 - X_2)}{2 \cdot d_1^{-1}} \right] \cdot d_2^{-1} - \frac{(X_2 - X_3)}{2 \cdot d_3} - \frac{(X_3 - X_4)}{2 \cdot d_4}$$

Where,

$$X_1 := \frac{Y_1}{K_1^2},\ X_2 := \frac{Y_2}{K_1^2},\ X_3 := \frac{Y_3}{K_2^2},\ X_4 := \frac{Y_4}{K_2^2},\ d_1 := e^{K_1 \cdot L_1},\ d_2 := e^{K_1 \cdot L_2},\ d_3 := e^{K_2 \cdot L_2},\ d_4 := e^{K_2 \cdot L_3},\ d_5 := e^{K_2 \cdot L_4}$$

REFERENCES

[1] Kumar M, Haldar S, Gupta M, Gupta R (2016a) Physics based analytical model for surface potential and subthreshold current of cylindrical Schottky Barrier gate all around MOSFET with high-k gate stack. Superlattices Microstruct 90:215–226.

[2] Shreya Nandy, Sanjana Srivastava, Sonam Rewari, Vandana Nath, R. S. Gupta (2019), "Dual metal and Schottky barrier asymmetric gate stack cylindrical gate all around MOSFET for improved analog performance", Springer-Verlag GmbH Germany, part of Springer Nature 2019

[3] Arunprathap S, Napolean A, Azariah C (2014) Fabrication of thin film transistor using high K dielectric materials. Int J Eng Comput Sci 3(3):5387–5391.

[4] ATLAS: 3D Device Simulator, SILVACO International, 2015.

[5] Rewari, Sonam & Haldar, Subhasis & Nath, Vandana & Deswal, S. & Gupta, R.S.. (2015). Numerical modeling of Subthreshold region of junctionless double surrounding gate MOSFET (JLDSG). Superlattices and Microstructures. 90. 10.1016/j.spmi.2015.11.026.

[6] Jiang, Chunsheng & Liang, Renrong & Wang, Jing & Xu, Jun. (2015). A two-dimensional analytical model for short channel junctionless double-gate MOSFETs A two-dimensional analytical model for short channel junctionless double-gate MOSFETs. AIP Advances. 4. 10.1063/1.4921086.

[7] Basak, Arighna & Sarkar, Angsuman. (2020). Drain Current Modelling of Asymmetric Junctionless Dual Material Double Gate MOSFET with High K Gate Stack for Analog and RF Performance. Silicon. 10.1007/s12633-020-00783-w.

[8] Pal, Arobinda & Sarkar, Angsuman. (2014). Analytical study of Dual Material Surrounding Gate MOSFET to suppress short-channel effects (SCEs). Engineering Science and Technology, an International Journal. 17. 10.1016/j.jestch.2014.06.002.

[9] Darwin, s & Samuel, Dr.T.S.Arun. (2020). A Holistic Approach on Junctionless Dual Material Double Gate (DMDG) MOSFET with High k Gate Stack for Low Power Digital Applications. Silicon. 12. 1-11. 10.1007/s12633-019-00128-2.

[10] Sonam Rewari, Vandana Nath, Subhasis Haldar, S. S. Deswal, R. S. Gupta. "Hafnium oxide based cylindrical junctionless double surrounding gate (CJLDSG) MOSFET for high speed, high frequency digital and analog applications", Microsystem Technologies, 2017.

[11] Gnani E, Reggiani S, Baccarani G (2007) Effect of high—k (HfO2) gate dielectrics in double—gate and cylindrical-nanowire FETs scaled to the ultimate technology nodes. IEEE Trans Nanotechnol 6(1):90–96.

[12] Sanjana Srivastava, Shreya Nandy, Sonam Rewari, R.S. Gupta, "Oxide Stack Engineered Double Surrounding Gate (OSE-DSG) MOSFET for Submillimeter Analog Application", 2018 IEEE Electron Device Kolkata Conference (EDKCON).

[13] Kang, S. Mo and Y. Leblebici. CMOS digital integrated circuits,Tata McGraw-Hill Education, 2003.

Sensing and Bio-Sensing Research 32 (2021) 100423

Contents lists available at ScienceDirect

Sensing and Bio-Sensing Research

journal homepage: www.elsevier.com/locate/sbsr

Synthesis and fabrication of 2D Tungsten trioxide nanosheets based platform for impedimetric sensing of cardiac biomarker

Deepika Sandil [a,b], Saurabh Srivastava [c], Ritika Khatri [a], Kanika Sharma [a], Nitin K. Puri [a,*]

[a] Advanced Sensor Laboratory, Department of Applied Physics, Delhi Technological University, Delhi 110042, India
[b] Department of Applied Physics, Bhagwan Parshuram Institute of Technology, Delhi 110089, India
[c] Department of Applied Science and Humanities, Rajkiya Engineering College Ambedkar Nagar (Dr. A.P.J. Abdul Kalam Technical University, Lucknow), Uttar Pradesh 224122, India

ARTICLE INFO

Keywords:
Biosensor
Cardiac biomarker
Impedimetry
Electrochemistry
Nanosheets
Tungsten trioxide

ABSTRACT

Herein, we report the development of a two-dimensional (2D) label-free matrix for cardiac biomarker detection. Tungsten trioxide nanosheets (WO_3 NS) were synthesized by a facile, environmentally friendly hydrothermal method and subsequently functionalized chemically using 3-aminopropyletriethoxysilane (APTES). Thin film of APTES functionalized WO_3 NS was deposited on indium tin oxide (ITO) substrate using electrophoretic deposition technique. The fabricated APTES/WO_3 NS/ITO electrode endows with large surface area for high loading of cardiac troponin I antibodies (anti-cTnI), good stability and strong adsorption platform for the biomolecules immobilization. The selective and quantitative detection of cTnI biomarker was accomplished by immobilizing anti-cTnI on APTES/WO_3 NS/ITO platform. The electrochemical sensing using anti-cTnI/APTES/WO_3 NS/ITO immunoelectrode was studied as a label-free immunoassay format using EIS (electrochemical impedance spectroscopy) technique. The performance of the sensor as investigated with the measurement of charge transfer resistance (R_{CT}) for variant concentrations of cTnI antigen demonstrated good sensitivity as 30.8 Ω mL ng^{-1} cm^{-2} for a linear detection range 0.1 to 100 ng mL^{-1} along with good stability up to 6 weeks. The competency of the biosensor was also tested by detecting cTnI antigen in spiked human serum samples.

1. Introduction

In recent times, sensitive and rapid detection of bio-entities has become a prime requisite for the development of biosensors, to be used in point-of-care (POC) testing in laboratories, military, clinics, and environmental applications [1]. Researchers are focussing towards the development of affordable and facile POC devices that can help in detecting the diseases at an early stage. In recent time, cardio vascular diseases (CVDs) have observed to be a major cause of early death in wide regions of the world [2]. And among different CVDs, ischemic heart disease (IHD) and ischemic stroke account for the majority of health issues related to CVD. Acute myocardial infarction or AMI is one such IHD, defined as the necrosis of myocardial cells resulting from the acute obstruction of the coronary heart arteries [3]. Although electrocardiogram (ECG) is in practice in identifying the existence of AMI, yet it is not a reliable test for early detection of AMI. On the other hand, learning the concentration of cardiac biomarkers can play a significant role in detecting AMI [4]. Cardiac troponin I biomarker (cTnI) is a muscle protein (molecular weight 24 KDa) that displays a unique molecular structure. It exhibits traces of amino acid residues on its N-terminal, which do not occur on its skeletal form. Consequently, this makes it a promising analyte for cardiac specificity [5,6]. The cTnI levels increase significantly with the onset of myocardial necrosis. Also, with the onset of AMI, cTnI releases into the blood serum from the death cell of cardiac muscles and remain elevated for 8–10 days. The unique amino sequence of cTnI makes it a potential diagnostic biomarker for early diagnosis of AMI. cTnI levels below 0.4 ng mL^{-1} are considered to be normal, whereas levels greater than 2 ng mL^{-1} can indicate a risk of serious heart-related diseases [7].

Different detection techniques have been employed for the cardiac biomarkers sensing including fluoro-immunoassay [8], ELISA (Enzyme-linked immunosorbent assay) [9], RIA (radioimmunoassay) [10], and chemiluminescence [11]. These assays, however, have limitations such as they are a time-consuming process, have low reliability, suffer from a complicated labelling process, and are labour-intensive with sluggish response times. Hence for rapid and facile diagnosis of AMI, a compact

* Corresponding author.
E-mail address: nitinkumarpuri@dtu.ac.in (N.K. Puri).

https://doi.org/10.1016/j.sbsr.2021.100423
Received 7 January 2021; Received in revised form 14 April 2021; Accepted 19 April 2021
Available online 21 April 2021
2214-1804/© 2021 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

analytical device such as "biosensor" can be considered as an efficient and sensitive platform. Electrochemical methods have been studied extensively as a transducer in biosensors due to their high sensitivity, simplification in processing, low cost, portability, and real-time analysis monitoring [12,13]. Electrochemical biosensors are gaining tremendous attention these days pertaining to the ease of miniaturization, low cost, and fast response which facilitates both qualitative and quantitative determination of biomolecules. The incorporation of advanced and functional nanomaterials in the fabrication of biosensors results in the development of novel devices with excellent sensing characteristics [14,15]. In recent time, two dimensional materials based platform has gained much interest and researchers have successfully demonstrated good sensing performance [15,16]. Among different nanomaterials, 2D metal-oxide nanosheets have emerged as a promising platform towards the development of electrochemical biosensors due to the layered structures which can promote high adsorption into interplanar sites and possess high quantum effects [17]. These materials demonstrate enhanced electron transfer kinetics, large specific surface area, and numerous active sites for the binding of biological species.

Tungsten trioxide (WO_3), n-type transition metal-oxide possesses multifunctional characteristics with photocatalytic, electro and opto-chromism and sensing applications [18–21]. Tungsten in WO_3 is found to exist in different oxidation states and W—O bond has the same bond length, imparting fast and reversible redox reactions at nanostructured WO_3 surface. This opens the door for use of nanostructured WO_3 in electrochemical sensing applications [22]. Also, WO_3 nanostructures with different morphologies have shown tremendous enhancement in the performance of the devices including nanoplates, nanorods, hollow sphere, and many others [22–24]. Interestingly, due to these distinctive characteristics, WO_3 nanostructures have become preferable candidate that exhibits great prospect in biological recognition events and signal transduction mechanism in biosensors. There are innumerable reports on the use of WO_3, in various forms for diverse applications [25–29]. However, utilization of the 2D form of WO_3 for detection of cTnI has not been reported so far, to the best of our knowledge. Henceforth, the employment of WO_3 nanosheets for the fabrication of an electrochemical biosensor for detecting cTnI in a wide linear range has been discussed in this work.

2. Experimental sections

2.1. Chemicals

N-hydroxysuccinimide (NHS) and N-[3-dimethylaminopropyl]-N-ethyl-carbodimidehydrochloride (EDC) were procured from Sigma Aldrich., hydrochloric acid (HCl), citric acid ($C_6H_8O_7$), sodium chloride (NaCl), sodium monophosphate (NaH_2PO_4), and sodium diphosphate-dihydrate ($Na_2HPO_4.2H_2O$), potassium ferrocyanide [$K_4\{Fe(CN)_6\}.3H_2O$], and Potassium ferricyanide [$K_3\{Fe(CN)_6\}$] were acquired through Merck. Sodium tungstate dehydrate ($Na_2WO_4.2H_2O$) and 3-aminopropyltriethoxy saline (APTES) were Sigma Aldrich product. Phosphate Buffer Saline (PBS, 50 mM) solution with a 7.4 pH was prepared with deionized water obtained from the Milli-Q water purification system. Anti-cTnI and human cardiac troponin I antigen were procured from Ray Biotech, Inc. (India).

2.2. Synthesis of WO_3 nanosheets (WO_3 NS)

WO_3 NS were synthesized using a wet chemical method employing $Na_2WO_4.2H_2O$ as a precursor as reported earlier with a modification [28]. Briefly, the preparation of solution was done by dissolving 0.5 g of $Na_2WO_4.2H_2O$ in 30 mL of deionized water (0.05 M). Further, 0.3 g of citric acid was added into the precursor solution and stirred well for 10 min. Thereafter, 6 M HCl was added drop-wise to the solution (for acidification) till the pH of the solution reached 1 and was stirred for further 30 min till the formation of a yellowish precipitate. The resultant precipitate was then poured into a Teflon-lined autoclave and kept in the oven for 20 h at 150 °C for a hydrothermal reaction. The autoclave was left to cool at room temperature after completion of the reaction period. The resultant yellow coloured product was washed thrice with deionized water and ethanol using centrifugation at 10000 rpm and then dried at 60 °C.

2.3. Functionalization and fabrication of APTES/WO_3 NS/ITO electrode

The functionalization of WO_3 NS was accomplished by using amino-silane coupling agent 3-Aminopropyl-triethoxysilane (APTES). This amino-silane provides an active platform for the bio-immobilization onto the surface of nanomaterials [30]. Firstly, a colloidal suspension of WO_3 NS was prepared in 2-propanol solution using ultrasonication. Thereafter, 250 μL of APTES was poured drop-wise to this solution and stirred at 280 rpm for 36 h at room temperature. The suspension was filtered and rinsed with DI water to remove the unattached APTES molecules and the final product APTES functionalized WO_3 NS was used for the fabrication of electrode.

Electrophoretic deposition (EPD) method was employed in the fabrication of APTES functionalized WO_3 NS thin film onto hydrolyzed indium tin oxide (ITO) substrate. For the film deposition, a colloidal suspension of APTES functionalized WO_3 NS (1 mg mL^{-1}) was prepared in acetonitrile and placed in two-electrode cell [31]. Acetonitrile being volatile in nature was used as a dispersion medium to carry out EPD [32]. Platinum wire acted as the auxiliary electrode and a hydrolyzed ITO acted as a working electrode. Both were aligned parallel to each other with 1 cm apart inside the cell containing colloidal suspension. The desired electrode (APTES/WO_3 NS/ITO) was fabricated at an optimized D·C potential of 50 V for 1.5 min. The resultant electrode was then washed with DI water and used after drying.

2.4. Fabrication of anti-cTnI/APTES/WO_3 NS/ITO immunoelectrode

Cardiac troponin I antibody (anti-cTnI) solution in PBS (pH 7.4) of concentration 50 μg mL^{-1} was used in the immobilization process. Before immobilization –COOH groups of antibodies were activated using EDC (0.4 M) and NHS (0.1 M) in 1:1 ratio. Later, 20 μL of this solution was evenly spread onto APTES/WO_3 NS/ITO surface and placed in a humid chamber at room temperature (25 °C) for 6 h followed by washing with PBS solution. The amino terminal ($-NH_2$) of APTES functionalized electrode was bound covalently with the –COOH group terminal of the anti-cTnI forming OC-NH amide bond. Further, to desist active sites, that were non-specific, the anti-cTnI/APTES/WO_3 NS/ITO electrode was treated with ethanolamine (EA, 0.01 M). To achieve this, 20 μL of EA was drop-casted over the electrode surface and kept undisturbed for 3 h and then rinsed with PBS solution. Finally, the resultant fabricated EA/anti-cTnI/APTES/WO_3 NS/ITO electrode was placed in a refrigerator when not in use (at 5 °C). Stock concentration 100 μg mL^{-1} of human cardiac troponin I (cTnI) was prepared in PBS (pH 7.4) that was used to prepare 100 ng mL^{-1}, 50 ng mL^{-1}, 10 ng mL^{-1}, 1 ng mL^{-1}, 0.1 ng mL^{-1} concentrations by further dilutions. The step-by-step fabrication of the electrode is schematically displayed in Fig. 1.

2.5. Characterization techniques

The synthesized WO_3 NS was investigated using X-ray Diffractometer (Bruker), Scanning Electron Microscope (SEM) and Transmission Electron Microscope (TEM), respectively. Chemical configuration of the electrode during functionalization and immobilization of antibody was examined using IR spectroscopy and Raman spectroscopy. The electrochemical characterization was conducted on an Autolab Potentiostat/Galvanostat (Metrohm, Netherland) using a three-electrode cell with the fabricated immuno-electrode acting as the working electrode; silver/silver-chloride (Ag/AgCl) acting as the reference electrode, and a 1 mm thick wire of platinum as the auxiliary electrode in a PBS (50 mM

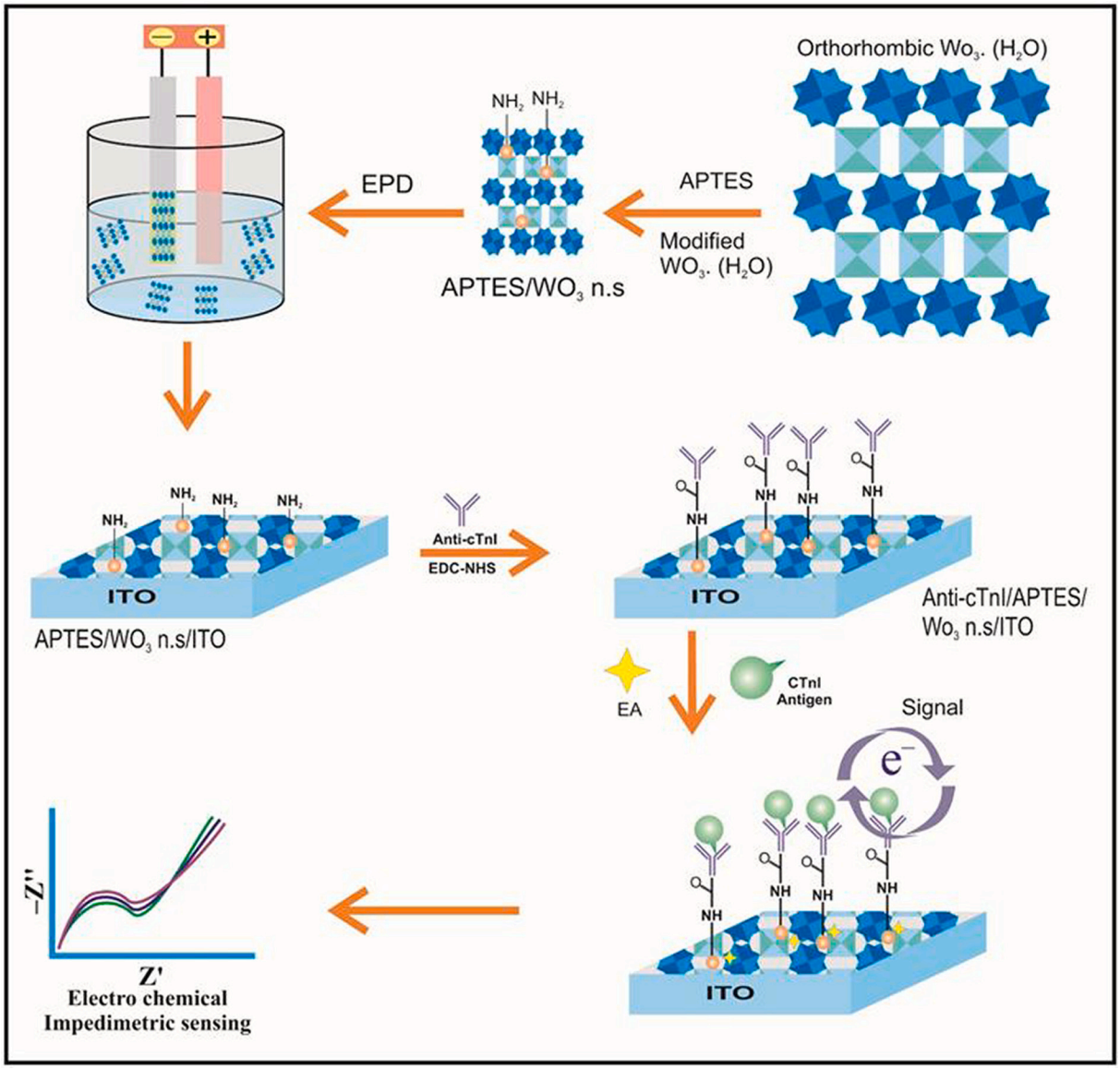


Fig. 1. Schematic of step-wise fabrication of WO_3 NS based immunoelectrode.

and pH 7.4) electrolyte solution containing 5 mM $[Fe(CN)_6]^{3-/4-}$ as redox agent.

3. Results and discussions

3.1. X-ray diffraction

Fig. 2 shows the X-ray diffraction (XRD) pattern obtained for the as-synthesized WO_3 nanostructures. The accentuated characteristic peaks observed at 16.5° and 25.6° correspond to the (020) and (111) reflected planes of $WO_3.H_2O$. The other diffracted peaks observed at 23.7°, 30.4°, 34.1°, 34.9°, 37.6°, 38.8°, 42.6°, 46.4°, 49.6°, 52.6°, 56.1°, 57.2°, 61.6°, and at 62.9° were well indexed to (120), (031), (200), (002), (140), (022), (122), (231), (202) (222), (311), (113), (331), and (260) reflection planes, respectively. The planes were well-matched with JCPDS card no.: 084–0886 indicating purity and crystallinity of the synthesized $WO_3.H_2O$.

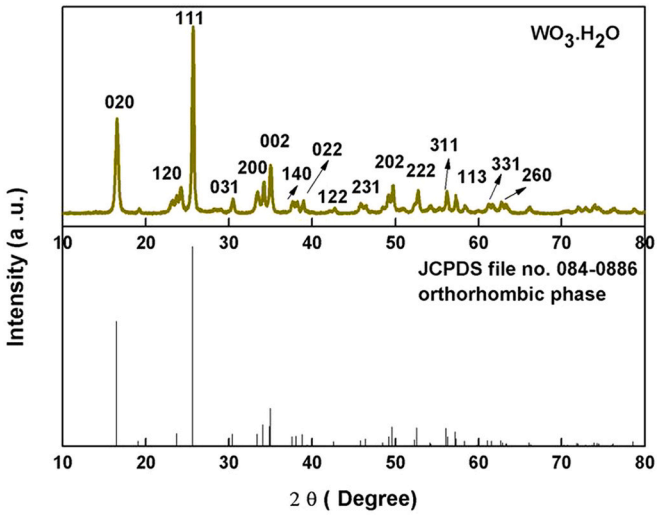


Fig. 2. X-ray diffraction of synthesized WO_3NS.

3.2. Electron microscopy

Fig. 3(a) represents the scanning electron micrograph of synthesized

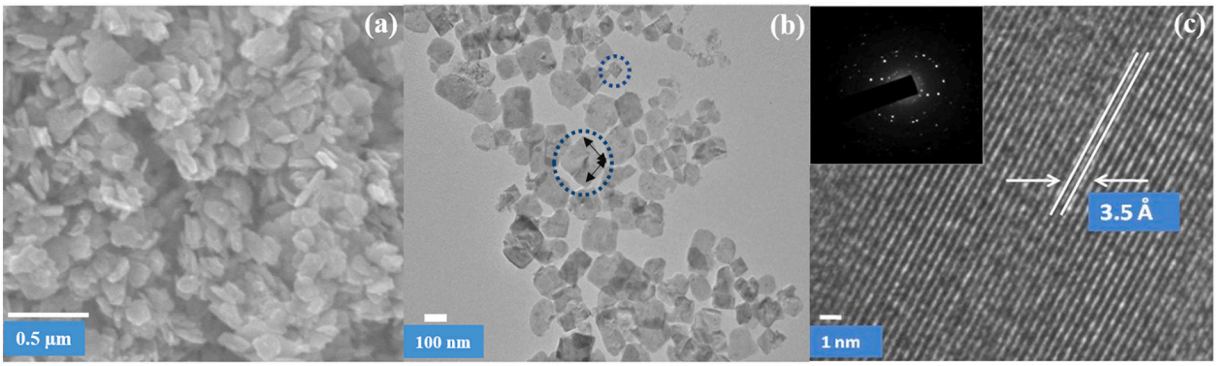


Fig. 3. (a) Scanning electron micrograph (SEM), (b) Transmission electron micrograph (TEM), and (c) High resolution transmission electron micrograph (HR-TEM) of WO_3nanosheet. (Inset SAED).

nanostructured Tungsten trioxide. Formation of nanoplates in the form of stacked nanosheets was clearly visible. Most of the nanoplates were found to be vertically oriented while some of them were horizontally aligned. Successful formation of the Tungsten trioxide nanosheets was confirmed by transmission electron microscopy [Fig. 3(b)]. TEM image revealed the existence of square-shaped nanosheets whose areas varies from 0.004–0.04 μm^2.The nanosheets were found to have lateral dimensions ranging from 90 to 180 nm. The high-resolution transmission electron microscopy (HR-TEM) micrograph [Fig. 3(c)] displays lattice fringes of WO_3 nanosheets with high crystallinity having inter-planar spacing of 3.5 Å [corresponding with (111) plane].

3.3. Spectroscopic investigation

Fig. 4 depicts the FT-IR spectra of the synthesized WO_3 NS before and after functionalization and immobilization process. The spectrum obtained for WO_3 NS [curve a] shows the characteristic band at 688 cm^{-1} corresponding to stretching ν(W—O_{inter}—W) vibration of the corner-sharing WO_6 octahedron in the WO_3 and the peak at 946 cm^{-1} corresponds to the W═O vibration. The bending vibration experienced at 1625 cm^{-1}concurs with the W—OH in-planes vibrations [33,34]. The sharp peak observed at 3425 cm^{-1} corresponds to the stretching mode of ν(O—H) due to the adsorption of water molecules. All these vibration features indicate that synthesized WO_3 NS contains no impurity. The peaks corresponding to the APTES [curve b] observed at 3400 cm^{-1}, and at 2935 cm^{-1} can be attributed to N—H stretching vibration and the presence of C—H bonds of aldehydes groups existent on the surface of APTES molecules. Further, the peaks observed at 1631 cm^{-1}, 1565 cm^{-1}, and at 1130 cm^{-1} correspond to N—H bending vibration of primary amide —NH_2 groups, bending vibration corresponding to secondary amide groups and stretching mode of Si—O—Si respectively [35]. After the immobilization of cTnI antibodies onto APTES functionalized WO_3 NS surface [curve c], the observed characteristic band at about 3460 cm^{-1} indicates the presence of amide A band of protein molecules. The peaks present at 1650 cm^{-1} and 1556 cm^{-1} correspond to amide I band associated with C═O stretching vibration and the amide II band associated with N—H bending vibration and C—N stretching vibration respectively [36]. Hence, FT-IR analysis confirms the successful surface functionalization with the presence of primary amide groups (—NH_2) and immobilization of antibody of cardiac troponin I with amide bond formation.

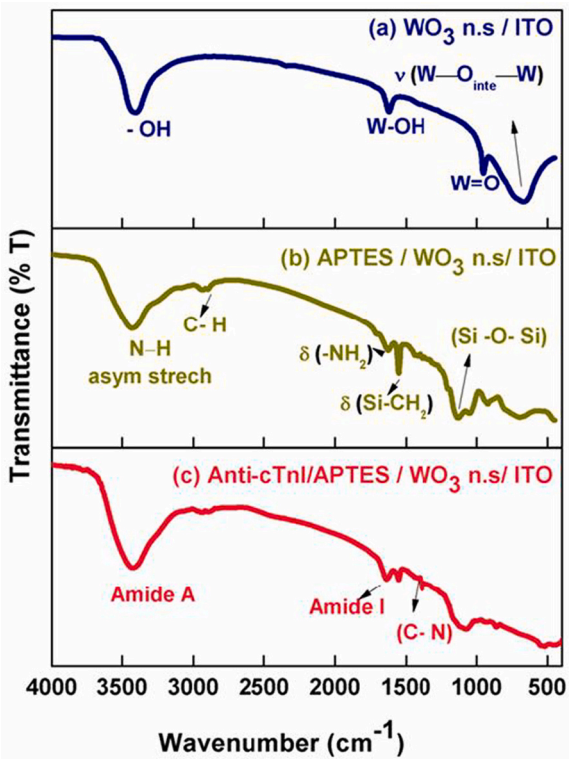


Fig. 4. Fourier transform infrared (FT-IR) spectra of (a) WO_3 NS/ITO, (b) APTES/WO_3 NS/ITO, and (c) Anti-cTnI/APTES/WO_3 NS/ITO.

The modification of WO_3 surface with antibody of cTnI was also investigated using Raman spectroscopy. Fig. 5 represents the Raman spectra of WO_3 NS and anti-cTnI/APTES/WO_3 NS obtained in the range of 100 cm^{-1} to 3500 cm^{-1} The intense peaks observed at 805 cm^{-1} and 715 cm^{-1} in WO_3 NS spectrum represent the stretching vibrations of O-W-O linked with longer and shorter W—O bonds [37]. Further, peaks observed at 265 cm^{-1} and 325 cm^{-1}marked the bending vibrations of O-W-O along with lattice vibration modes observed at 130 cm^{-1}. However, in anti-cTnI/APTES/WO_3NS spectrum, modification in the intensity of the peaks was observed along with new peaks and bands. The bands observed at 1638 cm^{-1} and at 1411 cm^{-1} indicates the presence of amine group NH_2 and CN of APTES molecules respectively whereas band observed at 2915 cm^{-1} attributed to the CH stretch modes of CH_2 groups present in APTES molecules confirming the silanization process onto WO_3 surface [38]. However, the peaks seen in the region of 1630 cm^{-1} and 850 cm^{-1} can be assigned to C═O stretching of carbonyl groups (amide I region) and amide IV respectively of protein antibodies, attributing towards successful immobilization of anti-cTnI onto APTES functionalised WO_3 NS surface.

3.4. Electrochemical investigations

An Autolab Potentiostat was employed for all electrochemical studies which were examined at room temperature. A conventional three-electrode cell with the fabricated immunoelectrode acting as the

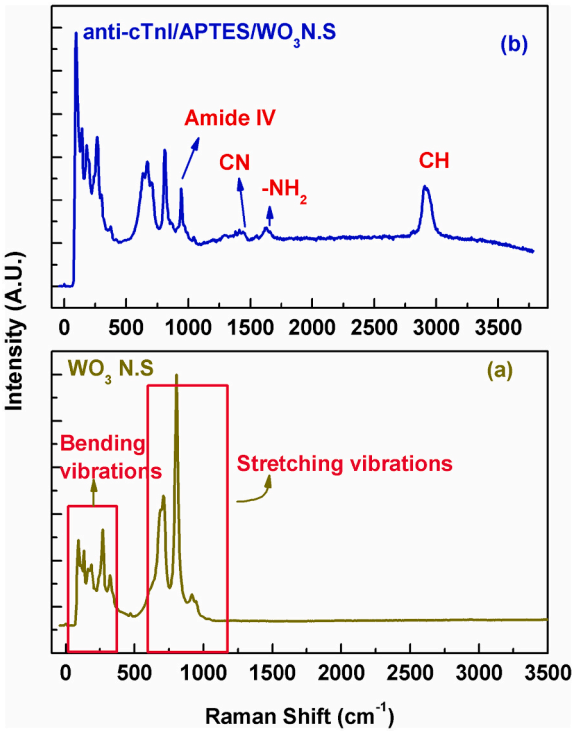


Fig. 5. Raman spectra of (a) WO_3 NS/ITO and (b) Anti-cTnI/APTES/WO_3 NS/ITO.

working electrode; platinum wire acting as a counter electrode, and silver/silver chloride acting as reference electrode was used. Standard phosphate buffer saline (7.4 pH) solution with 5 mM $[Fe(CN)_6]^{3-/4-}$ as redox agent was used as the electrolyte solution.$[Fe(CN)_6]^{3-/4-}$ was preferred as a redox couple due to its sensitive electron transfer kinetics towards the surface modifications.

3.4.1. *Cyclic voltammetry investigations*

Fig. 6(a) depicts the cyclic voltammetry (CV) evaluation of the APTES/WO_3 NS/ITO electrode at different stages of modification. The measurements were investigated in the scale of −0.8 V to +0.8 V at a scan rate 50 mVs^{-1}. The curves represented well-defined redox peaks due to the oxidation and reduction of $[Fe(CN)_6]^{3-/4-}$in the presence of the working electrode. The oxidation peak-current (I_a) and peak-to-peak potential separation (ΔE) of the APTES/WO_3 NS/ITO electrode were found as 0.546 mA and 0.433 V respectively. The ratio of oxidation peak current (I_a) to reduction peak current (I_c) was found to be 1.10 indicating the reversible process of the redox species. However, after the immobilization, the oxidation peak-current declined (0.480 mA) due to the insulating nature of the antibody. The current further declined to 0.411 mA with the application of EA on the antibody functionalized surface. This decrease indicates the fact that EA had blocked most of the un-immobilized active sites present on the immunosensor surface which might had hindered the permeability of redox agent $[Fe(CN)_6]^{3-/4-}$. Fig. 6(b) represents CV of the anti-cTnI/APTES/WO_3 NS/ITO immunoelectrode as a function of the scan rate (10–100 mV s^{-1}). A linear increase of the redox peak current with respect to square root of the scan rate was observed [Fig. 6(c)], indicating a quasi-reversible process and favourable electron transfer kinetics which follows Eqs. (1) and (2) respectively.

$$I_a\,[\mathrm{mA}] = 0.074\ \mathrm{mA} + \left[0.062\ \mathrm{mA}\ \left(\mathrm{s\ mV^{-1}}\right) \times \left(\text{scan rate}\ \left[\mathrm{mV\ s^{-1}}\right]\right)^{1/2}\right], R^2 = 0.999 \quad (1)$$

$$I_c\,[\mathrm{mA}] = -0.106\,\mathrm{mA} + \left[-0.0457\,\mathrm{mA}\ \left(\mathrm{s\,mV^{-1}}\right) \times \left(\text{scan rate}\ \left[\mathrm{mV\,s^{-1}}\right]\right)^{1/2}\right], R^2 = 0.994 \quad (2)$$

3.4.2. *Electrochemical impedance spectroscopy (EIS) investigations*

EIS technique was used to investigate electron transfer kinetics at bio-interface site using a three-electrode system in the frequency range of 0.01 to 10^5 Hz at a biasing potential of 10 mV. Fig. 6(d) depicts the Nyquist plot obtained for (i) APTES/WO_3 NS/ITO; (ii) anti-cTnI/APTES/WO_3 NS/ITO, and (iii) EA/anti-cTnI/ APTES/WO_3 NS/ITO electrodes in PBS (pH 7.4) containing $[Fe(CN)_6]^{4-/3-}$. The impedance spectra (Nyquist plot) consist of a semi-circle portion and a linear portion obtained at high frequencies and low frequencies region respectively. The value of charge transfer resistance (R_{CT}) at the electrode surface was analysed using the diameter of the semicircle in the Nyquist plot, whereas the linear portion signifies the mass transfer limited process.

The APTES/WO_3 NS/ITO electrode exhibited the R_{CT} value of 1.14 kΩ which increased to 1.31 kΩ and 1.79 kΩ respectively for anti-cTnI/APTES/WO_3 NS/ITO and EA/anti-cTnI/APTES/WO_3 NS/ITO electrode. The increased in R_{CT} value signified the insulating nature of antibody biomolecules and obstruction of non-specific active sites due to the application of ethanolamine (EA) onto antibody immobilized surface. All these results confirm the fabrication of the EA/anti-cTnI/APTES/WO_3 NS/ITO electrode.

Fig. 7(a) depicts the impedance spectra (Nyquist plot) of the fabricated immunoelectrode in response to different concentration of cardiac troponin I (cTnI, 0.1 to 100 ng mL^{-1}). All the measurements were carried out in the presence of a PBS solution containing 5 mM $[Fe(CN)_6]^{3-/4-}$ as a redox agent with an incubation time of 10 min before every measurement. An increase in semi-circle's diameter and hence the value of R_{CT} was observed with the rise of cTnI antigen concentration. The rise in R_{CT} value can be attributed to the formation of cTnI and anti-cTnI insulated immuno-complex which obstructs interfacial electron (e^-) transfer at the electrode surface. Fig. 7(b) reveals the linearity change in the value of R_{CT} with respect to change in the concentration of cTnI antigen (0.1 to 100 ng mL^{-1}). The calibration plot was fitted between cTnI concentration and the R_{CT} value obeying Eq. (3)

$$R_{CT}\ (k\Omega) = 2.09\ (k\Omega) + 0.010\ \left(k\Omega\ \mathrm{ng^{-1}\ mL}\right) \times \left[\text{concentration}\ \left(\mathrm{ng\ mL^{-1}}\right)\right] \quad (3)$$

with $R^2 = 0.982$

The sensitivity of the fabricated immunoelectrode (EA/anti-cTnI/APTES/WO_3 NS/ITO) as calculated from the slope of the calibration plot was found to be 30.8 Ω mL ng^{-1} cm^{-2} with regression coefficient as 0.982.

3.5. *Sensing performance studies*

The integrity of the immunoelectrode was verified through selectivity, reproducibility, and stability studies. To investigate the selectivity of the immunoelectrode, the interference study was conducted in the presence of various interfering biomarkers such as Carcinoembryonic (CEA, 20 ng mL^{-1}), C-reactive protein (CRP, 1000 ng mL^{-1}), Myoglobin (Mb, 100 ng mL^{-1}) and cardiac troponin I (cTnI, 50 ng mL^{-1}) as shown in Fig. 8(a). 10 min of incubation was preferred before each measurement. The value of the R_{CT} exhibited by the immunoelectrode in presence of the various interfering biomarkers was noted as 1.79 kΩ (blank), 1.81 kΩ (CEA), 1.88 kΩ (CRP), 1.92 kΩ (Mb), and 2.62 kΩ in presence of cTnI antigen. All these results mark the high specificity of the fabricated immunoelectrode towards cTnI antigen with negligible variation in R_{CT} value in the presence of interfering biomarkers.

The stability of the immunoelectrode was analysed for a period of a

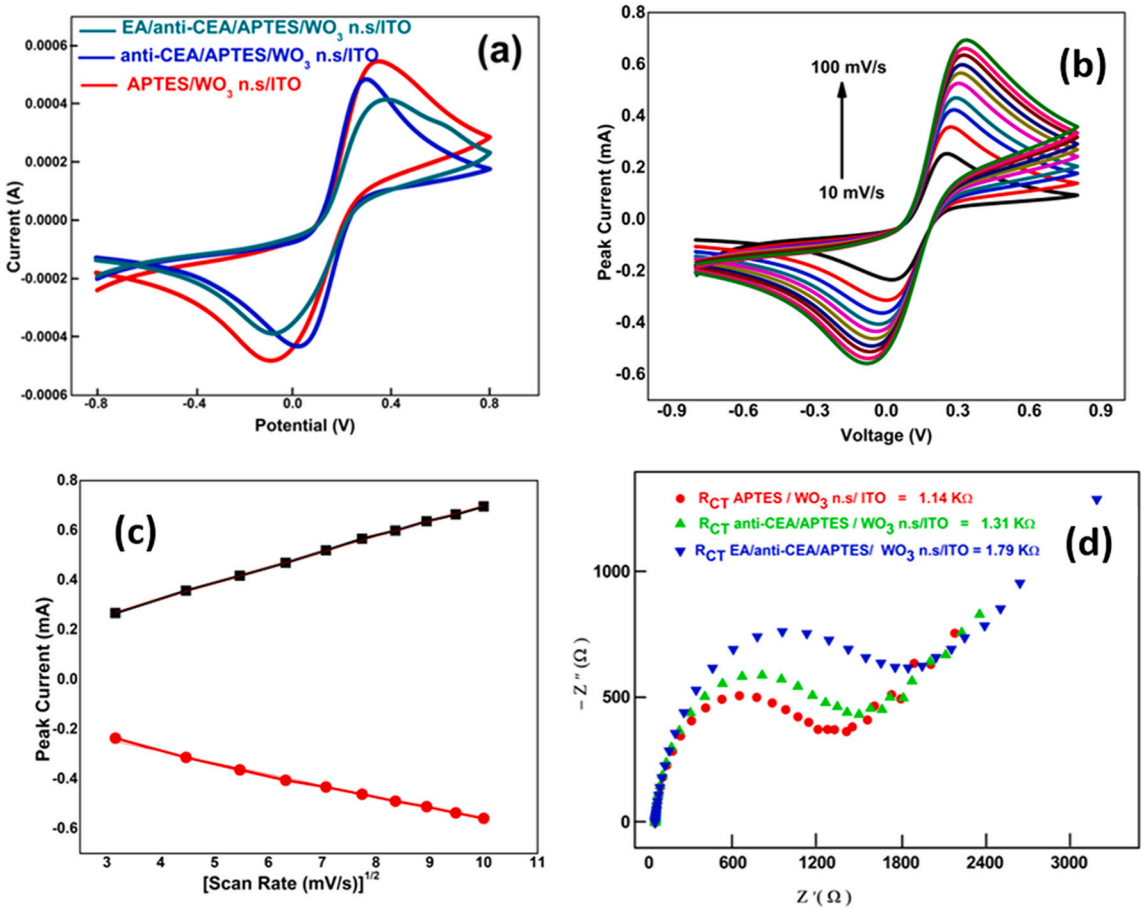


Fig. 6. (a): Cyclic voltammograms (CVs) of the fabricated electrodes at different stages of modification. (b) Cyclic voltammograms of the EA/anti-cTnI/APTES/WO_3 NS/ITO immunoelectrode as a function of the scan rate (10–100 mV s^{-1}). (c) The variation of redox peak current with respect to square root of scan rate. (d) Electrochemical impedance spectra (EIS) of fabricated electrodes obtained in a frequency range 0.01–10^5 Hz operated at a fix biasing potential 0.01 V (Inset: Randles equivalent circuit).

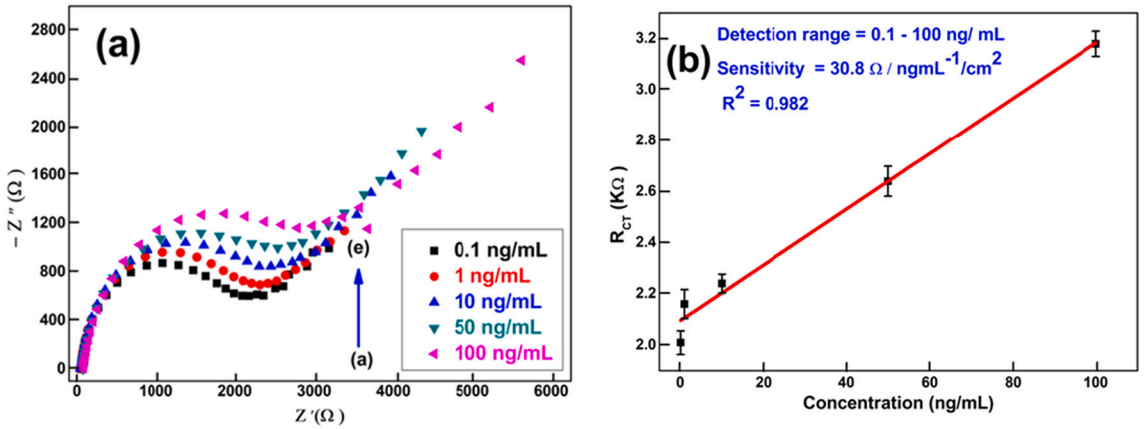


Fig. 7. (a) Electrochemical response studies of the EA/anti-cTnI /APTES/WO_3 NS/ ITO immunoelectrode as a function of cTnI concentration (0.1–100 ng mL^{-1}). And (b) Calibration curve plot of the immunoelectrode with respect to R_{CT} value as a function of cTnI concentration.

total 6 weeks and over regular intervals of one week. The EIS response was investigated in the PBS solution by investigating the change in R_{CT} value with 0.1 ng ml^{-1} concentration of cTnI antigen [Fig. 8(b)]. The immunoelectrode exhibits excellent stability up to 5 weeks with 90% response of R_{CT} value, after which the response decreases to 88% at the end of 6 weeks. The immunoelectrode displayed adequate stability till the end of the duration of analysis. The reproducibility of the immunoelectrode was affirmed by studying EIS measurements on 5 different immunoelectrodes fabricated under similar conditions. The measurements were investigated in the presence of cTnI (0.1 ng mL^{-1}) concentration with each measurement measured 4 times for each electrode and errors bars were included accordingly [Fig. 9(a)]. The low relative standard deviation% value (RSD %, 1.68%) showed good reproducibility of the fabricated immunoelectrode. Further to validate

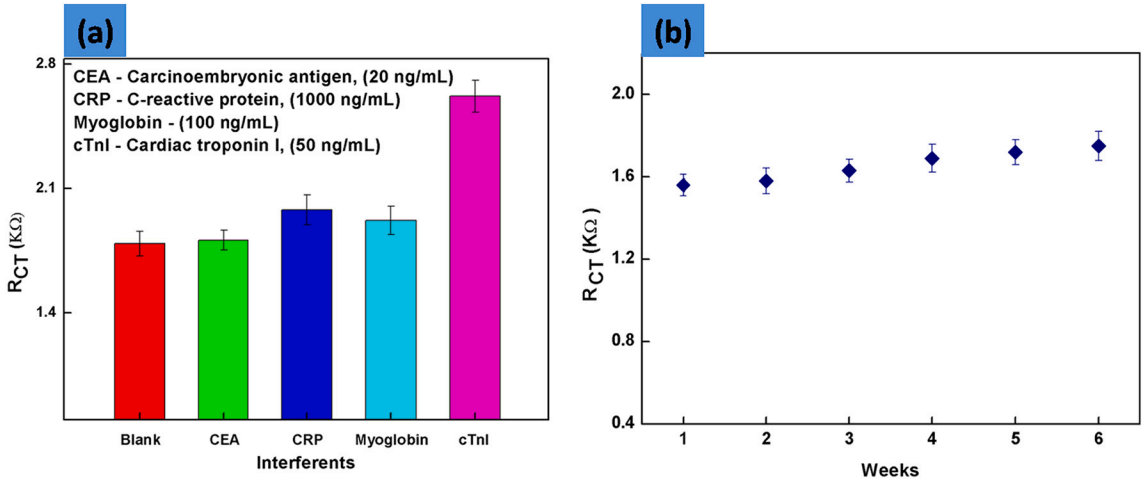


Fig. 8. (a) Selectivity study of the EA/anti-cTnI/APTES/WO_3 NS/ITO immunoelectrode in the presence of different interferents. And (b) Stability study of EA/anti-cTnI /APTES/WO_3 NS/ITO immunoelectrode in the presence of cTnI (0.1 ng mL^{-1}) obtained at a regular interval of 1 week.

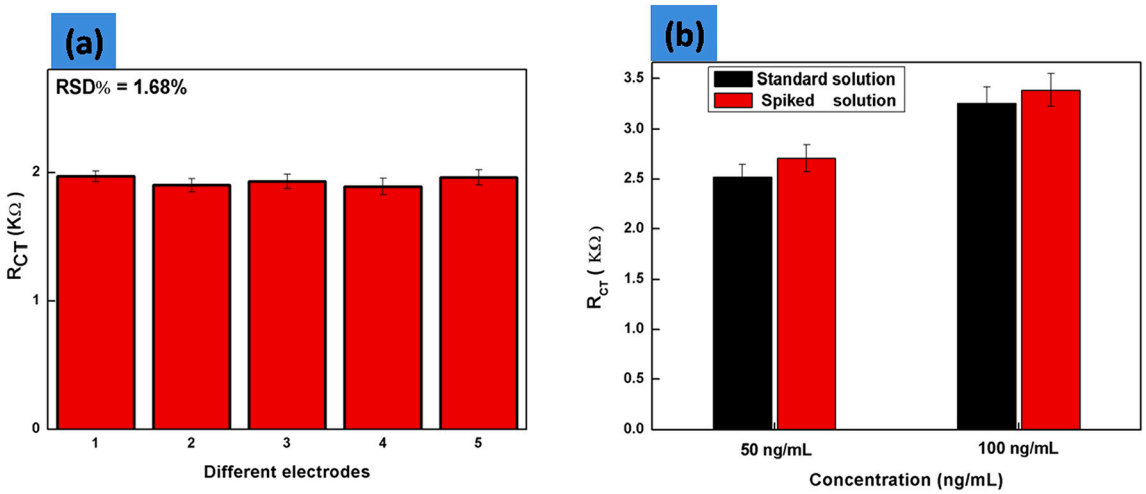


Fig 9. (a) EIS response of five different EA/anti-cTnI /APTES/WO_3 NS/ITO immunoelectrodes fabricated under similar condition in the presence of 0.1 ng mL^{-1}cTnI antigen. (b) Response study of the fabricated immunoelectrode in presence of spiked serum with cTnI antigen (50 and 100 ng mL^{-1}) was compared with standard sample solution.

the performance of the fabricated immunoelectrode in cardiac detection diagnosis, EIS response studies of the immunoelectrode was performed in human serum obtained through pathological laboratory (Dr. Lal path lab, Delhi). The R_{CT} levels of the fabricated immunoelectrode in the presence of spiked serum with cTnI antigen (50 and 100 ng mL^{-1}) were compared with a standard sample solution [Fig. 9(b)]. The results of the electrochemical response studies showed that there was no significant change in the R_{CT} value of the spiked serum sample and the standard solution. Hence, the developed immunoelectrode provides a promising platform for the diagnosis of cTnI biomarker in clinical immunoassays. The recovery (%) of spiked serum samples of cTnI (50 ng $mL^{-1}$and 100 ng mL^{-1}) was evaluated and results were indicated in Table 1. Based on the calibration curve (Eq. (3)), the recovery (%) of these concentrations of cTnI was evaluated. As shown in Table 1, the calculated recoveries varied from 112.92% to 118.74% that demonstrated the feasibility of applying the immunosensor in real serum samples. Thus, all these results validate that the as prepared anti-cTnI/APTES/WO_3 NS/ITO immunoelectrode can be a favourable diagnostic platform for the early detection of AMI. The sensing performance of the proposed sensor has been compared with some of those reported in literature (Table 2). Our sensor has been found superior in terms of wide linear detection range and high storage stability.

The label free detection, high sensitivity (30.8 Ω mL ng^{-1} cm^{-2}) in a wide linear detection range (0.1 to 100 ng mL^{-1}) is attributed to 2 D morphology of WO_3 NS that allows high loading of antibody molecules. The CV studies reveals quasi-reversible process and favourable electron transfer kinetics. This may be attributed to the fact that Tungsten in WO_3 NS is also found to exist in different oxidation states and W—O bond in WO_3 has the same bond length, imparting WO_3 fast and reversible redox reactions at its surface. Further, APTES functionalization of WO_3 NS enabled covalent attachment of antibody molecules on the nanosheets leading to high storage stability up to 6 weeks.

Table 1
Recovery of spiked cardiac troponin I (cTnI) in serum sample.

cTnI added (ng mL^{-1})	R_{ct} (kΩ)	cTnI detected (ng mL^{-1})	Recovery (%)
50	2.68 ± 0.1355	59.37	118.74
100	3.21 ± 0.1595	112.92	112.92

Table 2
Comparison of sensing performance of APTES functionalized WO_3 NS platform with some of the reported literature for cTnI detection.

Sensing platform	Technique for detection	Detection range (ng mL^{-1})	Linear detection range (ng mL^{-1})	Sensitivity	Shelf life	Reference
Conducting PANI coating on Filter Paper	CV	1–1000	1–100	5.5 μA/ng mL^{-1} cm^{-2}	–	[39]
Citrsate capped gold nanoparticles on screen printed electrodes	Capacitance	–	0.2–12.5	–	–	[40]
Carbon nanofiber based nanoelectrode arrays	EIS	0–10	–	–	–	[7]
Gold nanoparticles and graphene oxide on Commercial GC electrodes	SWV	0–6	0.05–3	–	30 days	[41]
APTES/WO_3 NS	EIS	0.1–100	0.1–100	30.8 Ω mL ng^{-1} cm^{-2}	6 Weeks (42 days)	Present Work

4. Conclusions

A facile and a sensitive electrochemical immunoelectrode was fabricated for the diagnosis of AMI by detecting a specific biomarker – cardiac troponin I (cTnI), through a 2D WO_3 NS platform. TEM studies displayed existence of square-shaped WO_3 nanosheets whose areas varies from 0.004–0.04 μm^2. Fourier transform infrared (FT-IR), Raman as well as electrochemical studies, revealed the successful functionalization with APTES and immobilization of anti-cTnI biomolecules onto WO_3 NS matrix. The electrochemical impedance spectroscopy (EIS) studies showed an enhancement in impedance signal of the immunoelectrode in the detection range 0.1 to 100 ng mL^{-1}. The immunoelectrode exhibited good sensitivity as 30.8 Ω mL ng^{-1} cm^{-2} along with high storage stability up to 6 weeks. These enhanced characteristics can be attributed to the high loading of cTnI antibodies on WO_3 NS matrix due to the enhanced active surface area provided by WO_3 NS. The proposed immunoelectrode platform also showed a good result with the spiked serum samples. Hence, owing to rapid electron transfer kinetics; strong adsorption capability; and high surface-to-volume ratio, the proposed platform (WO_3 NS) can be considered as an ideal candidate for the development of a portable and integrated biosensor for the electrochemical detection of cardiac biomarkers as a point-of-care (POC) diagnostic.

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.

Acknowledgements

We thank Vice-chancellor, Delhi Technological University, Delhi, India for providing research facilities. Research is supported by the funds received from SERB, DST, India (EMR/2016/007479) is gratefully acknowledged. S.S. acknowledges financial assistance from the DST, New Delhi, India (Grant No. DST/INSPIRE/04/2014/002540) and Dr. A.P.J. AKTU, Uttar Pradesh, India (Grant under Visvesvaraya Research Promotion Scheme).

References

[1] P.K. Drain, et al., Diagnostic point-of-care tests in resource-limited settings, Lancet Infect. Dis. 14 (3) (2014) 239–249.
[2] Z. Altintas, W.M. Fakanya, I.E. Tothill, Cardiovascular disease detection using biosensing techniques, Talanta 128 (2014) 177–186.
[3] K. Thygesen, et al., Third universal definition of myocardial infarction, Circulation 126 (16) (2012) 2020–2035.
[4] L. Babuin, A.S. Jaffe, Troponin: the biomarker of choice for the detection of cardiac injury, Cmaj 173 (10) (2005) 1191–1202.
[5] K. Thygesen, et al., Recommendations for the use of cardiac troponin measurement in acute cardiac care, Eur. Heart J. 31 (18) (2010) 2197–2204.
[6] X. Han, et al., Recent development of cardiac troponin I detection, Acs Sensors 1 (2) (2016) 106–114.
[7] A. Periyakaruppan, et al., Label-free detection of cardiac troponin-I using carbon nanofiber based nanoelectrode arrays, Anal. Chem. 85 (8) (2013) 3858–3863.
[8] J. Wang, et al., Carboxylated magnetic microbead-assisted fluoroimmunoassay for early biomarkers of acute myocardial infarction, Colloids Surf. B: Biointerfaces 72 (1) (2009) 112–120.
[9] C. Heeschen, et al., Evaluation of a rapid whole blood ELISA for quantification of troponin I in patients with acute chest pain, Clin. Chem. 45 (10) (1999) 1789–1796.
[10] B. Cummins, M.L. Auckland, P. Cummins, Cardiac-specific troponin-l radioimmunoassay in the diagnosis of acute myocardial infarction, Am. Heart J. 113 (6) (1987) 1333–1344.
[11] H. Qi, et al., Ultrasensitive electrogenerated chemiluminescence peptide-based method for the determination of cardiac troponin I incorporating amplification of signal reagent-encapsulated liposomes, Anal. Chem. 85 (8) (2013) 3886–3894.
[12] P. Skládal, Advances in electrochemical immunosensors, Electroanalysis 9 (10) (1997) 737–745.
[13] Y. Wan, et al., Development of electrochemical immunosensors towards point of care diagnostics, Biosens. Bioelectron. 47 (2013) 1–11.
[14] M. Holzinger, A. Le Goff, S. Cosnier, Nanomaterials for biosensing applications: a review, Front. Chem. 2 (2014) 63.
[15] S.A. Lim, M.U. Ahmed, Electrochemical immunosensors and their recent nanomaterial-based signal amplification strategies: a review, RSC Adv. 6 (30) (2016) 24995–25014.
[16] B. Ma, et al., A sensitive electrochemical nonenzymatic biosensor for the detection of H2O2 released from living cells based on ultrathin concave Ag nanosheets, Biosens. Bioelectron. 106 (2018) 29–36.
[17] A. Karimizefreh, et al., Impedimetic biosensor for the DNA of the human papilloma virus based on the use of gold nanosheets, Microchim. Acta 184 (6) (2017) 1729–1737.
[18] X. Lu, et al., Porous nanosheet-based ZnO microspheres for the construction of direct electrochemical biosensors, Biosens. Bioelectron. 24 (1) (2008) 93–98.
[19] S. Ansari, et al., WO3 decorated graphene nanocomposite based electrochemical sensor: a prospect for the detection of anti-anginal drug, Anal. Chim. Acta 1046 (2019) 99–109.
[20] S.H. Lee, et al., Crystalline WO3 nanoparticles for highly improved electrochromic applications, Adv. Mater. 18 (6) (2006) 763–766.
[21] R. Huang, et al., Effect of hydrothermal temperature on structure and photochromic properties of WO3 powder, Adv. Powder Technol. 23 (2) (2012) 211–214.
[22] M. Righettoni, A. Tricoli, S.E. Pratsinis, Si: WO3 sensors for highly selective detection of acetone for easy diagnosis of diabetes by breath analysis, Anal. Chem. 82 (9) (2010) 3581–3587.
[23] P.A. Shinde, et al., Facile synthesis of self-assembled WO3 nanorods for high-performance electrochemical capacitor, J. Alloys Compd. 770 (2019) 1130–1137.
[24] D. Chen, J. Ye, Hierarchical WO3 hollow shells: dendrite, sphere, dumbbell, and their photocatalytic properties, Adv. Funct. Mater. 18 (13) (2008) 1922–1928.
[25] V.K. Tomer, et al., Rapid acetone detection using indium loaded WO3/SnO2 nanohybrid sensor, Sensors Actuators B Chem. 253 (2017) 703–713.
[26] B. Zhang, et al., A novel Z-scheme ZnIn2S4/WO3 photocatalyst based photoelectrochemical immunosensor for the sensitive detection of prostate specific antigen, Sensors Actuators B Chem. 298 (2019) 126835.
[27] R. Ponnusamy, et al., Bifunctional WO3 microrods decorated RGO composite as catechol sensor and optical limiter, Appl. Surf. Sci. 536 (2021) 147669.
[28] F. Li, et al., Enhanced toluene sensing performance of gold-functionalized WO3·H2O nanosheets, Sensors Actuators B Chem. 223 (2016) 761–767.
[29] R. Ponnusamy, et al., Tuning the pure monoclinic phase of WO3 and WO3-Ag nanostructures for non-enzymatic glucose sensing application with theoretical insight from electronic structure simulations, J. Appl. Phys. 123 (2) (2018), 024701.
[30] S.K. Vashist, et al., Immobilization of antibodies and enzymes on 3-aminopropyl-triethoxysilane-functionalized bioanalytical platforms for biosensors and diagnostics, Chem. Rev. 114 (21) (2014) 11083–11130.
[31] M. Ata, Y. Liu, I. Zhitomirsky, A review of new methods of surface chemical modification, dispersion and electrophoretic deposition of metal oxide particles, RSC Adv. 4 (43) (2014) 22716–22732.
[32] A. Hosseinpour, H. Abdizadeh, M. Golobostanfard, Weighing acetonitrile against water as dispersing media for fabrication of graphene oxide films via electrophoretic deposition, Procedia Mater. Sci. 11 (2015) 480–485.

[33] W. Zhao, et al., Multifunctional Fe3O4@ WO3@ mSiO2–APTES nanocarrier for targeted drug delivery and controllable release with microwave irradiation triggered by WO3, Mater. Lett. 169 (2016) 185–188.
[34] M. Tong, et al., WO3 thin film prepared by PECVD technique and its gas sensing properties to NO2, J. Mater. Sci. 36 (10) (2001) 2535–2538.
[35] N. Majoul, S. Aouida, B. Bessaïs, Progress of porous silicon APTES-functionalization by FTIR investigations, Appl. Surf. Sci. 331 (2015) 388–391.
[36] S. Srivastava, et al., Protein conjugated carboxylated gold@ reduced graphene oxide for aflatoxin B 1 detection, RSC Adv. 5 (7) (2015) 5406–5414.
[37] C. Santato, et al., Crystallographically oriented mesoporous WO3 films: synthesis, characterization, and applications, J. Am. Chem. Soc. 123 (43) (2001) 10639–10649.
[38] M. Hiraoui, et al., Spectroscopy studies of functionalized oxidized porous silicon surface for biosensing applications, Mater. Chem. Phys. 128 (1–2) (2011) 151–156.
[39] K.K. Jagadeesan, S. Kumar, G. Sumana, Application of conducting paper for selective detection of troponin, Electrochem. Commun. 20 (2012) 71–74.
[40] V. Bhalla, et al., Gold nanoparticles mediated label-free capacitance detection of cardiac troponin I, Sensors Actuators B Chem. 161 (1) (2012) 761–768.
[41] G. Liu, et al., Nanocomposites of gold nanoparticles and graphene oxide towards an stable label-free electrochemical immunosensor for detection of cardiac marker troponin-I, Anal. Chim. Acta 909 (2016) 1–8.

Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Text Based Hate-Speech Analysis

Janak Sachdeva
Dept. of Information Technology
Delhi Technological University
Delhi, India
ischdeva@gmail.com

Kushank Kumar Chaudhary
Dept. of Information Technology
Delhi Technological University
Delhi, India
chaudhary.kushank@gmail.com

Harshit Madaan
Dept. of Information Technology
Delhi Technological University
Delhi, India
harshitmadaan98@gmail.com

Priyanka Meel
Assistant Professor
Dept. of Information Technology
Delhi Technological University
Delhi, India
priyankameel86@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9396013

***Abstract*- The definition of the term "hate speech" as per Oxford is 'a speech that might involve abusive or threatening words which can have or can express pre-bias against a special community / group. This pre-bias can be anything like religion, race or sexual orientation, caste.' Hatred is generally based on ethnicity, religion, disability, gender, caste, and sexual orientation the internet and social media has become a powerful tool for such propagandist to spread hate and reach new audience. The anonymity and flexibility that the internet offers allow such haters to easily and safely propagate hate without any fear. Lack of regulation and legal policy worsens the situation a bit more. The need of the hour is for automated state of the art and scalable methods for hate speech detection and classification.**

This paper introduces two ensemble-based models in which the first one is based on Linear SVC, Logistic Regression, Random Forest and other based on Random Forest, KNN, Logistic Regression. Also, few deep learning models using self-trained and pre-trained word embeddings have been introduced for Twitter hate speech classification systems. The proposed research work has attempted to classify the tweets & assign them in one of the 3 categories i.e., Racist, Sexist or None.

Keywords - Hate Speech, Natural Language Processing, Machine Learning, LSTM, CNN, Ensemble Models.

I. INTRODUCTION

Hate - a psychological state wherein somebody attempts to mortify an individual or gathering by dismissing or assaulting their confidence or other personality factors. Since the beginning, it is obvious that practically all hate violations are gone before by Hate speech. The destruction against the Tutsi in Rwanda began with hate speech. The holocaust didn't get from the gas chambers, it started some time before with Hate speech. What was found in Myanmar against the Rohingya populace, likewise sprang from hate speech. What's more, today what we are seeing the world over, when we see the manner in which transients and evacuees are being attacked, is a result of the ascent of fanatics all over the place. We need, along these lines, to bend over backward to address this hate speech.

For a very long time, social stages, for example, YouTube, Facebook and Twitter have been battling with the issue identified with hateful posts and remarks. Almost countless cash and an innumerable number of human asset are being contributed every year to handle this issue. In any case, they are neglecting to break it down primarily on the grounds that such measures generally include human mediation and manual audit of the online substance to recognize and to eliminate the hostile materials. This cycle is tedious, work concentrated and not versatile or doable in actuality.

The fundamental motivation behind the experimentation was to compare cutting edge neural network based models with ensemble based models. Thus, we tried to use ensemble-based models one based on Linear SVC, Logistic Regression, Random Forest and other based on Random Forest, KNN, Logistic Regression. Also, few deep learning models using self-trained and pre-trained word embedding have been introduced for Twitter hate speech classification systems.

II. DATA USED

For this project, the Dataset which we are using contains 16135 data points which are a list of various tweets tweeted by random twitter users. Initially, the tweets in the dataset have been classified into three categories i.e. "Racism", "Sexism" & "None". The tweets corresponding

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

to the "None" class are not offensive. The dataset is not skewed i.e. All the classes are balanced in terms of data points. There are various other information in the dataset, like location, time of the tweet, Hashtags used, the follow account of the person who has tweeted & obviously the textual data of the tweet itself. The main point of focus in this research remains on the textual data of the tweet. It has been published on github as well and has been used for the research prior to this. [6]

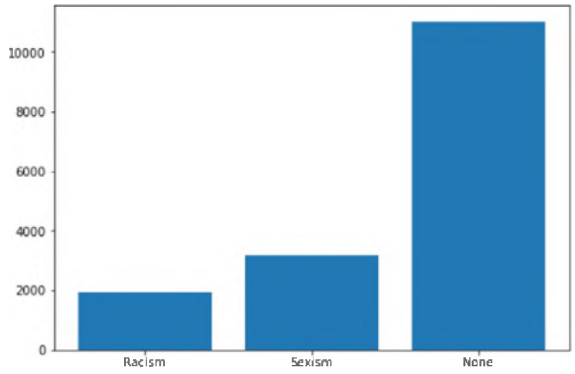


Fig 1. Class-wise Tweet Distribution

III. DATA PRE-PROCESSING

The dataset contained around 16135 tweets in which each tweet had no. of words in the average range of 16-24 words each. In order to have an efficient model, we tried to preprocess our data removing some words with negligible effect on the output or having certain amount of redundancy. Data cleaning was done. The data was going to be classified in three categories namely Racist, Sexist, None. We applied the following techniques to preprocess as well as to clean the given dataset:

- ***Tokenization***: This is done to split the text into tokens.
- ***Lowercasing***: All the words were lowercased.
- ***Stemming***: It is the process which is done as a part of linguistic normalisation, reducing derived words to their stem, base or root form.
- ***StopWords, Digits & Punctuation Removal***: There are few words in English language like 'the', 'his', 'her' etc. which are just used for grammatical purpose and do not add any information to the text for natural language processing tasks. Similar is the case with digits and Punctuations, therefore they are removed from the text.

Fig 2. Word Cloud Representation of tokens

IV. RESEARCH METHODOLOGY

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

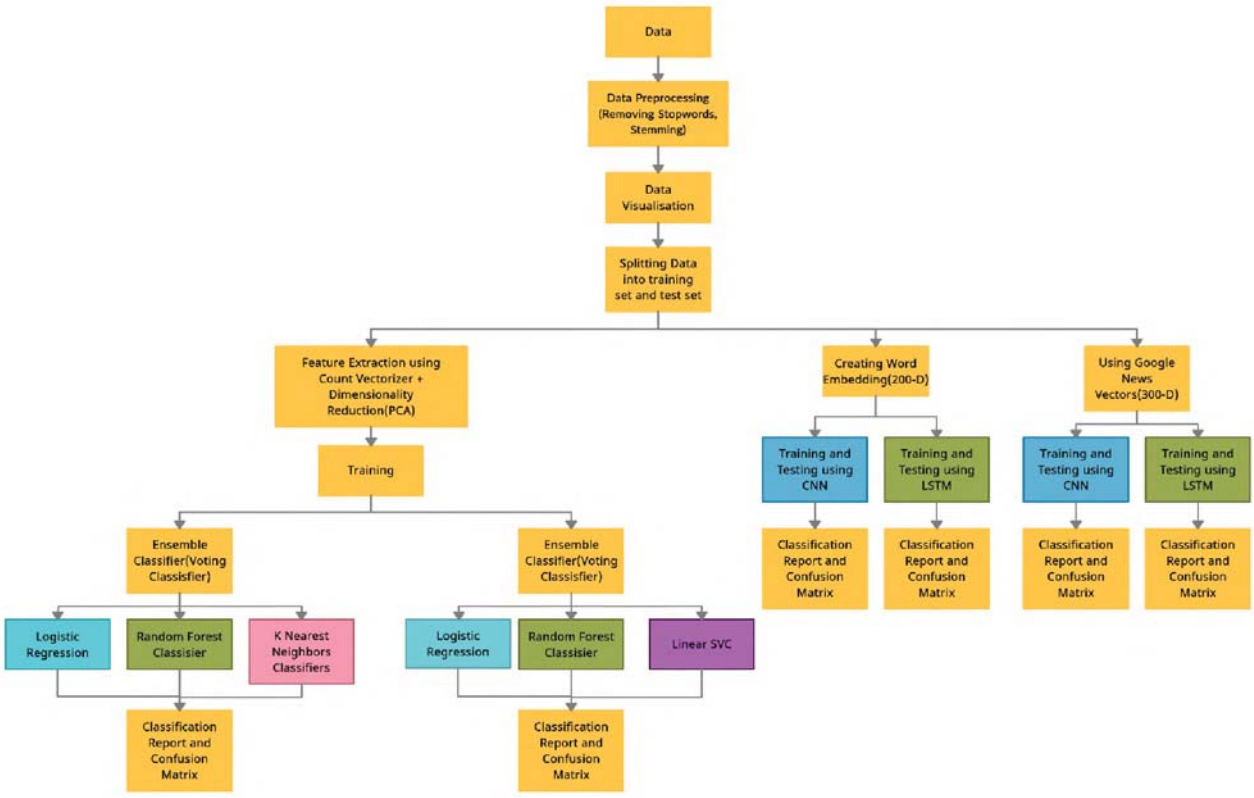


Fig 3. Flowchart Describing various Techniques Used

Various methodologies have been used and proposed to predict whether the tweet tweeted by a user contains hate or not. It has been a challenging issue to detect whether the tweet is hateful or not. Various social media platforms remove the posts which contain hate speech so that the fundamental rights of human beings are not violated.

In this paper we have proposed few methods to detect if given tweet contains hate speech or not, and if hate exists in the tweet then the tweet is further classified as "Racism based hate" or "Sexism based hate".

A. Ensemble Based Methods

Ensemble Learning is a method which is applied to solve computational intelligence problems. In this method more than one Machine Learning models are trained and are then used to predict the outcome. Each model predicts its own outcome and then the results of each of these models is combined together to get better results leading to a better prediction. Ensemble based methods generally tend to outperform single standalone Machine Learning based models because they are based on the collective decision of the models rather than on just the decision of one particular model.

A Voting Classifier has been used which is present in the sklearn. Ensemble module of the sklearn library. There are two types of voting schemes for Voting Classifier: 'hard' and 'soft'. In 'Hard' voting scheme Classifier uses class labels predicted by involved models for majority rule voting. In 'Soft' voting scheme the classifier predicts the class label based on the sums of the predicted probabilities for each class by involved models and the class with the highest probability sum is predicted.

In both the two Ensemble based models we have used a 'Hard' voting scheme. Equal weights have been assigned to each of the sub models i.e. each sub model has an equal say in prediction.

Feature Selection

After pre-processing the data bag of words approach has been used to select features from the tweets. Top 5000 most frequently occurring words in the tweet corpus have been selected as the columns using the CountVectorizer matrix which is present in the nltk library. The rows correspond to the tweet number in the matrix and the columns correspond to a token. A non-zero value is present in the cell (i, j) if the 'i'th tweet contains the 'j'th token. Since the matrix produced is a sparse matrix, therefore in order to reduce 5000 features to lesser number of features PCA has been used to reduce the number of features to 500. Principal Component Analysis (PCA) is a dimensionality reduction technique used to reduce the number of features and involves the use of Eigen values & Eigen vectors. The features extracted through PCA have been fed to the below mentioned ensemble models.

a) RFC, LR, KNN Based Ensemble Classifier

In this model RFC i.e. Random Forest Classifier, LR i.e. Logistic Regression, and k Nearest Neighbors have been applied and ensemble-based prediction based on the likelihood by the above said three models is made.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

i) Logistic Regression for hate speech detection

Logistic Regression is a very basic statistical model and fundamental classification method used in the domain of supervised machine learning for classification. Logistic Regression isn't very complicated and is fast. It can be applied to both two class classification as well as multiclass classification.

For a two-class classification it makes use of the sigmoid function which outputs a value between 0-1 and based upon a threshold value (usually 0.5) it predicts 1st class or 2nd class. For values less than equal to 0.5 it predicts 1st class else the prediction made is 2nd class.

The sigmoid function is given below:

$$h = \frac{1}{1 + e^{-z}}$$

$Z = \Theta X$

Where Z is the dot product of randomly initialized weight values denoted as Θ and input features denoted by X. In our case we have more than one feature (i.e. 500 for each tweet) so the formula is presented as:

$$h = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \cdots$$

Where Θ_0 denotes the bias term and X_1, X_2, X_3 ...X_{500} denote input features and. The target of this algorithm is to update these Θ values after each iteration so that our loss converges to a minima. The loss function without regularisation term is given as:

$$J = -\frac{1}{m}[\Sigma y^i log h^i + (1-y)\log(1-h^i)]$$

Here

'y' denotes the actual output label

'h' denotes the hypothesis or the predicted label

'm' denotes the no. of training data points

Here in our case we have employed a multinomial logistic regression approach with L2 regularization to prevent overfitting. We obtained an F1 score is the harmonic mean of Recall and Precision to be 0.843 when logistic regression was used as a standalone model for the three-class prediction of tweets in our case.

ii) Random Forest Classifier for hate speech detection

In random forest classifier various decision tree classifiers are constructed on various sub parts of the dataset and their average is used to control overfitting and to improve accuracy.

In a decision tree we start from a root node and the node is split into various other sub-nodes based on some feature. A decision tree can be described as a flowchart-like tree structure, where each test on an attribute is represented as an internal node, the outcomes of the test are denoted by branches, and the terminal node i.e. the leaf node represents the class label [2]. Given an input X, the attribute values of the input are checked in the decision tree classifier. A route is found from the root to a terminal node which states the class predicted for the input.[2] The no. of decision trees to be built for random forest have been kept as 10. The criteria used for deciding the feature to be split upon at a particular node can be Gini Index, Information Gain, and Gain Ratio. We have used Gini Index during our experimentation whose formula has been described below.

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Where p_i stands for the probability of an object belonging to a particular label.

While the decision tree construction, Gini Index is calculated for each feature and the split is performed upon that feature which has the least Gini Index. The max_features i.e. maximum number of attributes Random Forest is allowed to try in a single tree, was kept to be square root of (n_features) and n_features=500 in our case.

The **F1 score** obtained as a result of applying standalone random forest on the test data was **0.845**.

iii) K Nearest Neighbors for hate speech detection

K Nearest Neighbour is a classification technique that is based on the phenomenon that the things which are alike exist in close proximity meaning similar things are near to each other. KNN uses the phenomenon of proximity also known as similarity, distance or closeness. The proximity between two points X and Y in vector space is governed by the given formula called as Minkowski distance:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Here the i^{th} feature of point X is denoted by x_i

the i^{th} feature of point Y is denoted by y_i

the number of features is denoted by n

The value of p is specific to the type of distance we want to use. For Eg. When p=1 the proximity measure is called Manhattan distance and when p=2 it's called Euclidean distance.

There are various other similarity measures as well like cosine similarity, Jaccard Distance etc.

For classification of a new input, we calculate the K closest neighbours & the majority class among the neighbouring data decides the classification for that tuple[3]. Although the classifier is not much complex, in the classification task the value of 'K' plays a crucial role[3]. In our case

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

we have kept the value of K=5 and have employed Euclidean distance as proximity measure.

The **F1 score** obtained on standalone use of **KNN was 0.703**.

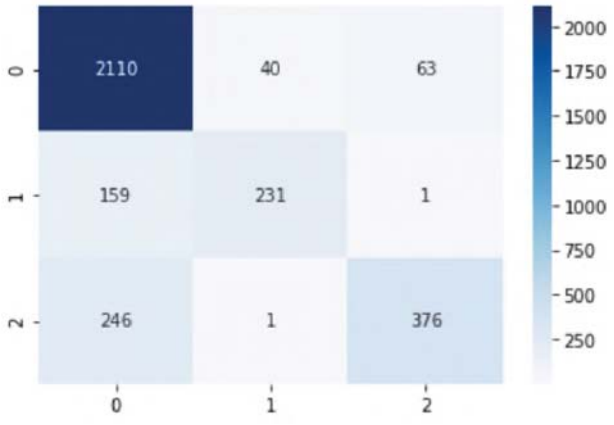


Fig 4. Confusion Matrix for Ensemble model which uses Logistic Regression, Random Forest & KNN.

b) LR, RFC, LSVC Based Ensemble Classifier

This model is composed of a similar Ensemble model except that the models used here are Random Forest (RFC), Logistic Regression (LR) and Linear Support Vector Classifier.

i) **LinearSVC-** It is a classification algorithm implemented in scikit learn library. It is very similar to support vector classifier with parameter kernel set as 'linear', but it is executed in terms of 'liblinear' rather than 'libsvc'. Linear SVC has greater adaptability in the selection of loss functions and penalties and it scales better to an enormous quantity of dataset. Generally, it is also faster than SVC for a large amount of data. The F1 score obtained when only Linear SVC was used was 0.832.

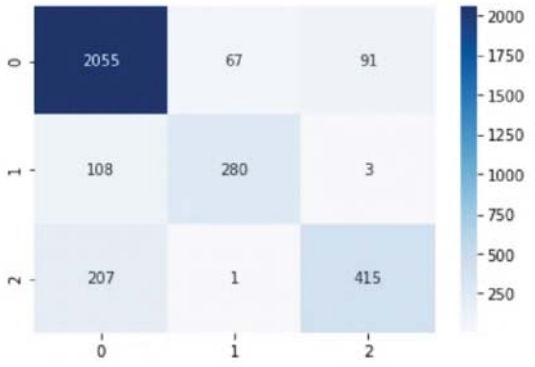


Fig 5. Confusion Matrix for Ensemble model which uses Logistic Regression, Random Forest & Linear SVC

B. Deep Learning & Word Embeddings Based Approach

a) Word Embeddings: There have been many classical approaches in NLP like Naive Bayes classifier using bag of words approach for supervised learning. Many NLP techniques at present treat words as single standalone units - there is no concept of similarity between them [4]. However, such approaches do not consider the context in which a word appears along with other words. In 2013 Mikolov [4](Google, Mountain View, CA) proposed techniques by which each word could be represented in the form of a n-dimensional vector based on the situation in which it appears along with other words. This enables us to generate word vectors such that similar words have similar word embeddings.

b) We have used two types of word embeddings for both CNN and LSTM based approach. In one case we have utilized 300-d pre trained word embeddings that were trained on Google News corpus (3 billion running words in English). In the other case we have trained 200-d word embeddings using Word2Vec approach, keeping window size to be 3 (i.e. no of words before and after a particular word to be considered as a context for generating embedding of a word). The window size can be varied depending upon the average length of sentences. A skip-gram approach, in simple words predicting the surrounding words given the current word, was utilized in our experimentation. However a Continuous Bag of Words approach, in simple words predicting the current word based on the context words, can also be used.

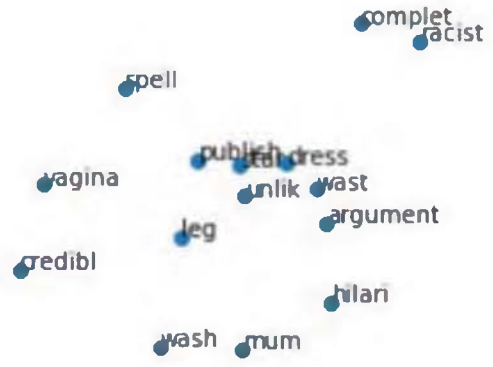


Fig 6. A 2d plot of some words after their embeddings have been generated after applying PCA.

c) CNN for Hate Speech Detection: Although CNN's have a great utility in the field of computer vision, they have subsequently been effective for NLP tasks as well. [5] Yoon Kim has presented CNN based approach for sentence classification. A similar kind of approach has been employed to classify tweets into 'Racism', 'Sexism' or 'None' by us. The model is fed with an embedding matrix consisting of word embeddings corresponding to every unique word found in the tweets. Since the maximum number of tokens in any tweet after text cleaning were

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

24, therefore Max Sequence Length is kept to be 25 and post padding has been applied meaning if a tweet is of shorter length than max sequence length then zeros would be padded corresponding to the remaining tokens.

Various 1d convolutional layers with 300 filters and kernel size=2,3,4,5,6 followed by max pooling were concatenated together having activation function as ReLU. Further few dense layers were added and dropout layers were also added to prevent overfitting. Finally, the model was compiled with binary cross-entropy loss and adam optimizer.

$$R(z) = \begin{cases} z & z > 0 \\ 0 & z <= 0 \end{cases}$$

ReLU function

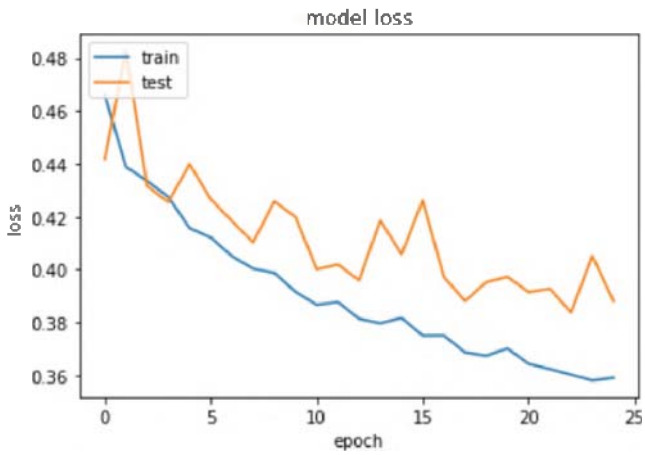


Fig 7. Graphical depiction of Loss vs Epoch for CNN model on self-trained word embeddings

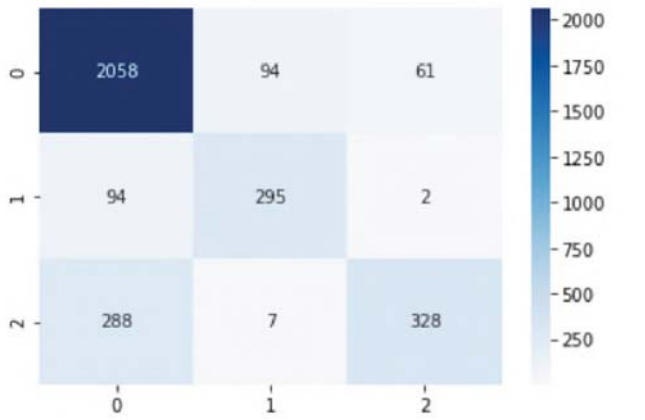


Fig 8. Confusion Matrix for CNN model using pretrained Google news embeddings

d) LSTM for Hate Speech Detection: LSTM stands for **L**ong **S**hort **T**erm **M**emory. The structure of an LSTM network is similar to RNN i.e. Recurrent Neural Network, where every cell does more operation. Long term dependencies in text are remembered by the LSTM network by enhancing operations in each cell. The basic difference between RNN and LSTM is that RNNs remember everything upto a limited depth, whereas LSTMs focus on what to forget and what to remember. Each LSTM cell has 3 gates:

i. **Forget Gate:** It decides whether to delete an info. if it is not worth remembering.
ii. **Input Gate:** The responsibility of input gate is to add a given piece of info. to the cell state.
iii. **Output Gate:** At a given step, it decides whether the given info. should be used & is relevant or not.

Every gate is a layer having a weight associated with it. At every iteration, it takes an input value, and a sigmoid function is applied that returns a value in the range 0-1, where 1 stands for letting everything through and 0 to let nothing through. Afterward, back-propagation mechanism is used to update each layer's value. After some time, this allows the gates to learn which info. is relevant and which is useless.

Similar procedure was followed for passing parameters to the embedding layer as described in the preceding section of CNN.

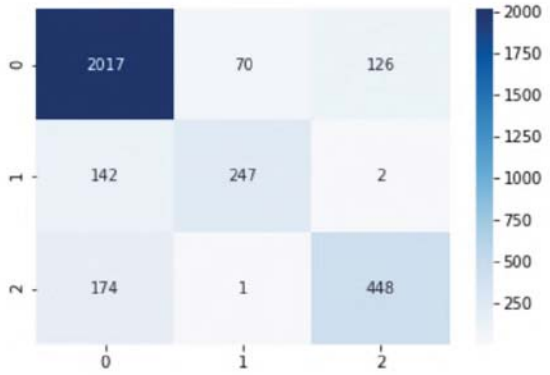


Fig 9. Confusion Matrix for LSTM model using self-trained word embeddings

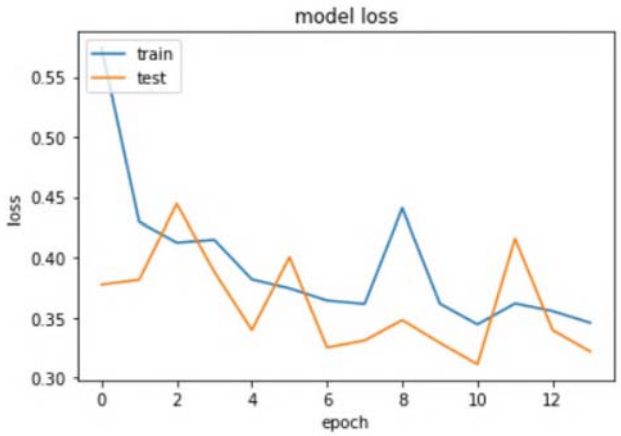


Fig 10. Graphical depiction of Loss vs Epoch for LSTM model on pre-trained Google news embeddings

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

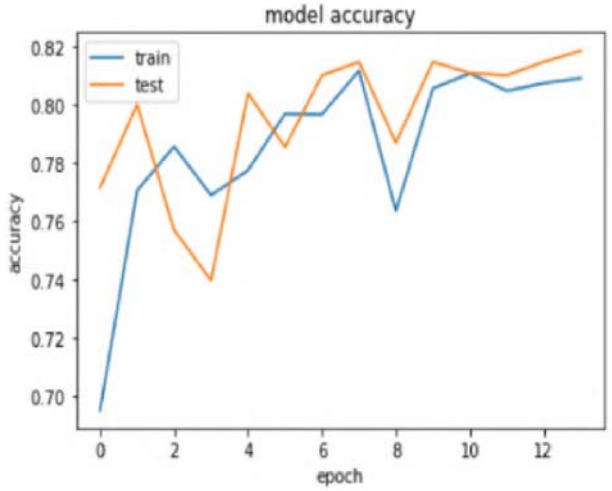


Fig 11. Graphical depiction of Accuracy vs Epoch for LSTM model on pre-trained Google news embeddings

Table 1. shows the results obtained by applying various classification algorithms and techniques for hate speech detection task.

S.No	Method	Accuracy	Precision	Recall	F1
1	Logistic Regression	0.848	0.845	0.848	0.843
2	Random Forest Classifier	0.848	0.845	0.848	0.845
3	K Nearest Neighbours	0.755	0.782	0.755	0.703
4	Linear Support Vector Classifier (LinearSVC)	0.834	0.831	0.834	0.832
5	Ensemble Classifier (using LR, KNN, RFC)	0.842	0.843	0.842	0.833
6	Ensemble Classifier (using LR, Linear SVC, RFC)	0.852	0.850	0.852	0.848
7	LSTM + Self-trained embeddings	0.840	0.837	0.840	0.837
8	CNN + Self-trained embeddings	0.771	0.779	0.771	0.742
9	LSTM + Google News Vectors(300-D)	0.814	0.814	0.814	0.808
10	CNN + Google News Vectors(300-D)	0.831	0.831	0.831	0.822

V. RESULTS & CONCLUSION

From the table, we can see that the Ensemble Classifier model build using RFC (Random Forest Classifier), LR (Logistic Regression), and Linear SVC (Linear Support Vector Classifier) provides a comparatively better result than the other models. On further observation, we can see that using self-trained embeddings provides a better result in the case of LSTM while for CNN we get better result on using pre-trained embeddings model (Google News Vectors 300-D). For other models, the results obtained using Logistic Regression and Random Forest Classifier are very similar for all the parameters, while the results of Linear SVC are slightly poorer than them. As expected, the K Nearest Neighbour Model performed the worst. We believe that the accuracy of the experiments performed could be improved by trying few more different approaches like using different embeddings approach i.e. BERT, FastText, ELMo, and Universal Sentence Encoder.

VI. FUTURE SCOPE

As hate speech keeps on being a cultural issue and sentiment analysis is turning continuously into a business just as a social necessity, the requirement for automated hate speech recognition, abusive content censorship and sentiment analysis system becomes more and more clear. So, in this paper we tried to introduce some accurate methodologies to address the problem of hate speech in a better way and also tried to present a new system that has reasonable accuracy. Anyways there are a couple of difficulties and future extension that were confronted that should be addressed:

- Hate Speech analysis for audio-video dataset.
- Hate speech analysis on images.
- Extending the research on other languages except English.
- Foreseeing utilizing a bigger number of features than simply the words present in the content.
- Classifying hateful content into more precise sub classes.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

- Another major challenge is that people try to get away with hate speech detection by intentionally misspelling the hateful words.
- Some users try to spread hateful messages by posting the content as images and not as text which causes a challenge to detect hateful content.
- Ultimately, there is always an ongoing trade-off between those who try to spread hate and those who try to detect & prevent it.

VII. REFERENCES

[1] Meel, Priyanka & Agrawal, Harsh & Agrawal, Mansi & Goyal, Archit. (2020). Analysing Tweets for Text and Image Features to Detect Fake News Using Ensemble Learning. 10.1007/978-981-15-0633-8_46.

[2] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

[3] Taunk, Kashvi & De, Sanjukta & Verma, Srishti & Swetapadma, Aleena. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 1255-1260. 10.1109/ICCS45141.2019.9065747.

[4] Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

[5] Kim, Yoon. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 10.3115/v1/D14-1181.

[6] Mansi Dhawan and Dr. M.L. Sharma, "HATE SPEECH DETECTION AND SENTIMENT ANALYSIS", International Research Journal of Engineering and Technology (IRJET) (2020)

[7] Roohani, & Rana, Tushar & Meel, Priyanka. (2019). Rumor Propagation: A State-of-the-art Survey of Current Challenges and Opportunities. 64-69. 10.1109/ICCT46177.2019.8969023.

[8] Waseem, Zeerak & Hovy, Dirk. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. 88-93. 10.18653/v1/N16-2013.

[9] Gambäck, Björn & Sikdar, Utpal. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. 85-90. 10.18653/v1/W17-3013.

[10] Badjatiya, Pinkesh & Gupta, Shashank & Gupta, Manish & Varma, Vasudeva. (2017). Deep Learning for Hate Speech Detection in Tweets. 10.1145/3041021.3054223.

[11] Zimmerman, Steven & Fox, Chris & Kruschwitz, Udo. (2018). Improving Hate Speech Detection with Deep Learning Ensembles.

[12] Fauzi, Muhammad & Yuniarti, Anny. (2018). Ensemble Method for Indonesian Twitter Hate Speech Detection. Indonesian Journal of Electrical Engineering and Computer Science. 11. 294-299. 10.11591/ijeecs.v11.i1.pp294-299.

[13] Meel, Priyanka & Vishwakarma, Dinesh. (2019). Fake News, Rumor, Information Pollution in Social Media and Web: A Contemporary Survey of State-of-the-arts, Challenges and Opportunities. Expert Systems with Applications. 153. 112986. 10.1016/j.eswa.2019.112986.

[14] Badjatiya, Pinkesh & Gupta, Shashank & Gupta, Manish & Varma, Vasudeva. (2017). Deep Learning for Hate Speech Detection in Tweets. 10.1145/3041021.3054223

[15] Zhang, Ziqi & Luo, Lei. (2018). Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. Semantic Web. Accepted. 10.3233/SW-180338.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 10:28:18 UTC from IEEE Xplore. Restrictions apply.

Journal of Contemporary Issues in Business and Government Vol. 27, No. 3, 2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.128

The Evolution, progression and Impact of ISO 9000 in the manufacturing organizations around the world and India: An Analysis

DEEPA PRIYANSHU[1], RUBINA LIYAKAT KHAN[2], MAHERA MAQDADI[3], MONISHA GUPTA VASHISHTA[4]

[1]Administrative Sciences Department, Community College, Imam Abdulrahman Bin Faisal University, Dammam 32422, Saudi Arabia
[2]Computer Science Department, Community College, Imam Abdulrahman Bin Faisal University, Dammam 32422, Saudi Arabia
[3]Financial Science Department, Community College, Imam Abdulrahman Bin Faisal University, Dammam 32422, Saudi Arabia
[4]University School of Management & Entrepreneurship, Delhi Technological University, New Delhi 110042, India
*Corresponding Author
Email: dpriyanshu@iau.edu.sa

Abstract: In order to survive in the combative market, the companies have to develop a kind of mechanism in order to be in the preferred positions from their competitors. They should identify mechanism to make sure about their superiority from their competitors. Superiority from others can be achieved through many factors. Quality Certification and Assurance has become an unprecedented way to differentiate from the lots. Quality Certification has also emerged as a demand of Customers. The meaning of Quality has perplexity but no one can deny its importance in creation of differentiation status for the companies. Quality has been undisputed factor of preference for customer satisfaction. This paper is a deliberate attempt to understand the evolution, reasons and benefits derived by implementing ISO 9000 in their organization.

Keywords: ISO 9000, Quality Development, Customer Satisfaction, Quality Management System

INTRODUCTION

Background of Quality, QMS and ISO 9000:2015

Technological innovations adoptions according to time have broken the country boundaries and it has resulted in more learned customers with higher expectations from the organization. Peer pressure, Business environment, and increased expectation of customers led the companies to improve efficiency while decreasing the cost of production and maintaining the Quality.

Quality is a very vulnerable concept and changed over a period of time, however the objective in all times has been same to conform to the standards of customer delight. The maintenance of quality in the functional area has been an uphill task since the management functions were getting more complex.

The International Organization of Standardization with American National Standard Institute (ANSI) worked as a technical committee (TC 176) to build up conventional overall norm to generic application in the year 1979.

The primarily focus of ISO 9000 has always been to promote the customer relations. For the effective Implementing of the ISO 9000 in an organization in all the functional department have to work together in integration.

The primary version in the standard of ISO 9000 came into light in the year 1987, and then further there were revisions and Updation continued in the year 1994, 2000, 2008 and 2015. Since the beginning of year 2004, all organizations had to follow ISO 9001:2000 for the goal of certification.

Presently, the Standard of ISO 9000 is supported by the national standards bodies from more than 165 countries. This support makes ISO 9000, a sensible choice for any organization that does international business or serves the customers who demand an International standard of quality.

American Society for Quality (ASQ) has characterized "ISO 9000 as a bunch of worldwide principles on quality administration and quality confirmation created to assist organizations with recording the quality framework components expected to keep a productive framework of Quality. They are not explicit to any one industry and can be applied to associations of any size."

Copyright © The Author(S) 2021. Published By Society Of Business And Management. This Is An Open Access Article Distributed Under The CCBY License. (Http://Creativecommons.Org/Licenses/By/4.0/)

By the research conducted by the author on the literature of ISO 9000 series and in samples of 100 papers, it came out that 70% of the samples are surveys, 17.5% were financial ratio analysis ,7% were the case studies and only 5.5% were either face to face interaction documented along with review of literature or Statistical data analysis.

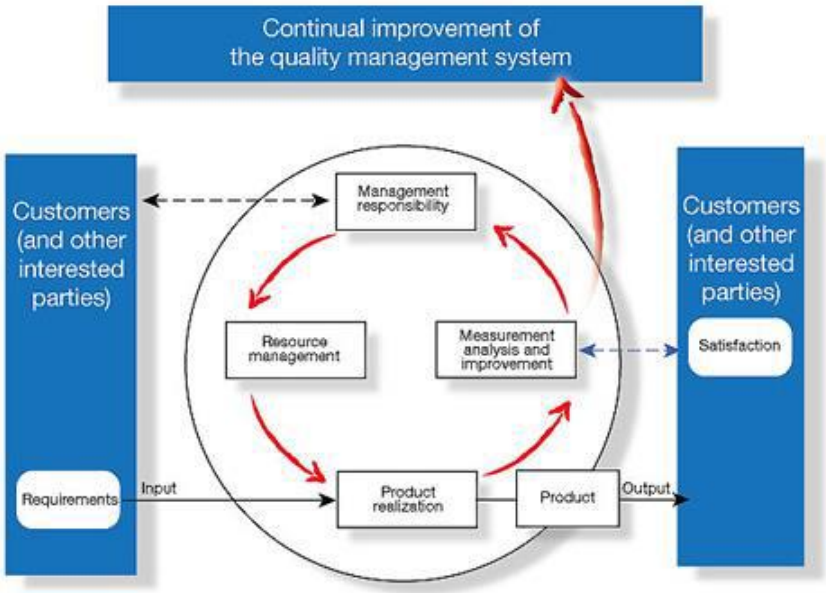
In the past there have been numerous examinations to comprehend the effect of execution of Quality Management System. As per Maguad (2006), the selection of QMS is basically subject to existing design and goals of the association. Dale (1994) in his book has written that the models of Quality Management provide the suggested guidelines or base that how an organization can work towards achievement of Quality. He additionally characterized how bit by bit little gradual changes can bring about perspective change from old management framework to new Quality Management framework.

1.2 Quality Management principles

The ISO 9000 documentation is based on seven quality management principles, they are:

QMP 1 – Customer focus
QMP 2 – Leadership
QMP 3 – Engagement of people
QMP 4 – Process approach
QMP 5 – Improvement
QMP 6 – Evidence-based decision making
QMP 7 – Relationship management

1.3 ISO 9000 Process Approach



The ISO 9000 process approach (source: www.iso.org)

The above model lays emphasis upon the significance of recognizable proof and comprehension of customer needs. The measurement of consumer satisfaction are subsequently utilized as criticism to assess and approve whether customer requirement have been accomplished or not. The management review will at that point give governance to strategic level management to alter the approval and advancement chances.
(IS / ISO 9001: 2008).

1.4 Need for Research

- It will assist with understanding the advantages and constraints of the Implementation of ISO 9000 worldwide as well as India.
- There is lack of published ISO 9000 literature availability in Indian Context.

2. The objective of the Research

- To study and appreciate the connectivity of QMS with ISO 9000.
- To study effects, benefits of ISO 9000 on the organization in World and India.
- Issues and challenges faced by organizations for the implementation of ISO 9000.

3. LITERATURE REVIEW

3.1 ISO 9000 and QM practices Connection

From the Survey by (Brown et al 1998) which happened in the year 1997 on Australia based ISO 9000 confirmed Small and medium enterprises revealed that ISO 9000 confirmation should be done like strategies for inside reasons, for instance, to improvement in customer help and enhance capability in spite of outside reasons to oversee more benefits .

Deepa 2014 quotes that ISO 9000 can act as a route to TQM, and at the same time they are complementary to one another. For the companies who are implementing TQM, installing ISO 9000 will be relatively straightforward. As indicated by the Kumar et al 2018 a definitive purpose of any association is to accomplish an "Innovation , Sustenance and Improvement " in each framework and processes so the ultimate objectives can be accomplished.

3.2 Effects experienced by Companies due to ISO certification

Since the advancement of ISO 9000 arrangement by the ISO, their norms are generally acknowledged by an enormous number of association across every enterprises non-related to their products and sizes. He further communicated that associations in agricultural nations are endeavouring to execute ISO 9000 to arrive at the international market.

ISO 9000 can't be seen as a TQM, anyway the suggestion was made that the version of ISO 9001:2015 is near to TQM considering the way that it relies upon the seven guidelines of QM that are more consolidated, as opposed to the eight norms of the past structure (2008).

Padma et al. (2008) in the investigation of Indian manufacturing firms have expressed that public firms show better market execution and inward execution than worldwide firms and conclude that accreditation adversely. Sometimes the expense of execution and upkeep is expensive than the potential advantages.

There have been many surveys conducted by various researchers during the past year to understand the motives behind ISO certifications by the companies. McTeer and Dale 1994 have opinion that the firms go for certifications to gain marketing benefits which is an external factor. Douglas et al 2003 cited that in order to anticipate future customer requirement the organizations do the ISO implementation. He further stated that to be a consideration for tenders , the company have ISO implementation.

Brown et al 1998 propounded that the reasons of ISO Certification and execution vary among large and small sized organizations. Since in small scale organizations there is no inside pressure from the administration to keep quality guidelines, for them time and cost associated with the certification turns out to be more significant. Terziovski et al, 2003, McTeer and Dale, 1994 concocted the primary justification execution as client request which is an outside factor. Basically there has been two thought processes in ISO 9000 execution, inside and outer Douglas et al. 2003 has likewise concocted a basic intention - ISO 9000 confirmation offers the chance to the organizations to open to new business sectors or get organizations they probably won't accomplish with its absence.

Since there is involvement of cost in the execution process , organizations try to decrease cost of production, efficiency and furthermore showcasing benefits (Douglas et al, 2003, Martinez-Costa et al, 2008.) There is an inescapable reasoning that the certification results might be immaterial. All things considered, ISO certificate set norms underway techniques with this there is decrease in inefficiencies. Whithers and Ebrahimpour (2000) have the discoveries that the assumptions for profitability and improvement are basically one of the best two purposes behind ISO 9000 execution. 66% of the organizations in their overview consented to similar. They additionally come further that pressure from competitors is also the motive apart from productivity improvement expectation. Till now, generally major findings designate that there is constructive and connection between motives and benefits accomplished from the ISO 9000 implementation.

3.3 Virtues of ISO 9000 Implementation

As indicated above an organization is internally motivated to implement ISO 9000, it can reap maximum benefits. However various researchers have segregated the positives of ISO 9000 implementation into dual aspects: Internal and External. Under the Internal benefits, it is achieved by internal stakeholders and environment while external benefits are related with external stakeholders including customers and suppliers.

Terziovski et al, 2003 have opinion that ISO certification can lead to production efficiency while McTeer and Dale 1994 concluded that it can lead to Quality awareness. They stated that the benefits of ISO implementation are reaped in years after certification but not immediately. Moreover firms need to understand that the certification requires money, time, energy and commitment and it is not a single day task.

Whithers and Ebrahimpour 2000 took samples from different European countries and came up with the results that with ISO 9000 implementation, the quality does improve. Terziovski et al 2003 opined that certification lead to error rates reduction.

According to the directions of the ISO standards, a company follow the procedures and can rectify the past mistakes. The employees will follow the directions provided will definitely lead towards the Quality improvement. Their discoveries are pointing that ISO 9000 arrangement's effect is more clear on net revenue rather than sales,cost reduction and easier training of new employees .These are the various internal benefits experienced by the companies due to ISO 9000 implementation. McTeer and Dale 1994 found considerable reduction in cost by small companies.

According to Brown et al 1998, the firm can have overseas market access due to certification. Increased sales and Marketing are external benefits which a company can experience. Mcteer and Dale 1994. Terziovski et al 2003 stated the market share of the company is increased due to implementation. McTeer and Dale 1994 in his research found that the 75% of the surveyed companies confirmed the increase in marketing opportunities however the actual measurement was missing. It express that numerous organizations expanded their positives after accreditation. Since then, there is still an element of doubt that the profitability has increased due to certification or any other factor. There is a consideration that ISO 9000 series is beneficial in manufacturing industry and certainly there is more literature availability on manufacturing industries.

Srivastav (2010) expressed that execution of ISO 9000 can't yield interior advantages for the association, except if there has been an inner inspiration for execution. He likewise expressed that ISO 9000 execution speed up organization advancement by improving the capacity of environment, and adapting technique which improves the human prosperity in the association. They furthermore coordinated an assessment which mirrors the positive association between ISO 9000 quality affirmation. The outcomes uncovered two solid and substantial builds: quality outcomes and operative outcomes.

3.4 Insights of Implementation of ISO 9000 in Indian context

According of ISO reports 2018 in the year 2018 India is on the fifth number on the ISO 9000 certifications with the figure of 31795 Certificates. Roughly 46% is the manufacturing industries have got the certifications.

In the same year, Out of the total 776608 ISO certification were issued to 161 countries.(ISO 2005 survey) However the latest survey of ISO in the year 2018 states that from the totality of 878664 certification of ISO 9000 issued, India had 31795 certifications in various industries .Apart from ISO 9000 there is another widely accepted ISO 14000 standard. The series 9000 refers to Quality Management and ISO 14000 refers to Environment Management. In current scenario, the two series seem to protrude each other.

Sachdeva et al 2007 quoted that attaining ISO authorization in today's competitive and excellence aware market has become elementary aspect for organization's presence in contrast with their global competitors. Very few researches in the past year have focused on empirically examining the influence of ISO 9000 certificate on Quality administration trainings. B Sarkar 1990 , led the overview on Indian businesses to comprehend the situation with Quality control in India.

He recognizes that organization approaches with respect to Quality are controlled from accomplishment in objectives as mentioned below:

(1) Customer anticipation
(2) Company status
(3) Profitability
(4) Market share

The research directed by Whithers and Ebrahimpour (2000) proposed the significant variances between four aspects (viz. management of technology, reasons for low quality, Cooperation in the quality projects and control of quality strategies utilized) of non-ISO and ISO affirmed associations in India.

Ray and Acharya (2000) conducted a comprehensive 1200 firm survey and come up with the analysis that exposed the benefit of ISO 9000 in comprehending the process well, responsibilities /authorities as well as linkage across the organization. They said that ISO 9000 accreditation facilitated the organizations in slightly refining the quality of their product. They further identified the critical success factors for implementation of QMS.

Dhar and Barua (2004) recognized that refining inward effectiveness and quality to expand service provided to customer is the foremost goal for ISO certificate through production and service providers in Indian province of Assam.

Bhat and Jagadeesh (2006) did a research in Karnataka and concluded that ISO 9000 confirmed organizations can accomplish the core advantages like better documentation and effectiveness in quality frameworks. In 2007, Sachdeva et al distinguished four significant aspects of estimating performance of the companies (excellence and cost of value, outer worth, time execution and buying) and was seen that associations have had the option to progress their exhibition on the whole the four aspects of study with level of progress changing from minimal to huge.

Bewoor et al (2010) conducted a survey on 220 companies to understand the associations between predictable and apparent profits through factual study. The concern of the organizations were for core motives as in refining procedures/efficiency or items/administrations. The peripheral reasons like pressure from clients or fake rivals

were given lesser significance. In the study conducted on quality practices which come up with the conclusion that their customers are asking for QMS Certifications.

Impacts on Firms performance

Well there are many contradictory views on firm profitability being improved through certification. There are studies that says there exist no undeniable connection of ISO 9000 certificate and advance in the sales or complete execution of the organization. Quazi and Padibjo in (1998) conducted survey on Singaporean firms came up that along with marketing gains, improvement in product quality was also experienced. They conducted an analysis on ISO 9000 enlistment influenced deals and benefit and came up with finding that no evidence proves the enhanced performance of the certified organization after enlistment in the 400 certified companies studied. Douglas et al, 2003 anticipated the sturdy association among the thought process behind the execution result and monetary enhancement for organization's execution. However on the contrary, (in 2003, Terziovski et al) 1000 firm's empirical study in New Zealand and Australia proposed the insignificance of optimistic association with companies performance. He analysed the ROI (return on Investment) between the year 1989 to 1995 and didn't found any significant improvement after the ISO 9000 certification. The organization use its ISO 9000 certification as the advertisement tool. This kind of attitude fail the purpose of implementation success and paperwork involved in the process. From the year 1990 to 1998, Wayhan et al.'s (2001) analysed the monetary after effects of 96 USA companies. It is found that there was no increase in the firm's sales revenue, but due to the reduction in internal cost has led to increased profitability. He also showed the partial effect of ISO 9000 accreditation on monetary execution, estimated with respect to return on resources, well this impact vanish rapidly over the long run.

It was proposed by Sachdeva et al 2007 that there are three scopes of monetary execution: working effectiveness – estimated with net revenue , sales advancement, and inclusive monetary execution estimated by income per share. The discoveries are more effective on net revenue than deals.

Martinez Costa et al 2008, researched in Spain to understand performance enhancement. He comprehend this scenario to be relatively pied along with substantial positive coordination among accreditation and commercial performance is not always affirmed. Also it had concluded solid evidence in association of ISO 9000 operation and enhancements in complete monetary operations such as quality responsiveness, efficiency, improved organizational performance, market share, client approval, and support of sales income.

To conclude the performance discussion, many researchers have proposed that the certainty lies in between. For the interpretation of the result with accuracy, one has to dig deeper. Few metrics are unquestionable like: reduced production cost, reduced errors, reduction in customer complaints and increase in sales. However insignificant variances were experienced in organizations rank.

Indecisions /Doubts

Till now a lot of discussion of effect and positives of ISO 9000 certification on the organizations is discussed. However few researchers also brought light to the doubts and negatives of the ISO 9000 implementation. Carr et al and Terziovski 2003 offered the rejection of the theory of improved financial performance. In the literature reviewed for this paper, two commonly cited disadvantages are : Cost involved in the process (Terziovski 2003) and large amount of paperwork involved in the process. Other factors that were also considered as disadvantages were time pressure(Whithers and Ebrahimpour 2000). Lack of knowledge from consultants and discouragement of thinking from employees point of view (McTeer and Dale 1994).

Cost is the most commonly cited disadvantage of ISO 9000 implementation by the small firms Terziovski et al 2003. It is likewise hard to precisely estimate the expense caused for implementation since it carries related cost also. Stevens and Barnes 2001 segregated the four foremost aspects including: time required by the processes, employees training which included lost man-hours, consultant payment and registration fees. Out of all these factors time needed and training hours are intangible but the consultant payment and registration fees can be specific in terms of money.

The need of numerous documents is stated as required for ISO 9000 certification. This is required for implementation and after too. Generally this is not an issue for the larger firm since they are carrying it in some way or other. Brown et al 1998 have brought into light that, large amount of documentation is needed in the audits. For elimination of errors and stabilization of procedures there is need of keeping the records.

When the paperwork is taken for granted the firm can face the disappointments mentioned in the literature. ISO 9000 serves as the advertisement and a powerful marketing tool for the companies.

Inferences and Scope for further research

The mapping of ISO 9000 Quality management System background, principles, process is conversed in this paper. The association of ISO 9000 and QM practices is studied. The effect, benefits of ISO 9000 on organizations is analysed. Also the doubts and disadvantages presented by various researchers are discussed. The insights about ISO 9000 in Indian context are taken. The current state of literature on the ISO 9000 QMS is

more concentrated on Asia and Europe whereas the developing countries like Latin America, Africa and India have very few studies conducted. Manufacturing sector was more focused in comparison to other sectors. Questionnaire is the most commonly used tool. The recurrent measurement method if found out to be Likert Scale along with the universal basis of information to be each respondent's perception.

Further, the future examination studies ought to sufficiently cover developed as well as developing economies in all continents and the types of organizations (manufacturing and service). Additionally the Quality Management frameworks being trailed by the associations (either independently or together) ought to be explored, with the goal that the research results are more complete and more perceptions can be inferred whether these scopes affect the research consequences.

The study to be led should depend on empirical information with longitudinal plan to comprehend the factual association among QMS and firm certification.

Research Limitations

The ISO 9000 QMS is taken in a generalized construct; other organizations which have implemented different QMS like TQM are not studied. Also there has not been clear cut differentiation for small scale industries and large industries. The environment to implement the ISO 9000 is not supposedly same for countries in Europe or Asia but no such consideration has been paid to the matter.

REFERENCES

1. Acharya, U.H. and Ray, S. (2000) 'ISO 9000 certification Indian industries: a survey', Total Quality Management, Vol. 11, No. 3, pp.261–266.
2. B. Sarkar (1990) Status of quality control in Indian industries—a survey, Total Quality Management, 1:1, 133-146, DOI: 10.1080/09544129000000014
3. Barua M and U.R. Dhar (2004) Identification of factors for ISO 9000 certification : A study of industries in Assam ,ICFAI Journal of Operations Management Volume 3- Page 17-27
4. Bewoor A.K. and Pawar M.S. (2010), "An empirical analysis of impact of QMS/ISO implementation on productivity/performance of Indian SMEs", International Journal of Industrial Engineering and Technology, Vol 2 (1), pp 85-109.
5. Bhat.S and Jagadeesh (2006), "ISO 9000 implementation in Indian industries: A study of its Reasons and Benefits", The ICFAI Journal of Operations Management- Vol 5(4), pp 53-63.
6. Ben A. Maguad (2006) The modern quality movement: Origins, development and trends, Total Quality Management & Business Excellence, 17:2, 179-203, DOI: 10.1080/14783360500450608
7. Brown A., Van de Wiele T. and Loughton K. (1998), "Smaller enterprises' experiences with ISO 9000", International Journal of Quality & Reliability Management, Vol. 15, No3, pp. 273-285
8. Douglas A., Coleman S. and Oddy R. (2003), "The case for ISO 9000", The TQM Magazine, Vol. 15, No5, pp. 316-324
9. Kumar, Pankaj, Maiti, Jhareswar & Gunasekaran, Angappa. (2018). Impact of Quality Management systems on firm performance. International Journal of Quality & Reliability Management. 35. 00-00. 10.1108/IJQRM-02-2017-0030.
10. McTeer M. M. and Dale B. G. (1994), "Are the ISO 9000 series of quality management system standards of value to small companies?", European Journal of Purchasing and Supply Management, Vol. 1, No4, pp. 227-235
11. Martinez-Costa M., Martinez-Lorente A. R. And Choi T. Y. (2008), "Simultaneous consideration of TQM and ISO 9000 on performance and motivation: An empirical study of Spanish companies", International Journal of Production Economics, Vol. 113, No1, pp. 23-39
12. Padma, P., Ganesh, L.S. and Rajendran, C. (2008), "A study on the critical factors of ISO 9001:2000 and organizational performance of Indian manufacturing firms", International Journal of Production Research, Vol. 46 No. 18, pp. 4981-5011.
13. Priyanshu, Deepa. (2015). Evolution of quality: a study of quality from primitive period to latest development in India. Pranjana: The journal of Management Awareness Volume17.Issue 2 DOI-10.5958/0974-0945.2014.00002.8.
14. Quazi, H.A. and Padibjo, S.R. (1998), "A journey toward total quality management through ISO 9000 certification - a study on small- and medium-sized enterprises in Singapore", International Journal of Quality & Reliability Management Vol. 15 No. 5, pp. 489-508. https://doi.org/10.1108/02656719810196225
15. Sachdeva, A., Bhardwaj, A. and Sharma, V.S. (2007) 'Impact of ISO 9000 certification on performance of SMEs: a study of Indian industry', Int. J. Management Practice, Vol. 2, No. 3, pp.226–239.
16. Srivastav, Avinash. (2010). Impact of ISO 9000 implementation on the organisation. International Journal of Quality & Reliability Management. 27. 438-450. 10.1108/02656711011035138.

17. Terziovski M., Power D. and Sohal A. S. (2003), "The longitudinal effects of the ISO 9000 certification process on business performance", European Journal of Operational Research, Vol. 146, No3, pp. 580-595
18. Wayhan, Victor & Kirche, Elias & Khumawala, Basheer. (2002). ISO 9000 certification: The financial performance implications. Total Quality Management & Business Excellence TOTAL QUAL MANAG BUS EXCELL. 13. 217-231. 10.1080/09544120120102450.
19. Whithers B. and Ebrahimpour M. (2000), "Does ISO 9000 certification affect the dimensions of quality used for competitive advantage?", European Management Journal, Vol. 18, No4, pp. 431-443

New Journal and we have not received input yet 23 (2021) 100925

Contents lists available at ScienceDirect

Thermal Science and Engineering Progress

journal homepage: www.sciencedirect.com/journal/thermal-science-and-engineering-progress

Thermo-economic analysis of the combined solar based pre-compression supercritical CO_2 cycle and organic Rankine cycle using ultra low GWP fluids

Yunis Khan *, Radhey Shyam Mishra

Department of Mechanical Engineering, Delhi Technological University, Bawana Road, New Delhi 110042, India

ARTICLE INFO

Keywords:
Thermal performance analysis
Solar power tower
Pre-compression sCO_2 cycle
Organic Rankine cycle
Ultra-low GWP fluids

ABSTRACT

Present study deals with the parametric and economic analysis of solar driven combined pre-compression supercritical CO_2 cycle and organic Rankine cycle using ultra low global warming potential HFO (hydro fluoro olefins) fluids and comparison with the R134a fluid. A numerical model was developed with the help of EES software to simulate the combined system. It was concluded from the results that HFO working performed better than the R134a. R1336mzz(Z) gave highest thermal and exergy efficiency and power output for the combined cycle by 55.02, 59.60%, 298.5 kW at 950 W/m^2 of solar irradiation respectively. Also highest waste heat recovery ratio was found 0.84 for the R1336mzz(Z) fluid while R134a gave only 0.099 at 0.95 heat exchanger effectiveness. At end, it was found that the R1336mzz(Z) and R1234ze(Z) estimated the lowest and highest specific investment cost i.e. 2234 and 2290 €/kWe respectively. It was also concluded from the results that if the fluid is the best for thermal performance, it is not necessary it would be best from the economic point of view.

1. Introduction

Due to industrial expansion and living standards, world energy demand is growing on a daily basis. Consequently, the use of fossil fuels is constantly growing, contributing to higher carbon emissions [1]. Fossil-fuel stocks are declining slowly, however. This leads to the challenge of seeking energy options that are renewable and secure [2]. Numerous renewable energy options, such as solar, wind, biomass and geothermal energy, are presently being used to produce clean and environmentally sustainable electricity. Among these alternative energy resources, solar energy is more appropriate for power generation due to its availability, low cost and noise-free operation [3].

To capture sunlight for cooling, heating, and power generation, solar collectors are employed [4]. Tian and Zhao [5] studied solar collectors and found that flat plate collectors can be used at low temperatures of up to 100 °C. For high temperature applications, however, concentrated solar collectors are used, such as parabolic trough collectors, dish collectors, heliostats, etc. Kalogirou [6] reported the working temperature range of the parabolic trough collector (60–500 °C), dish collector (750–1000 °C), and heliostat (150–1500 °C). Concentrated solar power (CSP) technology plays a vital role in resolving the existing and expected problem of power generation in tropical countries by using solar heat that is freely available in nature, amongst other solar technologies [7].

The solar power tower (SPT) is the latest technology, among the different CSP technologies. The SPT network consists of a set of complex sub-systems such as a receiver, a 75–150 m high tower, a thermal storage system (optional), a 50–150 m^2 heliostat field and an energy conversion system per heliostat field. Solar radiation is based on heliostat field receivers where it is used to generate high temperature heat for power generation across a high energy cycle or for the supply of industrial processes [8]. Several research have been carried out on SPT-driven cycles, such as the combined sCO_2 recompression cycle and *trans*-critical CO_2 cycle [9], sCO_2 Brayton cycle [10], sCO_2 recompression with and without intercooling main compressor [11], triple combined cycle [12], multi-generation hybrid cycle [13].

Supercritical CO_2 (sCO_2) is the condition of CO_2 above its critical values (7.38 MPa, 30.98 °C). In addition, the cycle of sCO_2 is the cycle that can use heat from different heat sources such as solar thermal energy, geothermal power, coal power, and natural gas [14]. A Few studies have been reported on the basis of sCO_2, A comparative study of the various sCO_2 cycle configurations, such as recompression, easy recovery, and partial cooling cycle, was carried out by Besarati and Goswami [15]. In order to recover the waste heat from all three cycles, they first compared the systems separately and then used ORC as the bottom cycle and compared it again. They claimed that the combined cycle obtained

* Corresponding author.
E-mail addresses: yuniskhan21@gmail.com (Y. Khan), rsmishra@dce.ac.in (R. Shyam Mishra).

https://doi.org/10.1016/j.tsep.2021.100925
Received 25 November 2020; Received in revised form 21 March 2021; Accepted 26 March 2021
Available online 8 April 2021
2451-9049/© 2021 Elsevier Ltd. All rights reserved.

Nomenclature

A_h	Single heliostat area (m^2)
Cp	Specific heat at constant pressure (kJ/kg-K)
$\dot{E}$	Exergy rate (kW)
$\dot{E}D$	Exergy destruction rate (kW)
$\dot{E}_{solar}$	Solar exergy inlet to combined cycle (kW)
f_{view}	Receiver view factor
G_b	Solar irradiation (W/m^2)
h	Specific enthalpy (kJ/kg)
h_{conv}	Convective heat loss coefficient (W/ m^2-K)
$\dot{m}$	Mass flow rate (kg/s)
N_h	Number of heliostat
$\dot{Q}$	Heat rate in (kW)
$\dot{Q}_h$	Actual solar heat received by heliostat field (kW)
$\dot{Q}_{loss,r}$	Heat loss from the receiver (kW)
$\dot{Q}_r$	Heat received by central receiver (kW)
$\dot{Q}_{solar}$	Solar heat received by heliostat field (kW)
s	Specific entropy (kJ/kg-K)
sCO_2	Supercritical carbon dioxide
T	Temperature (K)
T_R	Surface temperature of receiver (K)
$\dot{W}$	Power (kW)
η_{ex}	Exergy efficiency
η_h	Heliostat efficiency
η_r	Receiver thermal efficiency
η_{th}	Thermal efficiency

Abbreviations

CFC	Chlorofluorocarbon
Cond	Condenser
CR	Concentration ratio
GWP	Global warming potential
HFC	Hydro fluoro carbon
HFO	Hydro fluoro olefins
HTR	High temperature recuperator
HX1	Heat exchanger −1
HX2	Heat exchanger −2
IPMC	Inlet pressure of main compressor (MPa)
LTR	Low temperature recuperator
MC	Main compressor
MCT	Maximum cycle temperature (°C)
MT	Main turbine
ORC	Organic Rankine cycle
OT	ORC turbine
PC	Pre-compressor
PCIP	Pre-compressor inlet pressure (MPa)
SPT	Solar power tower
WHRR	Waste heat recovery ratio

Greek letters

β	Sun's subtended cone half angle(rad)
δ	Change in property
ζ	Thermal emittance
α	Solar absorbance
η	Efficiency
σ	Stephen Boltzmann constant (W/m)
ε	Effectiveness

Subscripts

0	environmental conditions
b	boiling
c	critical
e	exit
h	heliostat
i	inlet
ms	molten salt
r	receiver

more than 50% thermal efficiency and that the recompression cycle was the most efficient of the cycles considered. Kim et al [16] also carried out a comparative study of 12 different configurations of the sCO_2 cycle, such as simple recuperated cycle, single heated cascade cycle, As land filling gas turbine bottoming, recompression, pre-compression, dual cascade cycle, single heated cascade cycle with intercooler, partial heating, intercooled dual heated cascade cycle, dual heating and flow split cycles for triple heating cycle, dual expansion cycle and partial recovery cycle. One of the conclusions derived from their observations was that the pre-compression and recompression cycle had high thermal efficiency compared to other cycles considered. They also proposed combining both cycles for future study. Yu et al. [17] considered four sCO_2 cycle configurations for recovering the waste heat from the internal combustion exhaust, such as pre-compression, simple recuperated, recompression cycle, and split flow recompression cycle. Additionally, the ratio of exhaust heat recovery and thermal efficiency of the four systems was determined. The maximum exhaust heat recovery ratio for pre-compression and sCO_2 cycle recovery was found to be reached at 5.8 MPa and 7.65 MPa respectively. In addition, the best performing cycle with the highest recovery ratio of 24.75% was found to be the split flow cycle. The cost, design, and performance of the three sCO_2 cycle configurations, such as recompression cycle, simple cycle and partial cooling cycle, powered by the molten salt solar power tower, were examined by Neises and Turchi [18]. As a result, in the case of a recompression cycle followed by partial cooling and a simple cycle, the maximum thermal efficiency has been achieved. Because of the high demand for turbo machinery, partial cooling was found to be more expensive than other cycles. If come across the solar driven sCO_2 cycle, Khan and Mishra [1] performed a research on the solar parabolic trough collector coupled with partial heating of the sCO_2 cycle and the ORC. They also compared the integrated system's performance with and without the solar system. As a result, the thermal efficiency increased by 4.47% with the implementation of the ORC to the existing partial heating cycle. In addition, they explored the effect of solar radiation on the combined system's performance. They concluded that solar irradiation improved the thermal and exergy efficiency of the system and parabolic trough collector alone is responsible for 62.93% of the exergy destruction rate of the overall cycle. Abid, Khan and Ratlamwala [49] performed economic and comparative analysis of the solar driven of sCO_2 cycle with hydrogen generation. The overall energy and exergy efficiencies of the system were found 30.37% and 32.7%, respectively.

In addition, the pre-compression cycle uses the pre-compressor between the main compressor and the turbine in a simple Brayton cycle to make the turbine outlet pressure independent of the inlet pressure of the compressor. To prevent pinch issues, the pre-compression configuration needs one more heat exchanger. Additional compression is used for further regeneration when the temperature gap is lowered to a minimum in the recuperator [19]. The advantage of this design is the installation of an additional compressor that provides space for further regeneration [20]. Extra compression increases the temperature of the fluid, and also raises the specific heat of the fluid due to high compression. It is also possible to continue the regeneration process because, compared to the simple Brayton cycle, more heat is required for regeneration. Over the simple Brayton cycle, this results in an improvement in productivity [21]. As the pre-compression cycle is properly calibrated, its efficiency over the Brayton cycle can be increased by up to 6% [22].

In the current scenario, ORC is used as a bottoming system. There are a few studies to support this comment, such as a solar integrated

combined cycle research published by Shaaban [23], in which the ORC and steam Rankine cycle were used as a bottoming cycle for waste heat recovery. The authors mentioned 15 working fluids in the ORC and found that in terms of thermo-economic, safety and environmental requirements, R1234ze(Z) was the best fluid. Hoang [24] investigated the use of the ORC as the bottom cycle for the recovery of exhaust waste heat from diesel engines. They find that up to 25 percent thermal efficiency can be achieved by ORC, while up to 90 percent thermal efficiency is achieved by a combined system with a diesel engine. A performance and optimization analysis of the combined cycle of sCO_2 and ORC was carried out by Song et al. [25]. The effects of the recuperation ratio of the topping cycle, the pressure ratio on the bottom output, were examined. The thermal efficiency of the current sCO_2 cycle was improved by the addition of the ORC system. The ORC system has completely recovered the residual heat. Also solar driven ORC performance well such as Singh and Mishra [48] performed exergy and energy analysis on the solar driven supercritical ORC using five working fluids such as R600a, Toluene, R152a, isobutene, and cyclohexane. The exergy efficiency was found 96.09% at 0.95 kW/m^2 of solar irradiation.

Further from previous studies it was also found that the combined cycle performance also depends upon the proper working fluids selection for the bottoming ORC. In this direction several researcher selected the working fluids for the bottoming cycle such as Khan and Mishra [1] considered the six working fluids such as R1224yd(Z), R1234ze(Z), R1234yf, R1234ze(E), R1233zd(E) and R1243zf for the bottoming ORC of combined solar integrated partial heating sCO2 cycle and ORC as bottoming cycle. They concluded that among the all selected working fluids the R1233zd(E) gave the highest performance. Khan and Mishra [26] also considered the five working fluids (isopentane, R245fa, R236fa, isobutene, and R227ea) for the bottoming ORC of the combined solar power tower driven pre-compression sCO_2 and ORC. They found that the R227ea showed better performance than the other considered working fluids. Yu, Feng and Wang [27] considered eight working fluids for ORC driven by various heat sources. They found that the thermal performance of the system extremely depend upon the type of working fluids. They generally selected HFC (hydro fluoro carbon) working fluids such as R227ea, R236fa, R236ea, R245fa, R600, R600a, R601, and R601a. They finally concluded that the R601 gave the highest thermal efficiency.

From the literature survey, it was observed that performance of the combined sCO_2 cycle and ORC also depends on the working fluids selection for the bottoming ORC. ORC fluid should be high efficient, low GWP and zero ODP. Thermal performance of the combined pre-compression sCO_2 cycle and ORC can be further enhanced with selecting proper working fluid in ORC. Considering this research gap, present study deals with thermodynamic analysis of the combined solar power tower driven pre-compression sCO_2 cycle and ORC considering HFO low GWP fluids in ORC. Therefore, selection of the low GWP HFO working fluid for the current system denotes the novelty. Also, current study differs from previous studies due to selection of low GWP fluids based on both thermal and economic performance for combined SPT driven pre-compression sCO_2 cycle and ORC. Exergy efficiency, thermal efficiency, net output power, and waste heat recovery ratio were considered as performance parameters of the proposed system. A computer program in engineering equation solver (EES) software was made to simulate the system. Parametric analysis of the proposed combined cycle was conducted with different HFO working fluids and also performance was compared with the HFC high GWP fluid R134a. Apart from this economic analysis of the ORC system for each working fluid was also carried out. The effects of the system variable such as solar irradiation, maximum cycle temperature, inlet pressure of main compressor, inlet pressure of pre-compressor, heat exchanger-2 (it is the heat exchanger through which heat is transferred to ORC) effectiveness and low temperature recuperator effectiveness on the system performance were investigated. Analysis and design of solar power tower is out of scope of current study.

2. System description

The existing system of a combined cycle comprises of two thermodynamic cycles, such as the sCO_2 pre-compression cycle as a topping cycle and the ORC as a bottoming cycle for the recovery of waste heat from the topping cycle. Fig. 1 displays the schematic diagram of the new combined system powered by the solar power tower. Throughout the SPT field, heat transfer fluid (HTF) (molten salt) circulates. This HTF delivers heat via the heat exchanger-1 (HX1) (state a-b) to the topping cycle. The pre-compression cycle, via the heat exchanger-1 (state 9–1), takes heat from the SPT. Through the main turbine (MT), heated sCO_2 that circulates in the pre-compression cycle has extended to get power out (state 1–2). Expanded sCO_2 has a lot of heat that is recovered to heat the sCO_2 stream coming into the HX1 (state 2–3) in the high temperature recuperator (HTR). After the temperature of the HTR sCO_2 stream is decreased and the difference in temperature of the HTR streams is reduced. The sCO_2 stream is pre-compressed in the pre-compressor (PC) (state 3–4) to prevent pinch point issues. Then the sCO_2 stream goes to the low temperature recuperator (LTR) after the PC temperature and pressure has risen, where heat is recovered by the low temperature stream moving to HTR (state 4–5). Although the sCO_2 stream still has a lot of heat, it travels to the heat exchanger-2 (HX2) where HX2 (state 5–6) has provided this quantity of heat to the ORC. sCO_2 compressed in MC again to acquire MT inlet pressure (state 6–7) after the stream of MC sCO_2 continues to the LTR (state 7–8). After that the sCO_2 stream goes to regenerate the HTR (state 8–9). Eventually, to complete the cycle, the sCO_2 stream again goes to the HX1. Come on over to the ORC now. Via the waste of the topping cycle, ORC working fluids take heat via the HX2 and then expand in the organic turbine (OT) (state 10–11). Then the fluid stream moves to the heat rejecting condenser (state 11–12). Via the pump, fluid stream pressure again improved (state 12–13). Ultimately, it goes back to the HX2 to get the heat and also the cycle completed. T-s diagram for this model is shown in Fig. 2 according to corresponding states.

3. Thermodynamic analysis

3.1. Assumptions

To support the simulation of the current study following assumption has been taken. (1) All system components are in equilibrium and steady state conditions. (2) Friction and pressure loss has been ignored in piping. (3) All thermodynamic processes are polytropic. (4) Potential and kinetic energy for each component has been neglected. (5) Heliostat and receiver parameters have kept constant and assumed input data to support the simulation are given in Table 1. (6) Inlet to the HX1 the temperature of the molten salt has been taken as 700 °C [28]. (7) Due to thermal losses, inlet temperature of MT is 50 °C less than temperature of molten salt inlet to the HX1.

3.2. Thermal modeling of SPT

Taking into consideration of assumptions, modeling equation have been developed to simulate the proposed system. Modeling equations has been taken from previous studies Khan and Mishra [26] and Khatoon and Kim [9].

Direct solar heat incidence upon heliostat field is defined as [9,26]

$$\dot{Q}_{solar} = (G_b \cdot A_h \cdot N_h \quad (1)$$

Where, G_b is the solar irradiation in W/m^2 also termed as direct normal irradiation (DNI), A_h is single heliostat area (m^2) and N_h is the heliostats number. However, due to heliostat efficiency, some of that heat is lost in the surroundings. The amount of actual heat obtained through the heliostat field is therefore specified as [9,26];

$$\dot{Q}_h = \dot{Q}_{solar} \cdot \eta_h \quad (2)$$

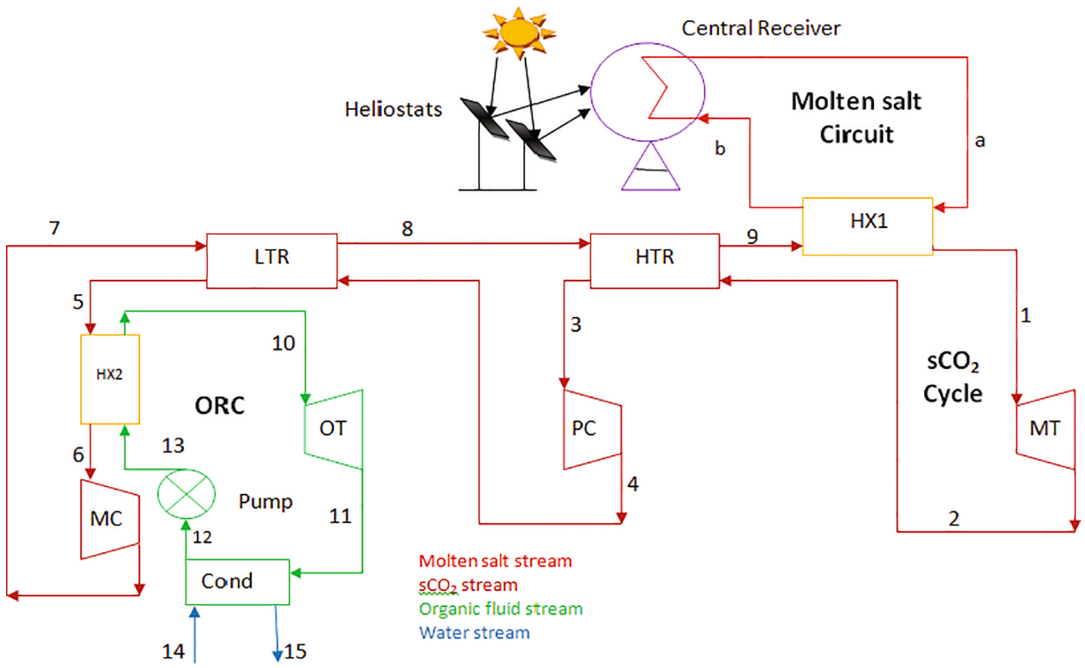


Fig. 1. Schematic diagram of solar based combined cycle [26].

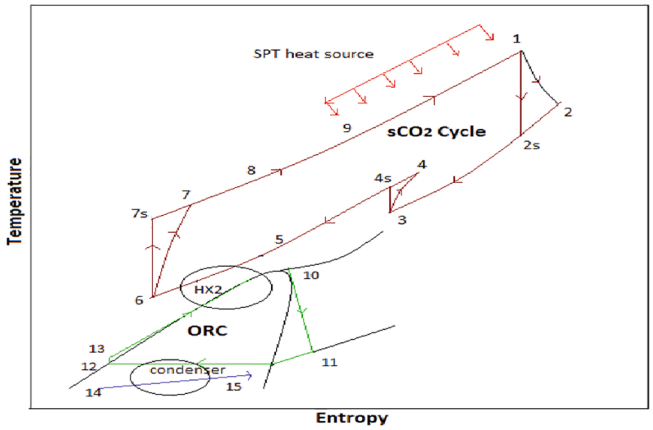


Fig. 2. Temperature - entropy (T-s) diagram of proposed system [26].

Where, η_h is the efficiency of heliostat. This amount of heat is directed to the solar receiver where the heat transfer fluid flows. But in the atmosphere a part of the heat is lost. The heat available at the solar center receiver is therefore determined as [9,12,26];

$$\dot{Q}_r = \dot{Q}_h \cdot \eta_r = \dot{Q}_h - \dot{Q}_{loss,r} \quad (3)$$

Where, η_r is the receiver thermal efficiency, is defined as [9,26];

$$\eta_r = \alpha - \frac{\zeta \cdot f_{view} \cdot \sigma \cdot T_R^4 + h_{conv} \cdot f_{conv} \cdot (T_R - T_{air})}{G_b \cdot \eta_h \cdot CR} \quad (4)$$

Where, T_R is the solar receiver surface temperature and CR is concentrated ratio. ζ is the solar emittance. To calculate heat loss, this can be approximated as [9,26];

$$T_R = T_1 + \delta T_R \quad (5)$$

Where, T_1 is the turbine's inlet temperature and δT_R is approach temperature of solar receiver in kelvin.

The operating and geometric parameters of the solar receiver and the heliostat field are listed in Table 1.

Furthermore, exergy of the any system can be explained as maximum work obtainable from the system when system is brought to its dead conditions. Control volume exergy balance equation can be determined as [29];

$$\sum \left(1 - \frac{T_0}{T_Q}\right) \dot{Q}_j - \dot{W}_{c.v} - \sum (\dot{m}_i E_i) - \sum (\dot{m}_e E_e) - \dot{E}D = 0 \quad (6)$$

Where, $\dot{E}D$ is the exergy destruction rate and subscript j refers to thermal property at particular state. Solar exergy inlet to the combined system is determined by Petela's formula [30];

Table 1
Input parameters for simulation of the proposed model [26].

Geometric and operating parameters for SPT	
Solar irradiation	970 W/m^2 [9]
Sun temperature	5800 K [4,33]
Solar's multiple	2.8 [9]
Heliostat efficiency	0.5871 [34]
Heliostat number	141 [35]
Total mirror area of heliostat	9.04 × 7.89 [9]
Difference in initial temperature	15 K [9]
Temperature approach of solar receiver	423.15 K [34]
Concentration ratio	900 [34]
Heat loss coefficient of convection	10 W/m^2-K [34]
Height of tower	74.62 m [35]
Heat loss factor in convection	1 [34]
Absorptance	95% [34]
View factor	80% [34]
Thermal emittance	85% [34]
Input parameters for combined cycle	
Inlet pressure of main turbine	25 MPa [36]
Inlet temperature of main turbine	650 °C [9,36]
Inlet pressure of Pre-compressor	5.6 – 6.8 MPa [28,36]
Inlet pressure of main compressor	7 –10.5 MPa
Inlet temperature of main compressor	32–38 °C [28,36]
Isentropic efficiency of main turbine	88% [28,36]
Isentropic efficiency of main compressor	85% [28,36]
Isentropic efficiency of Pre-compressor	85% [28,36]
Effectiveness of heat exchanger	95% [15]
LTR and HTR effectiveness	95% [15]
ORC turbine inlet pressure	3 MPa [1,26]
Isentropic efficiency of ORC turbine	80% [25]
Isentropic efficiency of ORC pump	70% [25]

$$\dot{E}_{solar} = \dot{Q}_{solar} \cdot \left[1 + \frac{1}{3}\left(\frac{T_0}{T_{su}}\right)^4 - \frac{4}{3}\left(\frac{T_0}{T_{su}}\right)\right] \quad (7)$$

Where, T_{su} is the sun temperature (5800 K) and T_0 is reference temperature. Further, in the receiver, useful exergy obtained by the molten salt is defined as

$$\dot{E}_r = \dot{m}_{ms} \cdot Cp_{ms} \cdot \left[(T_b - T_a) - \left(T_0 \cdot \ln\frac{T_b}{T_a}\right)\right] \quad (8)$$

Further chemical exergy of the system is constant throughout. After neglecting energy due to velocity and height, specific physical exergy at j^{th} point is defined as [4,26];

$$E_j = (h_j - h_0) - T_0(h_j - s_0) \quad (9)$$

3.3. Thermal modeling of combined cycle

Modeling equations for the thermal modeling of combined cycle have been taken from Khan and Mishra [26]. Heat received by the combined cycle from the SPT field is given by the heat balance equation in the heat exchanger (1);

$$\dot{Q}_r = \dot{Q}_{HX1} = \dot{m}_{ms} \cdot Cp_{ms} \cdot (h_b - h_a) = \dot{m}_{SCO2} \cdot (h_1 - h_9) \quad (10)$$

After considering the isentropic efficiency of main turbine, power output from the main turbine is determined as;

$$\dot{W}_{MT} = \dot{m}_{SCO2} \cdot (h_1 - h_{2s}) \cdot \eta_{MT} \quad (11)$$

Where, η_{MT} is the main turbine's isentropic efficiency and h_{2s} is the enthalpy at the outlet of the main turbine when expansion is done isentropically.

Heat transfer in the HTR is determined by effectiveness formula and energy balance equation;

$$\dot{Q}_{HTR} = \dot{m}_{sCO2} \cdot (h_2 - h_3) = \dot{m}_{sCO2} \cdot (h_9 - h_8) \quad (12)$$

Where, $\dot{m}_{sCO2}$is sCO_2 mass flow rate. Effectiveness of the HTR having same mass flow rate in hot side and cold side of HTR is defined as;

$$\varepsilon_{HTR} = \frac{T_2 - T_3}{T_2 - T_8} \quad (13)$$

Heat transfer in LTR is also defined by applying the heat balance and effectiveness formula;

$$\dot{Q}_{LTR} = \dot{m}_{sCO2} \cdot (h_4 - h_5) = \dot{m}_{sCO2} \cdot (h_8 - h_7) \quad (14)$$

Effectiveness of the LTR is defined as;

$$\varepsilon_{LTR} = \frac{T_4 - T_3}{T_4 - T_7} \quad (15)$$

Heat absorbed by the ORC through the heat exchanger-2 energy balance formula and the effectiveness formula can be given as;

$$\dot{Q}_{HX2} = \dot{m}_{SCO2} \cdot (h_5 - h_6) = \dot{m}_{ORC} \cdot (h_{10} - h_{13}) \quad (16)$$

Where, $\dot{m}_{ORC}$ is the ORC's working fluid mass flow rate

Effectiveness of the heat exchanger-2 is defined as;

$$\varepsilon_{LTR} = \frac{C_{ORC} \cdot (T_{10} - T_{13}) = \frac{C_{sCO2} \cdot (T_5 - T_6)}{C_{min} \cdot (T_5 - T_{13})}}{C_{min} \cdot (T_5 - T_{13})} \quad (17)$$

Where, C_{ORC} and C_{sCO2} are the heat capacity of ORC working fluid and sCO_2 respectively.C_{min} is the minimum of these two values.

Net power obtained from the individual bottoming ORC can be defined as;

$$\dot{W}_{net,ORC} = \dot{W}_{OT} - \dot{W}_{pump} \quad (18)$$

Exergy destruction and exergy in each component are determined by incorporating an exergy equilibrium Eq. (6) for each element after considering zero heat losses in the element [29].

After calculating the exergy destruction rate for each component, total exergy destruction rate for the combined cycle is calculated as;

$$\dot{E}D_{total} = \dot{E}D_{HX1} + \dot{E}D_{MT} + \dot{E}D_{HTR} + \dot{E}D_{PC} + \dot{E}D_{LTR} + \dot{E}D_{HX2} + \dot{E}D_{MC} + \dot{E}D_{OT} + \dot{E}D_{cond} + \dot{E}D_{pump} + \dot{E}D_{reciever} \quad (19)$$

3.4. Performance parameters

In the performance analysis of the solar power tower powered combined cycle, various mathematical relationships are used on the basis of thermal modeling.

The net output of power received from the combined cycle is specified as;

$$\dot{W}_{net} = \dot{W}_{MT} + \dot{W}_{OT} - \dot{W}_{PC} - \dot{W}_{MC} - \dot{W}_{pump} \quad (20)$$

Solar powered combined cycle's thermal efficiency is determined as;

$$\eta_{th} = \frac{\dot{W}_{net}}{\dot{Q}_{solar}} \quad (21)$$

Combined cycle exergy efficiency is determined as [4,26,30]

$$\eta_{ex} = 1 - \frac{\dot{E}D_{total}}{\dot{E}_{solar}} \quad (22)$$

or

$$\eta_{ex} = \frac{\dot{W}_{net}}{\dot{E}_{solar}} \quad (23)$$

The combined cycle's thermal efficiency can also be defined by the relationship between the combined cycle's thermal and exergy efficiency [26,29];

$$\eta_{th} = \eta_{ex} \cdot \eta_{Carnot} \quad (24)$$

Finally, Waste heat recovery ratio (WHRR) is a performance parameter that represents the ability to recover waste heat from the topping cycle by the bottoming ORC. It was defined as the ratio of the net power output (net power output of ORC) to the highest allowable waste heat to be recovered from the source of waste heat [31]. Therefore, WHRR was calculated as;

$$WHRR = \frac{\dot{W}_{net,ORC}}{\dot{m}_{sCO2} \cdot (h_5 - h_0)} \quad (25)$$

Where, h_0 is the enthalpy of waste heat of the topping cycle at environmental temperature.

In the engineering equation solver (EES) [32], modeling equations of

Table 2
Thermo-physical properties of molten salt ($MgCl_2$ + KCl) [41].

Parameters	Values
Density	1593 (kg/m^3)
Specific heat capacity	1.028 (kJ/kg-K)
Thermal conductivity	0.39 (W/m-K)
Solidification temperature	699 K
Stability limit	1691 K

the SPT driven combined cycle have been solved.

3.3. Selection for working fluids

The working fluids for any system should be chosen carefully because it affects the environment, economic feasibility and sustainability. A mixture of magnesium dichloride (MgC_{l2}) and potassium chloride (KCl) has been used as molten salt HTF in the receiver with mass fraction of 32% and 68% respectively [9]. The reason behind choosing this HTF is that, compared to solar salt and liquid sodium (Na), this is the cheapest alternative for the heliostat-driven sCO_2 cycle [37]. Thermo-physical properties of this molten salt are listed in Table 2. The choice of working fluid for the ORC is difficult since it destroys its chemical stability beyond its maximum temperature, but at optimum pressure and temperature, it obtains optimum thermo-physical properties [38]. Various parameters, critical point, including global warming potential (GWP), thermal stability and ozone depleting potential (ODP), and, were analyzed to select suitable fluids for the study. High GWP fluids have been omitted from the analysis, such as hydro fluorocarbons (HCFCs) and high ozone depleting potential (ODP) fluids, such as chlorofluorocarbons (CFCs). The ODP was restricted to<1. The GWP was restricted to <150, as constrained by regulations such as that of the European Union [39]. For the ORC system, working fluids are known as dry, isentropic, and wet fluid. Due to high-quality vapour at the expander outlet, dry and isentropic work is better appropriate than the other form of fluid. The waste heat supply also has a low temperature in the current analysis. In this analysis, ultra-low GWP nine HFO working fluids such as R1234ze(Z), R134a, R1224yd(Z), R1225ye(Z), R1233zd (E), R1234yf, R1243zf, R1234ze(E), and R1336mzz(Z) were therefore considered for the ORC analysis because of these reasons and low temperature applications. Table 3 includes the thermal properties, protection and environmental data of these working fluids. For each refrigerant, the protection category designation has two or three numeric values (e.g., B1 or A2L). The first character indicates toxicity and the numeral indicates flammability, with or without a suffix letter. For toxicity, there are two classes: lower toxicity (Class A and higher toxicity (Class B). There are four flammability classes: 1, 2L, 2 or 3 [40].

3.4. Validation of the proposed model

Before the thermal analysis of the combined model, it is important to validate the combined model with previous study to ensure the correct use of the modeling equations. Current model was validated with the previous studies at same baseline conditions. First topping pre-compression sCO_2 cycle was validated with the previous studies [16,17] as shown in Table 4 while bottoming ORC was validated with the previous study Clemente et al. [47] as Table 5 displays. Final combined cycle was validated with the Khan and Mishra [26] at same reference baseline conditions which shown in Table 6. It was observed that thermal efficiency of the cycles varied very close to the corresponding to references. Therefore, the current model is valid to further study.

4. Economic analysis

The purpose of this section is to propose an alternative analysis of the working conditions of the ORC: instead of performance, the specific investment cost (SIC) expressed in €/KWe is the chosen objective function for this analysis:

$$\mathrm{SIC} = \frac{Cost}{\dot{W}_{\mathrm{net,ORC}}} \tag{26}$$

Since, by definition, WHR sources are cost-free, optimizing this parameter is equal to optimizing the system's profitability if annual costs for maintenance and insurance are ignored. A cost correlation is used for each part of the system to obtain the overall investment cost and is given in Table 7. The cost of the expander is based on the cost of the same swept volume of hermetic compressors. Previous publications have also shown that the conversion of volumetric compressors (such as scroll compressors) into expanders is feasible with good efficiency [50].

5. Results and discussion

Parametric analysis of the solar power tower integrated combined pre-compression sCO_2 cycle and ORC using ultra low GWP HFO working fluids in bottoming ORC is conducted in current study. A computer program was made in the engineering equation solver (EES) to simulate the combined cycle's performance based on the assumptions in this study. Thermodynamic properties at the main stations have been listed in Table 8.

5.1. Effect of solar irradiation on system performance

As the current combined model is powered by a solar power tower, the effects of solar irradiation on the performance of the system must therefore be investigated. With solar irradiation, the energy efficiency of the combined cycle has been continuously improved. The explanation here seems to be that the solar concentrator field efficiently utilizes increased solar irradiation. This contributes to an improvement in the combined cycle's inlet exergy [1,26]. As solar irradiation increases, it can also be clarified that the inlet of solar energy to the combined cycle also increases according to the Eq. (7). The rate of total exergy destruction with solar irradiation however was not affected. The exergy efficiency was then increased in accordance with the Eq. (22), as can be seen in Fig. 3. Exergy efficiency increased from 37.8% to 59.60% on the basis of R1336mzz(Z) working fluid as solar irradiation shifted from 400 to 950 W/m^2. Among other considered working fluids, R1234ze(Z) produced the lowest exergy efficiency. All HFO working fluids except R1234ze(Z) can be observed to perform better than the HFC working fluids such as R134a in this analysis.

As can be seen from Fig. 4, with solar irradiation, the thermal efficiency of the combined cycle has improved. Among the other selected working fluids, R1336mzz(Z) gave the highest thermal efficiency. As solar irradiation raised from 400 to 950 W/m^2, the highest thermal efficiency was enhanced by 58.77 per cent. Thermal efficiency is directly

Table 3
Thermophysical and environmental properties of organic working fluids [1,42].

Working substance	P_c (MPa)	T_c (°C)	T_b[a] (°C)	Weight (Kg/Kmole)	Type	ODP	GWP	Lifetime (years)	Security group
R1234ze(Z)	3.53	150.1	9.8	114.04	Isentropic	0	<10	–	–
R134a	4.059	101	−26.1	102.03	Isentropic	0	1430	14	A1
R1224yd(Z)	3.33	155.5	14	148.5	Isentropic	0.00023	0.88	–	A1
R1225ye(Z)	3.335	106.5	−20	130.5	Isentropic	0.00012	0.87	–	–
R1233zd(E)	3.57	165.5	18.32	130.5	Isentropic	0.00024	1	–	A1
R1234yf	4.597	94.7	−30	114.04	Isentropic	0	<1	–	A2L
R1243zf	3.518	104.44	−25.41	96.05	Dry	0	<1	–	A2
R1234ze(E)	3.64	109.4	−19.0	114.043	Dry	0	6	0.025	A2L
R1336mzz(Z)	2.903	171.3	33.4	164	Dry	0	8.9	0.0602	A1

[a] Tb is corresponding to the atmospheric pressure.

Table 4
Validation of toping pre-compression sCO_2 cycle.

Cycle	References	Baseline conditions in reference	Thermal efficiency		Estimated error
			Reference	Current model	
Pre-compression sCO_2	Kim et al. [16]	P_1 = 27.18 MPa,T_1 = 384.85 °C,P_6 = 8.82 MPa, T_6=36.85 °C, η_{MC}=0.85, η_{MT}=0.9	31.42%	31.6%	0.05%
Pre-compression sCO_2	Yu et al. [17]	P_1 = 25 MPa, T_1=500 °C,P_6 = 7.5 MPa, T_6=32 °C, η_{MC}=0.85, η_{MT}=0.88	37.88%	37.91%	0.07%

Table 5
Validation of the bottoming ORC.

Working fluids	References	Baseline conditions in reference	Thermal efficiency		Estimated error
			Reference	Current model	
isopentane	Clemente et al. [47]	P_1 = 3.023 MPa,T_1 = 184.1 °C,P_{11} = 0.1515 MPa, η_{OT}=0.6, η_{Pump}=0.5,	12%	12.1%	0.83%
R245fa	Clemente et al. [47]	P_1 = 3.395 MPa,T_1 = 154.2 °C,P_{11} = 0.2504 MPa, η_{OT}=0.6, η_{Pump}=0.5	11%	11.02%	0.18%

related to the exergy efficiency [29]. Therefore, as exergy efficiency improved, thermal efficiency also improved by solar irradiation according to Eq. (24).

Apart from this, power production of the combine cycle also increased with solar irradiation as shown in Fig. 5. Enhanced solar irradiation raised the enthalpy at the turbine inlet. This results in increased work in turbine production. This allows the overall output power to increase. At 950 W/m^2 of solar irradiation with R1336mzz(Z), the highest output power was calculated at 298.5 kW. Corresponding to this maximum condition the mass flow rate for the sCO_2 and ORC were obtained 1.5 and 2.5 kg/s respectively.

5.2. Effect of maximum cycle temperature on system performance

The maximum cycle temperature (MCT) (main turbine inlet temperature) is mainly dependent on the temperature of the molten salt. With the temperature of the molten salt, MCT improves. The increased temperature of the molten salt will however, increase the loss of heat in the solar receiver. This leads to lower efficiency of receivers [9]. In the present analysis, however, the solar field parameters are kept constant and shown in Table 1. Therefore the key goal is to analyze the combined output of the cycle. In Figs. 6–8, the effects of MCT on system performance are shown. Other input parameters have been kept constant while examining the effects of MCT, as shown in Table 1. Fig. 6–8 indicate that MCT has improved exergy, thermal efficiency and power output. The theory behind improved efficiency and power output is that as MCT raises the difference in enthalpy across the MT, leading to increased power output However, performance and output power are in a different trend as MCT increases, as seen in Figs. 6, 7 and 8. Highest

Table 7
Cost correlations for each component [51].

Component	Dependent variable	Cost [€]
Expanders	Volume flow rate ($\dot{V}_{net}$)	1.5·(225 + 170·$\dot{V}_{net}$)
Heat exchangers	Heat exchange area A (m^2)	190 + 310·A
Working fluid pump	Electrical power ($\dot{W}_{pp}$)	900·($\dot{W}_{pp}$/300)^0.25
Working fluid	Mass (M)	20·M
Miscellaneous hardware	–	300
Control system	–	500
Labor	Total cost component (TCC)	0.3·TCC

Table 8
Thermodynamic properties at main stations of combined cycle.

State points	Working fluids	Temperature (°C)	Pressure (MPa)	Enthalpy (kJ/kg)	Entropy (kJ/kg-K)
1	sCO_2	650	25	651.1	0.09878
2	sCO_2	465.3	5.435	440.1	0.3691
3	sCO_2	119.6	5.6	53.07	−0.5719
4	sCO_2	129.9	6.216	61.27	−0.5688
5	sCO_2	97.63	6.216	23.42	−0.6667
6	sCO_2	32	7.5	−144.1	−1.206
7	sCO_2	95.73	23.48	−106.2	−1.1916
8	sCO_2	112.5	25	−75.42	−1.117
9	sCO_2	360.1	25	287.7	−0.3755
10	R1234yf	69.88	3	297.8	1.312
11	R1234yf	30.69	0.1504	281.8	1.328
12	R1234yf	25	0.1504	232.5	1.115
13	R1234yf	26.43	3	235.5	1.118

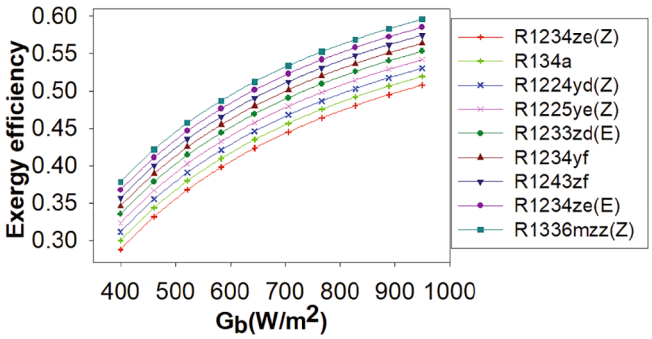


Fig. 3. Combined cycle's exergy efficiency variation with solar irradiation.

Table.6
Validation of combined pre-compression sCO_2 cycle and ORC.

Cycle	References	Baseline conditions in reference	Thermal efficiency		Estimated error
			Reference	Current model	
Pre-compression sCO_2 and ORC	Khan and Mishra [26]	P_1 = 25 MPa,T_1 = 650 °C,P_6 = 6.5 MPa,P_{10} = 3 MPa, T_6=32 °C, η_{MC}=0.85, η_{MT}=0.88	44.52%	44.53%	0.02%

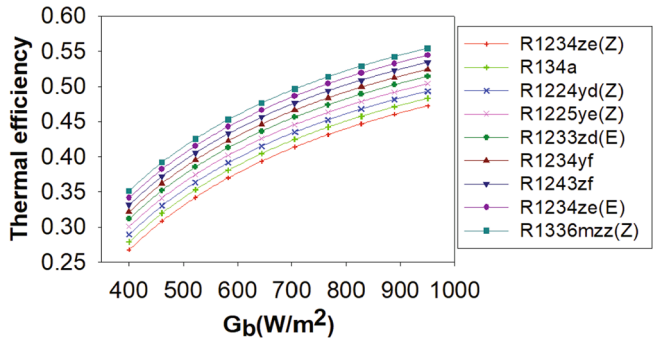


Fig. 4. Combined cycle's thermal efficiency variation with solar irradiation.

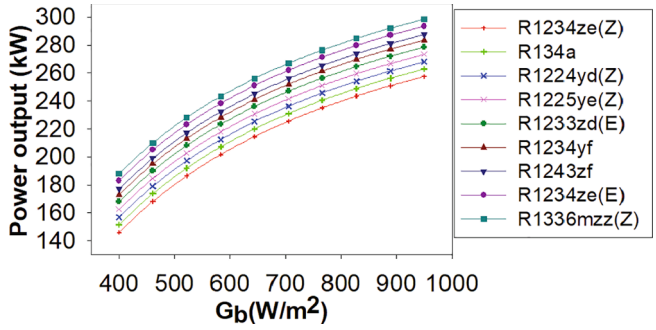


Fig. 5. Combined cycle's power output variation with solar irradiation.

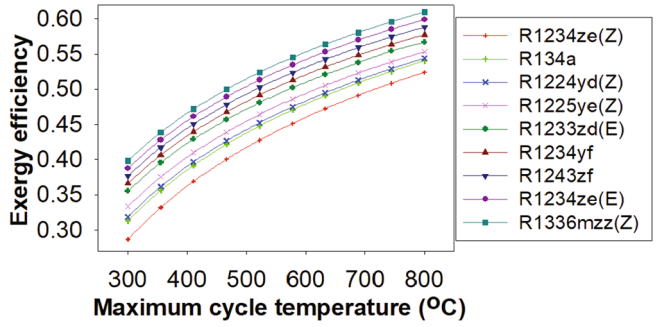


Fig. 6. Combined cycle's exergy efficiency variation with maximum cycle temperature.

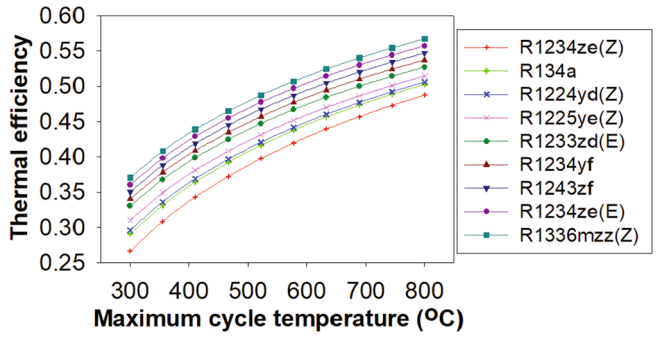


Fig. 7. Combined cycle's thermal efficiency variation with maximum cycle temperature.

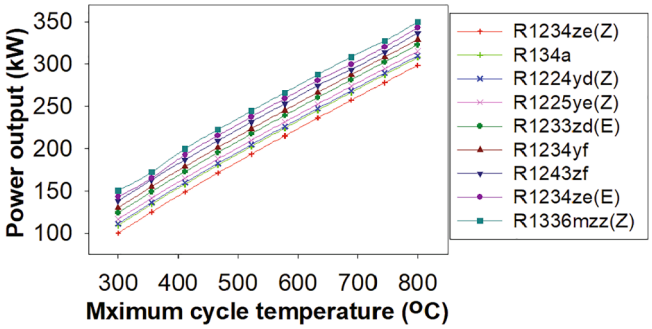


Fig. 8. Combined cycle's power output variation with maximum cycle temperature.

exergy and thermal efficiency and power output were achieved at 60.96%, 56.69 %and 349.7 kW respectively at 800 °C using R1336mzz (Z). Although R1234ze(Z) shows the lowest corresponding performance at 52.40, 48.74 and 298.4 kW respectively. All other working fluids output are between these two fluids. Efficiency and output power are therefore highest with higher input thermodynamic variables. However as a result of protection and material constraints, higher thermal efficiency cannot be achieved. Although the supercritical carbon dioxide cycle achieved greater efficiency at higher input turbine conditions than the superheated steam cycle at the same solar irradiation [43].

5.3. Effects of inlet pressure of main compressor on system performance

During in the examination of the influence of IPMC, MC inlet temperature, MT inlet pressure and temperature, PC inlet pressure were set at 32 °C, 25 MPa, 650 °C and 5.6 MPa respectively. It is found that the thermal and exergy efficiency increases with IPMC in the subcritical area where even the pressure is below the critical value of 7.38 MPa, thermal and exergy efficiency rise and then decline, beyond the critical pressure of CO_2. This implies that where net power output and efficiency have reached their highest values, there is an optimum pressure. The system has a bell-shaped curve for exergy and thermal efficiency and net output power, as can be seen in Figs. 9–11, respectively. The explanation behind this trend is that the highest CO_2 density contributes to lower compression power at critical pressure [44]. This corresponds to the maximum output of net power and therefore the highest efficiency of thermal and exergy. The highest thermal and exergy efficiency and output power were achieved by 58.33, 62.26% and 287.4 kW at an optimum pressure of 7.74 MPa using R1336mzz(Z). Figs. 11–13 also show that efficiencies and net output power have the same pattern due to the heat available at the combined cycle inlet being constant and only varying with the temperature of the solar field outlet [17]. It has also been shown that R1234ze(Z) has been investigated for the lowest thermal efficiency. It has also been found that the HFO working fluid works

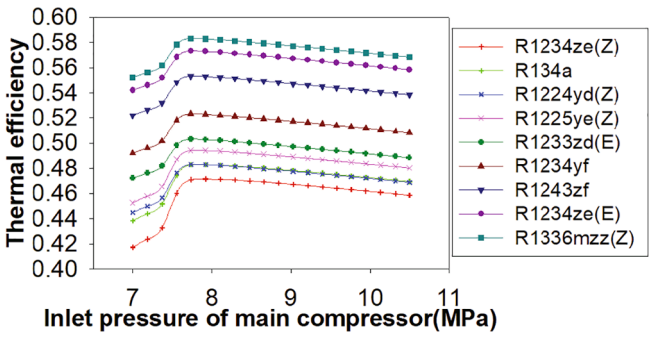


Fig. 9. Combined cycle's thermal efficiency variation with inlet pressure of main compressor.

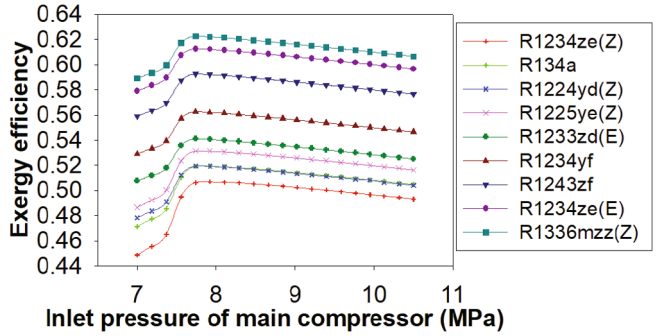


Fig. 10. Combined cycle's exergy efficiency variation with inlet pressure of main compressor.

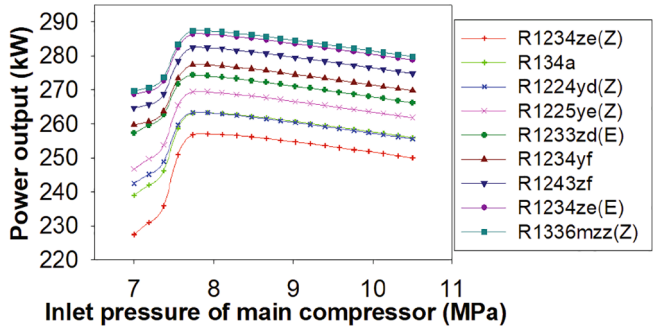


Fig. 11. Combined cycle's power output variation with inlet pressure of main compressor.

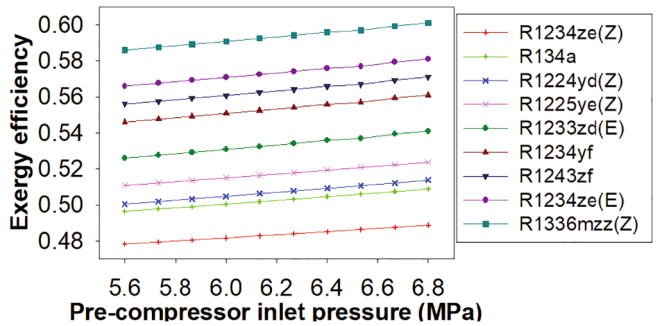


Fig. 12. Combined cycle's exergy efficiency variation with PCIP.

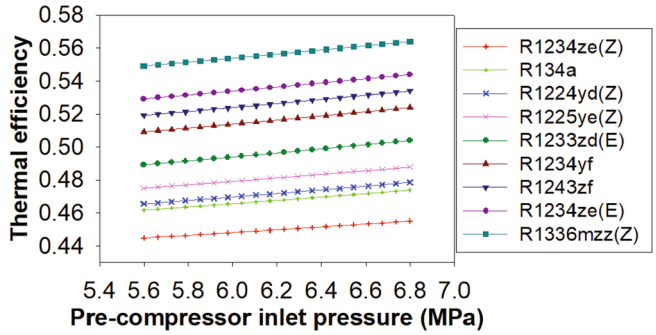


Fig. 13. Combined cycle's thermal efficiency variation with PCIP.

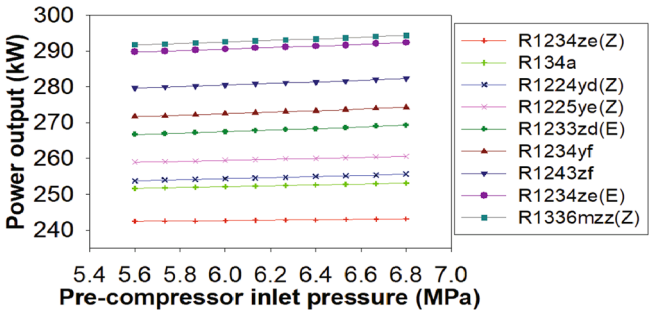


Fig. 14. Combined cycle's power output variation with PCIP.

better than the HFC working fluid.

5.4. Effect of inlet pressure of pre-compressor on system performance

Although the current study is all about investigating the thermal performance of the sCO_2 and ORC combined pre-compression cycle. The influence of the pre-compressor inlet pressure (PCIP) on the combined output of the system should therefore be investigated. Consequently, the impact of PCIP was analyzed to keep other parameters constant, as shown in Table 1. Between 5.6 and 6.8 MPa PCIP, the performance of the system was examined. Figs. 12–14 indicate that with PCIP, the combined system improved its exergy and thermal efficiency and net output power. Since the carbon dioxide density rises at low temperatures with a rise in pressure. This causes the input power of the PC to decrease. The combined cycle's net power output is therefore rising. This leads to increases in the system's thermal and exergy efficiency. It can be seen from Fig. 12 that exergy efficiency was enhanced from 58.6 to 60.1% using the R1336mzz(Z), while PCIP increased from 5.6 to 6.8 MPa. Although thermal efficiency and maximum net power output were expected to enhance from 54.92 to 56.4% and from 291.7 to 294.3 kW respectively, PCIP raised from 5.6 to 6.8 MPa with same R1336mzz(Z) working fluid as can be seen in Figs. 13 and 14 respectively.

Because of its thermo-physical properties, it can be shown that R1336mzz(Z) has the highest combined cycle efficiency among all of the selected working fluids. Its critical temperature and pressure are, as shown in Table 3, below the maximum ORC temperature. This implies that R1336mzz(Z) works in a super-critical environment. As is understood, when working fluid is near or above critical condition, the system provides the highest efficiency [45].

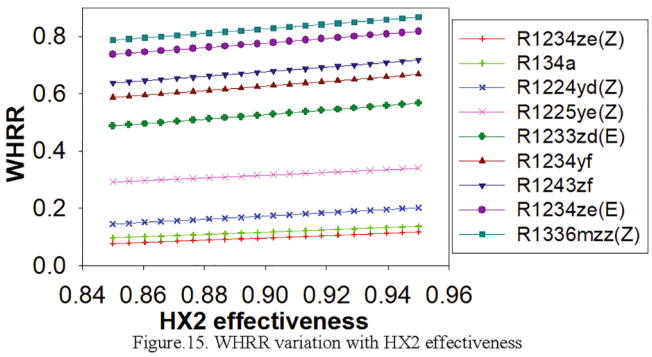


Figure.15. WHRR variation with HX2 effectiveness

Fig. 15. WHRR variation with HX2 effectiveness.

5.5. Effects of system variable on waste heat recovery ratio

The waste heat recovery ratio as defined above is the main parameter to define the performance of the bottoming cycle. It has already been identified in this research that the purpose of the present study is to add the bottoming cycle to the fundamental pre-compression cycle. It is therefore important to examine a parameter that could be used in the topping cycle to define the ORC application's utility. That's why the waste heat recovery ratio has been developed. Therefore, Fig. 15 indicates the variation of WHRR with HX2 effectiveness with various working fluids. With the effectiveness of HX2, WHRR continuously increased. The waste heat recovery in HX2 enhanced with effectiveness, as described in above section consequently greater enthalpy contributes to increased work output around the turbine at the inlet of the ORC turbine. According to the Eq. (25), WHRR increased as a result.

Using R1234ze(Z), R134a, R1224yd(Z), R1225ye(Z), R1234yf, R1234ze(E) and R1336mzz(Z) respectively, maximum WHRR at effectiveness 0.95 was found to be 0.1172, 0.1372, 0.2016, 0.3394, 0.5673, 0.6673, 0.7173, 0.8173 and 0.8673. It has also been noted that all working fluids have different rates of WHRR enhancement. This implies that choosing the correct working fluid for the bottoming ORC for the successful recovery of waste heat from the topping cycle is of major concern.

In addition, the effectiveness of LTR influences the WHRR in this segment. The effect of the LTR's effectiveness on the ORC's performance has also been investigated. With the effectiveness of the LTR, WHRR has been significantly increased, as shown in Fig. 16. The explanation behind this enhanced WHRR is that more heat is recovered by the sCO_2 cold stream when the effectiveness of the LTR is increased. This causes a decrease temperature of sCO_2 at the inlet of HX2. In other words, it is possible to assume that the low heat is at the inlet to the HX2. It decreases the ORC turbine's inlet temperature. As a consequence, in the event of organic working fluids, the lower inlet temperature of the ORC turbine raises the output power of the ORC turbine [46]. On the other hand, enthalpy also declines at state 5 due to lower heat at the inlet of HX2. As a consequence, with the LTR effectiveness according to the Eq. (25), WHRR increased. As a consequence of the study, the highest WHRR value was found to be 0.0997, 0.1197, 0.1775, 0.3352, 0.56, 0.66, 0.71, 0.8, 0.84 using R1234ze(Z), R134a, R1224yd(Z), R1225ye (Z), R1234yf, R1243zf, R1234ze(E) and R1336mzz(Z) respectively as shown in Fig. 16.

Further, results were compared with previous studies Kim et al. [16] and Yu et al. [17]. The thermal efficiency of the standalone pre-compression sCO_2 cycle in literature [16] and [17] were obtained 31.42% and 37.88% respectively at respective baseline conditions. However, in present study of combined cycle, after the integration of the ORC as the bottoming cycle in standalone pre-compression sCO_2 cycle, thermal efficiency was improved by 4.45% and 4.51% as compared to the Kim et al. [16] and Yu et al. [17] respectively at respective baseline conditions of literature. Further, combined cycle performance was compared with previous study Khan and Mishra [26], current system

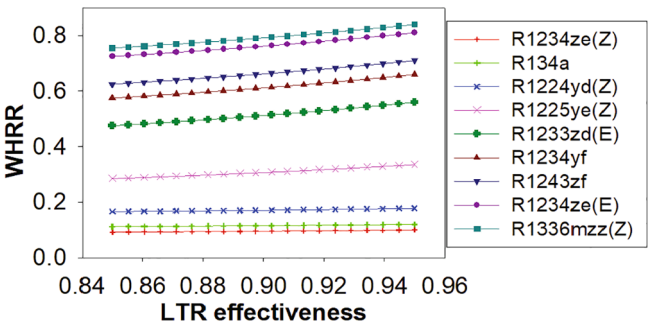


Fig. 16. WHRR variation with LTR effectiveness.

Table 9
Optimal results for the different working fluids corresponding to SIC in ORC.

Working fluids	HX2 temperature (°C)	η_{ORC}	$\dot{W}_{net,ORC}$(kW)	SIC [€/kWe]
R1234ze(Z)	150.1	11.09	41.43	2290
R134a	101	11.45	43.02	2280
R1224yd(Z)	155.5	12.01	43.34	2276
R1225ye(Z)	106.5	11.56	43.24	2265
R1233zd(E)	165.5	12.05	44.21	2254
R1234yf	94.7	11.12	40	2243
R1243zf	104.44	11.34	42.67	2233
R1234ze(E)	109.4	11.55	42.06	2245
R1336mzz(Z)	171.3	12.06	45.23	2234

achieved more thermal, exergy efficiency and power output with R1336mzz(Z) working fluid as compared to the previously R227ea. More waste heat was recovered by the current study with the HFO working fluid R1336mzz(Z) as compared to R227ea which was used in the previous study Khan and Mishra [26]. Therefore, R1336mzz(Z) can be selected as better fluid because it's lower GWP and zero ODP as compared to the R227ea.

5.6. Economic evaluation of different working fluids in ORC

For the cost analysis of the ORC, HX2 temperature, thermal efficiency of the ORC and net-work output of ORC are considered as independent parameters because affects the ORC performance. Optimal results for the different working fluids corresponding SIC in ORC have been listed in Table 9. It was found that for the particular working conditions selected for the work, both analysis lead to the selection of R1336mzz(Z) as best fluid. However, the "second-best fluid" differs for both analysis, as well as the next ones. Therefore, if the thermodynamic analysis can give a good idea of the best fluids, it won't necessarily lead to the selection of the optimal working fluid in terms of economical profitability.

6. Conclusions

Current study investigates the thermal and economic performance of the solar based sCO_2 and ORC using the ultra- low GWP HFO fluids. Results were obtained considering the maximum temperature and pressure of 650 °C and 25 MPa respectively and all other boundary conditions are listed in Table 1. Following conclusions were made from the results as discussed below.

1. Since current study deals with the performance investigation of SPT integrated combined system with the low GWP HFO fluids. Therefore, on the location of Mumbai (India), it was concluded that all ultra-low GWP HFO working fluids performs better than the R134 except R1234ze(Z) which shows the lowest thermal performance. R1336mzz(Z) shows the highest thermal performance among the all selected working fluids. Combined cycle's maximum thermal, exergy efficiency and power output were found using ultra low GWP HFO working fluid R1336mzz(Z) by 55.02, 59.6%, 298.5 kW at 950 W/m^2 of solar irradiation respectively.
2. Thermal efficiency and maximum net power output were expected to enhance from 54.92 to 56.4% and from 291.7 to 294.3 kW respectively, PCIP raised from 5.6 to 6.8 MPa with same R1336mzz(Z) working fluid. The R1336mzz(Z) performed better than other considered working fluids with present working conditions in current study.
3. Difference between WHRR for the different working fluids is much such as R1234ze(Z) and R1336mzz(Z) gave the 0.0997 and 0.84 respectively at 0.95 HX2 effectiveness. Highest waste heat was recovered by the HFO working fluid R1336mzz(Z). The highest WHRR value was found to be 0.0997, 0.1197, 0.1775, 0.3352, 0.56, 0.66,0.71, 0.8, 0.84 using R1234ze(Z), R134a, R1224yd(Z),

R1225ye(Z), R1234yf, R1243zf, R1234ze(E) and R1336mzz(Z) respectively. It means different working fluids shows different combined cycle performance. Also, careful design of HX2 is required to recover the waste heat efficiently.

4. Lowest and highest specific cost of invest were found for the R1336mzz(Z) and R1234ze(Z) respectively i.e. 2234 and 2290 €/kWe respectively. It was also found that thermodynamic analysis can give a good idea of the best fluids. It won't necessarily lead to the selection of the optimal working fluid in terms of economical profitability. Also current study is limited to the low temperature waste heat source, economic evolution of the SPT and sCO_2 cycle.
5. This study is beneficial in tropical areas such as India where solar irradiation are present almost throughout the year. This study developed environmental friendly power due to the application of SPT system and zero ozone depletion and low GWP working fluids.
6. Current study is limited to the thermo-economic analysis of the selected working fluids while exergoeconomic analysis is to be required in the future study. Also, detailed thermo-economic analysis of SPT system is to be performed with respect to combined power system in further study.

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.

References

[1] Y. Khan, R.S. Mishra, Parametric (exergy-energy) analysis of parabolic trough solar collector-driven combined partial heating supercritical CO_2 cycle and organic Rankine cycle, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects (2020), https://doi.org/10.1080/15567036.2020.1788676.

[2] L. Wei, Y. Zhang, Y. Mu, X. Yang, X. Hu, Simulation and experimental research of a low-grade energy conversion system using organic Rankine cycles, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 36 (5) (2014) 537–546.

[3] E. Kabira, P. Kumar, S. Kumar, A.A. Adelodund, K. Kime, Solar energy: Potential and future prospects, Renewable and Sustainable Energy Reviews 82 (2018) 894–900, https://doi.org/10.1016/j.rser.2017.09.094.

[4] F.A. Al-Sulaiman, Exergy analysis of parabolic trough solar collectors integrated with combined steam and organic Rankine cycles, Energy Conversion and Management 77 (2014) 441–449.

[5] Y. Tian, C.Y. Zhao, A review of solar collectors and thermal energy storage in solar thermal applications, Applied Energy 104 (2013) 538–553.

[6] S.A. Kalogirou, Solar thermal collectors and applications, Progress in Energy and Combustion Science 30 (3) (2004) 231–295.

[7] W. Fuqiang, C. Ziming, T. Jianyu, Y. Yuan, S. Yong, L. Linhua, Progress in concentrated solar power technology with parabolic trough collector system: A comprehensive review, Renewable and Sustainable Energy Reviews 79 (2017) 1314–1328, https://doi.org/10.1016/j.rser.2017.05.174.

[8] O. Farges, J. Bézian, M. El-Hafi, Global optimization of solar power tower systems using a Monte Carlo algorithm: Application to a redesign of the PS10 solar thermal power plant, Renewable Energy 119 (2018) 345–353, https://doi.org/10.1016/j.renene.2017.12.028.hal-01660563.

[9] S. Khatoon, M.-H. Kim, Performance analysis of carbon dioxide based combined power cycle for concentrating solar power, Energy Conversion and Management 205 (2020) 112416, https://doi.org/10.1016/j.enconman.2019.112416.

[10] A.S. Alsagri, A. Chiasson, M. Gadalla, Viability assessment of a concentrated solar power tower with a supercritical CO2 Brayton Cycle power plant, Journal of Solar Energy Engineering (2018), https://doi.org/10.1115/1.4043515.

[11] Y. Ma, T. Morozyuk, M. Liu, J. Yan, J. Liu, Optimal integration of recompression supercritical CO2 Brayton cycle with main compression intercooling in solar power tower system based on exergo-economic approach, Applied Energy 242 (2019) 1134–1154.

[12] A.K. Shukla, A. Sharma, M. Sharma, G. Nandan, Thermodynamic investigation of solar energy-based triple combined power cycle, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 41 (10) (2019) 1161–1179, https://doi.org/10.1080/15567036.2018.1544995.

[13] M. Abid, V.O. Adebayo, U. Atikol, Energetic and exegetic analysis of a novel multi-generation system using solar power tower, International Journal of Exergy 29 (2/3/4) (2019) 211, https://doi.org/10.1504/IJEX.2019.100364.

[14] Y. Ahn, S.J. Bae, M. Kim, S.K. Cho, S. Baik, J.I. Lee, J.E. Cha, Review of supercritical CO2 power cycle technology and current status of research and development, Nuclear Engineering Technology 47 (6) (2015) 647–661.

[15] S.M. Besarati, D.Y. Goswami, Analysis of advanced supercritical carbon dioxide power cycles with a bottoming cycle for concentrating solar power applications, Journal of Solar Energy Engineering 136 (2014), https://doi.org/10.1115/1.4025700, 010904-1-7.

[16] M.S. Kim, Y. Ahn, B. Kim, J.I. Lee, Study on the supercritical CO2 power cycles for landfill gas firing gas turbine bottoming cycle, Energy 111 (2016) 893–909.

[17] W. Yu, Q. Gong, D. Gao, G. Wang, H. Su, X. Li, Thermodynamic analysis of supercritical carbon dioxide cycle for internal combustion engine waste heat recovery, Processes 8 (2) (2020) 216, https://doi.org/10.3390/pr8020216.

[18] T. Neises, C. Turchi, Supercritical carbon dioxide power cycle design and configuration optimization to minimize levelized cost of energy of molten salt power towers operating at 650 °C, Solar Energy 181 (2019) 27–36.

[19] M. Kulhanek, V. Dostal. Supercritical carbon dioxide cycles thermodynamic analysis and comparison, Proc. scco2 Power Cycle Symp, 2011.

[20] F.A. Al-Sulaiman, M. Atif, Performance comparison of different supercritical carbon dioxide Brayton cycles integrated with a solar power tower, Energy 82 (2015) 61–71.

[21] V. Dostal, 2004. A supercritical carbon dioxide cycle for next generation nuclear reactors. Thesis submitted in Massachusetts institute of technology USA.

[22] G. Angelino, Carbon dioxide condensation cycles for power production. ASME Paper No. 68-GT-23, 1968.

[23] S. Shaaban, Analysis of an integrated solar combined cycle with steam and organic Rankine cycles as bottoming cycles, Energy Conversion and Management 126 (2016) 1003–1012.

[24] A.T. Hoang, Waste heat recovery from diesel engines based on Organic Rankine Cycle, Applied Energy 231 (2018) 138–166.

[25] J. Song, X.-S. Li, X.-D. Ren, C.-W. Gu, Performance analysis and parametric optimization of supercritical carbon dioxide (S-CO2) cycle with bottoming Organic Rankine Cycle (ORC), Energy 143 (2018) 406–416.

[26] Y. Khan, R.S. Mishra, Performance evaluation of solar-based combined pre-compression supercritical CO2 cycle and organic Rankine cycle, International Journal of Green Energy 18 (2) (2021) 172–186, https://doi.org/10.1080/15435075.2020.1847115.

[27] H. Yu, X. Feng, Y. Wang, Working fluid selection for organic Rankine cycle (ORC) considering the characteristics of waste heat sources, Industrial and Engineering Chemistry Research 55 (5) (2016) 1309–1321, https://doi.org/10.1021/acs.iecr.5b02277.

[28] M.A. Reyes-Belmonte, A. Sebastián, M. Romero. Optimization of a recompression supercritical carbon dioxide cycle for an innovative central receiver solar power plant. Energy 112 (2016):17–27.

[29] Y.A. Cengel, M.A. Boles, Thermodynamics an Engineering Approach, 5th edition, McGraw-Hill publication. New York, USA, 2004.

[30] R. Petela, Exergy analysis of the solar cylindrical-parabolic cooker, Solar Energy 79 (3) (2005) 221–233.

[31] Y.M. Kim, D.G. Shin, C.G. Kim, G.B. Cho, Single-loop organic Rankine cycles for engine waste heat recovery using both low- and high-temperature heat sources, Energy 96 (2016) 482–494.

[32] S.A. Klein, Engineering Equation Solver (EES), Academic Commercial V7.714. F-Chart Software, 2020, www.fChart.com.

[33] A. Bejan, D.W. Kearney, F. Kreith, Second law analysis and synthesis of solar collector systems, Journal of Solar Energy Engineering Transactions ASME 103 (1981) 23–28.

[34] C.K. Ho, B.D. Iverson, Review of high-temperature central receiver designs for concentrating solar power, Renewable Sustainable Energy Reviews 29 (2014) 835–846.

[35] X. Wang, Q. Liu, J. Lei, W. Han, H. Jin, Investigation of thermodynamic performances for two-stage recompression supercritical CO2 Brayton cycle with high temperature thermal energy storage system, Energy Conversion Management 165 (2018) 477–487.

[36] T. Neises, C. Turchi, A comparison of supercritical carbon dioxide power cycle configurations with an emphasis on CSP applications, Energy Procedia 49 (2014) 1187–1196.

[37] S. Polimeni, M. Binotti, L. Moretti, G. Manzolini, Comparison of sodium and KCl-MgC_{l2} as heat transfer fluids in CSP solar tower with sCO_2 power cycles, Solar Energy 162 (2018) 510–524.

[38] Y. Koc, H. Yagli, A. Koc, Exergy analysis and performance improvement of a subcritical/supercritical Organic Rankine Cycle (ORC) for exhaust gas waste heat recovery in a biogas fuelled Combined Heat and Power (CHP) 520 engine through the use of regeneration, Energies 12 (4) (2019) 575, https://doi.org/10.3390/en12040575.

[39] F. Moloney, E. Almatrafi, D.Y. Goswami, Working fluids parametric analysis for the regenerative supercritical organic Rankine cycle for medium geothermal reservoir temperatures, Energy Procedia 129 (2017) 599–606.

[40] J.M. Calm, Referegerant safety, ASHRAE Journal 367 (1994) 17–26.

[41] X. Xu, X. Wang, P. Li, Y. Li, Q. Hao, B.o. Xiao, H. Elsentriecy, D. Gervasio, Experimental test of properties of KCl–MgCl2 eutectic molten salt for heat transfer and thermal storage fluid in concentrated solar power systems, Journal of Solar Energy Engineering 140 (5) (2018), https://doi.org/10.1115/1.4040065.

[42] J. Navarro-Esbrí, F. Molés, B. Peris, A. Mota-Babiloni, K. Kontomaris, F. Molés, B. Peris, Mota-Babiloni Adriá, Experimental study of an Organic Rankine Cycle with HFO-1336mzz-Z as a low global warming potential working fluid for micro-scale low temperature applications, Energy 133 (2017) 79–89, https://doi.org/10.1016/j.energy.2017.05.092.

[43] R. Chacartegui, J.M. Muñoz, D. Escalona, D. Sánchez, B. Monje, T. Sánchez, Alternative cycles based on carbon dioxide for central receiver solar power plants, Applied Thermal Engineering 31 (2011) 872–879.

[44] M. Blanco, L.R. Santigosa. Advances in Concentrating Solar Thermal Research and Technology, Woodhead publishing series in energy (Elsevier), 2017. ISBN: 978-0-08-100516-3.
[45] C. Kalra, G. Becquin, J. Jackson, A. L. Laursen, H. Chen, K. Myers, A. Hardy, H. Klockow, J. Zia, High-potential power cycles & working fluids for next generation binary supercritical organic Rankine cycle for enhanced geothermal systems. Proceedings, Thirty-Seventh Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, 2012.
[46] Y. Dai, J. Wang, L. Gao, Parametric optimization and comparative study of organic Rankine cycle (ORC) for low grade waste heat recovery, Energy Conversion and Management 50 (3) (2009) 576–582.
[47] S. Clemente, D. Micheli, M. Reini, R. Taccani, Bottoming organic Rankine cycle for a small scale gas turbine: a comparison of different solutions, Applied Energy 106 (2013) 355–364.
[48] H. Singh, R.S. Mishra, Performance evaluation of the supercritical organic Rankine Cycle (SORC) integrated with large scale solar parabolic trough collector (SPTC) system: An exergy energy analysis, Environmental Progress & Sustainable Energy 37 (2) (2018) 891–899, https://doi.org/10.1002/ep.v37.210.1002/ep.12735.
[49] M. Abid, M.S. Khan, T.A.H. Ratlamwala, Comparative energy, exergy and exergo-economic analysis of solar driven supercritical carbon dioxide power and hydrogen generation cycle, International Journal of Hydrogen Energy 45 (9) (2020) 5653–5667.
[50] V. Lemort, S. Quoilin, C. Cuevas, J. Lebrun, Testing and modeling a scroll expander integrated into an Organic Rankine Cycle, Applied Thermal Engineering 29 (2009) 3094e3102.
[51] S. Quoilin, S. Declaye, B.F. Tchanche, V. Lemort, Thermo-economic optimization of waste heat recovery Organic Rankine Cycles, Applied Thermal Engineering 31 (14-15) (2011) 2885–2893.

Yunis Khan is a research scholar in Delhi Technological University, department of mechanical engineering, New Delhi India. His area of interest is waste heat recovery units, solar driven thermal power plant, refrigeration, heat transfer, co-generation and tri-generation systems.

Dr. Radhey Shyam Mishra is a professor and former HOD in Delhi Technological University (DTU), department of mechanical engineering New India since 1997. He completed his PhD from Indian Institute of Technology Delhi in solar energy. His area of interest is thermal power plant, solar energy, refrigeration and air conditioning, heat transfer, stir welding, alternative fluids, CFD.

ACADEMIA
Accelerating the world's research.

TRANSFORMING MANUFACTURING PROCESSES WITH THE HELP OF AI

IAEME Publication

IAEME PUBLICATION

Cite this paper

Downloaded from Academia.edu

Get the citation in MLA, APA, or Chicago styles

Related papers

Download a PDF Pack of the best related papers

EDUCATION AND THE FOURTH INDUSTRIAL REVOLUTION: CHALLENGES AND POSSIBILITIES F...
IAEME Publication

DIGITAL TRANSFORMATION AND ITS RELATIONSHIP WITH CUSTOMER EXPERIENCE: A COMPREHENSI...
IAEME Publication

AUTOMOTIVE SERVICE INDUSTRY NEEDS OF AUTOTRONIC COMPETENCIES AND THOSE PREPARED B...
IAEME Publication

International Journal of Mechanical Engineering and Technology (IJMET)
Volume 11, Issue 8, August 2020, pp. 8-20, Article ID: IJMET_11_08_002
Available online at http://www.iaeme.com/ijmet/issues.asp?JType=IJMET&VType=11&IType=8
Journal Impact Factor (2020): 11.2184 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6340 and ISSN Online: 0976-6359
© IAEME Publication

TRANSFORMING MANUFACTURING PROCESSES WITH THE HELP OF AI

Keshav Aggarwal
Department of Production & Industrial Engineering,
Delhi Technological University, Delhi India

Shaswat Garg
Department of Mechanical Engineering,
Delhi Technological University, Delhi India

Prakash Budhaulia
Department of Mechanical Engineering,
Delhi Technological University, Delhi India

ABSTRACT

In manufacturing businesses, productivity and cost savings are two important parameters. Technology ease out the load, increase efficiency and mitigate the financial pressure. Modern manufacturing is characterized by high level of automation and technology-driven processes. This digitization of manufacturing has given rise to Industry 4.0 – where cyber-systems, connectivity and data are the key. In this paper we attempt to highlight how each step in the manufacturing process can be changed and made more efficient with the help of technology. To reinforce this fact, examples of various companies have been mentioned which have transformed their supply chains with the help of technology. To better understand this transforming phase in manufacturing industry, we analyzed the role of artificial intelligence in industries, the major issue in implementation of artificial intelligence in manufacturing units, possible ways to overcome these issues and what are the different tasks that can be handled easily with implementation of AI in manufacturing units. With the help of this paper, the reader will get a holistic view of the transforming manufacturing trend and its associated applications in the real world. Certain challenges in the current fourth industrial revolution of manufacturing have also been identified upon which actions can be taken to further improve productivity, cost savings and efficiency.

Keywords: Artificial Intelligence, Manufacturing Industry, Smart Manufacturing, Supply Chain Management, Automation, Industry 4.0.

Cite this Article: Keshav Aggarwal, Shaswat Garg and Prakash Budhaulia, Transforming Manufacturing Processes with the Help of AI. *International Journal of Mechanical Engineering and Technology.* 11(8), 2020, pp. 8-20.
http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=11&IType=8

1. INTRODUCTION

In the age of globalization, the demands for unique and customized products by the customers are pushing the manufacturing industries to rely on automation and machine learning. This has led to the beginning of the fourth industrial revolution, the so-called Industry 4.0. It all began with the first industrial revolution in which transition was made from hand processes to the use of machines specifically from steam power, followed by mass production in second. Industry 4.0 inherits from the third which revolutionized the world with the discovery of computers.

Industry 4.0 makes it possible to gather and exchange data across machines which leads to production of high-quality goods at cheap rates more efficiently, flexibly and quickly. Computers now interconnected with each other can make decisions without any human intervention. Industry 4.0 is acting as a bridge between the digital and physical world. 4IR has a salient role to play towards sustainable growth of small and mid-size industries.

The technologies which form the building blocks of industry 4.0 include cyber-physical systems (CPS), the internet of things (IoT), industrial internet of things (IIOT), cloud computing, cognitive computing and artificial intelligence. Smart manufacturing, light out factories and industrial internet of things form the foundation of industry 4.0. But of all, artificial intelligence is favoured by the innovative manufactures to tackle many of the challenges.

1.1. Role of AI in Manufacturing and Automation

AI is basically the technology that allows the computer to think and reason like human beings. AI finds its roots in robotics, computer vision, natural language processing, cloud computing etc. Integration of AI in manufacturing can improve efficiency, flexibility, reduce errors and enable real time optimizing operations. A BCG analysis found that use of AI can reduce producers' conversion costs by up to 20%, with up to 70% of the cost reduction

resulting from higher workforce productivity. It allows the industries to meet the specific demands of the customers and deliver them at a much shorter time.

AI is also extensively used to bridge the distance between research and implementation by–

Harnessing useful data: With the help of AI industries are able to process huge pile of data which earlier was impossible to analyse. This helps them to improve customer service and be ahead in fast growing world.

Post-production support: AI allows 24/7 connected service that allows industries to show the customers how their products are to be used and also helps them to detect defects.

Quality Check: Industries use machine vision which make use of AI along with high resolution cameras to detect defects in products like circuit boards which are beyond human vision to detect.

Maintenance: smart factories are making use of AI to detect issues before they arise by tracking and processing the entire manufacturing process.

2. LITERATURE REVIEW

A considerable number of studies have investigated the Transformation AI is bringing in the manufacturing industry and the challenges that lie ahead.

Nancey Green Leigh [1] reviewed contributions of AI to the current jobs, their limitations and offered a modified approach. She came to a conclusion that since the Great recession ended, robots improved the employment in manufacturing industry in USA at the metropolitan level

Bo-hu LI [2] analysed the rapid development of core technologies in the new era of 'Internet plus AI' which is triggering development in the manufacturing industry. He proposed new models, means and forms of intelligent manufacturing, intelligent manufacturing system architecture based on the integration of AI technology with information communications, manufacturing, and related product technology.

Sameer Mittal [3] presented the building blocks frequently using for SM adoption. He found that for the manufacturers that are inclined toward adopting SM intelligent control, data analytics, IT-based production management systems (e.g., ERP, CAx, etc.), smart products/parts/materials are the most commonly used building blocks.

S.P. Leo Kumar [4] reviewed the applications of AI in Computer Aided Process Planning (CAPP) and manufacturing. He addressed the present review work by analysing existing review articles. Finally, they identified the research gaps ad presented the possible future research directions.

Kamran Mahroof [5] researched on the drawbacks and opportunities of AI within the warehouse of a major retailer. His findings revealed the challenges resulting from the shortage of both skill and mind-set of operational management.

Gideon Cohen [6] described the knowledge-base and inference-engine which were developed to determine various shop-floor operational strategies to achieve maximum efficiency of machines.

Simon Hagemann [7] presented a novel AI methodology capable of generating initial production configurations base. He concluded that to ensure high efficiency in the automotive industry AI techniques are vital.

Helen H. Lou [8] introduced a proactive Quality Control approach by resorting artificial intelligence and engineering fundamental. They came to the conclusion that Proactive QC is becoming more and more attractive in the manufacturing industries due to increased economic and environmental pressure.

S. Gallego García [9] reviewed the impact of new capabilities due I4.0 on production and maintenance management tasks in order to provide an impulse to current manufacturing companies.

Yuval Cohen [10] reviewed the challenges faced by the manufacturing and assembly systems and their readiness to adopt Industry 4.0 and. It also focuses on the most promising future developments and evolution of such production systems.

Pinar Urhal [11] presented and discussed the leading additive manufacturing methods that can be used with a robotic system. They came to the conclusion that current additive manufacturing systems have great potential to reduce time to market, increasing customisation, widening the design option compared to traditional methods.

Di Wu [12] tried to understand what knowledge is and what kinds of knowledge can be obtained and applied in a design. They came to the conclusion that to tackle challenging optimization problems and to further improve the efficacy of design optimization, it seems imperative to let go the assumption of black-box functions, but rather incorporating existing knowledge in optimization.

Max Busch [13] studied that various complexity in production planning and control, can be solved with the help of techniques of AI and digitization of manufacturing. For this he proposed a model named as maturity model. The maturity model is based on results from the research project WerkPriMa.

Mike Keen [14] reviewed that AI is helpful in manufacturing by improving quality, reducing cost, and to speed up design and manufacturing process. This can be done with the

help of Quality Design Expert System (QDES), Technical Information Engineering System (TIES), Vibrational Analysis in Rotating Machines, Trailer configure.

Stephan Matzka [15] reviewed that quality prediction can be done in manufacturing process using machine learning process. This can be using AI algorithm (in case of screw fastening process is determined using a torque versus rotation angle (or time diagram).

Ge Meng [16] researched on e-intelligence maintenance system for manufacturing industrial theory which is in progressing, the NET framework is progressing and the Web Services which is based on XML is applied widely.

Julian SK Tan [17] researched on how quality analysis can be achieved in big data supply chain. Data is collected using HIGH PERFORMANCE COMPUTING centers and use them to distribute HADOOP installations and how IoT devices at the point of data collection can be used to make fastening the process of analytics and advanced it.

Yifang Ma [18] researched on how artificial intelligence can be used in development of autonomous vehicles. This can be done by using various methods like big data, HPC, and high-resolution digital map for enhanced data collection and processing AR/VR which can be used to enhanced the process of constructing accelerated test scenarios, 5G for low latency connections among AV's.

B.F. Bashir, S.J. Steiner [19] researched on how an AI based simulation system to improve shop-floor efficiency and to reduce production cost this is done by using a method which has four sub systems. The four sub-systems of the model are as indicated: User interface, Process Routing, Simulation, and Control. Information arrives into the system, which can be in the form of product demand and/or feedback information.

Wesam Salah Alaloul, M.S. Liew [20] researched how industry 4.0 concept can be used in construction industry. Technologies such as BIM, cloud computing, and modularization have helped significantly in developing technologies such as augmented, virtual and mixed reality which can be used to influence the sustainability of the industry.

Roland W. Scholz [21] researched and explained how digital threats can be controlled. This can be done through a multi-level actor analysis, identifying strengths and weaknesses of organizations, constructing digital threat scenarios and provides judgment-based expert assessments on the organization's vulnerability

R. Satish [22] researched on how technology strategy can be used to improve safety management. Sensors collect vast amounts of data, and that information is then interpreted by artificial intelligence (AI). This optimizes the production line and creates synergies across the facility and provides a great thrust on Safety.

Hong Seok Park [23] researched on product quality. They used AI based on injection Moulding process in plastic industry. This can be done using real time AI based control of process parameters in injection moulding cycle.

T. Santos [24] researched how technology can be used for maintenance management in the food industry. It can be done with the help of the application of a methodology of Reliability Centred Maintenance (RCM), revision of Preventive Maintenance Plans (PMP), integration of critical asset in Computerized Maintenance Management System (CMMS) and many other.

Sidharth Baskaran [25] researched how siemens process can be used in automotive assembly line. A digital twin of a mobile robot designed to provide physical assistance to the human to perform the assembly operation is introduced.

Robert Joppen [26] researched on the optimization of production management process. It describes how the conflict of objectives can be analysed systematically and a reasonable operational status can be derived.

Muhammed Zekeriya Gunduz [27] researched cyber security, threats and potential solutions. The smart grid is one of the most significant applications of the Internet of Things (IoT) based smart grid systems are critical infrastructures, also they have complex architectures and include critical devices.

3. REALITY OF AI IN INDUSTRY

3.1. Gap between Ambition and Reality

Industries all around the world are looking forward to make use of AI in manufacturing units increase productivity and better product quality. Countries like China, India and Singapore on average are looking more forward to implement AI in industries as shown in Fig (1)[28].

Figure 1 Gap Existing between AI Ambition and Reality

Many companies know the benefits of AI in industries and are planning to bring this in the production factories but do not have proper roadmap to implement this. They do not have a proper planning, investments and implementation to meet up their ambitions. According to a survey 87% of companies want to implement AI in their industries within short coming years but, only 28% take a proper planning to implement it, while 72% do not have a proper planning on how to implement it while 27% have rough ideas on how to implement it while 13% do not consider AI importance in factories. Gap between ambition and reality in various industries has been depicted in Fig (2)[28].

Thus, there need to be a definite and effective planning to implement AI in industries. In past years only 50% of companies have successfully achieved AI in their industry while only 16% have adopted more than one AI cases in their industries. Surveys in 12 countries show that companies that early implement AI in their factories are 25% in United States of America, 23% in China, 19% in India,11% in Japan, 10% in Singapore, 10% in France, and in Germany 15%.

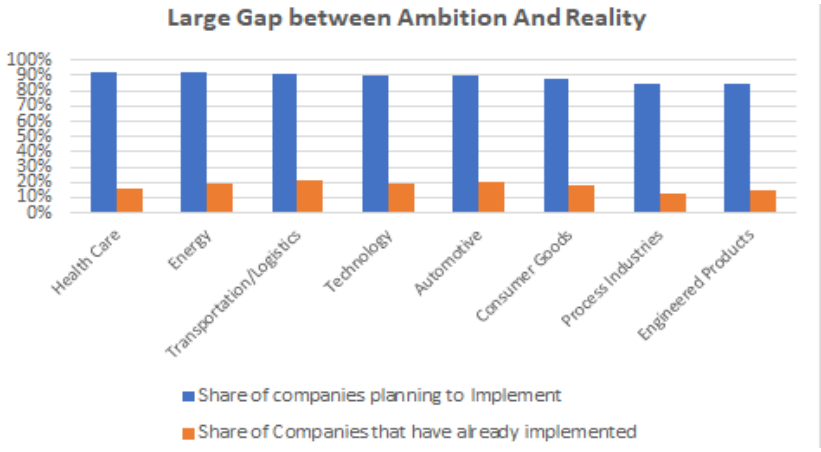


Figure 2 Gap Existing between AI Ambition and Reality In various sectors

3.2. Closing this Gap

Since the introduction of AI, every industry wants to adopt it and increase their effectiveness of productivity but small industries do take time to adopt it, due to high setup cost and not having proper roadmap how to implement it in their industry. Due to new technologies and development it has become easy to adopt this technology in manufacturing unit. There are some measures which can help in implementing AI in industry.

Strategies should be made on how to generate investments at each level and in which particular field of business AI should be adopted which result in ease of manufacturing. Company should have good employee competencies to handle the skills required work with AI, they should be skillful in data management, analytics and other competencies. For such skill employee can get it within the company or outside the company. From data around one in four (29%) of companies have increased the number of employees dedicated to AI and nearly half 47% will do the same in coming years. Testing should be done at small scale and then be implemented at large practices for successful and effective use of AI in industry. Among the different technologies of artificial intelligence, the one which are being usedthe most are extensive use of computer vision cameras which are used in machine inspection, implementation of Industrial Internet of Thing (IIoT), and the use of big data. With thehelp of other technologies too industries are revolutionizing.

Developments are done to improve the technologies in AI and make it more cost effective, more accurate and more productive. The upcoming technology will help in increasing production, bringing out the best quality for a product and helping in creating a better and safe environment of work.

4. IMPLEMENTATION OF AI IN SUPPLY CHAIN

Today AI is more accessible than ever. The manufacturing industry is beginning to embrace AI in their products and services. AI is finding its roots in all the segments of Supply chain like Product R&D, Resource Planning etc. Ways through which AI is transforming various sector of Supply chain are listed below-

4.1. Product R&D

AI is transforming the research and development (R&D) by accelerating the development of new products and services by continuous discovery, accessibility and application of data to improve the standard of living. Any company that innovates through R&D are bound to lead the market and even may monopolize it.

AI can transform the R&D process by accelerating the development of product by identifying the proper materials for the finished product, identifying any possible failure in parts in the near future, accelerating product data exploration.

4.2. Resource Planning and Sourcing-

The manufacturing industry is highly competitive and dynamic which makes it crucial to integrate it with integrated solution which can reduce cost, increase efficiency, sales and profitability but most importantly allowing the company to make accurate, informed and strategic decisions. This is where enterprise resource planning (ERP) steps in. It consists of hardware and software system working together to supervise the success of design and production of products followed by delivery to end user.

Integrating AI with ERP enable companies to improve their operational efficiency by using machine learning and natural language interface. SAP Leonardo is good example of AI driven ERP system which includes several microservices integrated with a cloud platform. AI improves ERP in the following ways [29]-

Advance Analytics: The ability of AI to work with massive amount of data allows real time, accurate data insights. For example- by analysing the behaviour of different type of customers can allow you to manufacture products and services that is meant for a certain audience.

Forecasting: By processing the old data AI can make predictions about the future.AI along with reducing the cost of forecasting can also make forecasting accurate. It reduces the risk of overproduction or underproduction when you manufacture the accurate number of products [30].

Financial Management: AI can improve transaction-processing efficiency, verify reports with accuracy and can also predict future financial results based on trends and market data [31].

Customer Service: AI based on performance evaluation and qualification of employees can assist with planning and schedule service calls.

Production processes: ERP integrated with AI can be used to detect inefficient processes and suggest ways to improve it thereby cutting cost and minimizing the wastage of resources.

4.3. Machining, Production, & Assembly

As the world have changed and so has manufacturing. As the standards of living are improving so is the cost of converting raw material to products. With the growing expectations of the customers for personalization and flexibility AI is poised to play a critical role. Therefore, to meet the needs of the customers cyber physical systems like industrial robots and 3D printers are becoming common in the modern industry.

Cobots (Collaborative robots) are most used. They learn first by being moved manually and then by copying the movement. They are collaborative because they work alongside humans. Airplane manufacture Airbus uses mobile robots which work alongside humans to drill thousands of holes into passenger jets. A new wave of technology called the Intelligent Edge System is taking the manufacturing industry by storm. It is a fast, real time control system for network linked robot and machine tools. By using deep learning AI technology these systems will learn and become smarter as linked machines manufacture products.

3D printing is a game changing technology constantly evolving itself. Combination of AI with 3D printing is now widely used by companies for manufacturing. AI can be used to create more efficient CAD models. Autodesk developed Dream catcher a tool allowing to work on regenerative designs. Hundreds of designs can be created within hours using this software. Also, AI can be used to detect defects in 3D model which could make it non-

printable. To spot microscopic cracks in machine parts General Electric's GE labs in New York began to developed a computer vision technology.

Therefore, by incorporating AI in assembly line can reduce error, make manufacturing efficient and improve the quality of product.

4.4. Quality Assurance

In the fast-evolving market the companies which can deliver quality products faster than its competitors have significant advantage over others. At the same time complexity and proneness to error has massively increased. Therefore, QA must evolve to meet the demands of the customers while maintaining the quality of service. AI can improve QA by imagining assembly lines as interconnected, data driven and autonomous networks. If even a slight deviation from the original value is detected by the sensors which they send to the users so that the issue can be dealt. The system can also learn as the time goes on and therefore making changes on its own.

Using QA with autonomous, self-learning AI can Spot early problems in production, identify slowdowns and potential failure before they occur, save resources and time thereby improving efficiency.

4.5. Warehousing

In the last few years, the warehouse industry has seen a lot of automation. By the year 2020 AI might be transforming the warehouse operations completely. Smart warehouse forms an ecosystem of interconnected technologies where AI governs the entire operation, from supply to delivery. To reduce error in warehousing, AI can be used to identify and sort, process, package goods and pull them for shipment [34].

Ways through which warehousing is expected to be affected by AI are [35][36]-

Communication: For smooth functioning of day-to-day jobs the workers should be able to communicate in an efficient manner in order to achieve targets and fulfil the task successfully. With the use of AI algorithms, the systems can communicate at a much faster rate than humans. By the means of Wireless, I-O Link and Fieldbus data communications all the elements of the system can communicate with each other.

Warehouse Logistics: AI can help in optimizing in logistics by calculating the number of packets needed to be handled on a given day and how much equipment is required. Earlier it all relied on variables such as operator skill level and SKU (stock keeping units). Introduction of machine learning algorithm has helped in stock movement forecasting and management of material. In this way operator error and processing time can be reduced thereby increasing productivity and efficiency.

Inventory: With the help of AI, money which was previously spent on inventory can now be invested in other business growth opportunities. Bar code scanners are now replaced by RFID (Radio frequency identification) for organization and control of inventory. Using AI speed and volume of order to be processed can be adjusted, therefore increasing the overall productivity.

Robots: Another way in which AI is transforming warehouse management is with the use of robots. They can carry heavy goods and transport them to the desired location in fraction of time as compared to human operatives. With the help of machine learning algorithms bots can efficiently pack the products efficiently, select the most efficient picking and slotting routes.

Ocado, a British online supermarket, designs highly automated warehouses. Crates are stacked in piles with the position of each decides by an algorithm- rare purchases are placed

on the bottom of the stack. Crates are grabbed by bots which manoeuvre around the warehouse for 24 hours while humans pack and unpack the crates.

4.6. Labour Augmentation & Management-

Since the birth of AI there has been a fear in the hearts of people "what effect AI will have on the job market". But rather than consuming jobs AI will improve the quality of skillset of workers. Impact of AI on jobs is depicted in Fig (3). A 2018 study by the accountancygiant PwC forecasted that machines would create as many new jobs in the UK as they destroyed over the next 20 years – although it said there would be "winners and losers" by industry sector, and many roles were likely to change [32]. Thus, in the future rather than AI taking jobs, humans and AI will work in unison.

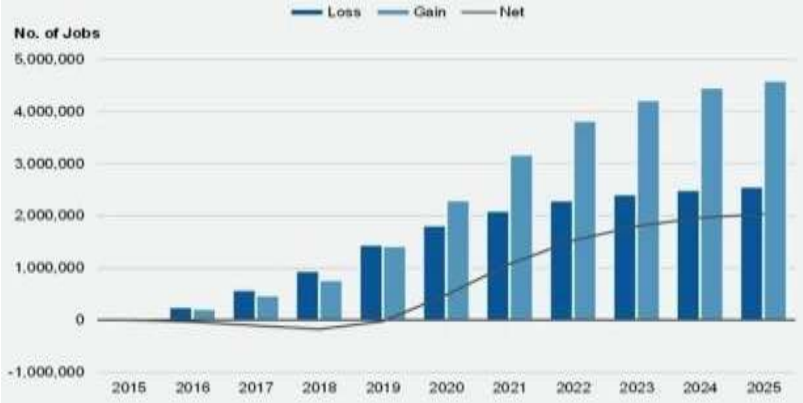


Figure 3 The impact of AI Automation on jobs

A very good example is of the exoskeletons. By developing a supportive techno-carapace, they have empowered people to lift weights and achieve performances that were initially unthinkable. In addition, chatbots are widely used by companies for answering customers queries and scheduling meeting [33].

4.7. Transportation and Logistics

This sector forms an integral part of manufacturing and AI is finding its roots to revolutionize this sector. Autonomous trucks are highly in demand for transportation of goods from suppliers to producers. Developments are done in designing such a system where truck can trace each other mark their presence with the sensors inbuilt in them and automate the acceleration and braking system, which increases the safety for the driver as well of the products from getting being damaged from accidents.

5. EXAMPLES OF INDUSTRIES WHERE AI IS IMPLEMENTED

In march 2016, Siemens launched Mindsphere. Mindsphere is a smart cloud for industry which allows machine manufacturers to monitor machines for fleets for service purposes throughout the world.

Lamborghini recently established a new production facility in Sant 'Agata Bolognese with an assembly line dedicated in Urus, a finishing department for all Lamborghini models and new office. Integration of new production technologies supports the collaboration of humans and robots thereby improving the assembly and maintenance activities. These robots are designed ergonomically for the workers using it. Also, tiny mobile robots called robotini are used for simple chores such as transporting components and tools to different workstations. It helped the company to increase its production and efficiency [37].

Adidas is investing on what it calls as 'Speed factory'. With the target of reducing the manufacturing and production period Adidas is embracing supply chain digitization by using 3D-design, 3D printing programs and state of the art robotics. With the help of speed factory, they will be able to shorten the time of prototyping from several months to a couple of days. The facility pairs human workforce with advanced technologies to manufacture "athlete data-driven designs, radical accelerated footwear production, open-source co-creation, and hyperflexible localised manufacturing".

Hirotec Group a Japanese automotive components manufacturer adopted machine learning and AI to tackle the perennial problem of unplanned downtime. To ensure continuous operation and to minimise unplanned downtime it combined HPE's Edgeline IoT technology, developed to crunch data out on the edge of IoT networks, with PTC's Thingworx IoT platform and KEPServerEX agent in order to collect and analyse data from the factory floor. The company is progressing towards light out manufacturing with the use of Yaskawa Motoman dual-arm robots mounted on Otto Motors self-driving vehicles in its production process as a way of realising that vision [38].

Nike collaborated Grabit to use their robot system electro adhesion technology to automate the task of fine material handling. With the help of Grabit's electroadhesive gripper, robots can stack materials like leather, composite fibre etc. Nike uses Grabit's Stackit material handling robot system to produce up to 600 pairs of shoes in an eight-hour shift. To innovate its supply chain Nike partnered with Flex, a world class manufacturer. Partnership between Nike and Flex will enable the product to reach consumers more quickly [39].

Bosch is striving to automate its production line. To assemble machines its production line is made up of three automatic and two manual modules. These modules use RFID scanners to determine the control to be produced. They make use Manufacturing Execution System (MES)-IT Shop floor Solutions to control their line operations. MES provides real time information about the work pieces and production steps for transparent analysis, planning and monitoring of assembly [40].

6. CONCLUSION

Industry 4.0 is one of the most emerging concepts used by the industries to enhance their production capacity, decrease their product cost, create a safer environment for work. In this paper we reviewed how implementing AI in industries can revolutionize the production line, supply chain and different ways of production in manufacturing units. We reviewed that implementing AI in industries is beneficiary but due to its high investment cost and requirement of a good knowledgeable team which helps in setup of technology in manufacturing units, it is currently limited to only big firms. We also reviewed that AI can immensely help manufacturing units by providing assistance in research and development as the entire world data can be accessed easily, it also helps production line by figuring out the best way to manufacture the product, it also helps in quality assurance by using different techniques to detect any defect in products, it also helps in creating a safer environment for work by keeping an eye check on all the machinery and systems and alert in case of any mishappening. We also figured out a lot of medium and small-scale industries want to implement AI in their industries but not having a proper roadmap on its implication they face difficulties. We also figured out what are gaps in implementation of AI and how these gaps can be filled. Thus, implementing AI in manufacturing units can help industries in various ways and bring out best from them and revolutionize the industries.

7. CHALLENGES AHEAD

The adoption of artificial intelligence in manufacturing industry can help in dealing great challenges with ease. For more implementation of artificial intelligence in industries, still a lot of work needs to be done. The biggest challenge for implementation of artificial intelligence was its cost for setup, work need to be done so that the cost of technologies can be reduced while maintaining high accuracy. Investment is a major cause which holds the hand of industrialist to implement AI in their industries. There is always a risk to invest a huge amount at any technology, small manufacturers always hesitate to do this as theywant return of investment faster at any technology that they have implemented in theirindustry.

Lack of knowledge about new technologies as the development is taking place at a faster pace in field of technology and companies need to be informed and updated about all the developments going around. This increase the burden on manufacturers and increase their problems to cope up with the advancement taking place in the technology

There is still a lot of space for AI in deciding the product for industry that could manufacture as per consumer choice, as AI will collect data from customers and by using algorithms will help in deciding factors. There is a lot of work to be done to create a safer environment for work with the help of AI.

Using AI in manufacturing units there should be a proper recording of all the information from all the stages and departments of manufacturing units so that pre decisions can be made with the help of algorithms of AI. Artificial intelligence can also be used to reduce the time, technique and steps for production of goods from the help of all the information which is collected to make decisions [41].

Implementation of AI in industry also deals with threats from cyber-attacks, so a more secure and advance software need to be developed to protect from this threat [42].

Thus, there is a lot of scope and work that need to be done in implementation of artificial intelligence in manufacturing units with fast development and progress in how AI can be implemented in industry we will surely be able to make reach of artificial intelligence in every industry.

REFERENCES

[1] Leigh N G, Kraft B, Lee H (2019), Robots, skill demand and manufacturing in US regional labour markets, Cambridge Journal of Regions, Economy and Society.

[2] Li B. hu, Hou B. cun, Yu W. tao, Lu X. bing, Yang, C. wei. (2017), Applications of artificial intelligence in intelligent manufacturing: a review, In Frontiers of Information Technology and Electronic Engineering, pp. 86–96, 18(1)

[3] Mittal S, Khan M A, Romero D, Wuest T (2019), Building blocks for adopting smart manufacturing, Procedia Manufacturing, pp. 978–985,34

[4] Leo Kumar S. P. (2017). State of The Art-Intense Review on Artificial Intelligence Systems Application in Process Planning and Manufacturing, Engineering Applications of Artificial Intelligence, pp. 294–329,65

[5] Mahroof K (2019), A human-centric perspective exploring the readiness towards smart warehousing: The case of a large retail distribution warehouse, International Journal of Information Management, pp. 176-190, 45

[6] Cohen G (1989), Using AI techniques to optimize manufacturing shop-floor operations, Engineering Applications of Artificial Intelligence, pp. 238-246, 2(3)

[7] Hagemann S, Sünnetcioglu A, Stark R (2019), Hybrid artificial intelligence system for the design of highly-automated production systems, Procedia Manufacturing, pp. 160–166,28

[8] Lou H H, Huang Y L (2003), Hierarchical decision making for proactive quality control: System development for defect reduction in automotive coating operations. Engineering Applications of Artificial Intelligence, pp. 237–250, 16(3)

[9] García G S, García G M (2019), Industry 4.0 implications in production and maintenance management: An overview, 8th Manufacturing Engineering Society International Conference 2019, pp. 415-422, Madrid, Spain

[10] Cohen Y, Faccio M, Pilati F, Yao X (2019), Design and management of digital manufacturing and assembly systems in the Industry 4.0 era, In International Journal of Advanced Manufacturing Technology, pp. 3565–3577, 105(9)

[11] Urhal P, Weightman A, Diver C, Bartolo P (2019), Robot assisted additive manufacturing: A review, Robotics and Computer-Integrated Manufacturing, pp. 335–345,59

[12] Wu D, Gary Wang G (2020), Knowledge-Assisted Optimization for Large-Scale Design Problems: A Review and Proposition, Journal of Mechanical Design, 142(1).

[13] Busch M, Schuh G, Kelzenberg C, de Lange J (2019). Short Paper: Development of Production Planning and Control through the Empowerment of Artificial Intelligence. 2019 Second International Conference on Artificial Intelligence for Industries (AI4I), pp. 115–118.

[14] Keen M. Successful Applications of AI in Manufacturing Industry.

[15] Matzka S (2018), Using Process Quality Prediction to increase Resource Efficiency in Manufacturing Processes, 2018 First International Conference on Artificial Intelligence for Industries, Laguna Hills, CA, USA

[16] Meng G, Yanli J (2011). Research on e-intelligence maintenance system for manufacturing industrials. Proceedings - 2011 International Conference on Internet Computing and Information Services, pp. 457-458, ICICIS 2011,

[17] https://ieeexplore.ieee.org/document/7848697

[18] Ma Y, Wang Z, Yang H, Yang L (2020). Artificial Intelligence Applications in the Development of Autonomous Vehicles: A Survey, 7(2)

[19] Bashir B F, Steiner S J, An A.I. based simulation system to improve shop-floor efficiency and to reduce production cost.

[20] Alaloul W S, Liew M S, Zawawi N A W A, Kennedy I B (2019). Industrial Revolution 4.0 in the construction industry: Challenges and opportunities for stakeholders. Ain Shams Engineering Journal.

[21] Scholz R W, Czichos R, Parycek P, Lampoltshammer T J (2020). Organizational vulnerability of digital threats: A first validation of an assessment method. European Journal of Operational Research, pp. 627–643, 282(2)

[22] Satish R, Murugabhoopathy K, Rajendhiran N, Vijayan, V. (2020). Technology strategy for improved safety management in steel industry. Materials Today: Proceedings.

[23] Park H S, Phuong D X, Kumar S (2019). AI based injection molding process for consistent product quality. Procedia Manufacturing, pp. 102–106, 28

[24] Santos T, Silva F J G, Ramos S F, Campilho R D S G, Ferreira L P (2019). Asset Priority Setting for Maintenance Management in the Food Industry. Procedia Manufacturing, pp. 1623–1633,38

[25] Baskaran S, Niaki F A, Tomaszewski M, Gill J S, Chen Y, Jia Y, Mears L, Krovi V (2019). Digital human and robot simulation in automotive assembly using siemens process simulate: A feasibility study. Procedia Manufacturing, pp. 986–994, 34

[26] Joppen R, von Enzberg S, Kuhn A, Dumitrescu R (2019). A practical Framework for the Optimization of Production Management Processes. Procedia Manufacturing, pp. 406–413,33

[27] Gunduz M Z, Das, R. (2020). Cyber-security on smart grid: Threats and potential solutions, Computer Networks, 169

[28] Bcg Homepage, https://www.bcg.com/en-in/publications/2018/artificial-intelligence-factory-future.aspx, last accessed 2020/2/12

[29] Panorama Consulting Group Homepage, https://www.panorama-consulting.com/how-artificial- intelligence-improves-erp-software/, last accessed 2020/2/12

[30] Infopulse Homepage, https://www.infopulse.com/blog/the-main-benefits-and-challenges-of-industry-4-0-adoption-in-manufacturing/, last accessed 2020/02/12

[31] The manufacturer Homepage, https://www.themanufacturer.com/articles/power-artificial-intelligence-manufacturing/, last accessed 2020/02/12

[32] Raconteur Homepage, https://www.raconteur.net/technology/ai-human-augmentation, last accessed 2020/2/12

[33] Horses for Sources Homepage, https://www.horsesforsources.com/gartner_fail_automation-AI_080418, last accessed 2020/2/12 ifr Homepage, https://ifr.org/ifr-press-releases/news/global-industrial-robot-sales-doubled-over- the-past-five-years, last accessed 2020/2/12

[34] Cbinsights Homepage, https://www.cbinsights.com/research/future-factory-manufacturing-tech- trends/, last accessed 2020/2/12

[35] Warehousing Logistics International Homepage, https://www.warehousinglogisticsinternational.com/how-will-ai-transform-the-smart-warehouse/, last accessed 2020/2/12

[36] Lamborghini Homepage, https://www.lamborghini.com/en-en/news/lamborghini- award-winning-factory-future, last accessed 2020/02/13

[37] Otto motors Homepage, https://ottomotors.com/blog/manufacturing-data-industry-4-0, last accessed 2020/02/13

[38] Nike Homepage, https://news.nike.com/news/nike-s-manufacturing-revolution- accelerated-by-new-partnership-with-flex, last accessed 2020/02/13

[39] Bosch Homepage, https://www.bosch.com/stories/industry-4-0-production-line/, last accessed 2020/02/13

[40] Datanami Homepage, https://www.datanami.com/2019/04/25/big-data-challenges-of-industry-4- 0/, last accessed 2020/02/12

[41] Challenges Homepage, https://www.challenge.org/industry-4-0-3-3/challenges/,last accessed 2020/02/12

Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021).
IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4

2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) | 978-1-6654-1960-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICICV50876.2021.9388531

Using Gaussian based Bayesian Optimization in Recommender Systems

Garvit Rana
Department Of Computer Engineering
Delhi Technological University
Delhi, India
garvitrana2000@gmail.com

Eklavya Bhatia
Department Of Computer Engineering
Delhi Technological University
Delhi, India
eklavyabhatia@gmail.com

Ejaz Ul Haq
Department Of Computer Engineering
Delhi Technological University
Delhi, India
ejazsaifi70@gmail.com

Shailender Kumar
Department Of Computer Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

***Abstract*—Recommender systems have evolved as a major component of many businesses across various domains. Recommender systems help users making quick decisions by providing relevant information based on the past history of the user. This research work is focussed to enhance the performance of a recommender system by doing hyper parameter tuning and optimization techniques using Grid Search and Bayesian Optimization. The dataset consists of ratings of various users on the product. The surprise package of Python has been used to build the recommender system. A Collaborative System type recommender system is used to predict the products for a user. Various classifiers like Baseline Only, KNN Basic, SVD, NMF, and Co-Clustering were used. RMSE is used as a metric to determine the efficacy of the recommender system. The results show that Bayesian Optimization performed better than grid search in terms of RMSE values and also in the number of iterations required. Tuning the parameters improved the RMSE values to a major extent as compared to default parameters.**

Keywords— Bayesian, Gaussian Process, Grid Search, Hyper Parameter, Matrix Factorization, Mean Square Error, Recommender Systems, Tuning.

I. INTRODUCTION

Data and users available on the internet are increasing at an exponential rate in the current century. With the growing demand of users, it is imperative to build intelligent systems that can suggest and direct users what to do based on their previous choice. Recommender systems address the above issue. Recommender systems are a part of an information filtering system that helps us to direct relevant content to the users based on their past preferences. Recommender systems enhance the user experience, increase the profits of the company by improving their sales and user activity, help users make decisions quickly, and give every individual personalized suggestion [1]. Thus, Recommender systems are useful to retain the existing customers and for the scalability of any company. One of the shortcomings of the Recommender system is that it can make a recommendation based on existing data only which can be less and therefore provide less accurate recommendations.

The recommender system finds its application in various companies. E-commerce companies like Amazon and Flipkart use user ratings and purchase history to suggest products to the users. Online video streaming and music companies like Netflix, Youtube, Spotify suggest video and music based on the history of the user. Professional networking sites like Linkedin provide job recommendations based on the previous job searches done by the user [2]. Social media sites like Facebook and Twitter provide content-based Recommendations to their users. These were some of the applications where recommender systems are being used to assist users in decision making.

Collaborative filtering works on the idea that users are going to engage in only those products in which they have done in the past. This method only requires past historical data. This is divided into two categories based on user preferences: The first method is an Explicit Rating. This rating is given directly by the user. It is a direct review by the user of a particular product. The second method is the Implicit Rating. It gives an indirect preference of a user for a particular product. For example view count, clicks, number of searches, and so on [3].

There are two methods used in Collaborative filtering: The first method is the Nearest Neighbourhood. In this approach, we find the closest users to our target user(nearest neighbors) and make predictions of our target user by weighting their ratings. The second method is Matrix Factorization: In this approach, we find how much a particular user associates with a set of attributes. This method is scalable and computationally efficient.

We have used Grid Search and Bayesian Optimization to get the optimal values of hyper-parameters. Grid search is a computationally exhaustive and complete search on the parameters explicitly specified by the user. Bayesian optimization uses Bayes theorem to align the search to find the maximum and minimum of a function. The domain space in grid search is discrete while as in Bayesian optimization it is a continuous probability distribution. Grid search tries all the combinations while Bayesian optimization searches along with space and thus does not try all the combinations.

The flow of the remaining paper is as: We have explained the related work and methodologies in the second section, in the third section we have discussed our proposed work, the fourth section is about our results. In the end, the paper has a section related to the conclusion and future work, and the last section is about references for the research papers referred to in making this research paper.

II. RELATED WORK

In today's world, industries especially which are either directly or indirectly influenced by social society on the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:04:24 UTC from IEEE Xplore. Restrictions apply.

world wide web are getting greatly impacted by the evolving nature of recommendation methodologies around the globe. Several industries correlated to movies on demand are in a continuous manner increasing their user base growth with the help of several new-age techniques. Reference [4] has depicted a similar scenario in which a procedure has been devised revolving around the idea of relative view time of several animated films and determining their percentile of viewership. To further emphasize the need for this process it was referenced that it is a great indicator of the user being interested in the content being served or not. Along with it, the performance of the devised strategy was compared and its efficiency was elaborated. As the concluding achievement, it was devised that the methodology was four times better as compared to conventional approaches.

In the approach and technique devised in [5], a collaborative filtering based method was analyzed for providing an accurate set or group of research manuscripts according to a particular area field. The outcomes of the method were put in either increasing or decreasing order with the support of the cosine similarity procedure. For evaluatory metric for the method, mean average fault detection was utilized. The outcome suggested contained very fine quality manuscripts in relevance to the provided query. One hindrance of lack of available data manuals was faced in the field of procedure was encountered and as a consequence, the data was taken from several users to be used for this manuscript.

Reference [6] devised and carried out a creative methodology revolving around the raw data in our data groups, considering it a complicated work to be handled given retrieving and analyzing of available raw data. It emphasized a procedure to convert the raw data in a way such that it can be stored with the utilization of verb taggers and after it is grouped using them. It devises a framework based on natural language processing. As in comparison to the baseline procedures, it was found more efficient and enabled easier data management procedures which overall resulted in much quicker retrieval for the model.

Reference [7] used the technique of collaborative filtering for building a recommender system that can be used in healthcare. The purpose of this research paper was to assign relevant doctors to patients. The main parameter of assigning doctors were past trust shown by the patient in a particular doctor. The dataset was obtained from a well-established hospital in Europe which consisted of 1.07 million consultations of more than four lakh patients. The model proposed was based on the idea that patients either continue with the same doctor or may sometimes switch the doctor. To find the hyper-parameters of the model cross-validation was performed. Hit Ratio and precision were used as the metrics to evaluate the recommendation list. This proposed model performed better than the heuristic baseline model by a huge amount. The trust measure further improved the results as compared to measuring only the number of times a particular patient has visited a doctor and therefore helped matching patients with relevant doctors.

Reference [8] enhanced the technique of collaborative filtering by choosing an optimal neighborhood for Recommender systems. Two different neighborhoods were taken into consideration. The first one is k-nearest neighbors and the second one is threshold-based neighbors. The results were obtained by choosing a varying number of neighbors and their corresponding threshold values. The dataset used was the MovieLens dataset which comprises more than one million ratings. Root Mean Squared Error(RMSE) was used as evaluation metrics. In the first method of N nearest neighbors, the nearest N neighbors were selected which gave the same rating given by the user. The initial value of n was taken to be 20. By comparing various values of nearest neighbors it was found that the least RMSE was obtained with 200 neighbors. The second method of threshold-based neighbors calculated the distance between the current user and its neighbors and then calculated the RMSE to find the best point. In the majority of the cases, RMSE was least for the threshold value of 0.1. A comparison of both the methods was done and it was found out that the latter method gives less RMSE and thereby providing better and accurate recommendations. Figure 1 depicts a comparison with Content-based and collaborative filtering methodology.

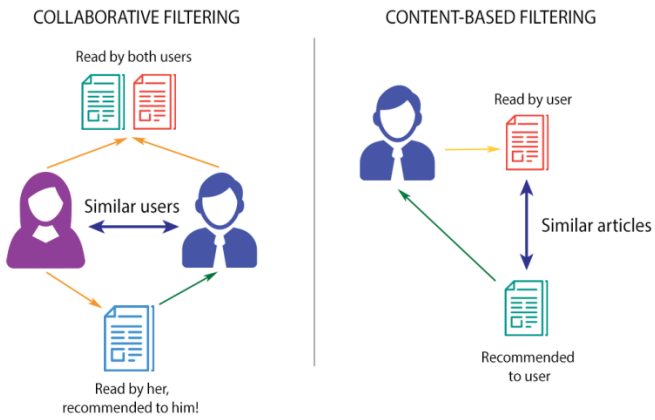


Fig. 1. A pictorial comparison between the two major techniques.

Reference [9] provided a content-based movie Recommender System that makes predictions based on user behavior. Dirichlet Process Mixture Model was used to understand user preferences. The profile of the user is defined by his activity. The dataset was taken from MovieLens. The activity signifies the content and access time. The model performed is user-centric in that it works differently for every user. For building the model of a particular user three main components are taken namely interest extraction, inferring of preferences, and prediction. To get the content information storyline and genre were fetched from IMDb. The textual information was processed using NLTK. Every user is modeled by taking into account the textual feature vector of rated movies. SVD++ was used to build the movie recommendation system. To evaluate the results precision, recall, f1-measure were used. These metrics were averaged for every user. The results show that the user preference model proposed by this research outperformed the previous recommender systems. One major advantage of this model is that it can recommend new movies based on their content.

III. PROPOSED WORK

In our proposed methodology, we have utilized the two kinds of tuning and optimization techniques that are Bayesian Optimization and Grid Search. The process of Hyper Parameter Tuning can be constructively partitioned into several procedures. Firstly, we need to identify the range of values of several parameters required to be optimized

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:04:24 UTC from IEEE Xplore. Restrictions apply.

within our problem domain. Secondly, we need to determine an unbiased heuristic function that can provide us a value that can be used for evaluating our parameter values in an optimized manner. Also, the major Crux part of Bayesian Optimization is that can diminish the total number of times our heuristic has to be called. In the next step, we need to define a feasibility model, here we are using Gaussian Process. Now, we determine a method that can select the next group of hyperparameters from the previous model. After all these aspects a log is created consisting of the several scores and values corresponding to various parameters along with the best one.

For Dataset, we have extracted and Utilized ten product Review data of several products belonging from categories like Health, Office items, Video Animations, Sports. Our datasets consist of a User identifier, Product identifier, timestamp, and rating of them. As the datasets were quite enormous, So to make them less indeterminate, we have considered those users who have given fifty reviews or more than that only. Figure 2 demonstrates an analysis of the most popular products in one of the datasets.

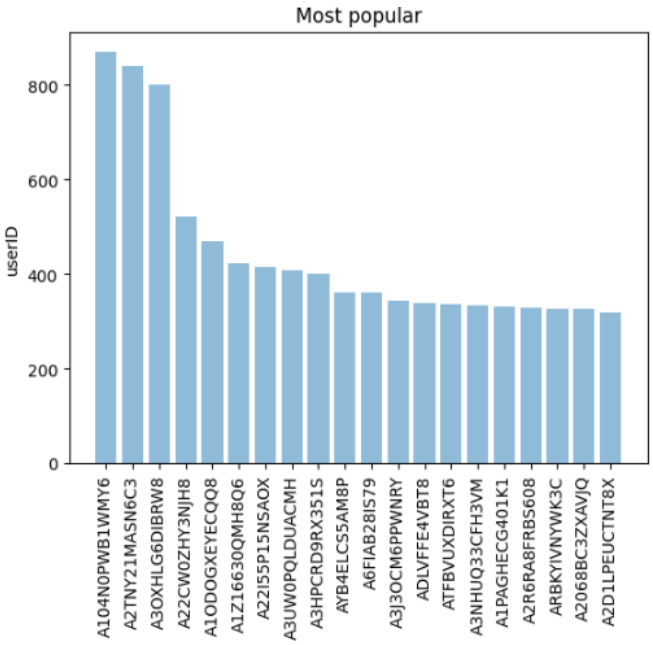


Fig. 2. Most popular products in Stationary Data

For the required procedure and the methodology, we are utilizing the Surprise Library which has been used for constructing our recommender system model. To perform tuning with the help of Bayesian Optimization, we are using the Bayes opt library. The Blackbox functions being used in this is determining the Root Mean Square Error values, and our optimization is minimizing it using the hyperparameter values by keeping a balance between both exploration which is searching in the unique value space, and exploitation through which the current values are further optimized to reach optimal condition. In the condition of imbalance between them, the former can lead to longer iterations without determining or reaching the required optimality and the latter can lead to local optimality rather than global optimality in the domain.

For Collaborative Filtering Based procedures and methods, we have used several types of classifiers, like BaselineOnly, KNNBasic, SVD, NMF, and CoClustering for our model. In SVD we are tuning two parameters that are respectively n factors in the range of [90,110] with its default value being 100 and n epochs in the range of [15,25] with its default value being 20. In NMF we are tuning two parameters that are n factors in the range of [10,20] with the default value being 15 and n epochs in the range of [45,55] with the default value being 50. In CoClustering we are tuning n epochs parameter in the range of [15,25] with the default value being 20 and n cltr in the range of [2,3,4] with the default value being 3. In KNNBasic we are tuning k in the range of [35,45] with the default value being 40 and min k in the range of [1,3] with the default value being 2. Figure 3 depicts the Pseudocode of the Complete Methodology consisting of input of set of Classifiers and Datasets and in output top 12 recommendations are provided.

1. Input: Set of Classifier {**A**}, Set of Datasets {**B**}

2. Output: Top 12 Recommendations for the particular User.

3. for all Dataset **x** in A **do**

4. Pre-Processing of the raw data (**x**).

5. To Make the data less condensed, only those entries are considered where the user has given more than 50 ratings.

6. Partitioning the data **x** in a ratio of 0.6 for the training and 0.4 for testing.

7. for all Classifiers **y** in B **do**

8. Constructing Collaborative Filtering model on **y** Classifier.

9. Performed parameter tuning on Classifier **y** using Grid search.

10. Performed parameter tuning on Classifier **y** using Bayesian Optimization.

11. Analyzing both the techniques of parameter

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:04:24 UTC from IEEE Xplore. Restrictions apply.

tuning by Graphs.

12. Compute RMSE, Precision, Recall after both the methods of parameter tuning on the given dataset **x**.

13. end for

14. end for

Fig. 3. Pseudocode of Devised Methodology

IV. EXPERIMENTS AND RESULTS

For achieving much preciseness in our procedure, numerous data have been extracted from several categories of products like Beauty, Toys, Health, Kindle store, and a lot more from Amazon. It contains User identifier, Item identifier, Timestamp, and rating field. As to make it less compact, only those users were considered who have given fifty or more than that ratings.

To determine the effectiveness, the RMSE metric has been utilized for comparison between tuned procedure with untuned technique. Along with it, the precision and recall metrics have been analyzed for fetching the amount of how well our model has been working. RMSE value is analyzed by determining the root of the mean square error value. MSE is determined by taking the mean of summation of squares of the net difference between the actual value with the predicted value. Equation (1) denotes the RMSE value. In (2) the variable y depicts the actual output and y' depicts the predicted output. Also, the variable N is the total number of values for calculating the mean.

$$RMSE = \sqrt{MSE} \tag{1}$$

$$MSE = \sum \frac{(y-y')^2}{N} \tag{2}$$

Precision value is denoted in (3), here true positive denotes the prediction in which the model has predicted positive value and the actual outcome is also positive. Equation (4) depicts the Recall value, here the predicted results are the total outcomes that have been achieved.

$$Precision = \frac{True\ Positives}{Actual\ Results} \tag{3}$$

$$Recall = \frac{True\ Positives}{Predicted\ Results} \tag{4}$$

In Table I, five different classifiers have been trained on ten Amazon Product Rating datasets and their RMSE values have been evaluated. In Table II and Table III, similarly, the Precision and Recall have been analyzed and formulated using their equations.

TABLE I. UNTUNED RMSE VALUES OF CLASSIFIERS

DATASET	CLASSIFIERS				
	SVD	KNN Baseline	Co-Clustering	NMF	KNN Basic
Video Games	0.9784	1.0481	1.0848	1.1538	1.1761
Toys Product	0.8693	0.937	0.9802	1.085	1.002
Home Decor	0.9392	1.0181	1.1123	1.1845	1.0815
Beauty Care	1.0176	1.089	1.1511	1.2329	1.1998
Stationary	0.8159	0.8755	0.8945	0.996	0.9277
Kitchenware	0.9483	1.0256	1.0679	1.1625	1.0858
Kindle Store	0.7399	0.7795	0.7858	0.8399	0.8851
Sports Tools	0.8528	0.9306	0.9712	1.0858	0.9933
Mobile Extras	0.8461	0.8747	0.9335	0.9882	0.9062
Health Care	0.9298	0.9676	1.0017	1.0614	1.0327

TABLE II. UNTUNED PRECISION VALUES OF CLASSIFIERS

DATASET	CLASSIFIERS				
	SVD	KNN Baseline	Co-Clustering	NMF	KNN Basic
Video Games	0.7628	0.7981	0.7854	0.7857	0.8097
Toys Product	0.8327	0.8627	0.8611	0.8359	0.8633
Home Decor	0.8335	0.8515	0.8622	0.8447	0.8523
Beauty Care	0.7222	0.7755	0.7773	0.7523	0.7846
Stationary	0.8467	0.8795	0.858	0.8317	0.8911
Kitchenware	0.827	0.856	0.856	0.844	0.851
Kindle Store	0.834	0.862	0.815	0.791	0.866
Sports Tools	0.853	0.879	0.881	0.858	0.876
Mobile Extras	0.839	0.872	0.861	0.879	0.874
Health Care	0.823	0.851	0.847	0.845	0.857

TABLE III. UNTUNED RECALL VALUES OF CLASSIFIERS

DATASET	CLASSIFIERS				
	SVD	KNN Baseline	Co-Clustering	NMF	KNN Basic
Video Games	0.4889	0.5301	0.5194	0.4464	0.5994
Toys Product	0.7661	0.7928	0.7614	0.6168	0.8291
Home Decor	0.7797	0.7964	0.7554	0.621	0.8549
Beauty Care	0.5582	0.5959	0.626	0.5086	0.7182
Stationary	0.6077	0.6397	0.6157	0.5327	0.676

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:04:24 UTC from IEEE Xplore. Restrictions apply.

Kitchenware	0.635	0.661	0.642	0.546	0.703
Kindle Store	0.712	0.728	0.649	0.561	0.773
Sports Tools	0.771	0.798	0.754	0.625	0.825
Mobile Extras	0.669	0.697	0.725	0.651	0.721
Health Care	0.532	0.566	0.557	0.503	0.631

After analyzing the untuned procedure metrics, the tuned methodology is compared with it. In Table IV the RMSE value has been determined after tuning the required variables in our methods with the Grid Search Tuning technique. Similarly, in Table V, the RMSE metric has been determined with the utilization of Gaussian process based Bayesian Optimization.

TABLE IV. TUNED RMSE VALUES OF CLASSIFIERS WITH GRID SEARCH

DATASET	**CLASSIFIERS**				
	SVD	KNN Baseline	Co-Clustering	NMF	KNN Basic
Video Games	0.6942	0.6336	0.7899	0.3109	0.7221
Toys Product	0.7606	0.6025	0.6103	0.3429	0.6969
Home Decor	0.6627	0.6905	0.6052	0.2981	0.8122
Beauty Care	0.8242	0.7434	0.8024	0.3724	0.8388
Stationary	0.5446	0.5916	0.5516	0.3284	0.6615
Kitchenware	0.6989	0.6368	0.6931	0.3962	0.7173
Kindle Store	0.5218	0.4917	0.6706	0.3816	0.5498
Sports Tools	0.6228	0.6203	0.5638	0.3232	0.7194
Mobile Extras	0.5677	0.7046	0.4922	0.3408	0.8058
Health Care	0.6973	0.6972	0.7649	0.4834	0.7633

TABLE V. TUNED RMSE VALUES OF CLASSIFIERS WITH BAYESIAN OPTIMIZATION

DATASET	**CLASSIFIERS**				
	SVD	**KNN Baseline**	**Co-Clustering**	**NMF**	**KNN Basic**
Video Games	0.6864	0.7466	0.8812	0.6935	0.831
Toys Product	0.654	0.6125	0.721	0.622	0.6487
Home Decor	0.6892	0.6398	0.767	0.6714	0.6823
Beauty Care	0.7514	0.7805	0.9133	0.7494	0.8381
Stationary	0.5241	0.5687	0.6362	0.5393	0.6067
Kitchenware	0.6574	0.6543	0.7888	0.6519	0.6884
Kindle Store	0.4901	0.5826	0.6855	0.5426	0.6536
Sports Tools	0.6131	0.5841	0.6709	0.5902	0.6277
Mobile Extras	0.5211	0.5342	0.6029	0.5271	0.5451
Health Care	0.6501	0.7285	0.8381	0.7081	0.7721

Figure 4 and Figure 5 clearly depicts the analysis of the metric values with ten different datasets with and without tuning. Figure 6 demonstrated a brief comparison of tuned with Bayesian and Grid Search procedures with methodology with default values of parameters.

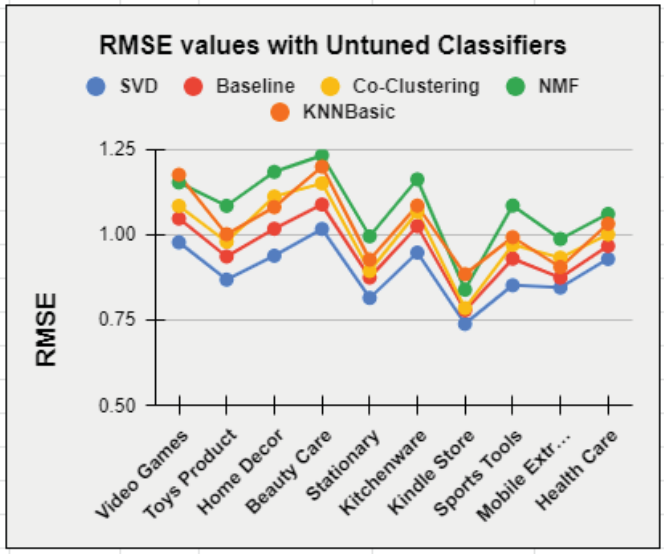


Fig. 4. RMSE value with Untuned Classifiers

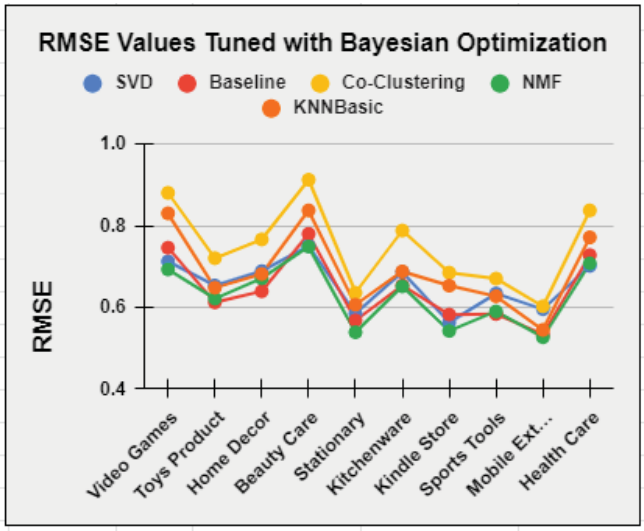


Fig. 5. RMSE Value on Tuning with Bayesian Optimization

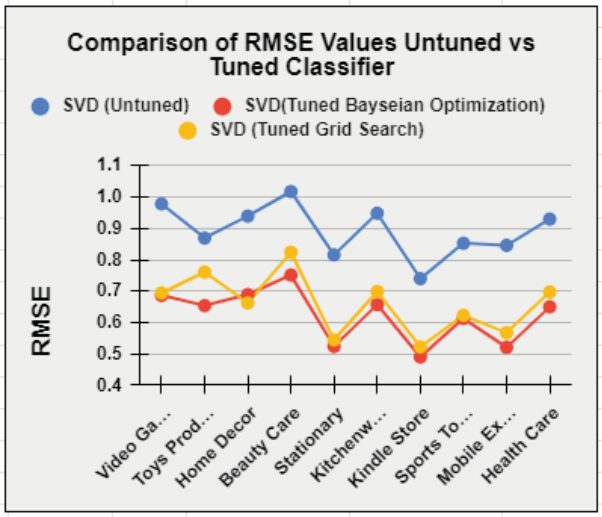


Fig. 6. Comparison among Tuned vs Untuned Procedures

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:04:24 UTC from IEEE Xplore. Restrictions apply.

V. CONCLUSION AND FUTURE WORK

Till now enormous innovations have been made in the field of advising the best set from a large volume of data and continue the creation of new methods, techniques and methodology is being carried out keeping in mind the requirement and necessity of social globe in this field. The majority of the procedures have been carried out on small or limited volumes of available data thus making them vulnerable to several kinds of drawbacks in industrial usage [10]. To provide a crystal clear view on the methodology, the procedures have been operated on a huge data comprising of several kinds of product domains like Toys, Cell Phones, home and kitchen, and a lot more, which further emphasizes the efficiency and accuracy in our methods making it suitable for future innovations [11]. A detailed and thorough comparison of both the procedures that is with untuned and with tuned variables in our model.

For it, two techniques have been analyzed and explored, first is Bayesian Optimization working on the probability procedure of the Gaussian process along with grid search which is the second one. A detailed and in-depth methodology has been formulated by considering a large number of datasets that are too related to a large space of varying fields, thus eliminating any chances of outcome failure. By analyzing both the methods we have identified that Bayesian Optimization based on Gaussian Process has outperformed Grid Search in both the fields of time requirements as well as iterations required to need a required minimum for our problem scenario [12]. It has even performed way too better in root mean square values as compared with the latter technique. Tuning has also improved the accuracy numbers by a large fraction as compared to outcomes with default values [13].

As a follow-up of comparing Hyperparameter tuning methods, implications can be fetched that tuning has improved the root mean square value metric by a large portion [14]. Although there has been a large set of innovators in the domain who are not oriented towards Tuning, whereas it can create huge improvements in the ongoing innovations as depicted in the methodology [15]. It has become the required necessity to include tuning in ongoing creative innovations and methods. In such scenarios where the required condition for solving a heuristic is to be utilized within a small time frame and with fewer heuristic calls, Bayesian Optimization should be taken in methodology [16].

For future work, the proposed model can be extended to build a content-based Recommender system that predicts the products based on their information. Also, the recommender system can be deployed for public use so that its usefulness in a real-world scenario can be tested and thus the model can be improved by taking user feedback [17]. Various other methods for parameter tuning can be explored to further enhance the prediction of the recommender systems.

REFERENCES

[1] C. Yu, Q. Tang, Z. Liu, B. Dong, and Z. Wei, "A Recommender System for Ordering Platform Based on an Improved Collaborative Filtering Algorithm," *ICALIP 2018 - 6th Int. Conf. Audio, Lang. Image Process.*, pp. 298–302, 2018, doi: 10.1109/ICALIP.2018.8455852.

[2] P. Venil, G. Vinodhini, and R. Suban, "Performance evaluation of ensemble based collaborative filtering recommender system," *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–5, 2019, doi: 10.1109/ICSCAN.2019.8878777.

[3] U. Kuzelewska and A. Kurylowicz, "Multi-clustering applied to collaborative recommender systems," *2018 13th Int. Conf. Digit. Inf. Manag. ICDIM 2018*, pp. 118–123, 2018, doi: 10.1109/ICDIM.2018.8847141.

[4] R. E. Nakhli, H. Moradi, and M. A. Sadeghi, "Movie Recommender System Based on Percentage of View," *2019 IEEE 5th Conf. Knowl. Based Eng. Innov. KBEI 2019*, pp. 656–660, 2019, doi: 10.1109/KBEI.2019.8734976.

[5] M. V. Murali, T. G. Vishnu, and N. Victor, "A Collaborative Filtering based Recommender System for Suggesting New Trends in Any Domain of Research," *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 550–553, 2019, doi: 10.1109/ICACCS.2019.8728409.

[6] S. Amara and R. R. Subramanian, "Collaborating personalized recommender system and content-based recommender system using TextCorpus," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 105–109, 2020, doi: 10.1109/ICACCS48705.2020.9074360.

[7] Q. Han, I. Martinez De Rituerto De Troya, M. Ji, M. Gaur, and L. Zejnilovic, "A Collaborative Filtering Recommender System in Primary Care: Towards a Trusting Patient-Doctor Relationship," *Proc. - 2018 IEEE Int. Conf. Healthc. Informatics, ICHI 2018*, pp. 377–379, 2018, doi: 10.1109/ICHI.2018.00062.

[8] S. J. Mehta and J. Javia, "Threshold based KNN for fast and more accurate recommendations," *2015 IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.*, pp. 109–113, 2015, doi: 10.1109/ReTIS.2015.7232862.

[9] B. R. Cami, H. Hassanpour, and H. Mashayekhi, "A content-based movie recommender system based on temporal user preferences," *Proc. - 3rd Iran. Conf. Signal Process. Intell. Syst. ICSPIS 2017*, vol. 2017-Decem, pp. 121–125, 2018, doi: 10.1109/ICSPIS.2017.8311601.

[10] M. V. V. R. Murali Krishna Rao, "A collaborative filtering recommender system with randomized learning rate and regularized parameter," *2016 IEEE Int. Conf. Curr. Trends Adv. Comput. ICCTAC 2016*, pp. 0–4, 2016, doi: 10.1109/ICCTAC.2016.7567331.

[11] N. Sivaramakrishnan and V. Subramaniyaswamy, "GPU-based Collaborative Filtering Recommendation System using Task parallelism approach," *Proc. Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2018*, pp. 111–116, 2019, doi: 10.1109/I-SMAC.2018.8653709.

[12] L. Al Hassanieh, C. A. Jaoudeh, J. B. Abdo, and J. Demerjian, "Similarity measures for collaborative filtering recommender systems," *2018 IEEE Middle East North Africa Commun. Conf. MENACOMM 2018*, pp. 1–5, 2018, doi: 10.1109/MENACOMM.2018.8371003.

[13] G. Suganeshwari and S. P. S. Ibrahim, "Rule-Based Effective Collaborative Recommendation Using Unfavorable Preference," *IEEE Access*, vol. 8, pp. 128116–128123, 2020, doi: 10.1109/ACCESS.2020.3008514.

[14] J. Zhang, Y. Wang, Z. Yuan, and Q. Jin, "Personalized real-time movie recommendation system: Practical prototype and evaluation," *Tsinghua Sci. Technol.*, vol. 25, no. 2, pp. 180–191, 2020, doi: 10.26599/TST.2018.9010118.

[15] Z. Cui *et al.*, "Personalized Recommendation System Based on Collaborative Filtering for IoT Scenarios," *IEEE Trans. Serv. Comput.*, vol. 13, no. 4, pp. 685–695, 2020, doi: 10.1109/TSC.2020.2964552.

[16] C. D. Wang, Z. H. Deng, J. H. Lai, and P. S. Yu, "Serendipitous recommendation in e-commerce using innovator-based collaborative filtering," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2678–2692, 2019, doi: 10.1109/TCYB.2018.2841924.

[17] J. Zhao, H. Wang, and H. Zhang, "A Regression-Based Collaborative Filtering Recommendation Approach to Time-Stepping Multi-Solver Co-Simulation," *IEEE Access*, vol. 7, pp. 22790–22806, 2019, doi: 10.1109/ACCESS.2019.2897486.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 15,2021 at 11:04:24 UTC from IEEE Xplore. Restrictions apply.

Colloid and Polymer Science
https://doi.org/10.1007/s00396-021-04831-8

ORIGINAL CONTRIBUTION

Zinc micronutrient-loaded carboxymethyl tamarind kernel gum-based superabsorbent hydrogels: controlled release and kinetics studies for agricultural applications

Khatri Khushbu[1] · Sudhir G. Warkar[1] · Nandkishore Thombare[2]

Received: 30 January 2021 / Revised: 11 March 2021 / Accepted: 22 March 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

The novel zinc-loaded superabsorbent hydrogel (ZSAH) was synthesized through in situ incorporation of zinc micronutrient in carboxymethyl tamarind kernel gum cross-linked network. The synthesized ZSAH was characterized by FTIR, SEM, TGA, swelling studies, etc. and found to have good water absorption capacity (410 g/g in distilled water) and with entanglement of zinc with weak physical forces. The ZSAH thus synthesized was assessed for controlled release application in soil as well as water, to test its application in the field of agriculture. The ZSAH exhibited slow and steady release pattern of zinc with 55% release within 72 h in water, and 83% release within 80 days in soil. Under the release kinetics studies, Korsmeyer-Peppas model was found to be best explaining the zinc release pattern through ZSAH. The ZSAH followed Fickian mechanism for release and it was observed that the diffusion was independent of the structure of the polymeric network. The time for zinc release from ZSAH, predicted as per the Olson's single exponential model, revealed significantly prolonged T_{99} (time for release of 99% Zn) in water (12.22 days) as also in soil (625 days). Hence, the novel synthesized ZSAH exhibited great potential as a zinc micronutrient delivery system in agriculture.

Keywords Carboxymethyl tamarind kernel gum · Zinc micronutrient · Superabsorbent hydrogels · Agriculture · Soil conditioner · Release kinetics

Introduction

The demand for the resurgence of industrial agriculture all around the world is increasing with the expansion of the global population, so as to keep up the pace with the growing inhabitants on earth. For achieving high agricultural productivity, an adequate supply of water and micronutrients is very important [1–3]. Moreover, the supply of micronutrients is very critical because limits of their minimum availability and toxicity in the soil are narrower as compared to major nutrients, and hence, they play a vital role in securing the productivity of crops by influencing their quality and yield [4–7]. Therefore, specific micronutrients should be made available to plants in a controlled fashion in order to attain an ideal performance. Conventionally, nutrients are applied to put a frequent and heavy dose of fertilizers to the crop. There are several drawbacks in the conventional means of micronutrient applications such as high labor cost, difficulty in control release, low nutrient efficiency, and environmental pollution [8, 9]. Different technological and scientific strategies have been developed over the years to overcome these hurdles, such as controlled or slow release systems. Various materials have been tested for this goal, among which superabsorbent hydrogels [10] with the added advantage of moisture retention, protrude as a promising matrix to supply agricultural nutrients [11].

Superabsorbent hydrogels are the tri-dimensional cross-linked polyelectrolyte polymeric structures that can imbibe more than 300 times its weight in water [12, 13]. It has fetched a lot of consideration in this context because of its controlled or slow-release properties [14]. They have been found effective in enhancing the physical properties of soil like water

✉ Sudhir G. Warkar
sudhirwarkar@gmail.com

[1] Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

[2] Processing and Product Development Division, ICAR- Indian Institute of Natural Resins and Gums, Namkum, Ranchi 834010, India

Published online: 30 March 2021

holding capacity, bulk density, and porosity and hence proved their potential in being an excellent aid for drought-prone areas [15–20]. Their utility as the carrier of fertilizers for their controlled release has also been reported [21–24] but there is little published work on the slow or controlled release of micronutrients from superabsorbent hydrogels [25].

Zinc is an essential micronutrient for crops as it plays a vital role in many physiological functions of plants. Zinc deficiency can affect plants through number of factors such as increasing crop maturity period, stunting its growth, decreasing the number of tillers, small leaves and chlorosis, spikelet sterility, and inferior quality of products [26]. Hence, zinc plays a notable role in plant metabolism by affecting the activities of carbonic anhydrase, hydrogenase, synthesis of cytochrome, and stabilization of ribosomal fractions [27] which further affects the integrity of cellular membranes, pollen formation, and protein synthesis and auxin synthesis in plants [28]. Zinc deficiency causes "Khaira" disease in rice grown in the calcareous soil of north India. Despite the need for this critical micronutrient, zinc is found to be the most deficient micronutrient in the soil. Hence, it has become imperative to do needful research in this direction. Also, very limited literature is available as utilization of hydrogels for nourishing the soil with zinc [14]. Keeping in view the research gaps and need to ensure steady supply of zinc micronutrient to the crops, to the best of our knowledge, CMTKG-based hydrogel is being used for the first time as a matrix for controlled release of zinc [29].

This study discusses the modification of carboxymethyl tamarind kernel gum (CMTKG)-based superabsorbent hydrogels by in situ incorporation of zinc source and its performance as a zinc micronutrient carrier vehicle for agricultural applications. The objectives of this study were to assess the following: (i) the swelling index of zinc-loaded superabsorbent hydrogel (ZSAH) at various pH solutions, saline solution, and distilled water; (ii) the release studies of zinc from ZSAH in water; (iii) the release studies of zinc from ZSAH in the soil; (iv) the release kinetic studies of ZSAH in both, water and soil.

Materials and methods

Materials

CMTKG (0.20°) was gifted by Hindustan Gum and Chemicals Ltd., Haryana, India. Acrylic acid (AAc, CDH, New Delhi), zinc chloride (ZC, Merck, Germany), potassium persulfate (KPS, Fischer Scientific, Mumbai), sodium hydroxide (Fischer Scientific, Mumbai), and N,N'-methylenebisacrylamide (MBA, Merck, Germany) were used as received. Distilled water was used for the preparation of solutions. Sodium acrylate was further prepared by mixing desired ratios of abovementioned acrylic acid (7 mL) and sodium hydroxide pellets (4 g).

Synthesis of ZSAH

Zinc was loaded by in situ incorporation technique to the superabsorbent hydrogel based on CMTKG polymer which was found out to be biodegradable by soil burial biodegradation studies [30] (Fig. 1a). All the polymerization and cross-linking reactions took place at 25°C. In order to synthesize ZSAH, CMTKG (1 g) was weighed and put in a beaker having distilled water (33.3 mL) and sodium acrylate (399.9 mmol). To this, a measured quantity of ZC (41.60 mmol) was added and stirred for 1 h. Afterward, a fixed amount of KPS (1.22 mmol) and MBA (0.37 mmol) were added and further stirred constantly for another 1 h. The whole mixture was transferred to the polyethylene tubes which were kept at 60°C in a hot water bath for 2 h. The polyethylene tubes were released to get the desired ZSAH, and then ZSAH was cut into small discs for further use. These discs were initially dried at room temperature for 3 days and afterward at 50°C in a vacuum oven, until the constant weight is achieved (Fig. 1b).

Swelling studies

Swelling studies of the synthesized ZSAH were done by the gravimetric method. Completely dried ZSAH samples (0.19 g) were weighed and immersed in solutions of pH 12, 9, and 4 (all the solutions were made by adding HCl or NaOH in dilute form), 0.9% NaCl solution (standard saline solution), and distilled water.

After a regular time interval, the swollen ZSAH were weighed post-wiping off the surface water with the help of a filter paper. All experiments were done in triplicate (Fig. 2).

The following equation [31] was used in calculating the swelling indexes.

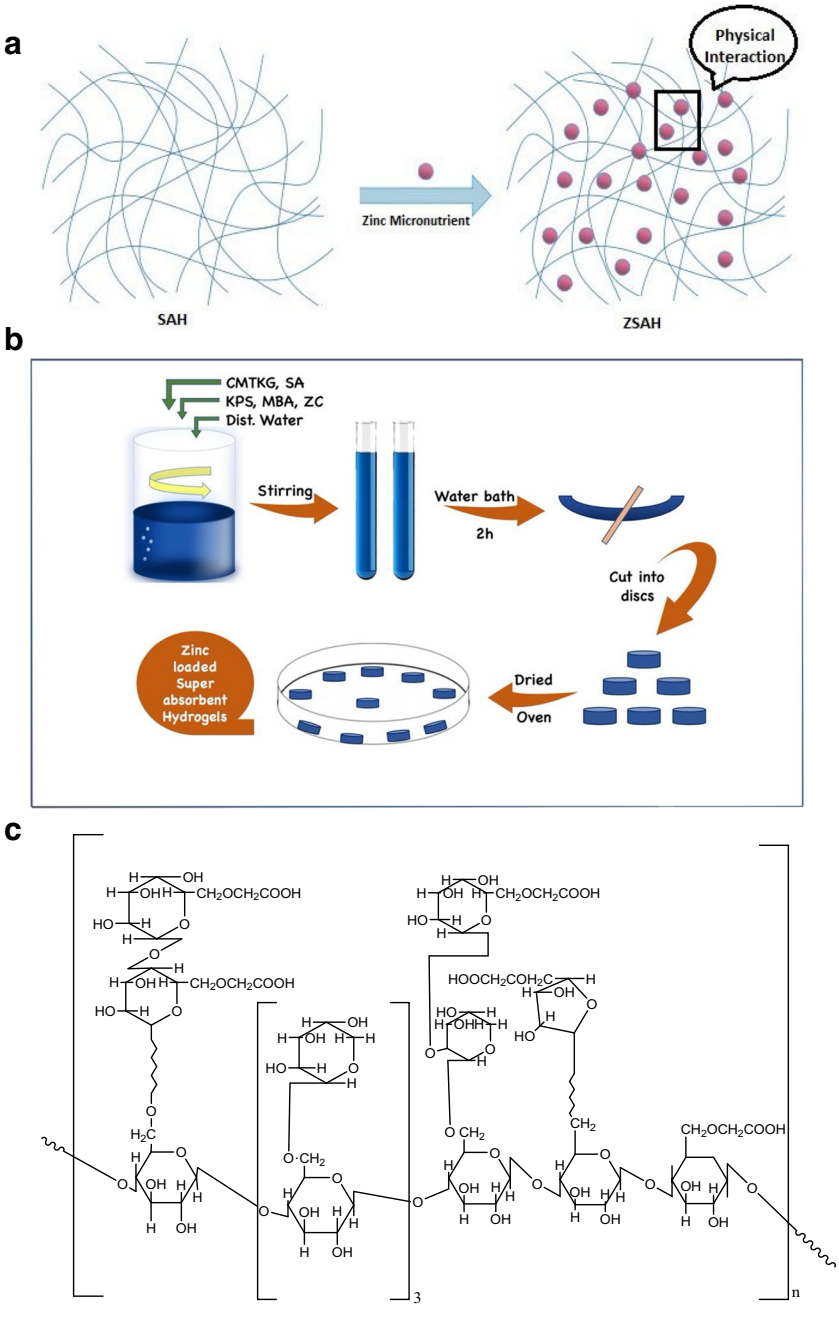
$$\text{Swelling index} = (W_{sh} - W_{dh})/W_{dh}$$

where W_{sh} is the weight of equilibrium swollen ZSAH and W_{dh} is the initial weight of the dried ZSAH

Characterization

FTIR spectroscopy

KBr pellet method was used for recording the FTIR (Fourier-transform Infrared) spectra of CMTKG, ZC, and ZSAH. It was done in the solid phase by using an FTIR spectrophotometer (Model: PerkinElmer spectrum version 10.5.3) between 400 and 4000 cm^{-1}.



Structure of CMTKG

Fig. 1 **a** Pictorial representation of the formation of ZSAH. **b** Schematic representation for the synthesis of ZSAH. **c** CMTKG: Chemical structure

Scanning electron microscopy (SEM)

Analysis of CMTKG and ZSAH's surface morphology was done by SEM (Model: JEOL JSM-6610LV).

Thermal analysis

PerkinElmer TGA was used for carrying out the thermogravimetric analysis (TGA) of CMTKG and ZSAH. For this purpose, the sample was heated from 25 to 900°C at a heating rate of 10°C/min under nitrogen atmosphere.

Applications of ZSAH as a zinc micronutrient carrier

The ZSAH synthesized was examined for the application in the sphere of agriculture for the controlled release of zinc micronutrient. It was established by studying and calculating the release behavior pattern along with the kinetics of ZSAH in water and as well in the soil for 72 h and 80 days, respectively. The composition of soil carried slit and clay 0.7% and sand 89% and showed bulk density 1.32 g/cm^3 and particle density 2.625 g/cm^3, which was collected from Delhi Technological University, New Delhi, India.

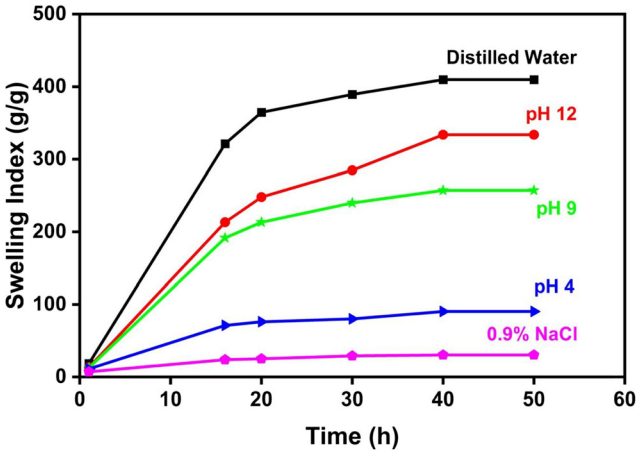


Fig. 2 Swelling studies of ZSAH

Release study of zinc in water

An accurately weighed ZSAH (200 mg, particle size: 100–150 mesh) was put into beakers containing 1 L of water under constant stirring. Two milliliters of solution was taken out after a regular interval of time, namely, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 24, 48, and 72 h. The sample solution was then analyzed for zinc concentration with the help of AAS (Spectrum ZXPRESS 8000). The results were then compiled with respect to time. The experiments were carried out in triplicates.

Release study of zinc in soil

An accurately weighed ZSAH (200 mg, particle size: 100–150 mesh) was taken in nylon sieve bags (Chiffon woven fabric: 200 mesh) and each was kept in soil (0.2 kg) in pots. Throughout the study, the moisture of soil was being maintained and samplings were done periodically (1, 3, 5, 10, 15, 20, 30, 40, and 80 days). The soil samples were taken out post-uniform mixing (after removing nylon bag) one by one at pre-fixed time intervals. These samples were dried in an oven at 100°C overnight and weighed. Ten grams of each soil samples was treated with 20 mL of DTPA (Diethylenetriamine pentaacetate) and shaken for 2 h. The supernatant solution was then filtered using Whatman filter paper-42 and analyzed in AAS. The experiments were carried out in triplicates.

Release kinetic studies of ZSAH in water and soil

Various models including the Korsmeyer-Peppas Model have been reported over time for estimating the release of a mobile agent from hydrogel as a function of time.

Zinc release data, in water as well as in soil, were fitted to different forms of kinetic models in order to understand zinc release mechanism, namely, the Korsmeyer-Peppas model [32], Higuchi model [33], first-order equation [34], and Olson model [35]. All these models are framed in Table 1. The model having the highest coefficient of determination (R^2) value has been considered as the best fit model.

Results and discussion

Synthesis of novel ZSAH

In an endeavor to fabricate ZSAH, zinc, in the form of ZC, was incorporated into the CMTKG cross-linked network (SAH) via in situ mechanism. The hypothetical representation incorporation of ZC ions in the three-dimensional network of SAH is illustrated through Fig. 1a whereas Fig. 1b is the schematic representation for the synthesis of ZSAH. Figure 1c tells about the chemical structure of CMTKG for better understanding.

Swelling studies

Swelling studies of ZSAH showed its immense capacity to swell by absorbing water into its porous framework. The swelling index of ZSAH for various solvents, viz., 0.9 % NaCl (standard saline solution), pH 12, 9, and 4, and distilled water is compiled in Fig. 2. ZSAH took 40 h in order to attain swelling equilibrium, where, among out of all used swelling media, the highest swelling index was observed by the distilled water (410 g/g).

For solutions of diverse pH, the water absorption capacity amplified significantly from pH 4 to 12. At pH 4, 90 g/g swelling index was observed by ZSAH which at pH 9 increased to 257 g/g and it got further increased to 334 g/g at pH 12. The lowest swelling was observed for acidic medium, the reason being improper ionization of the cross-linked network which resulted in a compact state of ZSAH. However, the spaces present in the network of ZSAH might have widened in case of alkaline conditions because of elevation in electrostatic repulsions produced by the carboxylate (COO^-) anions in ZSAH [36], whereas water absorbency of ZSAH was found to be dropped substantially in 0.9% NaCl solution case (30 g/g) which can be attributed to the existence of high ionic concentration which produced disparity in the surroundings of inner and outer phases of osmotic pressure and consequently entry of free water got obstructed inside the tridimensional network of ZSAH [37].

Characterization

FTIR spectroscopy

FTIR spectra of CMTKG, $ZnCl_2$, and ZSAH are shown in Fig. 3. CMTKG showed a broad band for -OH stretching vibrations at 3428 cm^{-1}. The asymmetric and symmetric

Table 1 Mathematical models used for the released kinetics from ZSAH

Model	Mechanism proposed in the model	Equation	Water		Soil		Ref.
			n	R^2	n	R^2	
Korsmeyer-Peppas	a. Fickian diffusion (n<0.45) b. Non-Fickian diffusion (0.89≥n≥0.45) c. Super case II (n>0.89)	$\frac{M_t}{M_\infty} = kt^n$ k=kinetic constant n=diffusion exponent	0.32	0.97	0.16	0.99	[32]
Higuchi	Fickian diffusion	$\frac{M_t}{M_\infty} = k_H t^{\frac{1}{2}}$ k_H= kinetic constant	-	0.88	-	0.91	[33]
First order	Rate is dependent on the concentration of one reactant only	$\log\left(1-\frac{M_t}{M_0}\right) = k_1 \cdot \frac{t}{2{\cdot}303}$ K_1=release rate constant	-	0.89	-	0.81	[34]

vibrations for -COO^- moiety are shown at 1631 cm^{-1} and 1415 cm^{-1}, respectively [38]. In ZSAH spectrum, a broad band (3418 cm^{-1}) is associated with -OH stretching vibration, whereas around 1568 cm^{-1} and 1410 cm^{-1} shows asymmetric and symmetric vibrations, respectively for -COO^- moiety. A sharp transmittance peak around 1568 cm^{-1} was observed for C=O stretching vibration of SA while the peak (1317 cm^{-1}) and a broad band (627 cm^{-1}) were linked with C-N and O=C-N groups of MBA [39]. The appearance of bands at 627 cm^{-1} and 1039 cm^{-1} can be associated with the CMTKG skeleton. It is notable mentioning that no significant peak was observed in the FTIR spectra of ZSAH as evident from the absence of any peak related to $ZnCl_2$. Considering this behavior, it can be concluded that only physical interactions (weak interactions) took place between zinc micronutrient and CMTKG cross-linked network. This further proved that the reversible retention of zinc micronutrient in the hydrogel matrix is unlikely to occur [14].

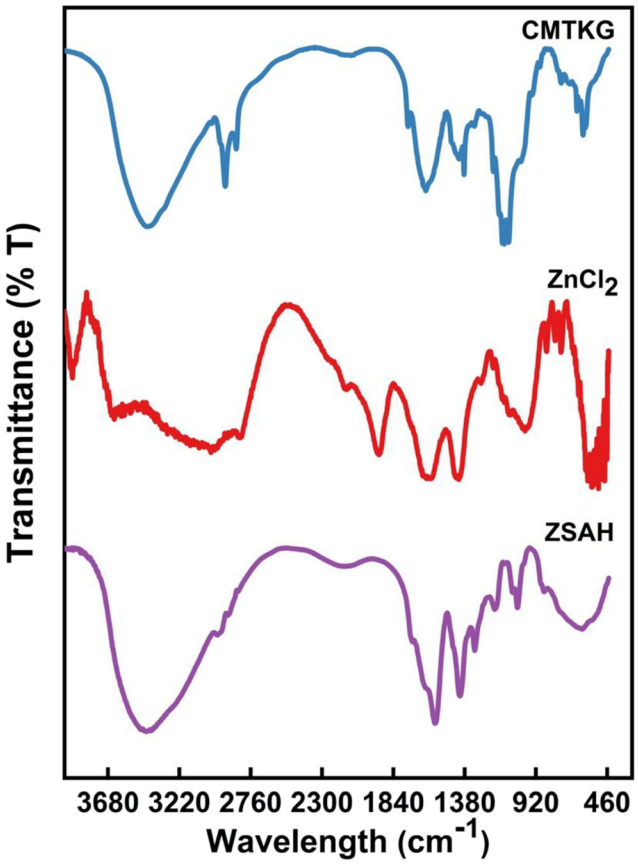


Fig. 3 FTIR spectra of CMTKG, $ZnCl_2$, and ZSAH

Scanning electron microscopy

SEM micrographs of CMTKG (a, b) and ZSAH (c, d), at different magnifications, are shown in Fig. 4. The morphological investigation of the ZSAH has been done so as to see the morphological changes in ZSAH as an effect of zinc incorporation. For a comparative study, SEM of CMTKG is also shown in Fig. 4. This analysis of CMTKG showed non-porous, rough, and compact surface whereas ZSAH showed spongy surface, interspatial voids, and distinct boundaries. However, the addition of zinc seemed to influence the morphology in the form of aggregated fractured appearance of ZSAH. This can be attributed to the intercalation of zinc into the matrix of cross-linked network of CMTKG [14].

Thermal analysis

To confirm the thermal stability of the CMTKG and ZSAH, thermogravimetric analysis (TGA) was done. Figure 5 shows TGA thermograms of CMTKG (Fig. 5a) and ZSAH (Fig. 5b). Weight loss was found in three distinct zones in the case of CMTKG. The very first zone (30–240°C) showed weight loss of 13.31% due to the removal of moisture content from CMTKG, whereas degradation of polymeric backbone could be attributed to the major weight loss of 47.38% for the second zone (240–602°C) followed by the third and the final weight loss (21.49%) in the range of 602–900°C [40].

However, in the case of ZSAH, the weight loss pattern was found out in four zones; the weight loss of 15.34 %, at an initial zone of 30–262°C, is due to the removal of moisture

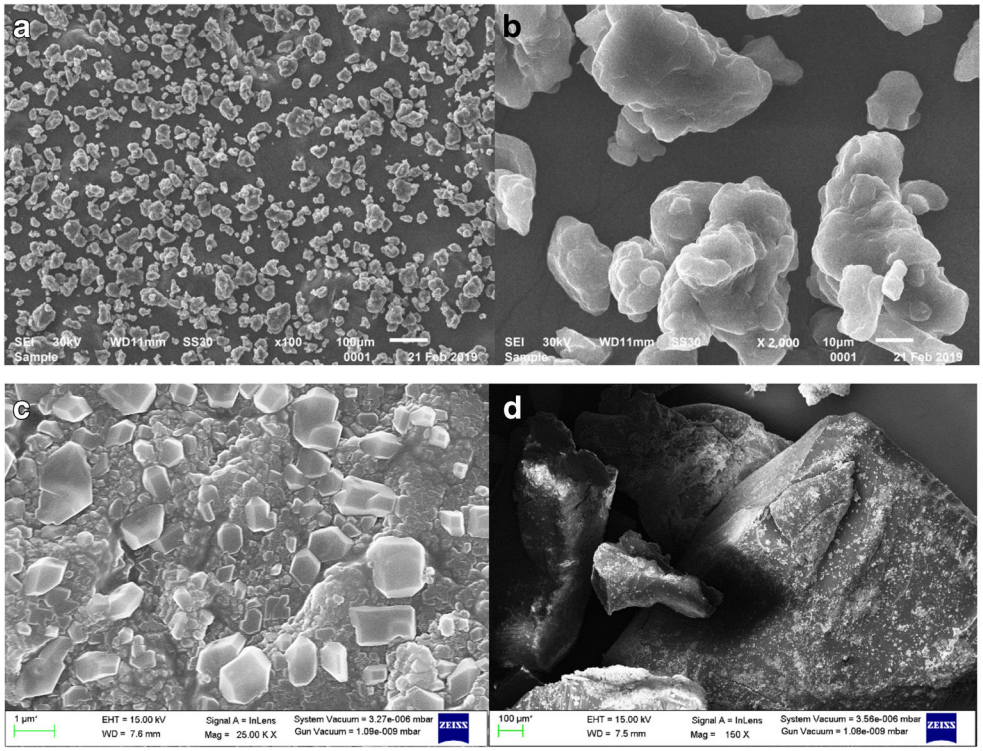


Fig. 4 SEM micrographs of CMTKG (**a**, **b**) and ZSAH (**c**, **d**)

from the cross-linked network of ZSAH. The second zone of 262–379°C showed a weight loss of 5.65% which was for degradation of cross-linked bonds between hydroxyl groups of the polymer backbone and cross-linker present in the network of ZSAH. Major weight loss of 24.97% was observed for the third zone (379–494°C) which can be attributed to the degradation of the main polymeric backbone, namely, carboxymethyl and hydroxyl group of CMTKG [41]. This was lastly followed by the fourth zone of complete combustion from 494 to 900°C with 18.53% weight loss. It can be clearly seen that 50% of ZSAH degraded at 598°C while for CMTKG, 50% of degradation observed at 365°C. It may also be noted here that the residual mass left was also found out to be more than 35% which was far more than the residual mass left in the case of CMTKG (18%). Hence, one may conclude ZSAH to be highly thermally stable than CMTKG.

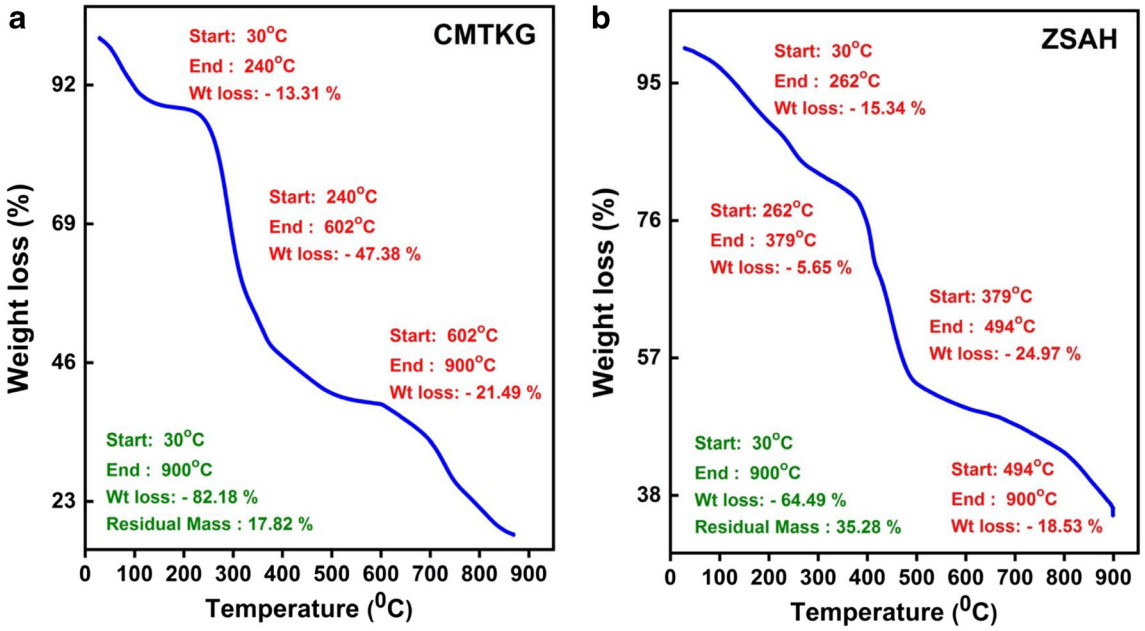


Fig. 5 TGA thermogram of CMTKG (**a**) and ZSAH (**b**)

Agricultural applications of ZSAH

Zinc is primarily the utmost essential micronutrient for plants but even than earth is still suffering from the valuable micronutrients and zinc being the number one on the most deficient micronutrient. In this work, ZSAH was fabricated in order to provide zinc to the plants. Figure 6 depicts the release rate of zinc with respect to time while Fig. 7 depicts the release kinetics of ZSAH.

Release study of zinc in water and soil

The periodic release of zinc from ZSAH in water is shown in Fig. 6a. This study was done for the duration of 72 h, and the release of zinc was observed to follow a linear trend. Initially at the time period of 30 min, zinc release percentage was found out to be just 12.2%. With time, the release percentage got steeply increased and in 72 h, the release percentage reached up to 54.2%.

In the case of soil also, the periodic release was analyzed to see the release of zinc from ZSAH which is depicted in Fig. 6b. This study was done for the duration of 80 days. In this case also, a linear trend was shown by ZSAH in releasing zinc. In soil, ZSAH exhibited 59.28% for the 10th day which notably enhanced to 83.68% on the 80th day. Hence, in the case of soil also, ZSAH showed a gradual increase in releasing zinc. It is significant to mention that the obtained result from the cumulative release of zinc through ZSAH in water was found out to be greater in comparison to the release of zinc in soil. For a time period of 72 h, the release was found to be 54.2% in water in contrast to just 42.28% in soil. The controlled release of zinc in the soil could be explained through three phases as seen in Fig. 6b. At first, the release occurred as a result of the wetting time of ZSAH whereas later on, zinc happened to release from the external layers of ZSAH due to the diffusion mechanism followed by, zinc diffused through inward layers by setting equilibrium with the external network of ZSAH. Thus, it can be inferred, in the soil case also, the ZSAH displayed a moderate increment in the pattern of releasing zinc.

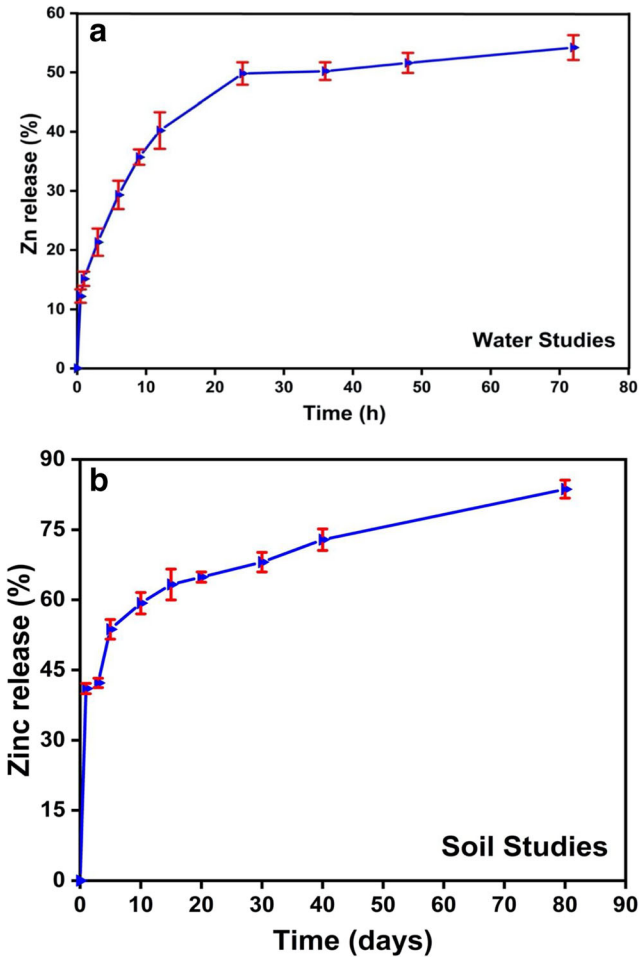


Fig. 6 Cumulative release percentage of zinc with respect to time in water (**a**) and in soil (**b**)

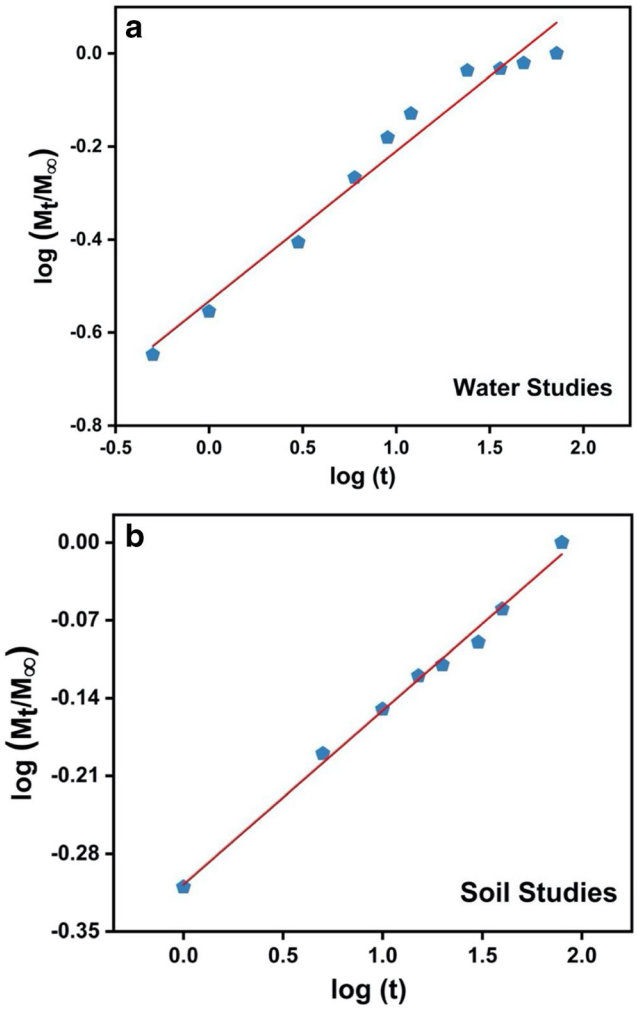


Fig. 7 Release kinetics of zinc by Korsmeyer-Peppas model in water (**a**) and in soil (**b**)

Table 2 Relative time period for the released quantity of zinc from ZSAH

Media	Olson model equation	k	T_{50} (days)	T_{95} (days)	T_{99} (days)
Water	$y = 22.391e^{0.401x}$	0.017	1.69	7.33	12.22
Soil	$y = 49.749e^{0.008x}$	0.008	86.63	375	625

Release kinetics study of zinc in water and soil

Release data were examined as per the various kinetic equations to describe the kinetics of zinc micronutrient release from the ZSAH matrix. All the release data showed the coefficient of determination values (R^2) nearer to one for the Korsmeyer-Peppas model and hence ZSAH followed the Korsmeyer-Peppas model (Table 1).

As per the value of R^2 given in Table 1, the Korsmeyer-Peppas model was suggested to be the best fit model which statistically narrated the release kinetic behavior of zinc from ZSAH. Zinc release in the water had 0.32 as diffusion exponent value (n) and 0.97 as R^2, which clearly suggested Fickian diffusion that happens due to the simple diffusion method through the network of hydrogel (Fig. 7a).

Interestingly, the release of zinc from ZSAH in soil showed its n value to be 0.16 and 0.99 as R^2 (Fig. 7b). This clearly demonstrates that the release mechanism was solely Fickian diffusion, which implies that zinc micronutrient is considered to be diffused through the framework of the hydrogel network simply by diffusion processes [42].

The anticipated time (in days), according to the Olson equation [35] to release 50 %, 95 %, and 99 %, that is (T_{50}, T_{95}, and T_{99}), of zinc from developed ZSAH, is recorded alongside with different parameters in Table 2. Zinc release in soil from ZSAH had a higher half-life (T_{50}), and T_{95}, T_{99} values, 86.63, and, 375, 625 days respectively than the release of zinc in water (1.69, 7.33, and 12.22 days in a similar fashion). This distinction can be ascribed to the fact that the water is a constraining component for the release in water, though in the soil, it can be attributed to the presence of moisture.

Conclusions

Novel ZSAH was synthesized by using *the* in situ technique, where zinc was doped into the CMTKG cross-linked network. The effect of various swelling media, viz. distilled water, pH 4, 9, 12, and standard saline solution, on the water absorption capacity of ZSAH was studied. The highest swelling index was observed up to 410 g/g, in the case of distilled water. Release studies of zinc were tested in water and as well as in soil with respect to time. It is significant to note that in soil, ZSAH exhibited 59.28% zinc release for the 10th day which notably enhanced to 83.68% on the 80th day. Release kinetic studies were also analyzed in both water and soil in which the Korsmeyer-Peppas model fitted the best as its R^2 values were nearer to one. It also revealed that the release of zinc from ZSAH, in the soil as well as in water, was highly independent of the structure of the polymeric network as it followed the Fickian diffusion mechanism. The half-life (T_{50}), T_{95}, and T_{99} values have been calculated by utilizing the Olson's single exponential model, which unmistakably demonstrated the zinc release in soil (86.63, 375, and, 625 days, respectively) had higher amount than the release of zinc in water (1.69, 7.33, and 12.22 days in a similar manner).

These desirable characteristics of ZSAH make it a potent candidate in the field of agriculture not only as a water-conserving agent but most importantly as a zinc micronutrient vehicle.

Acknowledgements The authors would like to thank the senior management of Delhi Technological University for constant encouragement and support.

Data accessibility This article has no additional data.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. An Z, Yan J, Sha J, Ma Y, Mou S (2021) Dynamic simulation for comprehensive water resources policies to improve water-use efficiency in coastal city. Environ Sci Pollut Res:1–22. https://doi.org/10.1007/s11356-020-12191-z
2. Dhakal MP, Ali A, Khan MZ, et al. (2021) Agricultural water management challenges in the Hunza River Basin: is a solar water pump an alternative option?*. Irrig Drain ird.2563. https://doi.org/10.1002/ird.2563
3. He G, Geng C, Zhai J, Zhao Y, Wang Q, Jiang S, Zhu Y, Wang L (2021) Impact of food consumption patterns change on agricultural water requirements: an urban-rural comparison in China. Agric Water Manag 243:106504. https://doi.org/10.1016/j.agwat.2020.106504
4. Suhr N, Widdowson M, Kamber BS (2021) The role of pedogenesis and natural fertiliser as vectors for essential metal content in agricultural topsoils, Central India. SN Appl Sci 3:40. https://doi.org/10.1007/s42452-020-03982-7
5. Mohammed S, Alsafadi K, Enaruvbe GO, Harsányi E (2021) Assessment of soil micronutrient level for vineyard production in southern Syria. Model Earth Syst Environ 1:3. https://doi.org/10.1007/s40808-021-01104-9
6. Baroni-nezhad H, Karimi M, Motaghian H, Direkvand-moghadam F (2021) Response of stevia (stevia rebaudiana) to copper, iron and zinc. J Plant Nutr 44:875–884. https://doi.org/10.1080/01904167.2021.1871753

7. El-Desouky HS, Islam KR, Bergefurd B et al (2021) Nano iron fertilization significantly increases tomato yield by increasing plants' vegetable growth and photosynthetic efficiency. J Plant Nutr:1–15. https://doi.org/10.1080/01904167.2021.1871749
8. Li J, Zhuang X, Font O, Moreno N, Vallejo VR, Querol X, Tobias A (2014) Synthesis of merlinoite from Chinese coal fly ashes and its potential utilization as slow release K-fertilizer. J Hazard Mater 265:242–252. https://doi.org/10.1016/j.jhazmat.2013.11.063
9. Bortolin A, Serafim AR, Aouada FA, Mattoso LHC, Ribeiro C (2016) Macro- and micronutrient simultaneous slow release from highly swellable nanocomposite hydrogels. J Agric Food Chem 64: 3133–3140. https://doi.org/10.1021/acs.jafc.6b00190
10. Hosseini SH, Niyungeko C, Khan S, Liang X (2021) Effects of superabsorbent polyacrylamide hydrogel and gypsum applications on colloidal phosphorus release from agricultural soils. J Soils Sediments 21:925–935. https://doi.org/10.1007/s11368-020-02819-5
11. Azeem B, Kushaari K, Man ZB et al (2014) Review on materials & methods to produce controlled release coated urea fertilizer. J Control Release 181:11–21. https://doi.org/10.1016/j.jconrel.2014.02.020
12. Ramli RA (2019) Slow release fertilizer hydrogels: a review. Polym Chem 10:6073–6090
13. Thombare N, Mishra S, Siddiqui MZ, Jha U, Singh D, Mahajan GR (2018) Design and development of guar gum based novel , superabsorbent and moisture retaining hydrogels for agricultural applications. Carbohydr Polym 185:169–178
14. Sarkar DJ, Singh A, Mandal P, Kumar A, Parmar BS (2015) Synthesis and characterization of poly (CMC-g-cl-PAam/zeolite) superabsorbent composites for controlled delivery of zinc micronutrient: swelling and release behavior. Polym-Plast Technol Eng 54: 357–367. https://doi.org/10.1080/03602559.2014.958773
15. El-Hady O, And SA-S-IJ of A (2006) U (2006) Conditioning effect of composts and acrylamide hydrogels on a sandy calcareous soil. II-Physico-bio-chemical properties of the soil. Int J Agric Biol 8:876–884
16. Bhardwaj AK, Warrington D (2007) Water retention and hydraulic conductivity of cross-linked polyacrylamides in sandy soils climate change mitigation and adaptation strategies for salt affected soils View project . Soil Sci Soc Am J 71:406–412. https://doi.org/10.2136/sssaj2006.0138
17. Johnson MS, Leah RT (1990) Effects of superabsorbent polyacrylamides on efficiency of water use by crop seedlings. J Sci Food Agric 52:431–434. https://doi.org/10.1002/jsfa.2740520316
18. Johnson MS (1984) Effect of soluble salts on water absorption by gel-forming soil conditioners. J Sci Food Agric 35:1063–1066. https://doi.org/10.1002/jsfa.2740351004
19. Kumar R, Parmar B, Kumar A et al (2007) Performance of a new superabsorbent polymer on seedling and post planting growth and water use pattern of chrysanthemum grown under controlled environment. Acta Hortic 742:43–50. https://doi.org/10.17660/ActaHortic.2007.742.5
20. Anupama K, Jat M, Management BP-J of W (2005) U (2005) Performance of a new superabsorbent polymer on crop and water productivity of summer mung bean (Phaseolus radiatus). J Water Manage 13:1–5
21. Rudzinski WE, Dave AM, Vaishnav UH, Kumbar SG, Kulkarni AR, Aminabhavi TM (2002) Hydrogels as controlled release devices in agriculture. Des Monomers Polym 5:39–65. https://doi.org/10.1163/156855502760151580
22. Bettini R, Catellani P, Santi P et al (2001) Translocation of drug particles in HPMC matrix gel layer: effect of drug solubility and influence on release rate. J Control Release 70:383–391
23. Shavit U, Shaviv A, Shalit G, Zaslavsky D (1997) Release characteristics of a new controlled release fertilizer. J Control Release 43: 131–138
24. Jamnongkan T, Kaewpirom S (2010) Controlled-release fertilizer based on chitosan hydrogel: phosphorus release kinetics
25. Thombare N, Mishra S, Shinde R, Siddiqui MZ, Jha U (2021) Guar gum based hydrogel as controlled micronutrient delivery system: mechanism and kinetics of boron release for agricultural applications. Biopolymers. 112. https://doi.org/10.1002/bip.23418
26. Hafeez B, Khanif YM, Saleem M (2013) Role of zinc in plant nutrition- a review. Am J Exp Agric 3:374–391
27. Tisdale SL, Nelson WL, Beaton JD (1985) Soil fertility and fertilizers. Soil fertility and fertilizers. Collier Macmillan Publishers, In, pp 382–391
28. Marschner H (1995) Mineral nutrition of higher plants
29. Khushbu WSG (2020) Potential applications and various aspects of polyfunctional macromolecule- carboxymethyl tamarind kernel gum. Eur Polym J 140:110042. https://doi.org/10.1016/j.eurpolymj.2020.110042
30. Khushbu WSG, Kumar A (2019) Synthesis and assessment of carboxymethyl tamarind kernel gum based novel superabsorbent hydrogels for agricultural applications. Polymer (Guildf) 182: 121823. https://doi.org/10.1016/j.polymer.2019.121823
31. Warkar SG, A.P. G (2015) Synthesis, characterization and swelling properties of poly (Acrylamide-cl-carboxymethylguargum) hydrogels. Int J Pharm Bio Sci 6:516–529
32. Korsmeyer RW, Von Meerwall E, Peppas NA (1986) Solute and penetrant diffusion in swellable polymers. II. Verification of theoretical models. J Polym Sci Part B Polym Phys 24:409–434. https://doi.org/10.1002/polb.1986.090240215
33. Higuchi T (1963) Mechanism of sustained-action medication. Theoretical analysis of rate of release of solid drugs dispersed in solid matrices. J Pharm Sci 52:1145–1149. https://doi.org/10.1002/jps.2600521210
34. Serra L, Doménech J, Biomaterials NP (2006) U (2006) Drug transport mechanisms and release kinetics from molecularly designed poly (acrylic acid-g-ethylene glycol) hydrogels. Biomaterials 27: 5440–5451
35. Olson JS (1963) Energy storage and the balance of producers and decomposers in ecological systems. Ecology 44:322–331. https://doi.org/10.2307/1932179
36. Chang C, Duan B, Cai J, Zhang L (2010) Superabsorbent hydrogels based on cellulose for smart swelling and controllable delivery. Eur Polym J 46:92–100. https://doi.org/10.1016/j.eurpolymj.2009.04.033
37. Zhao Y, Kang J, Tan T (2006) Salt-, pH- and temperature-responsive semi-interpenetrating polymer network hydrogel based on poly(aspartic acid) and poly(acrylic acid). Polymer (Guildf) 47: 7702–7710. https://doi.org/10.1016/j.polymer.2006.08.056
38. Trivedi JH (2013) Synthesis, characterization, and swelling behavior of superabsorbent hydrogel from sodium salt of partially carboxymethylated tamarind kernel powder- g -PAN. J Appl Polym Sci 129:1992–2003. https://doi.org/10.1002/app.38910
39. Reddy BV, Rao GR (2008) Vibrational spectra and modified valence force field for N,N'-methylenebisacrylamide. Indian J Pure Appl Phys 46:611–616
40. Pal S, Sen G, Mishra S, Dey RK, Jha U (2008) Carboxymethyl tamarind: synthesis, characterization and its application as novel drug-delivery agent. J Appl Polym Sci 110:392–400. https://doi.org/10.1002/app.28455
41. Dias RJ, Havaldar VD, Mali KK et al (2017) Sustained delivery of aceclofenac Interpenetrating networks of carboxymethyl tamarind gum and chitosan for sustained delivery of aceclofenac. https://doi.org/10.12991/mpj.2017.20
42. Ritger P, Release NP (1987) A simple equation for description of solute release I. Fickian and non-fickian release from non-swellable devices in the form of slabs, spheres, cylinders or discs. J Control Release 5:23–26

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